

https://dspace.org/

Exploring the extent and impact of genetic mosaicism in African populations

Evans, Jonathan

2025
University of the Witwatersrand, Johannesburg
https://hdl.handle.net/10539/49974
Evans, Jonathan. (2025). Exploring the extent and impact of genetic mosaicism in African populations [PhD thesis, University of the Witwatersrand, Johannesburg]. WIReDSpace. https://hdl.handle.net/10539/49974

Downloaded from a DSpace repository. https://dspace.org/



Exploring the extent and impact of genetic mosaicism in African populations

Jonathan Evans

Supervisors:

Prof. Zane Lombard

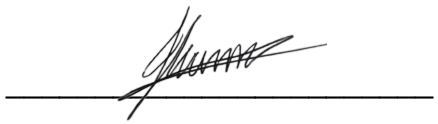
Dr. Nadia Carstens

A Thesis submitted to the Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Doctor of Philosophy.

Johannesburg, September 2025.

Declaration

I, **Jonathan Evans**, declare that this Thesis is my own, unaided work. It is being submitted for the Degree of Doctor of Philosophy at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.



Jonathan Evans

Signed on the 30 September 2025, in Melrose Estate, Johannesburg, South Africa.

Dedication

This thesis is dedicated to my family.

To my mother, Kathleen Evans, for supporting me unconditionally every day. You have been there at every step of this PhD journey, holding my head above the water, pushing me to the finish line through sheer willpower. Thank you for showing up for every concert, play, match, and PhD.

To my sister, Sarah Evans. Thank you for your kind heart, and for reminding me to be gentler on myself. You have always championed me. Thank you for looking after me, and for always having my back.

To my niece, Georgia Mary Evans, who brags about her “scientist” uncle whenever she can. You have grown so much over the course of my PhD. I hope that this example of where curiosity can take you. Thank you for the incessant questions, and willingness to learn about the world around you. In truth, I have learned as much from your gentle soul, and your bravery, as I have from this PhD. Stay curious.

To the Skippers, across the sea, who I can always count on to greet me with an enthusiastic J! Thank you for being so invested in the outcome of my PhD, and in the outcome of our lives. I will never forget the countless times you have supported us.

To the memory of Mary and Graham Fuller. I miss you every day.

To my little friend, Bilbo, who has watched over me like hawk while writing this thesis.

Research outputs

Oral Presentations:

The University of the Witwatersrand Faculty of Health Sciences Research Day - Johannesburg, South Africa. Evans, J., Mpangase, P. T., Carstens, N., Lombard, Z. Characterising mosaic chromosomal alterations in an African cohort.

The 14th International Congress of Human Genetics - 25th February 2023. Cape Town, South Africa. Evans, J., Mpangase, P. T., Carstens, N., Lombard, Z. Characterising mosaic chromosomal alterations in an African cohort.

Poster presentations:

The 14th International Congress of Human Genetics, Young Investigators Forum - 21st February 2023. Cape Town, South Africa. Evans, J., Mpangase, P. T., Carstens, N., Lombard, Z. Characterising mosaic chromosomal alterations in an African cohort.

Mutations in Time and Space – 25th April 2024. Edinburgh, United Kingdom. Evans, J., Mpangase, P. T., Carstens, N., Lombard, Z. Characterising mosaic chromosomal alterations in an African cohort.

Abstract

Post-zygotic mutations (PZMs) accumulate inevitably over a lifetime across all tissues. When PZMs are maintained in genetically distinct clones, it is known as genetic mosaicism (GM). Recent technological advances have facilitated accurate characterisations of GM in large datasets, revealing a complex relationship with normal aging. However, the contribution of GM across a spectrum of debilitating disorders has also become increasingly clear. GM is chronically under investigated in continental African cohorts and our understanding of the scale and impact of GM in healthy and disease affected Africans remains wholly inadequate. This thesis aimed to address these gaps by implementing state-of-the-art bioinformatics tools to sensitively detect GM in two distinct African cohorts. The first objective was to investigate the contribution of GM to developmental disorders (DD) in South Africans by leveraging 62 whole-exome trios from the Deciphering Developmental Disorders in Africa cohort to screen for candidate pathogenic PZMs. Screening successfully identified 5 proband PZM candidates, 3 germline *de novo* variants, and 5 parental PZM candidates with evidence of gonadosomatic mosaicism. In total, 12 novel variants were classified as likely pathogenic and recommended for future validation. The second objective was to detect and characterise mosaic chromosomal alterations (MCAs) associated with clonal haematopoiesis in healthy adults from the Africa Wits–INDEPTH Partnership for Genomic Studies (AWI-Gen) study. MCAs were successfully characterised in 10.1% of 9917 AWI-Gen participants, and were more prevalent in the elderly, in males, and in Western Africans. The genomic distributions of distinct MCA loci differed significantly between participants from Eastern, Western, and Southern Africans. Finally, novel genes were identified as candidate drivers of MCA expansion in Africans. This is the largest characterization of MCAs in a diverse cohort of continental Africans to date, and the first study to investigate the contribution of GM to a South African DD cohort. These results significantly contribute to our understanding of GM in Africans.

Acknowledgments

I need to firstly acknowledge my supervisors Prof. Zane Lombard and Dr. Nadia Carstens for their support and patience. I will forever be grateful to them for going above and beyond to support me through this time.

I want to thank Prof. Michelle Ramsay for allowing me to have access to the AWI-Gen dataset and for being so supportive and excited to help me with this analysis. It is much appreciated.

I want to acknowledge the help of Dr. Phelelani T. Mpangase who helped me with running the MoChA code in Chapter 3. Thank you for all your time and for setting up the pipeline.

Thank you to Dr. C. Cheng and Dr JT. Brandenburg for all your advice and help with figuring out Chapter 3.

List of abbreviations

3D	Three-dimensional
5mC	5-methylcytosine
A	Adenine
AA	African ancestry
ACGS	Association for Clinical Genomic Science
ACMG	American College of Medical Genetics and Genomics
ALT	Alternate allele
AMP	Association for Molecular Pathology
AoU	All of Us Research Program
ASD	Autism Spectrum Disorder
AWI-Gen	African Wits-INDEPTH Genomic Study
BAM	Binary Alignment Map
BCF	Binary Call Format
BER	Base Excision Repair
BIR	Break-Induced Replication
BL	Burkett's Lymphoma
BM	Bone Marrow
bp	Base Pair
BPM	Bead Pool Manifest
C	Cytosine
CH	Clonal Haematopoiesis
CNV	Copy Number Variant
CRAM	Compressed Reference-oriented Alignment Map
CRE	Cis-Regulatory Element
DD	Developmental Disorder
DDD-Africa	Deciphering Developmental Disorders in Africa
DDG2P	Developmental Disorder Genotype-to-Phenotype database
DDR	DNA Damage Response
DECIPHER	DatabasE of genomiC varIation and Phenotype in Humans using Ensembl Resources
DNA	Deoxyribonucleic Acid
DNM	De Novo Mutation
DP	Depth (sequencing coverage)
DSB	Double-Strand Break
EA	Eastern Africa
EGT	Illumina cluster file
EM	Expectation-Maximization
IDAT	Illumina Data File
IGV	Integrative Genomics Viewer
IL-1	Interleukin-1
IL-6	Interleukin-6
ILCOO	International Lung Cancer Consortium
INDEL	Insertion/Deletion Variant
IRR	Incidence Rate Ratio
Kb	Kilobase (1,000 bases)
LCR	Low Copy Repeat
LOF	Loss of Function
MAF	Minor Allele Frequency
Mb	Megabase (1 million bases)
MCA	Mosaic Chromosomal Alteration
MMR	Mismatch Repair
MoChA	Mosaic Chromosomal Alterations caller (software)
MSK-IMPACT	Memorial Sloan Kettering - Integrated Mutation Profiling of Actionable Cancer Targets
Mtb	Mycobacterium tuberculosis
MVP	Million Veteran Program
NDD	Neurodevelopmental Disorder
NHEJ	Non-Homologous End Joining
OMIM	Online Mendelian Inheritance in Man
OR	Odds Ratio
PGC	Primordial germ cell
pLI	Probability of Loss-of-function Intolerance
PM2.5	Particulate Matter <2.5µm (air pollution exposure)
post-odds	Posterior odds
PZM	Postzygotic Mutation
QC	Quality Control
REF	Reference allele
REPRIEVE	Randomized Trial to Prevent Vascular Events in HIV
ROS	Reactive Oxygen Species
SA	Southern Africa
SCD	Sickle Cell Disease
SD	Standard Deviation
SE	Standard Error
SegDup	Segmental Duplication

ESCC	Esophageal Squamous Cell Carcinoma
FDR	False Discovery Rate
FS	FisherStrand
G	Guanine
GATK-HC	Genome Analysis Toolkit-HaplotypeCaller
GDD	Global Developmental Delay
GDP	Gross Domestic Product
GENCODE	GENCODE Gene Annotation Project
GHS	Geisinger Community Health Initiative
GM	Genetic Mosaicism
gnomAD	Genome Aggregation Database
GTC	Genotype Call File (Illumina)
GWAS	Genome-Wide Association Study
HIV	Human Immunodeficiency Virus
HR	Homologous Recombination
HSC	Hematopoietic Stem Cell
IAAP	Illumina Array Analysis Platform
ID	Intellectual Disability
SNP	Single Nucleotide Polymorphism
SOR	Strand Odds Ratio
SSA	Sub-Saharan Africa
SSB	Single-Strand Break
SV	Structural Variant
T	Thymine
TAD	Topologically Associating Domain
TB	Tuberculosis
TOPMed	Trans-Omics for Precision Medicine
trio-WES	Trio-based Whole Exome Sequencing
TSG	Tumour Suppressor Gene
UCSC	University of California, Santa Cruz (Genome Browser)
VAF	Variant Allele Fraction
VCF	Variant Call Format
VEP	Variant Effect Predictor
VQSR	Variant Quality Score Recalibration
VUS	Variant of Uncertain Significance
WA	Western Africa
WSI	Welcome Sanger Institute

Table of contents

DECLARATION I

DEDICATION II

RESEARCH OUTPUTS III

ABSTRACT IV

ACKNOWLEDGMENTS V

LIST OF ABBREVIATIONS VI

TABLE OF CONTENTS VIII

LIST OF FIGURES XIII

LIST OF TABLES XVI

PREFACE XVIII

CHAPTER 1: LITERATURE REVIEW AND RATIONALE 1

1.1 Genetic mosaicism 2

1.2 Mechanisms of PZM formation 8

1.2.1 Mosaic SNPs and INDELs 10

1.2.2 Mosaic chromosomal alterations 14

1.2.3 DNA damage response pathways are efficient 18

1.2.4 DNA damage response pathways are imperfect 19

1.2.5 Mutational signatures 20

1.3 Biological factors impacting observable PZM frequencies in GM 22

1.3.1 Mechanistic factors impacting PZM rates 23

1.3.2 Factors impacting PZM fitness and clonal dynamics 25

1.4 Detecting PZMs from bulk genotyping data in single- sample normal tissues 26
1.4.1 Detecting MCAs from bulk genotyping data in non-matched single samples 28
1.4.2 Detecting mosaic small variants from bulk sequencing data in non-matched single samples 31

1.5 Genetic mosaicism across the human lifespan 35
1.5.1 Early genetic mosaicism: developmental disorders 36
1.5.2 Somatic mosaicism in adulthood: clonal haematopoiesis 41

1.6 Genetic mosaicism in Africans 49
1.6.1 Rare disorder mosaicism in Africa 49
1.6.2 Clonal haematopoiesis in Africa.....54

1.7 Thesis rationale 73
1.7.1 Research questions.....74
1.7.2 Hypothesis 75
1.7.3 Aim and objectives 75

CHAPTER 2: SCREENING FOR PATHOGENIC GENETIC MOSAICISM IN THE DECIPHERING DEVELOPMENTAL DISORDERS IN AFRICA STUDY 76

2.1 Introduction 76

2.2 Methodology 78
2.2.1 Study design and ethics 78
2.2.2 The DDD-Africa cohort.....78
2.2.3 DDD-Africa whole exome sequencing.....79
2.2.4 Candidate variant prioritization strategy.....81
2.2.4.1 Candidate variant quality control and filtering strategy 82
2.2.4.2 Statistical determination of mosaicism evidence 84
2.2.4.3 Candidate variant interpretation and clinical curation 87

2.3 Results..... 96
2.3.1 Study demographics.....96
2.3.2 Proband variant prioritization results 97
2.3.3 Parental gonadosomatic variant prioritization results 104
2.3.4 Candidate variants passing clinical curation and variant interpretation.....107
2.3.5 Clinical curation of selected candidate variants 111
2.3.5.1 Proband 1: *CNOT1*:c.2654G>T and *CHM*:c.434C>A.....111
2.3.5.2 Proband 17: *PMP22*:c.448G>C.....115
2.3.5.3 Proband 18: *MTOR:*c.5734G>T 118
2.3.5.4 Proband 23: *KIF5C*:c.1288G>T 121

2.3.5.5 Proband 41: *CDKL5*:c.978-14T>A 123
2.3.5.6 Proband 53: *NIPBL*:c.2584A>G 126
2.3.5.7 Proband 66: *RFX7*:c.812-3C>A 129
2.3.5.8 Proband 112: *NFE2L2*:c.1688-1690del 132
2.3.5.9 Proband 130: *KMT2A*:c.3576A>C 135
2.3.5.10 Variants of uncertain significance 137
2.3.6 Diagnostic yield potential in a best-case scenario 138

2.4 Discussion 140
2.4.1 Estimated DNM rates in DDD-Africa 141
2.4.2 Technical considerations 143
2.4.3 Further validation of candidates is essential 149
2.4.4 Challenges in mosaic variant interpretation and curation 151
2.4.5 Feasibility of trio-WES mosaic screening in African clinical settings 153

CHAPTER 3: CHARACTERISING MOSAIC CHROMOSOMAL ALTERATIONS IN AN AFRICAN COHORT 154

3.1 Introduction 154

3.2 Methodology 156
3.2.1 Study design and ethics 156
3.2.2 Study cohort: AWI-Gen 156
3.2.3 Genotyping with the H3Africa SNP array 156
3.2.4 Detecting MCAs from SNP array data: MoChA 157
3.2.5 Pipeline for detecting autosomal MCAs with MoChA 157
3.2.5.1 Converting Illumina H3Africa SNP array intensity data files for MoChA input. 160
3.2.5.2 Applying AWI-Gen GWAS quality control. 161
3.2.5.3 Additional quality control as recommended by MoChA for redundancy. 162
3.2.5.4 Statistical haplotype phasing the multisampling VCF with SHAPEIT4. 163
3.2.5.5 Calling candidate autosomal MCAs with MoChA. 164
3.2.5.6 Post-processing and filtering the putative autosomal MCA call set. 164
3.2.6 Statistical analysis 165
3.2.6.1 Initial univariate analyses. 165
3.2.6.2 Multivariate logistic regression analysis for odds of autosomal MCA occurrence. 166
3.2.6.3 Visualization of autosome wide MCA distribution with Circos plots 167
3.2.6.4 Identifying plausible candidate genes underlying the selection of commonly occurring autosomal MCAs. 168
3.2.6.5 Determining statistically significant MCA distributions 170
3.2.6.6 Phase switch error analysis 170

3.2.6.7 Determining MCA distribution differences between population cohorts 171

3.3 Results 172

3.3.1 Study cohort demographic characteristics 172

3.3.2 Autosomal MCAs were successfully detected in the AWI-Gen cohort 175

3.3.3 Participant demographic variables were significantly associated with the odds of autosomal MCA occurrence in the AWI-Gen cohort 176

3.3.3.1 The odds of autosomal MCA occurrence were significantly greater in male participants. 178

3.3.3.2 The odds of autosomal MCA occurrence increased significantly with increasing participant age. 179

3.3.3.3 The odds of at least one autosomal MCA occurring differed significantly between participants from Southern Africa, Western Africa, and Easten Africa. 182

3.3.4 The sizes and cell fractions of autosomal losses, gains, and CN-LOHs differed significantly in the AWI-Gen cohort 189

3.3.5 The distribution of 1245 detected MCAs across 998 African autosomes in the AWI-Gen cohort 194

3.3.5.1 MCAs were non-randomly distributed across the autosome. 196

3.3.5.2 Novel candidate drivers of common autosomal gain and loss loci. 200

3.3.5.3 Differential autosomal MCA distributions were associated with participant sex and age. 205

3.3.5.4 Autosomal MCA distributions differed across regions of recruitment. 210

3.3.6 Phase switch errors in autosomal MCAs differed by region of recruitment 220

3.3.7 Comparison of detected MCAs between cohorts from Africa, Europe, and Asia 228

3.3.7.1 Autosomal MCA characteristics differed between the AWI-Gen, UKB, and BBJ cohorts. 229

3.3.7.2 Autosomal MCA frequencies differed between the AWI-Gen, UKB, and BBJ cohorts. ... 231

3.3.8 Participants with multiple distinct autosomal MCAs 236

3.3.8.1 Multiple autosomal MCAs were less common in AWI-Gen 236

3.3.8.2 Participants with multiple sub-clonal CN-LOHs 237

3.3.8.3 Co-occurring MCAs in the AWI-Gen cohort 239

3.4 Discussion 244

3.4.1 Study strengths 245

3.4.2 Technical considerations 246

3.4.3 Future methodology improvements 250

3.4.4 Biological factors impacting MCA distributions in AWI-Gen 251

CHAPTER 4: CONCLUSION 256

4.1 Key findings 257

4.2 Study limitations 260

4.3 Future directions 263

REFERENCES 265

APPENDICES 300

Appendix A: PhD ethics certificate 300

Appendix B: DDD-Africa ethics certificate 301

Appendix C: Plagiarism declaration 302

Appendix D: Genetic mosaicism in African rare disease review papers 303

Appendix E: DDD-Africa recruitment criteria and demographic distribution 307

Appendix F: Variants identified in a parallel DDD-Africa study 308

Appendix G: Table of variants classified as VUS 309

Appendix H: IGV plots for candidate VUS variants 310

Appendix I: Stratified MCA phenotypes for bubble graph 311

Appendix J: Multiple MCA association analysis results 314

Appendix K: Expanded autosomal MCAs 315

Appendix L: Turnitin Report 316

List of figures

Figure 1.1: A simple representation of genetic mosaicism in human development from embryogenesis to adulthood. 3

Figure 1.2: A visualisation of a newly proposed system of classifying clinacally relevant genetic mosaicism in humans. 5

Figure 1.3: Common genotoxic agents produce characteristic DNA lesions which are repaired by diverse DNA damage repair pathways. 9

Figure 1.4: Illustratation of the small variations and large structural variations observed as somatic mutations. 10

Figure 1.5: DNA damage and erroneous repair pathways resulting in mutations. 13

Figure 1.6: Illustrative representation of diverse mosaic chromosomal alterations....14

Figure 1.7: Major pathways leading to SV formation. 16

Figure 1.8: A summary representation of mutational signature features of somatic mutations....21

Figure 1.9: Hypothetical impact of mechanistic factors and selective factors on observed PZM frequencies in tumour cells. 22

Figure 1.10: Environmental factors that shape genomic diversity within Africa. 60

Figure 2.1: Workflow of variation prioritization strategies deployed to identify candidate mosaic variants from trio-WES in the DDD-Africa cohort. 95

Figure 2.2: Results from the proband *de novo* variant prioritization pipeline for 62 DDD-Africa trios. 98

Figure 2.3: Candidate SNPs passing the EM-mosaic FDR-minimum alternate read depth threshold. 99

Figure 2.4: EM-Mosaic generated variant allele fraction density plots of likely mosaic and likely germline candidate SNPs. 100

Figure 2.5: Overdispersion plot of calculated VAF variance across observed read depth values. 101

Figure 2.6: Candidate proband SNPs predicted to be mosaic versus germline *de novo* by EM-Mosaic with posterior odds >10. 102

Figure 2.7: Mosaic detection power curves. 103

Figure 2.8: The parental gonadosomatic variant prioritization pipeline for 62 DDD-Africa trios outlining the results at each step. 105

Figure 2.9: The *CNOT1*:c.2654G>T PZM candidate in Family 1. 113

Figure 2.10: The mosaic *CHM*:c.434C>A stopgain SNP detected in Family 1. 115

Figure 2.11: The *PMP22*:c.448G>C gonadosomatic SNP candidate detected in Family 17. 117

Figure 2.12: The *MTOR*:c.5734G>T gonadosomatic SNP candidate detected in Family 18. 120

Figure 2.13: The *KIF5C*:c.1288G>T germline *de novo* missense SNP detected in proband 23. 122

Figure 2.14: The *CDKL5*:c.978-14T>A *de novo* splice polypyrimidine tract SNP detected in proband 41. 125

Figure 2.15: The *NIPBL*:c.2584A>G gonadosomatic missense SNP in Family 53 ..129

Figure 2.16: The *RFX7*:c812-3C>A mosaic splice altering SNP candidate in family 66. 131

Figure 2.17: The *NFE2L2*:c.1688-1690del gonadosomatic in-frame INDEL in Family 112 134

Figure 2.18: The *KMT2A*:c.3576A>C germline *de novo* missense SNP detected in proband 130. 136

Figure 3.1: The MCA detection workflow for the AWI-Gen dataset. 159

Figure 3.2: The age distributions of male and female AWI-Gen participants in this study. 173

Figure 3.3: Univariate analysis results showing significant differences in the demographic makeup of participants from each region of recruitment. 174

Figure 3.4: Univariate analysis showing the impact of participant sex on the odds of autosomal MCA occurrence in AWI-Gen. 178

Figure 3.5: Increased age was significantly associated with greater odds of at least one autosomal MCA in both male and female AWI-Gen participants. 181

Figure 3.6: The proportion of participants with at least one autosomal MCA differed between both region of recruitment and study site in AWI-Gen. 183

Figure 3.7: Multivariate logistic regression analysis showing the combined impact of region of recruitment, age, sex, and batch on differential odds of autosomal MCA occurance. 186

Figure 3.8: Multivariate regression anaylsis showed that particiapnts across study sites had differential odds of autosomal MCA occurance when adjusting for sex, age, and batch. 188

Figure 3.9 Univariate analysis showing CN-LOHs, gains, and losses differing in size and estimated cell fraction distributions. 191

Figure 3.10: Scatterplots showing negative correlations between MCA size (Mb) and cell fraction stratified for losses, gains, and CN-LOHs. 193

Figure 3.11: The proportion of MCAs detected in each chromosome in the AWI-Gen cohort 195

Figure 3.12: *Autosomal distribution of* 1245 MCA *loci* detected *in 998* AWI-Gen *participants*. 197

Figure 3.13: Circos plot showing the mean phased heterozygous SNP probe coverage per 1Mb window in the AWI-Gen cohort achieved with the H3Africa SNP array. 199

Figure 3.14: Recurrent focal loss loci in the AWI-Gen cohort commonly impacting candidate driver genes. 203

Figure 3.15: Bubblegraph showng autosomal MCA phenotype proportions by mean age and fraction male sex distributions. 205

Figure 3.16: Autosomal MCA loci with evidence of sex biased selection in the AWI-Gen cohort. 208

Figure 3.17: Autosomal MCA phenotype frequency correlations by region of recruitment. 211

Figure 3.18: Autosomal loss loci with evidence of region of recruitment biased selection in the AWI-Gen cohort. 215

Figure 3.19: Autosomal gain loci with evidence of region of recruitment biased selection in the AWI-Gen cohort. 217

Figure 3.20: Autosomal CN-LOH loci with evidence of region of recruitment biased selection in the AWI-Gen cohort. 219

Figure 3.21: Univariate analysis of factors associated with phase flip counts in the AWI-Gen cohort. 222

Figure 3.22: The effect of MCA size, type, and region of recruitment on phase flip counts.....224

Figure 3.23: Scatter plots of MCA phase flip counts with corresponding MCA size (Mb) stratified by type and region of recruitment. 226

Figure 3.24: Phase switch errors are more common on the terminal ends of autosomes. 227

Figure 3.25: Autosomal MCA frequency correlations between AWI-Gen, UKB, and BBJ cohorts. 232

Figure 3.26: Comparison of autosomal MCA frequencies between participants from the AWI-Gen, U.K Biobank and Biobank Japan cohorts. 234

Figure 3.27: Five distinct but overlapping sub-clonal CN-LOH expasions in a 46-year-old male from Dikgale. 239

Figure 3.28: Estimated fraction of cells containing co-occurring autosomal MCAs .240

Figure 3.29: Circos plot of coocuring MCAs in the AWI-Gen cohort 242

Supplementary figures:

Figure A. 1: IGV plots for VUS candidate varaints.....310

Figure A. 2: Univariate analysis of multiple MCAs by age, sex, and region of recruitement for 998 participants with at least one autosomal MCA.....314

Figure A. 3: The cell fraction distribution of detected autosomal MCAs by type, size, and chromosomal location 315

List of tables

Table 1.1: A selection of bioiformatic tools for detecting GM from SNP-array and NGS data. 27

Table 1.2: Clonal haematopoeisis studies with African ancestry participants. 56

Table 2.1: Probands with excessive DNM counts flagged by EM-Mosaic. 86

Table 2.2: Calibrated *in silico score* thresholds selected for this study. 90

Table 2.3: Summary of DDD-Africa demographics 96

Table 2.4: Probands with prior genetic diagnoses. 97

Table 2.5: A final list of candidate variants recomended for furthur validation. 108

Table 2.6: A summary of clinical phenotypes for DDD-Africa trios with potentially pathogenic candidate variants....109

Table 3.1 Study cohort demographics stratified by African region and site of recruitment. 172

Table 3.2: Summary statistics of 1245 autosomal MCAs detected in the AWI-Gen cohort. 176

Table 3.3: A summary of the demographic makeup of 998 AWI-Gen participants with at least one detected autosomal MCA. 177

Table 3.4: The number of male and female AWI-Gen participants with at least one autosomal MCA stratified by increasing age groups. 179

Table 3.5: A logistic regression model showing the combined effect of participant age and sex on autosomal MCA occurrence odds. 180

Table 3.6: Univariate logistic regression models showing the impact of region of recruitment on the odds autosomal MCA incidence in AWI-Gen. 182

Table 3.7: A fixed-effects multivariate logistic regression of sex, age, region of recruitment, and sample batch on autosomal MCA occurrence odds. 185

Table 3.8: The characteristics of 1245 autosomal MCAs stratified by MCA type. 190

Table 3.9: Parameter estimates for a power curve fitted regression of MCA event size predicted by MCA event cell fraction with the formula $Y=aX^b$. 192

Table 3.10: Candidate genes driving the selection of commonly deleted loci scored by the gene approach. 201

Table 3.11: Candidate genes driving the selection of commonly deleted loci scored by the window approach. 202

Table 3.12: Candidate pan-cancer and leukaemia associated driver genes recurrently duplicated in common loci scored by 1 Mb windows in the AWI-Gen cohort. 204

Table 3.13: Autosomal window loci with evidence of differential distributions between 521 male and 477 female participants with at least one MCA. 206

Table 3.14: Autosomal windows with evidence of differential MCA distributions between 483 SA, 362 WA, and 153 EA participants. 213

Table 3.15: Fixed-effects multivariate negative binomial regression of MCA size (Mb), MCA type, region of recruitment, and batch on MCA phase flip counts. 223

Table 3.16: Autosomal MCA characteristics summary across the AWI-Gen, U.K Biobank, and Biobank Japan cohorts. 229

Table 3.17: Autosomal MCA counts that were significantly different between the AWI-Gen and UKB or BBJ cohorts....233

Table 3.18: Number of distinct MCAs detected in 998 participants....237

Table 3.19: Six participants with evidenc of multiple overlapping sub-clonal CN-LOH expasions 238

Table 3.20: Pairwise enrichemnt of cooccuring MCAs in AWI-Gen....241

Supplementary tables:

Table A. 2: DDD-Africa inclusion and exclusion criteria categories. 307

Table A. 1: Reasons for recruitment in 62 probands. 307

Table A. 3: Pathogenic small variants previously detected in parallel DDD-Africa analysis 308

Table A. 4: Candidate proband PZMs and parental PZMs classified as variants of uncertain significance. 309

Table A. 5: The characteristics of 82 MCA phenotypes stratified by chromosomal arm and type 311

Table A. 6: Multivariate logistic regression model for multiple (three or more) MCAs in 845 SA and WA participants 314

Preface

DNA is well-suited to its dual evolutionary role as the repository of hereditary genetic information accumulated over eons, and as the primary substrate for natural selection (Herrick, 2024). While it cannot be static as a source of variation for adaptation, DNAs fidelity must be maintained to preserve conserved sequence over successive generations (Lynch et al., 2016). To this end, we have evolved several responses to DNA damage, including DNA damage repair pathways (DDRs), cell-cycle checkpoints, damage tolerance pathways, and apoptosis pathways (Buoninfante et al., 2023). However, the combination of the reactive chemical composition of DNA and its organisation in the 3D genome make it vulnerable to damage (Chatterjee and Walker, 2017, Swenberg et al., 2011). Moreover, cell division is a complex process prone to DNA replication errors and mitotic segregation errors (Currie et al., 2022). While DDR pathways are efficient, they are also imperfect, often introducing errors of their own or compounding existing damage (Chatterjee and Walker, 2017, Currie et al., 2022, Saini et al., 2021). DNA mutations are inevitable and accumulate with age from the very first post-zygotic division during embryogenesis (Coorens et al., 2021, Lynch et al., 2016, Moore et al., 2021).

Post-zygotic mutations (PZMs) are subject to selection and genetic drift, with most having no impact on cell function or fitness (Loewe and Hill, 2010, Watson et al., 2020). Selectively neutral PZMs can become fixed in cell clones through random drift alone as passenger mutations (Martincorena and Campbell, 2015). PZMs with a net selective advantage for cells in their dynamic niches, drive clonal expansions over time (Challen and Goodell, 2020, Watson et al., 2020, Watson and Blundell, 2023). Consequently, multicellular organisms are composed of heterogeneous cell clones, each with slightly different genomes (Biesecker and Spinner, 2013, Martínez-Glez et al., 2020). When genetically distinct clones are observed in an organism having developed from a single zygote, it is known as genetic mosaicism (GM) (Biesecker and Spinner, 2013, Martínez-Glez et al., 2020).

Genetic mosaicism has long been recognised as a contributor to malignancy, skin disorders, and neurodevelopmental diseases (Alissa and Walsh, 2018, Happle, 1993, Martincorena and Campbell, 2015). However, the full extent and impact of GM in human health and ageing have historically been underestimated due to the technical challenges of its detection(Dou et al., 2018, Martínez-Glez et al., 2020). Over the past

decade, significant advancements in sampling protocols, genotyping and sequencing technologies, and bioinformatics tools have enabled more accurate detection of GM in non-malignant or histologically “normal” tissues (Abascal et al., 2021, Ellis et al., 2021, Yang et al., 2023, Loh et al., 2020, Hsieh et al., 2020). Additionally, the establishment of biobanks and tissue repositories has enabled the detection of GM, both prospectively and retrospectively, in large cohorts with significantly greater statistical power (Jakubek et al., 2023b, Loh et al., 2020, Terao et al., 2020, Uddin et al., 2022, Zekavat et al., 2021, Bick et al., 2020b, Kessler et al., 2022, Niroula et al., 2021). These technological breakthroughs have reinvigorated research interest in GM, across tissues, in both diseased and healthy cohorts (Evans and Walsh, 2022). Perhaps the most surprising findings from these large characterizations is that GM is more common in non-malignant tissues than previously thought, with PZMs being the norm rather than the exception as we age (Abascal et al., 2021, Coorens et al., 2021, Ellis et al., 2021, Herms and Jones, 2023, Loh et al., 2020, Martincorena and Campbell, 2015, Moore et al., 2021, Bick et al., 2020b, Jakubek et al., 2023b, Kessler et al., 2022, Mitchell et al., 2022, Terao et al., 2020).

Pathogenic PZMs have also continued to be linked to aberrant phenotypes and disease pathologies (Martínez-Glez et al., 2020, Mustjoki and Young Neal, 2021). Pathogenic PZMs that arise early in embryogenesis can expand to significant proportions of developing tissues, leading to somatic or gonadosomatic mosaicism (Martínez-Glez et al., 2020). This is an established mechanism underlying non-malignant diseases, including developmental disorders, autism spectrum disorders, and congenital malformations (Dou et al., 2017, Eberhardt et al., 2023, Lim et al., 2017, Sherman et al., 2021, Wright et al., 2019). Several studies have leveraged trio whole exome sequencing (WES) to accurately detect both *de novo* somatic mosaicism in affected probands and inheritable parental mosaicism in these conditions (Domogala et al., 2021, Gambin et al., 2020, Lim et al., 2017, Wright et al., 2019). Thus, improving diagnostic yields within existing datasets.

Selectively advantageous PZMs in adult stem or progenitor cells may lead to clonal expansions, contributing to a greater proportion of daughter cells which can be detected in mature tissues. The best-characterized clonal expansions are in the haematopoietic system, which is also known as clonal haematopoiesis (CH) (Bick et al., 2020b, Jakubek et al., 2023b, Kessler et al., 2022, Loh et al., 2020, Terao et al., 2020). Blood samples can be collected noninvasively, and large cohorts with

microarray genotyping or bulk sequencing data facilitate the retrospective detection of PZMs in CH with minimal additional cost (Evans and Walsh, 2022). The Mosaic Chromosomal Alterations (MoChA) caller was developed to sensitively detect mosaic chromosomal alterations from SNP-array data commonly generated for large genome-wide association studies (GWAS) (Loh et al., 2018, Loh et al., 2020). MoChA has become the gold standard tool for detecting MCAs and has been used to successfully characterize MCAs in several large cohorts (Jakubek et al., 2023b, Loh et al., 2020, Niroula et al., 2021, Terao et al., 2020, Uddin et al., 2022, Zekavat et al., 2021). These studies showed that MCAs are both a normal feature of aging, and a significant risk factor of malignancy, infection, and mortality (Jakubek et al., 2023b, Loh et al., 2020, Niroula et al., 2021, Terao et al., 2020, Uddin et al., 2022, Zekavat et al., 2021).

Technological breakthroughs have advanced GM research exponentially within a short period of time. The full extent of GM is rapidly being uncovered (Evans and Walsh, 2022, Mustjoki and Young Neal, 2021, Alissa and Walsh, 2018). However, there is significant gap in our understanding of GM. There is a paucity of GM data from continental African participants in both population genetics and clinical genetics cohorts. This mirrors the overall underrepresentation of continental African data in genomics datasets (Fatumo et al., 2022). African based genomics groups may not have access to the appropriate datasets, expertise, or infrastructure required to detect and analyse GM effectively. This requires significant funding which is limited in low to middle income countries. Conversely, research groups outside of Africa may have the requisite resources and expertise, but not access to continental African data. Additionally, the African ancestry (AA) data from admixed African American cohorts, that is currently in the literature, may not be a fair reflection of GM in continental Africans.

This thesis aimed to address this gap through an exploratory investigation into the extent and impact of GM in continental African datasets. A central hypothesis of this thesis was that GM can be retrospectively characterised in continental African datasets by integrating state-of-the-art GM detection tools into existing analysis pipelines. Additionally, that this can be achieved at minimal additional cost in a resource-constrained setting. This was achieved by leveraging access to two continual African datasets. A clinical cohort and a population genetics epidemiological cohort. Cohort specific GM analysis strategies were applied to each to maximise utility. The two analyses strategies presented in this thesis allowed for the characterisation of GM in

Africans across a spectrum of age, demography, and health status. The data presented herein both contributed to the global GM knowledgebase, while also aligning with the specific objectives of the studies from which the data derived. The following is a general overview of the structure of the thesis.

Chapter 1 is a comprehensive literature review of GM, how it is defined and classified, how it arises, and the impact it has in human health, aging and disease. The review is extensive and summarises the challenges in GM detection and some innovative strategies which have been developed to overcome them. The review contextualises GM in line with the analyses in Chapters 2 and 3. An overview of GM in continental African cohorts is then presented, and the gap in the literature is highlighted. Finally, the rationale for the thesis is explored, and the aims and objectives are outlined in depth.

Chapter 2 details the first objective of this thesis, a screening strategy for detecting pathogenic PZMs in South African children with DDs from the Deciphering Developmental Disorders in Africa (DDD-Africa) study (DDD-Africa, https://h3africa.org/index.php/ddd-africa/). It centres on an adapted variant prioritization pipeline, which incorporated the EM-Mosaic tool in trio-WES data, to screen for *de novo* PZMs in affected probands and inherited parental PZMs (Hsieh et al., 2020, Wright et al., 2019). A list of candidate mosaic small variants is recommended for future validation.

Chapter 3 details a retrospective characterisation of MCA driven CH in the Africa Wits-INDEPTH Partnership for Genomic Research (AWI-Gen) study. This chapter details the largest and most diverse autosomal MCA characterisation in continental Africans to date (Loh et al., 2018, Loh et al., 2020, Mulder et al., 2018). An exploratory analysis of detected MCAs is detailed, highlighting the diversity of autosomal MCA distributions across participant sex and region of recruitment. Finally, novel candidate drivers of MCA selection in continental Africans are identified.

The thesis concludes with Chapter 4, a final discussion of the analyses detailed in Chapters 2 and 3. Chapter 4 highlights the limitations of both chapters, outlines future perspectives, and discusses the success of this thesis in achieving its aims.

Chapter 1: Literature review and rationale

In his defining work, *On the Origin of Species*, Darwin formulated the elegant theory of natural selection, which acts on heritable variation to drive the adaptive evolution of species through time (Darwin, 1859). Key to this theory, was Darwin's observation that interindividual variation is passed down through successive generations. However, how this variation arose or the mechanisms behind its inheritance was not known.(Bromham, 2009). The discovery of DNA and the gene as the molecular basis of heritable variation was a revelation, validating Darwin's insight, and ushering in a golden age of biological discovery in modern genetics (Brenner et al., 1961, Flemming, 1882, Franklin and Gosling, 1953, Miescher, 1871, Morgan, 1910, Nirenberg and Matthaei, 1961, Watson and Crick, 1953).

We now understand how natural selection acts on the genetic variation to drive evolution through adaptation, a process that is manifested from the cellular level throughout the hierarchical organization of life (Gregory, 2009, Kovuri et al., 2023).. The foundational principles of Darwinian biology remain at the heart of modern population genetics theory; where the forces of mutagenesis, selection, gene flow, and genetic drift shape human allele frequencies through time and space (Kern and Hahn, 2018, Lynch et al., 2016, Simon and Coop, 2024). These same population-level principles have adapted to explain the observable patterns of somatic mutations in GM and to model clonal dynamics within tissues (Maher et al., 2016, Martincorena and Campbell, 2015, Martincorena et al., 2017). Technological breakthroughs over the past decade have led to significant advancements in GM research in genomic medicine, aging, cancer research, and genomic epidemiology (Evans and Walsh, 2022, Lee et al., 2023, Mohiuddin et al., 2022, Vijg, 2021). Central to these advancements is the enduring relevance of Darwinian biology, facilitated by our modern understanding of DNA.

The information encoded within DNA is critical for cellular function, homeostasis, and development. It must be maintained through successive cell divisions with high fidelity. However, as genetic variation also serves as a fundamental substrate for natural selection, DNA sequence cannot be entirely static. A balance must be maintained between conserving the DNA sequence that is essential for short-term cell survivability, and the generation of sufficient variation for long-term adaptation to dynamic niches (Lynch et al., 2016). This duality in evolutionary function is facilitated by DNA's reactive

composition and structure. However, DNAs reactivity also means that it is vulnerable to damage from endogenous and exogenous sources (Chatterjee and Walker, 2017). While natural selection acts to minimize mutation rates, DNA repair mechanisms are not infallible (Lynch et al., 2016). Ultimately, DNA mutations are inevitable and universal (Herms and Jones, 2023).

1.1 Genetic mosaicism

Recent work has conclusively shown that genetic variations arise inevitably in all studied tissues, steadily accumulating from the very first post-zygotic cell division until death (Coorens et al., 2021, Mitchell et al., 2022, Moore et al., 2021, Park et al., 2021, Ren et al., 2022). All PZMs are acted on by the forces of selection and genetic drift at the cellular level, which govern the clonal dynamics of the cells in which they arise (Loewe and Hill, 2010, Mustjoki and Young Neal, 2021). Most PZMs are benign or selectively neutral, with no observable impact on cell function or fitness. These low effect mutations are often maintained in successive cell generations through drift alone as passenger mutations (Loewe and Hill, 2010, Mustjoki and Young Neal, 2021, Mitchell et al., 2022). Deleterious mutations in the germline are often lost at the population level through purifying selection; however, the effect of purifying selection is much weaker in somatic cells, and deleterious PZMs are much less likely to be purged if their fitness effect is low (Yadav et al., 2016). The last outcome of a new PZM is where it is selectively advantageous to the cell, resulting in clonal expansions proportional to its fitness effect (Mitchell et al., 2022, Mustjoki and Young Neal, 2021).

Thus, newly arising unrepaired PZMs in somatic progenitor cells are relatively likely to be maintained across successive cell divisions with very few being lost. The result is that multicellular organisms are not composed of cells with identical genomes but are instead made up of a mosaic of heterogenous cells with very slightly different genomes (Biesecker and Spinner, 2013). This phenomenon is known as genetic mosaicism, formally defined as the presence of two or more genetically different cell clones which are derived from a single zygote (Biesecker and Spinner, 2013). The term has its roots in clinical genetics, where it most often describes clinically relevant PZMs which have been detected in significant proportions of cells (Martínez-Glez et al., 2020). However, the definition also technically encompasses PZMs across a spectrum of impact or phenotype (Martínez-Glez et al., 2020). Thus, for the purpose of this thesis, GM will be

used to describe all PZMs which originate in a single cell and are propagated to detectable levels in all tissues (Figure 1.1).

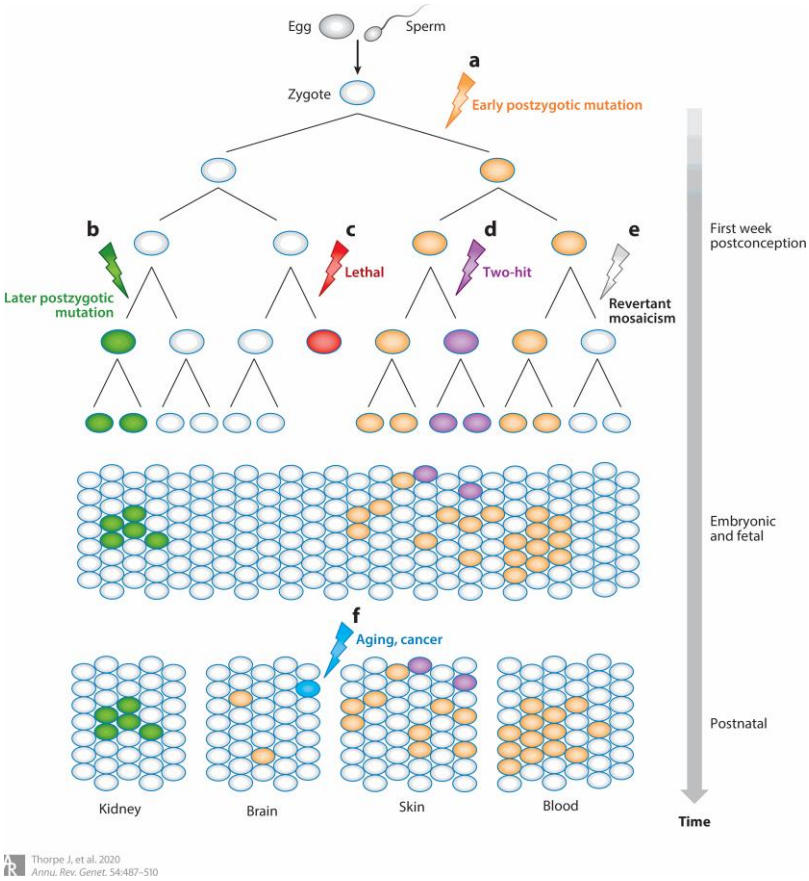


Figure 1.1: A simple representation of genetic mosaicism in human development from embryogenesis to adulthood.

Lightning bolts represent postzygotic mutation (PZM) events. Individual ovals represent cells. Colours represent cells resulting from different GM scenarios. a) An early PZM is maintained through successive cell divisions and is present across multiple adult tissues at relatively high cell fractions. b) A PZM arising later in development is maintained through cell divisions is present at a lower cell fraction and is restricted to a single adult tissue. c) a significantly deleterious PZM arising later in development results in cell death and is not propagated further. d) an additional independent second-hit PZM arises in a cell lineage already impacted by a previous PZM leading to pleotropic modulation of its phenotypic expression. e) an additional independent PZM arises in a cell lineage already impacted by a mutation which changes the genotype back to the wild type in somatic genetic rescue. f) new PZMs continue to accrue in adult tissues due to endogenous and exogenous genotoxic exposure and age which are subject to Darwinian selection and can result in clonal expansions in a confined tissue environment. Image from (Thorpe et al., 2020) with the permission of Annual Reviews, Inc., conveyed through Copyright Clearance Center, Inc.

The earliest conceptualisation of GM in humans likely dates to 1914, when Theodor Boveri speculated that cancer is a cell-based disease resulting from somatic alterations to chromosomes (Boveri, 1914). Supporting evidence for this hypothesis emerged later in subsequent studies (Knudson, 1971, Nowell, 1976). Beyond the broad scope of cancer, human “mosaics” were identified in cases of sex chromosome aneuploidies as early as the 1960s (Ford et al., 1959, Hirschhorn et al., 1960). It is now a well characterised contributor across the spectrum of disease pathologies (Alissa and Walsh, 2018, Happle, 1993, Martincorena and Campbell, 2015). Importantly, GM is not limited to a disease context and is also recognized as an inevitable age-related phenotype (Vijg, 2014). The first links between aging and PZMs were hypothesised in the late 1950s (Szilard, 1959, Failla, 1958). A direct association between aging and the accumulation of PZMs has also been consistently reported in normal non-malignant tissues in multiple investigations (Coorens et al., 2021, Moore et al., 2021, Park et al., 2021, Loh et al., 2020, Mitchell et al., 2022, Terao et al., 2020, Watson et al., 2020).

The breadth in GM research scope has led to variability in the terminology used to describe similar phenomena. A six-attribute classification of GM was proposed to standardize the nomenclature of PZMs classification, aiming to provide a unified framework that can guide future research and clinical interpretations (Figure 1.2) (Martínez-Glez et al., 2020). The proposed classifications were centred on the specific factors influencing the phenotypic impact of PZMs in the context of reported clinically manifesting GM. These factors include the type of affected tissues, the risk of heritability, the distribution pattern of the PZM in affected tissues, the direction of pathogenicity, the developmental mechanism of the PZM related phenotype or disease, the aetiology or class of PZM, and the proportion of tissue affected by the PZM (Martínez-Glez et al., 2020). While the specific classifications themselves still relate to clinical contexts, they are nevertheless useful in understanding the potential impact of a PZM across the board.

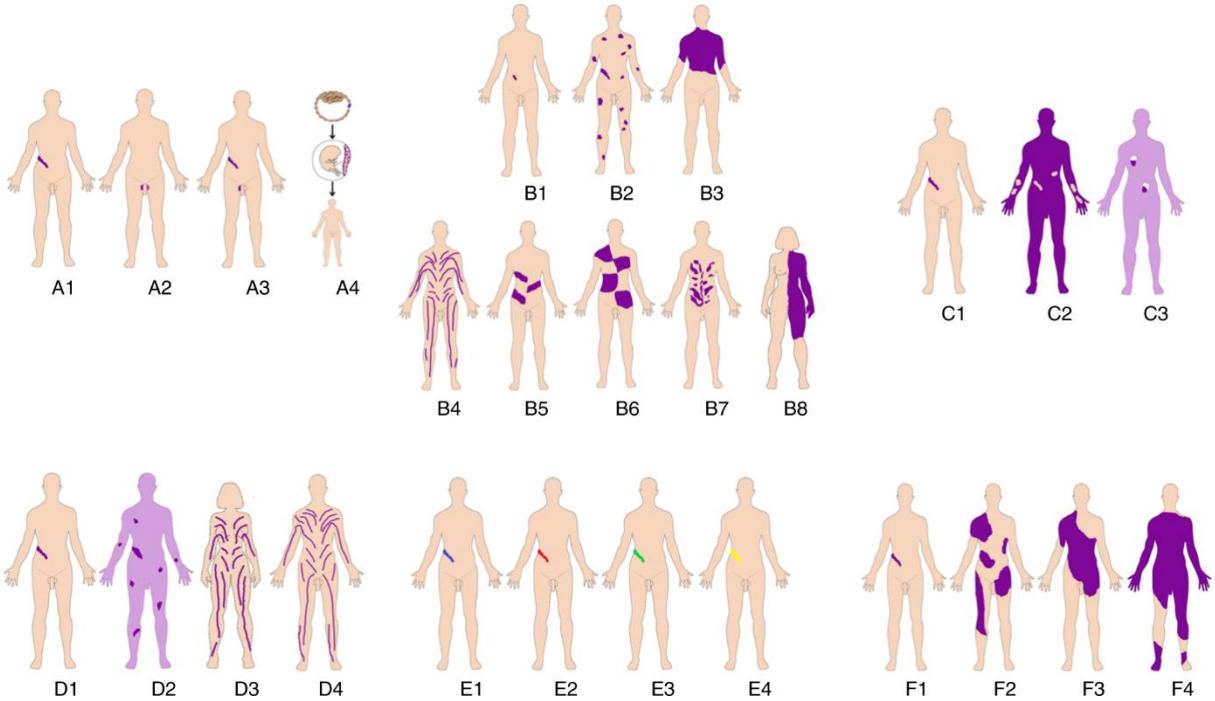


Figure 1.2: A visualisation of a newly proposed system of classifying clinacally relevant genetic mosaicism in humans.

A new system consisting of six genetic mosaicism attributes has been proposed to standardise nomenclature in the clinical setting. These classifications are useful in understanding the contexts in which GM arises and what its impact may be in humans. Broadly defined genetic mosaicism can be defined by its (A) affected tissues at the germline, somatic, gonadosomatic, or confined placental. These broad tissues classes can be further delineated by the (B) body pattern of affected tissue, as segmental, single point, and segmental. The impact of the mutation is dependent on (C) the direction of the change in realtion to the wild type genome, as wither benign to pathogenic, pathogenic to benign or as revertant mosaicism. Clincal impact can be further defined by (D) its development mechanism, which is linked to the kinds of disorders observed in genetic mosaics. Heterozygous dominant, a second-hit to an existing mutation, functional X-chromosome linked in for males, and mutations that can only exist as mosaics which are lethal otherwise. The (E) eitiology of GM is by mutations of all sizes and classes from small varaints to anueplodies, as well as epigenetic changes. The severity of GM is linked to (F) the proportion of cells and tissues in which it is found. Image from (Martínez-Glez *et al.*, 2020) with permission from Springer Nature.

Post-zygotic mutations arise across all stages of development, as well as in mature cell lineages in adulthood (Martínez-Glez et al., 2020). The developmental timing of PZM origin significantly impacts which tissues are subsequently affected. There are four broad classes of GM which are entirely dependent on the tissue types in which a PZM of interest is detected: somatic mosaicism, gonadal mosaicism, gonadosomatic mosaicism, and confined placental mosaicism (Figure 1.2) (Martínez-Glez et al., 2020). PZMs that arise very early in embryogenesis can expand into a significant proportion

of the resulting daughter cells and may be detected in multiple tissues (Figure 1.1). PZMs that present in only somatic cells are classified as somatic mosaicism, whereas PZMs that are present in both somatic and germ cells are known as gonadosomatic mosaicism (Martínez-Glez et al., 2020). PZMs can also emerge later in postnatal development or adulthood as *de novo* clonal events that are confined to a single tissue. *De novo* PZMs originating in germ cells are classified as gonadal or germline mosaicism (Martínez-Glez et al., 2020). *De novo* PZMs arising postnatally in a developed somatic tissue is classed as somatic mosaicism. This also encompasses clonal expansions such as PZM-driven clonal haematopoiesis (CH) (Jaiswal and Ebert, 2019). Confined placental mosaicism occurs when a PZM arises in the mother's placenta that is not present in the developing foetus (Toutain et al., 2018). Only germline and gonadosomatic mosaicism can be passed on to future generations (Martínez-Glez et al., 2020). These are inherited as constitutional germline variation in the proband. It is important to note that *de novo* germline variation arises as PZMs in parental germ cells, either as a singular event, or as appreciable gonadal/gonadosomatic mosaicism (Yu et al., 2024).

The timing of PZM origin also impacts the body pattern or distribution in affected tissues (Figure 1.1) (Figure 1.2). This body pattern classification reflects the patterns and extent to which the GM phenotype can be seen across affected tissues (Martínez-Glez et al., 2020). There are two main categories of body pattern classifications; nonsegmental and segmental patterns (Martínez-Glez et al., 2020). Nonsegmental patterns are more common and include single point mosaicism where a PZM arises in a single tissue location postnatally (Martínez-Glez et al., 2020). Single point mosaicism encompasses single tissue tumours as well as clonal expansions in confined tissue structures such as colon crypts. Nonsegmental patterns also include disseminated mosaicism, where PZMs impact multiple tissues such as those dealt with in Chapter 2 (Martínez-Glez et al., 2020). Segmental patterns are most often used to describe mosaicism that is directly observable on the skin, such as Blaschko lines in broad or narrow bands (Martínez-Glez et al., 2020). Other known segmental distributions include checkerboard, block-like, phylloid, and lateralization patterns which are often indicative of an underlying mosaic disorder (Martínez-Glez et al., 2020).

Most instances of GM involve a PZM that changes a wild-type allele to a functionally impactful or pathogenic variant (Martínez-Glez et al., 2020). In these cases, the directionality of the PZM is classed as benign to pathogenic. However, the reverse

scenario has also been observed, where the directionality of the PZM is from pathogenic to benign (Martínez-Glez et al., 2020). This is known as somatic genetic rescue or revertant mosaicism, where the deleterious impact of a pathogenic germline mutation is partially or completely offset by the introduction of a PZM (Revy et al., 2019). Somatic genetic rescue can be direct reversion of the pathogenic PZMs to the wild type, or indirectly through PZMs arising in linked genes in a relevant pathway (Revy et al., 2019). While the reversion may improve selective advantage at the cell level, it does not guarantee clinical improvement in a patient with a Mendelian disorder (Revy et al., 2019).

The third classification of GM centres on its developmental mechanism (Martínez-Glez et al., 2020). These include type 1 segmental mosaicism, which occurs through dominant nonlethal PZM (Martínez-Glez et al., 2020). Type 2 segmental mosaicism occurs in individuals with constitutional heterozygous pathogenic variants that are modulated by a newly arising PZM in *trans* (Martínez-Glez et al., 2020). In autosomal recessive conditions, the constitutional variant does not cause the phenotype on its own, and a newly arising second-hit PZM results in the aberrant phenotype in parts of the tissue in which it is present. In autosomal dominant conditions the second-hit PZM results in a more severe phenotype in the tissue in which it is present (Martínez-Glez et al., 2020). Random X-chromosome inactivation in female embryos can manifest as functional X-chromosome mosaicism with or without X-linked disorder traits that are often lethal in males (Martínez-Glez et al., 2020). Obligatory mosaic disorders result from PZMs that are lethal if constitutional (Martínez-Glez et al., 2020).

The final two classifications are related to the fraction or proportion of cells in which the PZM is detected, and the PZM class. The fraction of cells harbouring the PZM is often linearly correlated with the severity of its phenotypic expression (Martínez-Glez et al., 2020). This ranges from mild involvement in which less than 10% of the tissue is impacted, to very severe involvement in which more than 50% of the tissue is impacted (Martínez-Glez et al., 2020). The aetiology of GM classification is directly related to the class of the variant, which can differ in size and mechanism of formation from single-nucleotide polymorphisms (SNPs) to aneuploidies (Chatterjee and Walker, 2017, Coorens et al., 2021, Martínez-Glez et al., 2020).

1.2 Mechanisms of PZM formation

DNA is under near constant attrition by endogenous and exogenous sources of damage (Figure 1.3) (Chatterjee and Walker, 2017). Additionally, mitotic cell division is a significant source of mutation due to the complexity of DNA replication and errors in chromosome segregation (Levine and Holland, 2018, Mustjoki and Young Neal, 2021). Several mechanisms are in place to reduce the impact of DNA damage, including DNA damage response (DDR) pathways, DNA damage tolerance pathways, mitotic cell-cycle checkpoints, and apoptosis pathways (Chatterjee and Walker, 2017). DDR pathways are diverse and include base excision repair (BER), nucleotide excision repair (NER), mismatch repair (MMR), repairing of double stranded breaks through homologous recombination (HR), non-homologous end joining (NHEJ), microhomology-mediated end-joining (MMEJ) (Burssed et al., 2022, Chatterjee and Walker, 2017).

Many DDR pathways are substrate or lesion specific, with each having specialised roles in repairing distinct DNA lesions (Figure 1.3). There can be overlap, depending on several tightly regulated cellular processes. While they are usually efficient at repairing most lesions, they are not infallible, and erroneous DDR can lead to the formation of PZMs (Burssed et al., 2022, Chatterjee and Walker, 2017). PZMs can be maintained and propagated through DNA replication and cell division when DDR mechanisms or cell cycle checkpoints are inactive, when additional mutations dysregulate DDR proteins, or when the mutation load overwhelms DDR capacity (Tubbs and Nussenzweig, 2017).

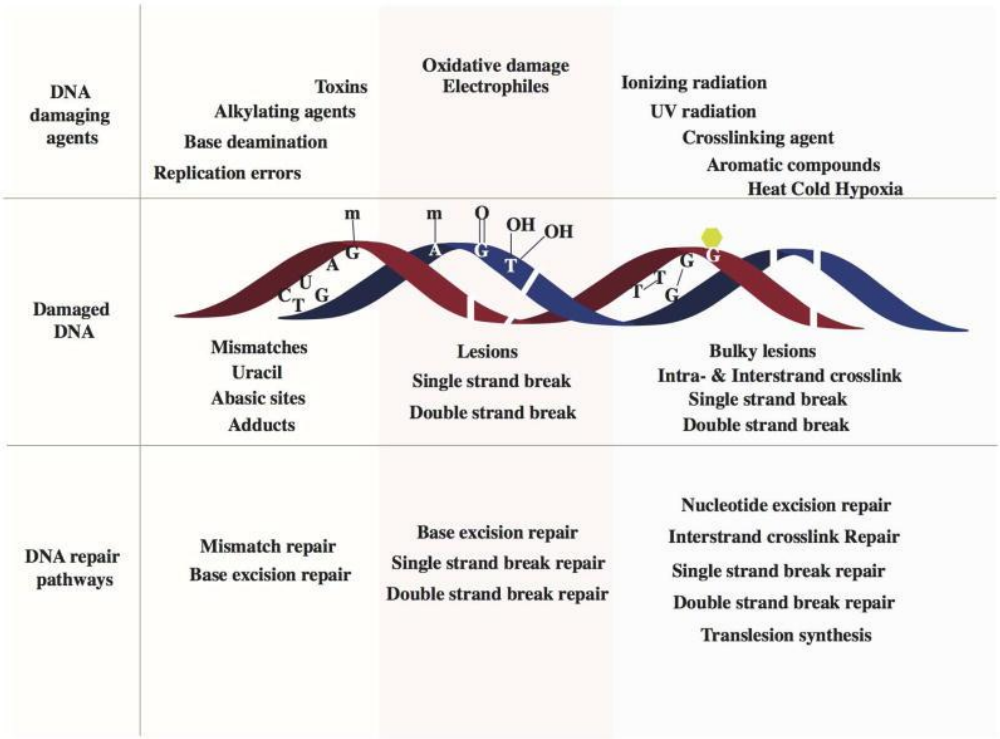


Figure 1.3: Common genotoxic agents produce characteristic DNA lesions which are repaired by diverse DNA damage repair pathways.

Image from Chatterjee and Walker (2017) with permission of John Wiley and Sons through Copyright Clearance Center, Inc.

Genetic variations of all shapes and sizes have been detected as PZMs, including epigenetic variations, SNPs and small insertions or deletions (INDELs), retrotransposons, as well as larger copy number variations, structural variations and aneuploidies (Coorens et al., 2021, Loh et al., 2020, Mitchell et al., 2022, Park et al., 2021, Terao et al., 2020, Duffy et al., 2021, Macia et al., 2011, Nam et al., 2023). Each class of PZM can be formed through distinct mutational mechanisms involving DNA damage and DDR, (Chatterjee and Walker, 2017, Nesta et al., 2021, Tubbs and Nussenzweig, 2017). The following is a summary of the mechanisms behind the formation of mosaic small variations (SNPs and INDELs) and MCAs, the two classes of PZMs dealt with in this thesis (Figure 1.4).

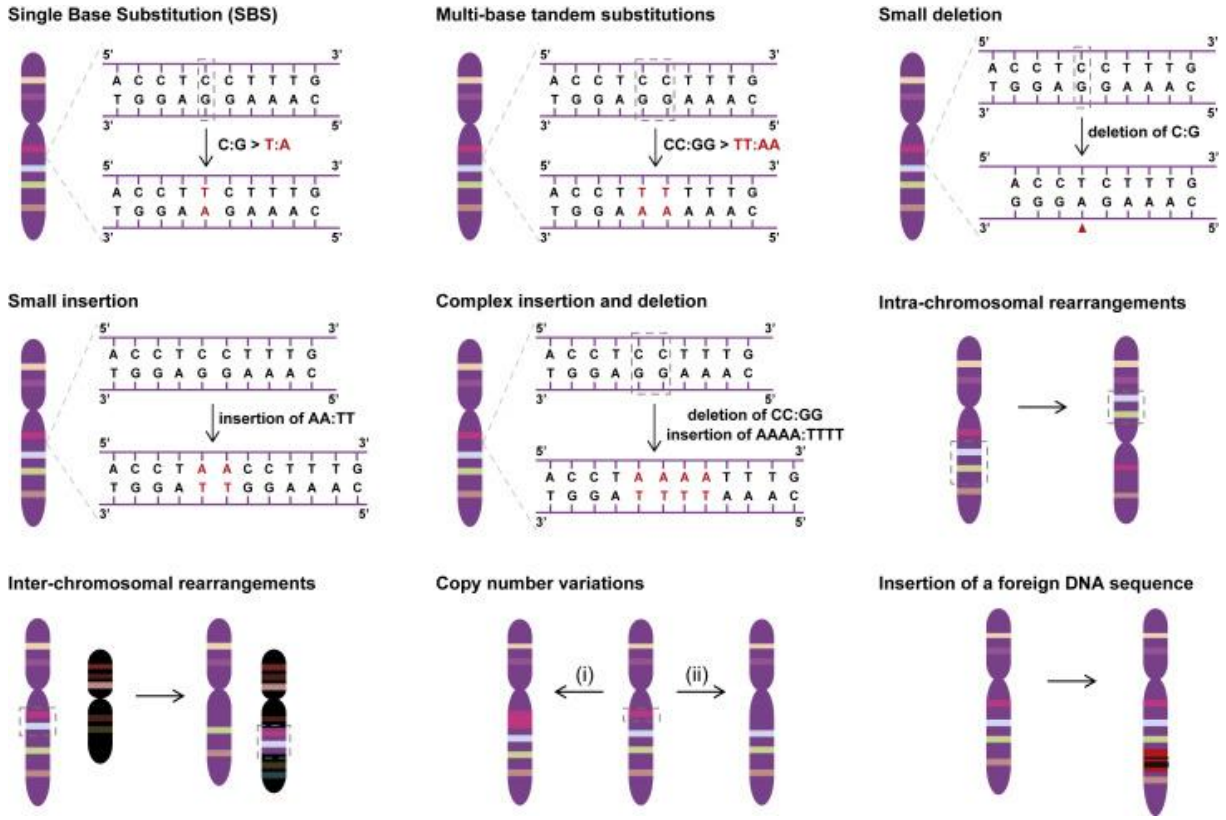


Figure 1.4: Illustratation of the small variations and large structural variations observed as somatic mutations.

Image from (Alexandrov and Zhivagui, 2019) with permission from Elsevier.

1.2.1 Mosaic SNPs and INDELs

The most observed PZMs are small variants such as SNPs and INDELs (Figure 1.4) (Alexandrov et al., 2013). Post-zygotic SNPs and INDELs are primarily formed through endogenous DNA damage and subsequent erroneous DDR pathways. However, DNA replication is a major endogenous source of both SNPs and INDELs. DNA replication is facilitated by a suite of DNA polymerases with varying degrees of fidelity and 3′-5′ exonuclease proofreading capabilities (Kunkel, 2009). Despite this inbuilt error correction mechanism, base mismatch SNPs may be introduced when an incorrect nucleotide is incorporated (Figure 1.5) (Kunkel, 2009). DNA polymerase replication slippage at repetitive sequences can lead to INDEL formation (Kunkel, 2009, Viguera et al., 2001, Montgomery et al., 2013). Microsatellites are abundant in the human genome and are known hotspots for INDEL formation (Montgomery et al., 2013, Nesta et al., 2021). It has been estimated that as much as 70% of INDELs are concentrated

in regions prone to replication slippage (Montgomery et al., 2013). SNPs and INDELs that escape the MMR pathways may be maintained as PZMs in future replications (Chatterjee and Walker, 2017, Kunkel, 2009). Additionally, individuals with deleterious variations in DNA polymerase or MMR genes have been shown to accumulate PZMs at greater rates (Lee et al., 2022, Robinson et al., 2021).

Spontaneous base deamination is a significant source of endogenous damage and occurs frequently in resting cells (Chatterjee and Walker, 2017). During this process Cytosine (C), adenine (A), guanine (G), and 5-methyl cytosine (5mC) lose an exocyclic amine, changing to uracil, hypoxanthine, xanthine and thymine (T), respectively (Lindahl, 1993). The most common deamination occurs at C and 5mC bases, leading to C>T or CpG>TpG substitutions (Figure 1.5) (Alexandrov and Stratton, 2014). Exposure to exogenous sources of DNA damage including ultraviolet (UV) radiation and intercalating agents can induce further deamination (Moyer et al., 1993, Pfeifer et al., 2005). CpG islands are hotpots for PZMs through deamination of 5mC (Tomso and Bell, 2003). However, it has been shown that SNPs in CpG islands are ~7 times less frequent than SNPs in other CpG dinucleotide motifs (Tomso and Bell, 2003).

Break excision repair is employed to repair deamination through the action of DNA glycosylases and DNA polymerase β (Dianov and Hübscher, 2013). In normal BER, DNA glycosylase removes the deaminated base, producing an abasic site which is cleaved along with 1 to 13 bp to creates a single-stranded DNA break (SSB), allowing access for DNA polymerase β to resynthesis the strand (Figure 1.5) (Dianov and Hübscher, 2013). DNA polymerase β lacks exonuclease proofreading, leading to an increased risk of incorrect base incorporation (Chatterjee and Walker, 2017). Abasic sites are unstable and prone to SSB conversion (Dianov and Hübscher, 2013). When deamination occurs within a repetitive sequence, the risk of DNA polymerase β slippage and subsequent INDEL formation increases. Additionally, unrepaired SSBs can progress to double-strand breaks (DSBs) through replication fork collapse, potentially resulting in further mutation through error-prone NHEJ repair (Chatterjee and Walker, 2017, Dianov and Hübscher, 2013, Caldecott, 2024).

Reactive oxygen species (ROS), a highly reactive group of molecules derived from molecular oxygen, are a significant source of DNA damage (Chatterjee and Walker, 2017). ROS are formed endogenously by several metabolic enzymatic reduction-oxidation reactions, including the electron transport chain (Sies and Jones, 2020). ROS are also introduced exogenously through ionizing radiation, UV radiation, certain

chemotherapeutics, alcohol consumption, air pollution, and other environmental exposures (de Almeida et al., 2022). Excess ROS react with DNA, damaging nucleotide bases, and inducing SSBs and DSBs (Figure 1.5) (Tudek et al., 2010). The formation of 8-oxoguanine (8-oxoG), which incorrectly pairs with A instead of C during DNA replication leads to a G>T substitution in subsequent rounds of replication (Figure 1.5) (Cheng et al., 1992, Tudek et al., 2010). Repair of DSBs by NEHJ is error-prone, often resulting in small INDELs at the site of the repaired breakpoints (Burssed et al., 2022). Other potent exogenous sources of DNA damage include UV radiation, alkylating agents, aromatic amines, polycyclic aromatic hydrocarbons, microorganism toxins, and environmental factors such as heat, cold, and oxidative stress (Figure 1.4) (Chatterjee and Walker, 2017).

Newly arising SNPs and INDELs can alter protein function and expression by disrupting coding sequences, splicing, mRNA stability, or regulatory elements. Their effects range from subtle changes in translation efficiency to frameshifts and premature stop codons that truncate proteins or abolish their function (Ramírez-Bello and Jiménez-Morales, 2017, Lin et al., 2017).

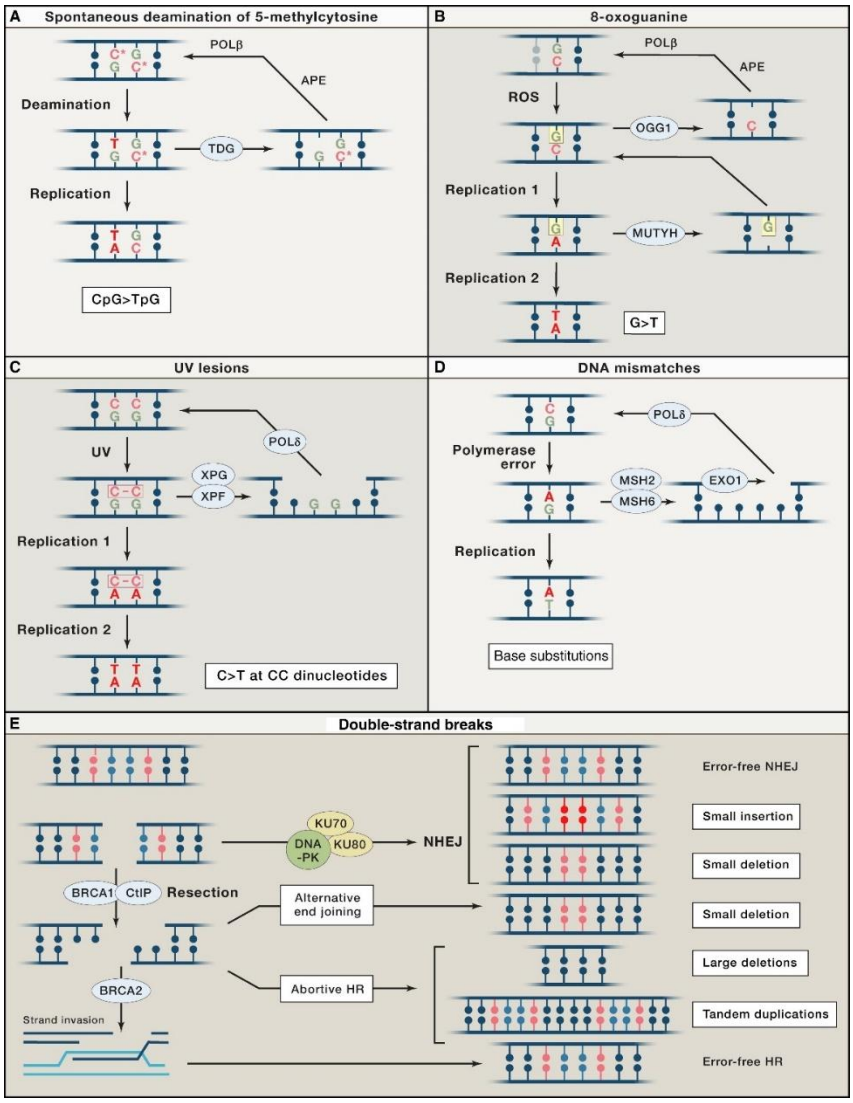


Figure 1.5: DNA damage and erroneous repair pathways resulting in mutations.

A) Spontaneous deamination of 5-methylcytosine (C∗) generates an aging signature, characterized by CpG>TpG transition mutations. 5-methylcytosine deamination converts C>T, which is recognized by thymine DNA glycosylase (TDG). Apurinic/apyrimidinic endodeoxyribonuclease (APE) excises the abasic nucleotide, which is replaced by polymerase beta (POLβ). (B) Reactive oxygen species (ROS) oxidizes guanine to form 8-oxoG. 8-oxoG can pair with adenine and may generate G>T mutations if 8-oxoG:C pairing is not detected by OGG1 before replication or if the 8-oxoG:A pairing is not recognized by MUTYH after the first round of replication.(C) Exposure to UV light generates cyclobutane pyrimidine dimers, which are recognized by nucleotide excision repair (NER) factors and excised by XPG and XPF flap endonucleases. Polymerase delta (POLδ) synthesizes DNA to restore C:G pairing. Failure to complete NER results in C>T transition mutations at CC dinucleotides. (D) DNA polymerase errors induce mismatch repair (MMR) through MSH2/MSH6. Exonuclease I (EXO1) removes a short patch of DNA that is resynthesized by POLδ. Failure to complete MMR results in base substitutions. (E) DNA DSBs are repaired using non-homologous end joining (NHEJ) or homologous recombination (HR). Normal NHEJ may result in small insertions or deletions, as does an inherently error-prone form of end-joining (alternative end joining), which relies on short homologous sequences (microhomologies) for base pairing. Normal HR is an error-free repair mechanism, but abortive HR may yield chromosomal rearrangements, such as tandem duplications or large deletions, which are common in HR-deficient cancers. BRCA1 and CtIP promote DNA end resection during HR, whereas BRCA2 promotes strand invasion to facilitate normal HR. Reprinted from (Tubbs and Nussenzweig, 2017), with permission from Elsevier.

1.2.2 Mosaic chromosomal alterations

Mosaic chromosomal alterations (MCAs) are postzygotic copy-number variants, structural variants and aneuploidies that arise within a subset of cells. They include interstitial deletions and duplications, whole-arm deletions/duplications, mosaic aneuploidies and polyploidies, copy-neutral loss of heterozygosity (CN-LOH), and more complex rearrangements such as inversions and translocations (Figure 1.6). typically span ≥50 bp and frequently extend to kilobase (Kb) or megabase (Mb) scale regions, giving them much larger genomic footprints than SNPs or INDELs (Liu et al., 2021). Importantly, MCAs are an umbrella category. Thus, these larger mosaic variants are often grouped together in analyses, but they arise by distinct mechanisms and have different functional consequences (Dubois et al., 2022, Liu et al., 2021). Sub-chromosomal MCAs are characterized by physical alterations of DNA sequence between two breakpoints, whereas aneuploidies affect whole chromosomes or large chromosome arms (Figure 1.6) (Dubois et al., 2022, Liu et al., 2021).

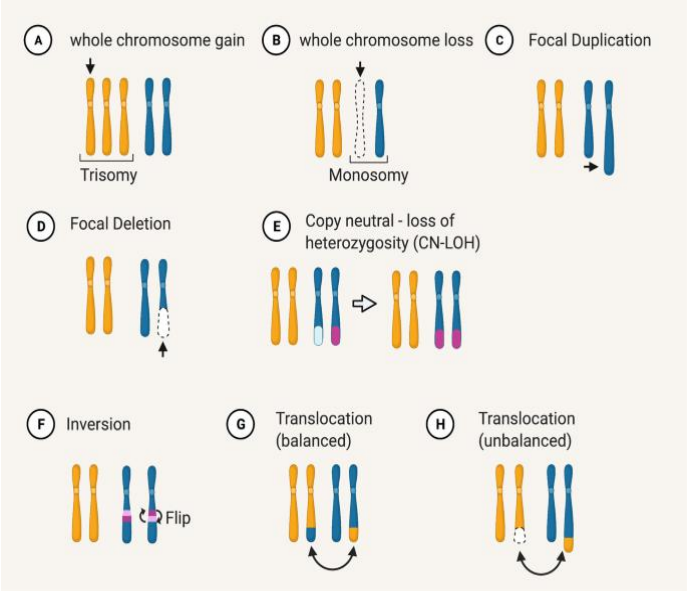


Figure 1.6: Illustrative representation of diverse mosaic chromosomal alterations.

(A-E) MCAs with copy number changes. (F-H) MCAs of abnormal chromosomal size or morphology. Reproduced from (Liu et al., 2021) with permission from Nature Springer.

Mosaic aneuploidies arise from mitotic segregation errors, most commonly due to anaphase lagging, nondisjunction, endoreplication, cell fusion, telomere dysfunction,

or cytokinesis failure (Andriani et al., 2017, Campbell et al., 2015). Mosaic aneuploidies arise through anaphase lagging, non-disjunction, endoreplication, cell fusion, telomere dysfunction, merotelic attachments and cytokinesis errors during mitosis (Andriani et al., 2017, Campbell et al., 2015). These errors typically reflect broader defects in genome maintenance and are further exacerbated by mutations in genes regulating chromosome condensation, cohesion, kinetochore assembly, spindle dynamics, cytokinesis, and spindle assembly checkpoint signalling (Holland and Cleveland, 2012). Mosaic aneuploidy likelihood is further modulated by interplay between sequence and chromatin level determinants. These include centromere and kinetochore composition, cohesin function and telomere integrity, higher-order 3D genome organization, integrity of mitotic cell-cycle machinery, and environmental stresses (Devillers et al., 2024, Baker et al., 2024).

Sub-chromosomal MCAs form predominantly through errors of DSB repair (Jackson and Bartek, 2009). These repair mechanisms differ and can lead to diverse MCA outcomes (Figure 1.7) (Jackson and Bartek, 2009). The primary repair pathway involved is NHEJ, which directly ligates DNA ends without needing extensive homology and is available throughout the cell cycle. However, NHEJ is prone to errors. Alternative repair mechanisms such as MMEJ operate via short homologous overhanging sequences, often deleting DNA sequence. Both pathways are known to erroneously join distant ends when multiple DSBs are present, leading to interstitial deletions, duplications, and translocations (Figure 1.7) (Pös et al., 2021, Cosenza et al., 2022). Double-stranded DNA breaks repaired by erroneous non-allelic homologous recombination (HR) can form recurrent MCAs in the form of inversions, deletions, and translocations, which are most often seen in regions of the genome with high sequence similarity such as segmental duplications (Pös et al., 2021, Cosenza et al., 2022).

Replication stress, which occurs when DNA replication forks stall, induces single-stranded DNA breaks that are repaired by break induced replication or template-switching mechanisms, produces more complex structural rearrangements, nonreciprocal translocations and tandem duplications (Cosenza et al., 2022). Telomere attrition can lead to the deletion or fusion of the terminal ends of chromosomes which form variety of dicentric-chromosome variations, isochromosomes, ring chromosomes, and initiate breakage–fusion–bridge cycles (Cosenza et al., 2022). More complex MCAs can form in single catastrophic events such as chromothripsis, chromoanasynthesis, and chromoplexy (Figure 1.7). These MCAs involve dozens of

breakpoints in highly complex patterns, far beyond simple end-joining repair models (Cosenza et al., 2022).

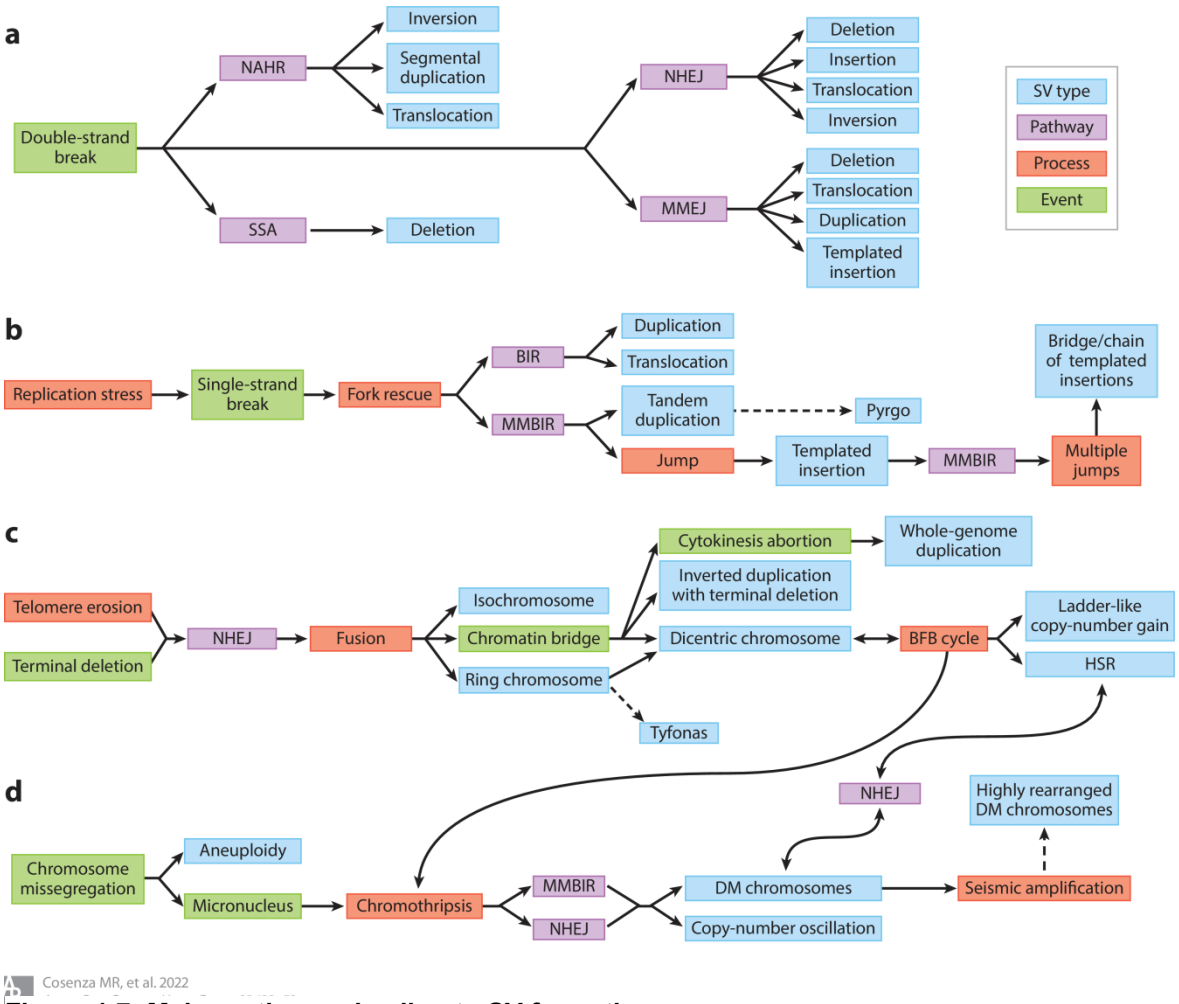


Figure 1.7: Major pathways leading to SV formation.

A) erroneous repair of DSBs by four main DNA repair pathways leads to simple SVs. B) Replication stress induced SSBs are erroneously repaired by BIR leading to characteristic MCAs of intermediate complexity. C) Telomere erosion and terminal duplications lead to chromosome end fusions resulting in dicentric-chromosome SVs, which can be remodelled by BFB cycles into complex SVs. D) Mitotic chromosome segregation errors result in aneuploidy or micronuclei prone to chromothripsis. BFB, breakage–fusion–bridge; BIR, break-induced replication; DM, double minute; HSR, homogeneously staining region; MMBIR, microhomology-mediated break-induced replication; MMEJ, microhomology-mediated end joining; NAHR, nonallelic homologous recombination; NHEJ, nonhomologous end joining; SSA, single-strand annealing; SV, structural variant. Image adapted from (Cosenza et al., 2022).

A commonly observed MCA is copy-neutral loss of heterozygosity. In diploid cells replace one parental allele with the other without a net change in copy number, leaving

the affected region homozygous. CN-LOH may encompass entire chromosomes (Uniparental disomy) but more commonly affects terminal stretches of chromosome arms. Intrachromosomal and interstitial CN-LOH segments are rarer because they require additional recombination events or gene conversions (O'Keefe et al., 2010). Mechanistically, uniparental disomy arises through meiosis error followed by trisomy rescue, whereas segmental or terminal CN-LOH usually arises from erroneous mitotic HR or BIR which copy the homolog from the breakpoint to the telomere or breakpoint to breakpoint (gene conversion) (Dutta and Schacherer, 2025). CN-LOH has important functional consequences as it converts heterozygous loci to homozygosity, thereby unmasking recessive alleles, amplifying the effect of pathogenic or driver SNPs and INDELs, and altering imprinted epigenetic states (Dutta and Schacherer, 2025).

Local chromatin and sequence context, replication timing, 3D architecture and repetitive elements such as segmental duplications or fragile sites bias where DSB occur and which DDR pathway predominates, thereby modulating the likelihood of MCA formation (Dubois et al., 2022, Baker et al., 2024). Thus, MCA formation is not entirely stochastic. MCAs influence cell function primarily by altering gene dosages and disrupting 3D genome folding, which together change patterns of gene expression (Dubois et al., 2022, Sreenivasan et al., 2025). These effects can occur in *tran*s, by disrupting chromatin regulators or topologically associating domains (TADs) (Sreenivasan et al., 2025). However, they most often act directly by disruption CRE interactions, altering the regulatory control of target genes (Dubois et al., 2022).

Importantly, the functional impact of MCAs depends on the fraction of cells in which they occur, as not all cells within a tissue harbour them. Constitutional structural variations and aneuploidies that would be incompatible with viability can be tolerated when restricted to a mosaic fraction, with their consequences further dependent on tissue context and developmental timing (Martínez-Glez et al., 2020). This cellular heterogeneity allows otherwise deleterious events to persist as mosaics, where they disrupt tissue function or confer a selective advantage (e.g. malignancy) proportionate to cell fractions. The same alterations in the germline are typically uniformly pathogenic or embryonically lethal (Mustjoki and Young Neal, 2021). This is exemplified most extremely by aneuploidies, where germline autosomal monosomies are universally non-viable, yet they can occur and persist as mosaic events, often with severe functional consequences (Toral-Lopez et al., 2012).

1.2.3 DNA damage response pathways are efficient

The rates at which PZMs form across the genome differs between SNPs, INDELs and MCAs, is dynamic across development and age, and are tissue-specific (Ren et al., 2022). At the cellular level, SNPs have been consistently observed about 10 times more frequently than INDELs across tissues, whereas MCAs are comparatively rare (Ren et al., 2022). Humans have evolved efficient DDR pathways and cell cycle checkpoints to ensure that most DNA lesions are repaired, and that chromosomes are successfully segregated into daughter cells during mitosis (Matthews et al., 2022). These pathways survey the genome throughout the cell cycle and can delay progression when damage is detected, allowing repair to occur (Matthews et al., 2022). The DDR and checkpoint pathways are interconnected through signalling networks, are tightly regulated, and involve numerous essential genes, which are extensively reviewed elsewhere (Matthews et al., 2022, Tubbs and Nussenzweig, 2017).

Cells and tissues are exposed to a diverse array of exogenous genotoxins and, more importantly, to near-constant endogenous DNA damage from reactive byproducts of essential metabolic reactions (De Bont and van Larebeke, 2004, Swenberg et al., 2011). Estimates suggest that each cell accrues around 10,000 - 100,000 DNA lesions per day (De Bont and van Larebeke, 2004, Swenberg et al., 2011). The most frequent lesions, aside from polymerase errors, are abasic sites and SSBs, while other common damage includes deamination, alkylation, and oxidation (De Bont and van Larebeke, 2004, Swenberg et al., 2011). Of course, the relative burden of specific lesions varies by cell type and tissue environment. The high probability of lesion exposure necessitates efficient DDR pathways that coordinate with cell cycle checkpoints, apoptosis, senescence, as well as other important cellular processes such as transcription and DNA replication (Vijg, 2021). Indeed, most DNA lesions are quickly repaired upon detection by DDR pathways in the span of a few minutes to hours (Hamilton et al., 2001, Swenberg et al., 2011). Under normal physiological conditions, DDR pathways are highly efficient in recognising and repairing most DNA lesions (Vijg, 2021, De Bont and van Larebeke, 2004, Swenberg et al., 2011).

Not all DNA lesions are equivalent in their frequency, repair kinetics, or potential to disrupt cellular homeostasis. This variability is thought to be linked to the relative burden of DNA lesion exposure and/or toxicity (Spencer Chapman et al., 2025). Base excision repair rapidly resolves single lesions within minutes, though more complex lesions may require several hours (Dianov and Hübscher, 2013). Mismatch repair has

the highest fidelity, with repair rates depending on replication status (Hsieh and Zhang, 2017). High fidelity NER rates are dependent on chromatin state, with transcription coupled repair taking several hours and global repair taking up to two days (Adar et al., 2016).

Double-stranded breaks are rarer, occurring 10–50 times per cell per day, but are more challenging to repair (Tubbs and Nussenzweig, 2017). NHEJ resolves most DSBs within an hour, whereas HR is more accurate but may take several hours due to its dependence on template homology (Cho et al., 2014, Scully et al., 1999, Okayasu and Iliakis, 1994, Löbrich and Jeggo, 2017, Scully et al., 2019). Despite the impressive efficiency of the DDR response pathways, many DNA lesions and errors of replication go unrepaired, evading the surveillance of the cell cycle checkpoints and becoming fixed as permanent PZMs in subsequent cell divisions.

1.2.4 DNA damage response pathways are imperfect

Despite the efficiency of DDR pathways and cell cycle checkpoints, PZMs are ubiquitous across all tissues and species studied to date (Cagan et al., 2022, Moore et al., 2021). These findings have been consistently observed in both malignant and non-malignant tissues, from population-scale biobank studies to single-cell analyses (Abascal et al., 2021, Cagan et al., 2022, Jakubek et al., 2023b, Li et al., 2021, Loh et al., 2020, Moore et al., 2021, Martincorena et al., 2018, Rodin et al., 2021, Coorens et al., 2024, Lee-Six et al., 2019, Ng et al., 2021, Aaltonen et al., 2020). This raises a key question. Given their potential to disrupt cell function and fitness, why are somatic mutation rates not zero?

There are several interconnected hypotheses that plausibly explain this phenomenon. Darwinian selection acts on genetic variation to drive evolutionary adaptation. At the cellular level, this translates to genetic variation promoting cellular heterogeneity (Goldman et al., 2019). Slight variability in the morphology or function of cells from the same tissue source is essential for rapid adaption to changing environmental dynamics in cellular niches (Li et al., 2016). Cellular heterogeneity is a multifaceted phenomenon that results from complex interactions between environmental, pathological, stochastic, and genetic factors. Genetic factors including PZMs (Goldman et al., 2019). While there is a risk of PZM driven cellular dysregulation, PZMs also contribute to tissue plasticity and improved fitness.

Overall, mutagenesis is a stochastic process and DNA damage is induced randomly across the genome. It is estimated that only ~1-2% of the human genome is protein coding, and that most small variants are functionally benign (Hatje et al., 2019, Martincorena et al., 2017). It follows that the most frequently observed PZMs are selectively neutral, having no impact on cellular function. DNA repair is costly, requiring time, cellular resources, and energy (Vijg, 2021). If a cell were to repair all the thousands of lesions it was exposed to a day, this cost may be too high leading to depletion of resources and cell death (Sniegowski et al., 2000). Additionally, cells would constantly be pausing at DNA damage cell cycle checkpoints to facilitate repair which would interfere with replication and transcription programmes (Vijg, 2021). Cells tightly regulate and limit DDR pathways through several mechanisms, including preferentially limiting physical access to DNA lesions across the organisational structures of the 3D genome. This is supported by mathematical modelling which shows that DDR limiting stem cells are more fit (Madden-Hennessey et al., 2022). Functionally normal somatic cells direct specific DDR pathways to the more functionally important and constrained regions of the genome, where they efficiently repair the most damaging DNA lesions most of the time (Spencer Chapman et al., 2025). This strategy of cell survival by tolerating PZM allows the cell to regulate DDR cost and provides a substrate for the selection of cellular heterogeneity (Roux et al., 2021).

Lastly, there is the practical consideration that DDR pathways, cell cycle checkpoints, DNA replication, and mitosis are biological processes that are imperfect. While they are tightly regulated and efficient, they are not error free. Given the large numbers at play in development and maintaining homeostasis over the average human lifetime, it should not be surprising that PZMs occur even stochastically.

1.2.5 Mutational signatures

Each mutational process leaves behind characteristic patterns of base substitutions, small indels, or MCAs known as mutational signatures (Figure 1.8) (Alexandrov et al., 2020a). Distinct signatures have been linked to both endogenous and exogenous sources of DNA damage, as well as to specific DDR errors (Alexandrov et al., 2020a, Li et al., 2020, Coorens et al., 2021, Moore et al., 2021, Park et al., 2021, Li et al., 2021). Importantly, distinct tissues harbour diverse profiles of mutational signatures, which also show considerable interindividual variability among donors. This diversity

reflects the wide range of mutational mechanisms and exposures shaping PZM formation across the human body (Coorens et al., 2021, Li et al., 2021, Moore et al., 2021).

These characteristic signatures have been used to uncover sources of PZM accumulation, furthering our understanding of the scale and impact of PZMs in normal tissues and elucidating the aetiologies of cancers (Figure 1.8) (Alexandrov et al., 2020a, Li et al., 2020). In addition to the finding that PZM rates differ across normal tissues, studies have found that certain genomic loci are more susceptible to PZM formation than others (Supek and Lehner, 2019). Recurrent PZMs are more frequently seen in common loci often reflecting known drivers in cancer associated genes, while recurrent MCAs often share common breakpoints in separate samples (Loh et al., 2020). Mutational signature analysis has shown that not only do distinct mutational processes differ in burden between individuals, but they also accumulate at different rates (Chen et al., 2025).

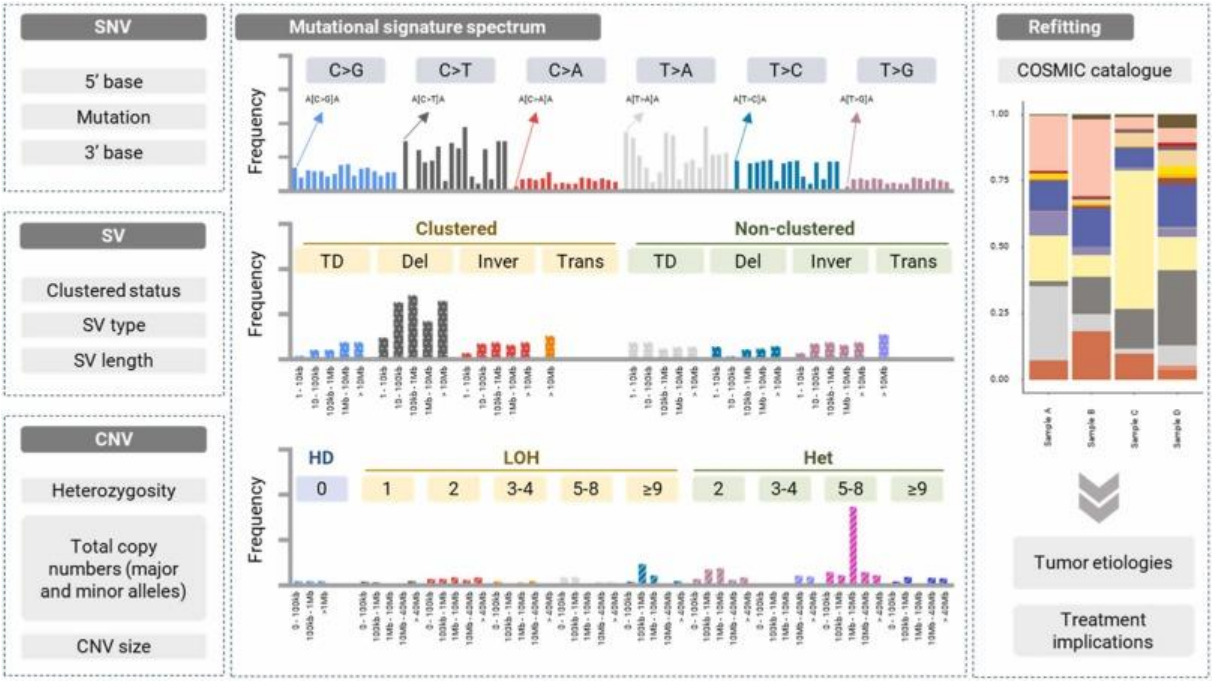


Figure 1.8: A summary representation of mutational signature features of somatic mutations.

SNV signatures consider trinucleotide combinations with the mutation in the centre. SV signatures consider SV types, the distance between two breakpoints, and the clustering status. CNVs signatures consider heterozygosity, copy number, and size. The COSMIC catalogue uses fitting to determine combinations of signatures that identify specific cancer aetiologies. Image adapted from (Ma et al., 2024) under CC by 4.0.

1.3 Biological factors impacting observable PZM frequencies in GM

The frequency at which distinct PZMs are observed in an analysis is the product of biological factors and platform specific technical factors impacting detection sensitivity and resolution. Mechanistic biases influence mutation rates, and selection factors impact the likelihood of being maintained in the cell population (Figure 1.9) (Watson et al., 2020, Watson and Blundell, 2023). PZMs with high mutation rates form more frequently, while those with low mutation rates form less frequently as a function of time (Figure 1.9). Those with large negative fitness effects may be purged and go undetected. Those with neutral fitness effects may be maintained as passengers through the effects of genetic drift, being detected by chance (Baker et al., 2024). Those with positive fitness effects may expand to large clone sizes and are more likely to be detected (Figure 1.9) (Baker et al., 2024). Thus, the frequency of PZM detection at any given loci is determined by how often it mutates, and its relative effect on cell function impacting its odds of being maintained over successive divisions (Figure 1.9).

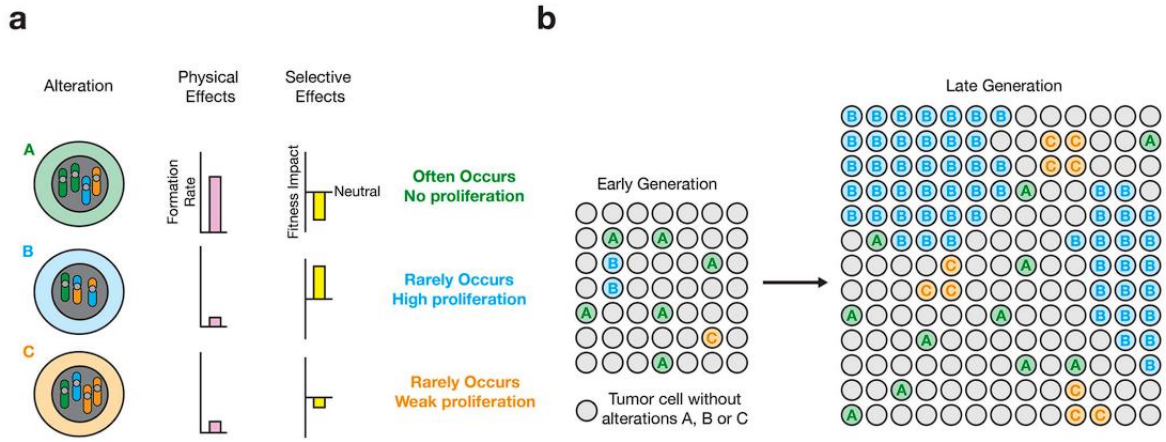


Figure 1.9: Hypothetical impact of mechanistic factors and selective factors on observed PZM frequencies in tumour cells.

A) Three hypothetical PZM with differing mutation rates and fitness effects. B) The potential impact of each PZM on observed frequencies is illustrated. PZM A with a high mutation rate will arise frequently in early individual tumour cells but will not expand to large clones with overall negative fitness effects. PZMs B and C have similar mutation rates, but B has a higher positive fitness effect and will expand to a higher proportion of the late tumour clones than C. Image adapted from (Baker et al., 2024)with permission from Nature Springer.

Importantly, while there is some overlap, the biological factors impacting mutation rates and fitness effects may differ for SNPs, INDELs, and the various MCA classes (Watson et al., 2020, Watson and Blundell, 2023). Each has its own mechanism of formation and functional consequence. They also differ across tissues and may change over time, resulting in the differences in mutation rates and distributions observed across tissues with age (Yu et al., 2024, Ren et al., 2022). These biological factors have mostly been characterised in cancer samples and have only recently started to be investigated in 'normal' tissues in depth (Coorens et al., 2024, Yu et al., 2024). While cancers are essentially large single cell derived clonal expansions, and Darwinian selection models most accurately fit cancerous clonal dynamics, we now known that most clonal expansions of normal cells do not progress to malignancy (Evans and Walsh, 2022). The physiological background in which the mechanistic and selection effects occur in cancers versus normal cells is not identical. However, similarities in mutation signatures, mutation hotspot prediction, and driver mutations between matched tumour and normal tissue samples have been reported (Alexandrov et al., 2020b, Coorens et al., 2021, Moore et al., 2021). Thus, cancer data still useful in understanding PZMs and clonal dynamics in normal tissues. Ultimately, it is important to understand that DNA mutation is not an entirely stochastic process, with non-uniform probability across the genome. The biological factors impacting PZM rates and selection effects are discussed briefly here.

1.3.1 Mechanistic factors impacting PZM rates

Variable mutation rates are the product of predisposition to DNA damage and/or differential access to DNA repair (Supek and Lehner, 2019). The mechanistic factors that impact PZM likelihood at any locus are physical genome features, such as local sequence composition, epigenetic marker makeup, organisation in the 3D genome, and the cell cycle stages (Supek and Lehner, 2019, Seplyarskiy and Sunyaev, 2021). The physical genome features scale from trinucleotide composition to chromosomal territory position, with differential effects on the probability of SNP, INDEL, or MCA formation (Supek and Lehner, 2019).

The 3D organisation of the genome is linked to mitosis and the replication timing (RT) programme, the regulation of transcription and DNA repair efficiency, and the facilitation of interaction between linearly distant genomic loci (Dubois et al., 2022).

DNA is a physical molecule that must be packaged efficiently into nanometre-sized nuclei and organised in such way to allow for gene expression regulation. This is achieved by ordered, non-random, and self-organising iterative folding and compaction of DNA at multiple levels across each chromosome into a fractal globule 3D conformation (Misteli, 2020). The role of 3D genome architecture is particularly impactful in MCA formation and selection, as MCAs span large genomic regions and are more likely to disrupt 3D conformation (Dubois et al., 2022). The 3D genome organisation is expertly reviewed in detail elsewhere (Misteli, 2020). Together, local sequence level, 3D genome level, and cell cycle level features shape non-random PZM rates across the genome by biasing loci to DNA damage and DDR access (Supek and Lehner, 2019).

Somatic SNPs and INDELs are strongly shaped by local sequence context (Seplyarskiy and Sunyaev, 2021). Regions rich in GC-content, particularly methylated CpG dinucleotides, are enriched for C>T transitions, while INDELs frequently occur within repetitive DNA such as microsatellites, homopolymer runs, and short tandem repeats that are prone to polymerase slippage during replication (Tubbs and Nussenzweig, 2017). Chromatin accessibility, replication timing, and 3D genome organization, influence both the susceptibility of a locus to DNA damage and the accessibility of DDR pathways (Seplyarskiy and Sunyaev, 2021, Tubbs and Nussenzweig, 2017). Early replicating euchromatin, characterized by high transcriptional activity and open chromatin, is efficiently surveyed by MMR and NER due to increased accessibility and shows lower rates of SNPs and INDELs (Tubbs and Nussenzweig, 2017). In contrast, late-replicating heterochromatin is enriched for SNPs and INDELs, particularly G>T deamination events (Tubbs and Nussenzweig, 2017).

Mosaic chromosomal alterations commonly form at DSB hotspots and are significantly impacted by 3D genome features and RT (Dubois et al., 2022). DSBs are enriched in repetitive motifs found at fragile sites, near the centromeres, and at the telomeres. DSBs also cluster around TAD anchors such as cohesin sites (Lezaja and Altmeyer, 2021). Replication and transcriptional stress further promote DSB enrichment, with topoisomerase II activity during transcription maintenance providing a major source of DSBs in euchromatic regions (Dubois et al., 2022). Unlike small variants, MCAs are paradoxically enriched in early-replicating and actively transcribed regions despite the presence of efficient NHEJ and HR activity (Tubbs and Nussenzweig, 2017). Thus, the enrichment of MCAs in these regions may be due to increased abundance of DSBs.

Larger aneuploidies are impacted by chromosomal features such as centromere integrity, telomere maintenance, spindle assembly, and 3D nuclear organization, which influence mis-segregation risk (Devillers et al., 2024, Baker et al., 2024). Thus, PZM rates are not random are a product of local DNA sequence context, 3D genome architecture, replication timing, and DDR accessibility, with each class of PZM shaped by distinct yet overlapping mechanistic biases.

1.3.2 Factors impacting PZM fitness and clonal dynamics

Most adult tissue homeostasis is maintained through asymmetric stem cell/progenitor cell divisions within spatially defined niches (Acha-Sagredo et al., 2022). According to Darwinian biology, heterogenous cell clones compete for space and resources within these niches. Fitter clones are more likely to divide, contributing proportionally more progeny to the niche. Accumulated PZMs alter individual cell fitness through functional impact (Martincorena et al., 2017). The direction and strength of the fitness effect dictate the clonal dynamics of the host cell (Watson et al., 2020, Watson and Blundell, 2023). Positively selected PZMs confer a competitive advantage through enhanced proliferation, survival, altered differentiation, or self-renewal, driving clonal expansions (Martincorena et al., 2017). Although Darwinian selection traditionally models' malignant clonal evolution, it is increasingly recognized as a suitable framework for non-malignant clonal dynamics, where the mechanisms of fitness advantage often mirror those that drive malignancy (Watson et al., 2020, Watson and Blundell, 2023). Many classic malignancy driver genes have also been associated with clonal expansions in normal tissues (Kapadia and Goodell, 2024). However, most clonal expansions occur in the absence of known malignancy driver mutations (Kapadia and Goodell, 2024). It is likely that selective advantage is conferred through other means, modulated by the niche in a tissue-specific and temporally dynamic manner. Aging, exogenous exposure, and niche remodelling continuously shape clonal adaptations to PZM fitness effects (Herms and Jones, 2023).

The clonal dynamics within distinct niches are shaped by a complex interplay of intrinsic PZM-driven fitness adaptations and extrinsic niche factors that change over time. Therefore, PZM fitness effects are very rarely uniform across the lifespan. PZMs which originate in early embryogenesis, are not subject to the selection pressures as in adult tissues due to the requirements of development (Kapadia and Goodell, 2024,

Park et al., 2021). These PZMs, if non-lethal, are fixed by neutral drift across differentiating tissues and acted upon by selection as niche conditions evolve. Fitness effects of shared drivers change with age and differ across tissues, reflecting shifting niche conditions (Kapadia and Goodell, 2024). Exposures to chemicals, radiation, inflammation, and infection further remodel niches, altering the fitness landscape within the host environment (Herms and Jones, 2023). Finally, most fit clones are constrained from perpetual expansion by clonal competition, physical niche architecture barriers, and immune surveillance (Herms and Jones, 2023). Expansions in spatially restricted tissues such as the colon or liver are constrained by physical barriers inherent to the tissue structure. Additionally, clones expand until encountering neighbours within a shared niche that are equally as fit (Herms and Jones, 2023).

Thus, the observed frequency of PZMs in tissues reflects both their mutation rate and fitness effects, as clones continuously compete and adapt within dynamic niches to maintain tissue homeostasis over the human lifespan. The final factor shaping observed PZM frequencies is not biological but technical, reflecting the limitations of current genotyping platforms.

1.4 Detecting PZMs from bulk genotyping data in single- sample normal tissues

Accurately detecting PZMs in single-sample, non-malignant tissues is technically challenging (Dou et al., 2018). Unlike drivers of malignancies, where large clonal populations allow confident identification of PZMs against a matched germline reference, PZMs in normal tissues are usually confined to small fractions of cells and can be difficult to distinguish from sequencing errors or genotyping artefacts (Dou et al., 2018). Detection is further complicated by biological context, particularly the timing of mutation origin and its tissue distribution (Martínez-Glez et al., 2020). Early embryonic PZMs are often shared across multiple developed tissues, confounding the choice of suitable controls for matched-sample analysis, whereas later-arising variants may be tissue-specific and would remain undetected if the relevant tissue is not sampled (Dou et al., 2018, Martínez-Glez et al., 2020).

Thus, detecting PZMs in normal tissues demands careful study design and specialised analytical tools beyond those used in matched-tumour comparisons (Ha et al., 2023). Considerations must include mechanisms of PZM formation, cohort demographics and

disease context, as well as the genotyping platform choice. Because GM encompasses diverse forms of PZM, no single strategy has been able to account for all GM contexts. The history of PZM detection reflects this. Over time, these challenges have progressively been addressed through the development of increasingly sophisticated detection strategies, evolving alongside advances in genotyping technologies (Ha et al., 2023, Dou et al., 2018). From early cytogenetic approaches capable of identifying only large aneuploidies to modern strategies that sensitively resolve low fraction SNPs at high resolution, technological breakthroughs have been forthcoming. Currently, multiple experimental platforms with accompanying analytical tools are available for detecting PZMs, each differing in sensitivity, resolution, and applicability (Table 1.1). Broadly, these can be grouped into cytogenetics, SNP-array based genotyping, bulk sequencing, and single-cell sequencing approaches.

Table 1.1: A selection of bioiformatic tools for detecting GM from SNP-array and NGS data.

Tool	Data	Method	Reference
MCAs			
MAD-GADA	SNP-array	GADA segmentation on BAF deviation	(González et al., 2011)
MONTAGE	SNP-array	Allelic imbalance in sliding windows	(Glessner et al., 2021)
MoChA	SNP-array/WGS	Vertibi HMM captures phased BAF deviations	(Loh et al., 2020)
MrMosaic	WES	WES implementation of MAD-GADA algorithm	(King et al., 2017)
HaHMMR	RNA-seq	Haplotype aware HMM in Numbat	(Gao et al., 2023)
Small variants			
GATK-HC	WES/WGS	Haplotype-based cardiant caller with modified ploidy setting	(Van der Auwera et al., 2013)
Mutect2	WES/WGS	Allele frequency analysis	(Benjamin et al., 2019)
MosaicHunter	WES/WGS	Bayesian genotyping	(Huang et al., 2017)
MosaicForecast	WES/WGS	Random forecast CNN	(Dou et al., 2020)
DeepMosaic	WES/WGS	Image-based CNN	(Yang et al., 2023)
EM-Mosaic	WES/WGS	Expectation-maximization algorithm	(Hsieh et al., 2020)
RePlow	Ultra-deep TGS	Joint analysis of replicates	(Kim et al., 2019)

GADA - Genome Alteration Detection Analysis; BAF – B-allele fraction; WGS – Whole Genome Sequencing; WES – Whole Exome Sequencing; RNA-seq – RNA sequencing; HMM – Hidden Markov Model; MAD – Mosaic Alteration Detector; CNN – Convolutional Neural Network; TGS – Targeted Genome Sequencing

This brief review will focus on approaches most relevant to this thesis. That being the detection of MCAs from SNP-array data and small variants from conventional bulk sequencing (Table 1.1). Expert review of single-cell methods can be found elsewhere (Huang and Lee, 2022, Valecha and Posada, 2022).

1.4.1 Detecting MCAs from bulk genotyping data in non-matched single samples

Detecting MCAs is particularly challenging, as they are typically rarer than small variants and often present at low cellular fractions (Liu et al., 2021). Historically, detection was limited to large aneuploidies using cytogenetic techniques. G-banded karyotyping, the long-standing standard for identifying chromosomal alterations, can detect mosaicism but is constrained to events larger than ~5 Mb (Steventon-Jones et al., 2022). Karyotyping examines metaphase cells and is therefore prone to sampling bias. It also is prone underestimation of low-level mosaicism, requiring 30 cell counts to detect GM at 10%, and 500 cell counts to detect GM at 1%, which is a significant time investment (Steventon-Jones et al., 2022). Fluorescence *in situ* hybridization (FISH) improved both resolution (~50-100 kb), sensitivity, and turnaround time through targeted probe-based analysis (Steventon-Jones et al., 2022). FISH requires 100 cells to detect GM at 5% fractions, but it requires prior knowledge of the locus and cannot provide genome-wide coverage or screening (Vijg et al., 2023, Steventon-Jones et al., 2022). Moreover, *in vitro* culturing can introduce spurious alterations that resemble true MCAs, and both methods are labour intensive, require extensive training, and are unsuitable for large scale characterizations (Steventon-Jones et al., 2022).

The advent of SNP-arrays represented a major advance for detecting both germline CNVs and MCAs, as they provide genome-wide coverage by measuring fluorescent signal intensities across hundreds of thousands to millions of SNP probes. Two key intensity metrics are generated for each probe, the $\log_2$R ratio (LRR) and the B-allele frequency (BAF) (Steventon-Jones et al., 2022). The LRR is a normalized ratio of observed signal intensity to expected signal intensity for each SNP probe and is a measure of total allelic intensity (Steventon-Jones et al., 2022). The BAF is the normalized ratio of the frequency of the signal intensity of the less-common B allele for a SNP with two alleles named A and B and is a measure of relative allelic intensity. Homozygous probes have BAFs centered around 0 and 1, while heterozygous probe

BAFs center around 0.5 (BAF ratios: AA=0, AB=0.5, BB=1) (Steventon-Jones et al., 2022). Although SNP-arrays were initially designed for SNP genotyping in GWAS, they have since been adapted for CNV and MCA detection by computationally detecting deviations in LRR and BAF (Steventon-Jones et al., 2022, Balagué-Dobón et al., 2022).

Chromosomal alterations disrupt allelic balance, producing characteristic mean BAF deviations at heterozygous probes within breakpoints while leaving surrounding regions unchanged (Steventon-Jones et al., 2022). Germline duplications, deletions, and CN-LOHs cause distinct BAF and LRR deviation patterns (Laurie et al., 2012). Deletions decrease in LRR, duplications increase it, and CN-LOH alters BAF without affecting LRR. In contrast, MCAs affect only a fraction of cells, resulting in intermediate deviations in both BAF and LRR, proportional to the impacted cell fraction (Laurie et al., 2012).

Several computational methods have been developed to detect BAF and LRR deviations with segmentation algorithms or Hidden Markov Models (HMM) (Balagué-Dobón et al., 2022). Popular segmentation algorithms include MAD-GADA and MONTAGE (Table 1.1) (González et al., 2011, Glessner et al., 2021). Currently, the gold standard tool for MCA detection from SNP-array data is the Mosaic Chromosomal Alteration Caller (MoChA) (Loh et al., 2018, Loh et al., 2020).

The MoChA package can sensitively and accurately call MCAs from SNP array and high-coverage WGS data at estimated cell fractions as low as 1% (Loh et al., 2018, Loh et al., 2020). A key innovation of MoChA is its incorporation of long-range haplotype phase information, which substantially improves sensitivity. Statistical phasing enhances mosaic variant detection by exploiting the fact that true PZMs arise on a single haplotype, producing the expected three haplotypes when co-observed with a germline heterozygote (Loh et al., 2018, Loh et al., 2020, Dou et al., 2018). In contrast, sequencing or genotyping artifacts often appear on both alleles, generating two or four haplotypes, which can be confidently rejected (Dou et al., 2018). MoChA applies this principle at a chromosomal scale, using phased SNP haplotypes within a HMM to distinguish genuine allelic imbalance from spurious noise. True MCA events alter the balance of parental genomic material in a cell population, causing systematic deviations in phased BAF values that the HMM captures (Loh et al., 2018, Loh et al., 2020). Importantly, MoChA intentionally ignores global LRR variation during initial detection, as it is highly sensitive to technical artifacts, instead relying on allelic

imbalance as the primary signal (Loh et al., 2018, Loh et al., 2020). LRR is only incorporated later to classify detected events. This approach improves on unphased BAF segmentation by effectively increasing the number of informative markers, as long-range haplotype phasing aggregates subtle deviations across a parental haplotype, thereby amplifying the signal-to-noise ratio. This allows MoChA to sensitively detect low cell-fraction MCAs and large CN-LOHs that would be difficult to distinguish from noise using segmentation of unphased BAF values alone (Loh et al., 2018, Loh et al., 2020).

Estimating sensitivity and specificity of MCA-calling tools remains challenging. MoChA reports a false discovery rate (FDR) between 4.5–8.6%, although these estimates lack orthogonal experimental validation (Loh et al., 2018, Loh et al., 2020). Independent benchmarking by the MONTAGE authors reported a sensitivity of 0.975 and false positive rate (FPR) of 0.344 for MONTAGE, 0.920 sensitivity and 0.598 FPR for MoChA, and 0.280 sensitivity with 0.627 FPR for MAD-GADA (Glessner et al., 2021). These comparisons should be interpreted cautiously, as comprehensive, standardized benchmarking of MCA detection tools is still limited.

However, MoChA has consistently demonstrated the ability to detect MCAs at higher resolution and lower cell fractions than earlier tools such as MAD-GADA (Dou et al., 2020, Eberhardt et al., 2023, Machiela et al., 2015). Its performance, though, is influenced by probe density, local heterozygosity rates, and the accuracy of phasing Probe density, in particular, is a well-recognized determinant of SNP-array resolution and sensitivity, with higher-density arrays and WGS consistently outperforming lower-density platforms (Steventon-Jones et al., 2022, Dou et al., 2020). Despite these limitations, SNP-arrays generally outperform cytogenetic approaches in both sensitivity and resolution. They also offer several practical advantages. SNP-arrays assess far more cells across different cell cycle stages, avoid artifacts introduced by cell culture, require less labour, provide genome-wide coverage in a single assay, are relatively cost-effective, and uniquely allow detection of CN-LOH events (Steventon-Jones et al., 2022).

Beyond SNP-array data, several methods have been developed to detect MCAs from WGS, WES, and RNA-seq (Table 1.1). MoChA can be applied to high-coverage (~30X) WGS data, detecting MCAs through allelic imbalance derived from informative reads

spanning heterozygous sites (Dou et al., 2020). Compared with SNP-arrays, WGS provides a far greater density of informative sites, enhancing sensitivity and breakpoint resolution (Jakubek et al., 2023b). MoChA cannot be applied to WES due to sparse and uneven probe distribution across the exome and its inability to model reference allele bias (Dou et al., 2020). For WES, MrMosaic was specifically developed to identify MCAs similarly to MAD-GADA, though its resolution and sensitivity remain limited due to uneven read coverage and low informative marker density (King et al., 2017). More recently, RNA-seq–based tools such as HaHMMR have been used to detect mosaic CNVs by leveraging allelic imbalance in expressed transcripts, but this approach is highly dependent on expression levels and remains less comprehensive than DNA-based methods (Gao et al., 2023). Overall, WGS provides the most powerful platform for MCA detection, with uniform coverage, high resolution, and superior sensitivity compared with SNP-arrays, but remains considerably more costly and computationally intensive. SNP-arrays therefore continue to be widely used, offering a cost-effective balance of resolution, sensitivity, and throughput, especially in large-scale studies.

1.4.2 Detecting mosaic small variants from bulk sequencing data in non-matched single samples

In bulk sequencing data, the read support for reference and variant alleles at any site follows binomial sampling, with the expected variant allele fraction (VAF) depending on zygosity and total sequencing depth (Huang and Lee, 2022). The VAF is an indication of the proportion of cells in which the variant allele is present. Thus, constitutional heterozygotes in diploid cells have expected VAFs of 0.5, while homozygotes approach 1.0. By contrast, PZMs deviate from these distributions because they are present in only a subset of cells (Dou et al., 2018). Theoretical VAF expectations of PZMs depend on when it originated, fitness effect, and cell division (Dou et al., 2018). A PZM arising in the first cell division of embryogenesis would be present in 50% cell fractions, corresponding to a VAF of ~0.25 (Dou et al., 2018). The expected VAF of a PZM would, thus, halve with each subsequent symmetric division (Dou et al., 2018). This principle underpins strategies for parsing PZMs from the germline, as deviations from the expected binomial distribution can be formally tested using two-tailed binomial tests (Hsieh et al., 2020). However, embryogenesis is an asymmetrical process, PZMs are subject to tissue specific selective pressures and clonal dynamics, and technical

artifacts from sequencing strongly resemble PZMs with similar VAF distributions (Dou et al., 2018, Martincorena et al., 2017, Ju et al., 2017). Early arising PZMs are expected to have higher VAFs than newly acquired mutations, whereas most PZMs accumulate later in life and thus occur at very low VAFs (often less than 1%), making them difficult to distinguish from sequencing errors (Huang and Lee, 2022). These factors can also distort the observed VAFs of germline heterozygous variants away from the expected 0.5, leading to false classifications (Huang and Lee, 2022). Common artifacts from DNA contamination, DNA damage, misaligned reads, and platform-specific errors frequently confound variant calling (Dou et al., 2018). Thus, candidate PZMs are often enriched for false positives and require stringent filtering and careful review in the absence of matched tissue controls.

Recent advances have substantially enhanced the sensitivity of PZM detection in NGS by combining multiple complementary strategies to overcome these challenges. Sequencing depth is a critical determinant of PZM detection sensitivity, with increased depth substantially enhancing the ability to identify low-frequency variants, whereas insufficient depth can lead to under-detection and biased frequency estimates, ultimately impacting biological and clinical interpretations. Empirical studies indicate that ~30X sequencing is largely powered to detect VAFs>0.1, whereas deeper sequencing ~400X is required to reliably capture variants with VAFs in the 0.02–0.05 range (Bick et al., 2020b). Increasing read depth from ~100X to >200X has been shown to improve recall for VAF<0.1 by up to 44%, though precision declines slightly due to accumulation of sequencing artifacts (Chen et al., 2020). Whole-genome sequencing provides more uniform read coverage across the genome, whereas whole-exome sequencing is subject to capture biases that generate uneven read distributions, particularly at exon boundaries, which can compromise detection power uniformity and sensitivity (Ha et al., 2023). While increasing sequencing depth is a powerful strategy to enhance PZM detection sensitivity, it also comes at substantially higher costs proportional to genome coverage needs. Choice of platform should therefore be guided by study objectives. The chosen platform then dictates appropriate variant calling tools and quality control strategies.

A diverse array of computational tools is available for PZM detection across bulk sequencing data, each with distinct methodological approaches and performance characteristics that must be carefully considered (Table 1.1) (Ha et al., 2023). In addition to in-built caller-specific features, a variety of quality control measures can be

applied to further enrich variant calling for true PZMs (Ha et al., 2023). These include leveraging phasing information, filtering against a matched panel-of-normals (PON) to remove recurrent germline variants and platform-specific artifacts, and excluding common variants present in large population databases (Dou et al., 2018). Variants consistent with known platform-specific mutational signatures, as well as those located in regions which map poorly or known low complexity, can be excluded to significantly reduce false positives (Dou et al., 2018). Standard sequence-level feature filters, such as minimum total read depth, base quality, strand bias metrics, and minimum VAF thresholds, are recommended as well (Dou et al., 2018). Together, these strategies provide a multi-layered framework that complements variant calling algorithms and improves confidence mosaic small variant calling. While some of these filtering strategies are integrated into specific variant callers, others were developed as tool-specific features. Importantly, recent benchmarking studies demonstrate that combining the most effective quality control features across multiple callers substantially improves both sensitivity and specificity, underscoring the value of integrative approaches (Ha et al., 2023).

Popular strategies employ somatic paired-sample or germline variant callers that have been adapted for PZM detection. Mutect2 in tumour-only mode uses haplotype-based Bayesian genotyping to sensitively detect somatic SNPs and INDELs (Benjamin et al., 2019). The GATK HaplotypeCaller (GATK-HC) can also be used with increased ploidy settings for improved sensitivity of low VAF SNPs and INDELs (Van der Auwera et al., 2013). Specialized PZM variant callers have also been developed. MosaicHunter utilizes Bayesian genotyping and beta-binomial modelling with stringent error filtering for low sensitivity and high specificity calling (Huang et al., 2017). MosaicForecast leverages machine learning on read-based phasing with over 30 read-level features to differentiate true PZMs from artifacts (Dou et al., 2020). DeepMosaic introduces a convolutional neural network-based framework that directly learns image-like representations of read pileups, enabling robust detection of mosaic SNPs (Yang et al., 2023). Recent benchmarking of the above tools demonstrated that no single PZM caller is best in all contexts. Tools vary substantially in sensitivity and specificity, with many showing reduced accuracy for INDELs relative to SNPs, and even the best-performing methods fail to detect very VAF<0.04 INDELs (Ha et al., 2023). Moreover, concordance of call sets between different algorithms is low, and overlap-based

consensus approaches are not recommended. Instead, best practice is to deploy complementary callers targeted to their respective strengths (Ha et al., 2023).

Beyond single-sample approaches, cohort-based trio whole-exome sequencing (trio-WES) can be uniquely leveraged to improving PZM calling (Wright et al., 2019). This design is most often implemented in clinical cohorts, where sequencing of affected probands alongside both parents allows stringent filtering of candidate variants (Wright et al., 2019). By excluding variants with any read support in either parent, as well as the expectation that true PZMs are very unlikely to be shared across the cohort, trio-WES substantially reduces cohort and platform specific false positives (Dou et al., 2018). Trio-WES can also reveal evidence of parental gonadosomatic mosaicism, which has direct implications for recurrence risk counselling (Wright et al., 2019). However, trio-WES alone does not guarantee complete accuracy, as its performance remains constrained by WES limitations such as sequencing depth and the uneven coverage intrinsic to exome capture (Dou et al., 2018). Consequently, additional statistical and filtering strategies are required to further increase confidence in PZM detection.

Observed VAFs in WES datasets display greater variance than expected under a simple binomial model, reflecting overdispersion introduced by exome enrichment in library preparation, capture efficiency, and study-specific artifacts (Hsieh et al., 2020, Huang and Lee, 2022). The VAF overdispersion complicates downstream estimation of PZM fractions and destabilizes correction. To address this challenge, EM-Mosaic was developed as a statistical framework that explicitly models VAF overdispersion within WES data by fitting VAFs to a study-specific beta-binomial model, applying an expectation–maximization (EM) algorithm to iteratively refine cohort-wide mosaic fraction estimates and assign posterior probabilities for individual variants being mosaic over germline (Hsieh et al., 2020).

First, EM-mosaic pools all candidate *de novo* SNPs and fits a beta-binomial model for the alternate read counts, estimating an overdispersion parameter from the data. It also estimates a global prior mosaic fraction from the cohort using VAF distributions under the beta-binomial model. Then in each EM iteration, EM-mosaic computes the likelihood of each variant's read counts under two models. A germline *de novo* model and a mosaic model. By combining these likelihoods with the prior, the algorithm computes a posterior odds (post-odds) of mosaicism for each site. Variants exceeding a chosen FDR threshold on the post-odds (>10) are reported as high-confidence

mosaic SNPs. This framework allows EM-Mosaic to explicitly account for cohort-level mosaicism rates, data-specific variability, and provides statistical evidence for mosaicism (Hsieh et al., 2020). The EM-Mosaic authors reported a confirmation rate of 88% and state that at depths of 60X the sensitivity for VAF<0.05 is low, but that it would be over 80% at 500X (Hsieh et al., 2020). A major limitation of EM-Mosaic is that it can only be applied to *de novo* SNPs and not INDELs (Hsieh et al., 2020). Furthermore, EM-Mosaic is not a variant caller itself and requires as input a set of SNPs that have already been confirmed as *de novo*. The method also does not incorporate additional filtering steps, necessitating that input data be rigorously quality-controlled prior to analysis (Hsieh et al., 2020). Thus, it is most appropriately used in conjunction with trio-WES in which PZM filtering is applied.

Even with careful modelling, independent validation by ultra-deep resequencing or digital droplet PCR is essential to confirm PZMs (Dou et al., 2018). At the very least read support must be visualised with tools such as Integrative Genomics Viewer (IGV) to reduce the odds of false positives. Even the most sophisticated PZM variant callers require high sequencing depth to power low-VAF PZM detections and rely on large cohorts to fit the models. Overall, detecting PZMs from bulk sequencing of single-sample tissues remains technically challenging, with no one-size-fits-all approach. Cost and computational demands further constrain routine application, as WGS and ultra-high-depth validation across multiple tissues are expensive. Emerging technologies are poised to substantially improve PZM detection sensitivity and accuracy (Yu et al., 2024). Collectively, these technological advances have made it possible to characterize PZMs across tissues at unprecedented scale, rapidly expanding our understanding of the landscape of genetic mosaicism in both healthy and diseased contexts.

1.5 Genetic mosaicism across the human lifespan

The true landscape of GM in normal tissues has only recently begun to emerge, driven by advances in tissue sampling techniques, the establishment of biobank-scale cohorts, and reductions in sequencing error rates (Acha-Sagredo et al., 2022, Ren et al., 2022). Unlike cancer genomes where mutations are clonally expanded and readily detectable, PZMs in normal tissues typically exist at single-cell resolution at low VAF, requiring specialized detection approaches with sequencing error rates below 10^{-07}

errors per base pair (Acha-Sagredo et al., 2022, Ren et al., 2022). High-fidelity single-cell sequencing, combined with laser capture microdissection and advanced computational algorithms, now allows detailed mutational analyses across diverse tissues and developmental stages. These approaches have revealed tissue-specific mutational processes, burdens, and phylogenetic trajectories that can be traced from early embryogenesis through to old age (Coorens et al., 2021, Moore et al., 2021, Abascal et al., 2021).

We now know that PZMs and clonal expansions accumulate across all tissues. Some mutational processes, such as the deamination of 5mC, are universal “clock-like” signatures, whereas others are tissue-specific and shaped by local exposures and DNA repair environments (Coorens et al., 2021, Moore et al., 2021, Abascal et al., 2021). Most PZMs are SNPs, with INDELs occurring ~10 fold less frequently, while MCAs are much rarer (Coorens et al., 2021, Moore et al., 2021, Abascal et al., 2021). Considerable inter-individual variation also exists in mutation rates across tissues, reflecting both intrinsic and environmental influences. The highest mutation rates are observed in the colon, liver, and oesophagus (~50 SNPs/cell/year), while sperm cells maintain the lowest rates (~3 SNPs/cell/year), consistent with strong evolutionary pressures to maintain genome integrity in heritable lineages (Coorens et al., 2021, Moore et al., 2021, Abascal et al., 2021). Overall, mutation load, signatures, and selective pressures differ markedly between embryogenesis, mature tissues, and aging tissues, making GM a temporally and spatially dynamic phenomenon with distinct consequences in early development and in aging contexts.

This section explores GM in two distinct spatio-temporal contexts that form the focus of subsequent chapters. First, pathogenic PZMs in embryogenesis that underlie developmental disease. Second, PZMs that drive clonal haematopoiesis as an aging phenotype.

1.5.1 Early genetic mosaicism: developmental disorders

The transition from a single-celled zygote to fully developed adult comprised of trillions of cells involves a cascade of rapid and tightly regulated cell divisions against a backdrop of transcriptional programs, differentiation, and continuous PZM accumulation (Hatton et al., 2023, Waldvogel et al., 2024). The PZM landscape of early embryogenesis exhibits distinct temporal patterns. The earliest cell divisions, occurring

between the 2-4 cell stage, experience elevated mutational rates of ~3 SNPs per cell division, reflecting suppressed expression of DDR pathways, cell cycle checkpoints optimized for rapid division, and delayed zygotic genome activation (Coorens et al., 2021, Munisha and Schimenti, 2021). The tolerance for DNA damage favours timely embryonic differentiation over genomic fidelity. Prior to genome activation, occurring between the 4-8 cell stage, embryos rely predominantly on maternally deposited mRNAs and proteins, operating with minimal transcriptional activity (Munisha and Schimenti, 2021).

Zygotic genome activation represents a critical shift from maternal to zygotic control of embryogenesis, with decreased rates ~1 SNPs per division endogenous DNA repair systems become fully operational (Coorens et al., 2021, Munisha and Schimenti, 2021). Chromosomal segregation errors are also unusually common during the first mitoses, exceed those observed in mature tissues, and decline from the 2-4 cell stage onwards (Currie et al., 2022). These errors result in MCAs that surprisingly remain compatible with live birth. The spatial control of first mitosis demonstrates vulnerability to maternal age effects, with declining fidelity contributing to increased mosaicism frequencies in pregnancies of older mothers (Gruhn et al., 2019).

The timing of a PZM is therefore critical in shaping its distribution and phenotypic impact. PZMs arising during pre-gastrulation (~ 10-15 cell divisions) are seen at higher VAFs across multiple organ systems and are more likely to involve loss-of-function (LOF) tolerant genes, whereas later-arising events are progressively restricted to single tissues or organs (Waldvogel et al., 2024). Early embryonic PZMs expand through stochastic drift and asymmetrical cell divisions and with observable VAFs that are later governed by complex interactions involving tissue-specific fitness effects and developmental bottlenecks (Waldvogel et al., 2024). While most embryonic PZMs are tolerated, occasionally a damaging variant in a gene essential for development can disrupt organogenesis or neurodevelopment, manifesting as a developmental disorder (Wright et al., 2019).

Developmental disorders (DDs) are a heterogeneous group of conditions characterized by impaired growth, cognition, behaviour, or physical development arising during prenatal or early postnatal periods (Solarsh and Hofman, 2006, Srour and Shevell, 2014, Holm, 1989). Clinical presentations of DDs range from isolated physical malformations to complex multi-system phenotypes involving transient global developmental delay (GDD), chronic intellectual disability (ID), congenital

abnormalities, and autism spectrum disorders (ASD) (American Psychiatric Association and American Psychiatric, 2013, Srour and Shevell, 2014). Affected individuals present with symptoms along a spectrum of severity and often share comorbidities with other DDs, suggesting common underlying molecular aetiologies (Hormozdiari et al., 2015, Parenti et al., 2020). A range of interacting environmental and genetic factors are involved in DD aetiology (Solarsh and Hofman, 2006). While early estimates attributed ~50% of DDs to genetic causes, more recent studies suggest the proportion may be closer to 80% (Srour and Shevell, 2014, Copeland et al., 2021).

The clinical complexity observed in DDs is driven by underlying genetic heterogeneity, pleiotropy, and gene-environment interactions (Parenti et al., 2020). Phenotypic outcome is modulated by gene specific vulnerability, mutation load, and epistatic interactions (Parenti et al., 2020). Many developmental genes are dosage-sensitive and highly intolerant to loss-of-function variants, explaining the prominent role of *de novo* mutations (DNMs) in DD aetiology and highlighting precise gene dosage requirement during embryogenesis (Wright et al., 2023). DD associated genes are often enriched for DNMs, and DD cohorts experience a greater burden of deleterious DNMs compared to unaffected controls (McRae et al., 2017, Wilfert et al., 2017). DNMs in monogenic DD associated genes had the highest diagnostic yield in the Deciphering Developmental Disorders UK (DDD-UK), with 76% of trios having an underlying pathogenic DNM (Wright et al., 2023). Even in apparently monogenic cases, phenotypic expressivity and penetrance can vary significantly, likely reflecting contributions from additional mutations in the same or other loci, consistent with the two-hit hypothesis (Karaca et al., 2018, Ribierre et al., 2018).

The classical two-hit model, originally described in cancer biology, applies to DDs where complete gene inactivation requires a PZM interacting with a germline recessive variant within a single recessive gene (Parenti et al., 2020). Oligogenic inheritance expands this concept to encompass scenarios where multiple genetic variants, each individually insufficient to cause disease, collectively exceed pathogenic thresholds (Parenti et al., 2020). These models account for cases where apparent monogenic explanations fail to fully explain observed phenotypic severity or familial segregation patterns, encompassing more complex relationships involving gene networks, protein complexes, and regulatory cascades (Parenti et al., 2020).

Most apparent DNMs originate as stochastic single event PZMs in the parental germline, which are inherited by a single offspring as constitutional heterozygote

variant (Mohiuddin et al., 2022). In this scenario, a single germ cell acquires a single PZM, which then forms part of the zygote as a heterozygous. However, recent expansions of GM research in clinical settings show a significant proportion of apparent single-event DNMs arise post-zygotically in the proband, with estimates suggesting >10% of DNMs represent PZMs at significant VAFs constituting somatic mosaicism (Hsieh et al., 2020, Krupp et al., 2017). Somatic mosaicism often underpins some of the most severe DD phenotypes, especially where pathogenic mutations in essential genes would be embryonically lethal in the constitutional state (Moog et al., 2020). These obligatory mosaic DDs include McCune-Albright syndrome, Focal cortical dysplasia type II, Pallister–Killian syndrome, and Proteus syndrome amongst others (Waldvogel et al., 2024). Many well characterized DDs with constitutional aetiologies such as Beckwith–Wiedemann syndrome, Alport syndrome, and Dravet syndrome, also have recognised mosaic aetiologies, with phenotypic severity correlating with proportionality of affected tissues (Waldvogel et al., 2024).

The contribution of GM to rare genetic disorders is likely underestimated due to technical challenges in detection, as many large clinical cohorts still rely on bulk NGS data with limited read depth and insufficient sensitivity for low-level mosaicism. Additionally, low VAF PZMs often mimic sequencing artifacts and are therefore excluded by standard variant-calling pipelines, leading to the systematic loss of clinically relevant variants (Martínez-Glez et al., 2020, Huang and Lee, 2022). Trio-WES has been a particularly powerful approach for uncovering pathogenic PZMs, as demonstrated by the Deciphering Developmental Disorders (DDD-UK) study, where trios were 4.7 times more likely to receive a diagnosis compared to singletons or duos (Wright et al., 2023). Mosaic diagnoses accounted for ~1% of DDD-UK cases, a yield mirrored in a Korean DD cohort (~1.2%), with even higher contributions reported in ASD, where up to 10% of diagnoses involve GM (Kim et al., 2023, Lim et al., 2017, Wright et al., 2023). Finally, PZMs arising later in development which may be confined to tissues may not be present in blood from which most DNA is sampled, thus confined somatic mosaicism may not be detected unless the representative tissue has been sampled.

An additional benefit of trio-WES is the ability to identify parental gonadosomatic mosaicism, which represents a critical yet underdiagnosed contributor to DDs. It arises when PZMs occur before primordial germ cell (PGC) specification, resulting in both somatic tissues and gametes inheriting the mutation (Mohiuddin et al., 2022).

Apparently unaffected or mildly affected parents may carry pathogenic PZMs in their gametes. In contrast, germline mosaicism arises after PGC specification and in adulthood, when mutations are restricted to gametes (Mohiuddin et al., 2022). Maternal oogenesis involves only ~20 post-PGC divisions followed by prolonged meiotic arrest, limiting further mutational accumulation primarily to non-replicative events such erroneous DSB repair. Conversely, paternal spermatogenesis continues throughout life (Mohiuddin et al., 2022). This continuous replication creates opportunities for both age-dependent PZM accumulation and selfish spermatogonial selection.

Gonadosomatic mosaicism has been reported across several well-known DDs including Duchenne muscular dystrophy, Tuberous sclerosis, and Neurofibromatosis type 1, and has been estimated to account for ~4% of DDs and up to 22% ASD cases (Rahbari et al., 2016, Breuss et al., 2020, Geiger et al., 2024). A recent study demonstrated that as many as 1 in 15 men harbour pathogenic variants in ~5% of their sperm cells, with sperm showing relatively stable clonal fractions while matched blood samples from the same individuals exhibited evidence of clonal expansion (Yang et al., 2021). This divergence highlights that blood may not be the optimal tissue for detecting gonadosomatic mosaicism, as fitness effects and clonal dynamics differ across tissues, necessitating multi-tissue sampling strategies including buccal swabs (ectoderm), blood (mesoderm), and urothelial cells from urine (endoderm).

Recurrence risk is not a single fixed number but a continuum determined by the timing of the PZM, the proportion of parental gametes that carry it, the parent-of-origin, and the type of mosaicism (Rahbari et al., 2016, Bernkopf et al., 2023). Early embryonic events that include the germline can place a substantial fraction of gametes at risk and therefore may approach high recurrence probabilities with a theoretical upper bound ~50%. By contrast, true one-off DNMs confined to the proband have essentially negligible recurrence probability (Rahbari et al., 2016, Bernkopf et al., 2023). Recent empirical quantification using high-depth long-read sequencing and sperm analysis estimated that paternally derived DNMs carry recurrence risk of ~0.3% while paternal gonadosomatic and germline mosaicism demonstrate direct correlation with sperm VAF, ranging from 0.2-16.6% with mean levels of 4.7% (Bernkopf et al., 2023). Maternal recurrence risks, estimated at ~1.1% for all categories combined, cannot be directly measured due to oocyte inaccessibility but are inferred from post-PGC mutational modelling (Bernkopf et al., 2023). Gonadosomatic mosaicism exhibits highly variable gametic burden (0.6-10.2% VAF range) rather than uniformly high risk

(Bernkopf et al., 2023). While advanced paternal age increases overall burden of DNMs, younger fathers with evidence of mosaicism in the absence of sperm sequences have increased risk of recurrence due to increased likelihood of gonadosomatic patterns (Bernkopf et al., 2023). This evidence-based stratification enables precision counselling, fundamentally transforming genetic counselling practice from generic 1-2% estimates toward individualized, couple-specific risk determination (Bernkopf et al., 2023).

1.5.2 Somatic mosaicism in adulthood: clonal haematopoiesis

Haematopoiesis is the lifelong process of proliferation, differentiation, and turnover of all blood cell lineages (Jagannathan-Bogdan and Zon, 2013). The production of all mature blood and immune cell lineages in adults is maintained by haematopoietic stem cells (HSC) in spatially defined bone marrow (BM) niches (Morrison and Scadden, 2014). The adult haematopoietic system is comprised of more than 10 blood cell types, with diverse cell morphologies and functions. Broadly defined, these include erythrocytes for tissue oxygenation, thrombocytes for hemostasis, and leukocytes which are responsible for innate and acquired immunity (Jagannathan-Bogdan and Zon, 2013). Haematopoietic homeostasis requires continuous replenishment of these cells for health and survival (Jagannathan-Bogdan and Zon, 2013). It is estimated that adults produce approximately $200x10^9$ erythrocytes, $150x10^9$ thrombocytes, and $140x10^9$ leukocytes daily to offset the loss of short-lived lineages to apoptosis (Fliedner et al., 2002). Haematopoiesis tightly regulated and adaptive, with production increasing by more than 10X in response to haematopoietic stressors such injury and infection (Baldridge et al., 2010, Kaushansky, 2006, Cheshier et al., 2007).

Haematopoiesis is highly efficient, being maintained by approximately 20,000 – 200,000 multipotent HSCs with significant self-renewal potential (Mitchell et al., 2022). These multipotent HSCs can remain quiescent, self-renew to sustain the stem cell compartment, differentiate along haematopoietic lineages as required, migrate from the bone marrow niche in response to signalling cues, or undergo apoptosis (Kulkarni and Kale, 2020). The regulation of these HSC trajectories is determined by a dynamic interplay between cell-intrinsic factors, including epigenetic, metabolic, and transcriptional programs, and cell-extrinsic cues derived from the bone marrow niche (Pinho and Frenette, 2019). HSCs rarely enter the cell cycle, with only 1% actively

participating in the replenishment of progenitor cells each day and only outputting new progeny every 2-4 months (Busch et al., 2015). However, a balance between differentiation and self-renewal must be maintained (Pinho and Frenette, 2019). Quiescence serves to protect HSCs from extrinsic damage, the accumulation of replication-based DNA damage, and prevents premature depletion of the HSC compartment (Arai et al., 2004, Cheng et al., 2000). Despite their high degree of quiescence, HSCs continually replenish specialized progenitors, which expand and differentiate in a hierarchical cascade to generate the full spectrum of mature blood and immune cells (Cheng et al., 2020).

The modern model of haematopoiesis has downstream progenitors as transitory states between biased HSC pools and mature cells, reflecting the molecular and functional heterogeneity of HSCs and multipotent progenitors (MPP) (Haas et al., 2018). Homeostatic production is driven by MPP differentiation, which generates intermediate biased progeny for myeloid and lymphoid lineages (Cheng et al., 2020). Notably, a substantial fraction of megakaryocytes bypasses MPP intermediates and derives directly from HSCs, indicating that lineage choice is not stochastic but rather reflects early HSC bias (Rodriguez-Fraticelli and Camargo, 2021). Cell fate mapping and single-cell transcriptomics have resolved HSCs into differentiation-inactive, multilineage, myeloid-erythroid restricted, and megakaryocyte-restricted lineages that are set early in embryogenesis and remain remarkably stable into adulthood (Bowling et al., 2020, Pei et al., 2020). Nonetheless, extrinsic cues from the BM niche remain essential in maintaining HSC integrity and balancing differentiation trajectories (Baryawno et al., 2017, Kandarakov et al., 2022).

Haematopoietic stem cells do not have the capacity for infinite differentiation. The haematopoietic niche ages, experiencing a gradual loss in mature cell production and function (Fujino et al., 2022). Haematopoietic stem cell aging is a ubiquitous phenomenon that is characterized by both quantitative and qualitative changes in the HSC compartment (Fujino et al., 2022). Although the magnitude of these changes varies among individuals, common hallmarks include mitochondrial dysfunction, genomic instability, epigenetic drift, reduced intercellular communication, senescence, and altered proteostasis (Schmauck-Medina et al., 2022). A key feature of aging is the expansion of the HSC pool relative to total bone marrow cellularity (Pang et al., 2011). However, these enlarged pools exhibit diminished proliferative capacity, reduced self-renewal, and loss of quiescence, leading to replicative stress and epigenetic alterations

that compromise the long-term functionality (Pang et al., 2011, Rundberg Nilsson et al., 2016, Flach et al., 2014).

Aging also alters the lineage output of HSCs, with a skew toward myelopoiesis at the expense of lymphopoiesis, and an increase in megakaryocyte-biased HSCs (Young et al., 2016, Grover et al., 2016). The myeloid cells produced are functionally inferior, and reduced lymphopoiesis impairs de novo production of naïve B and T cells, weakening both innate and adaptive immunity and increasing susceptibility to infection and myeloid malignancies in older individuals (Bleve et al., 2023, Linton and Dorshkind, 2004, Kogut et al., 2012, Weiskopf et al., 2009, Zeidan et al., 2019). HSC aging is driven by complex interactions between intrinsic and extrinsic factors (SanMiguel et al., 2020). Age-related changes in the bone marrow niche, including altered stromal support and increased inflammatory cytokine production, contribute to chronic low-grade inflammation, termed “inflammaging” (Franceschi et al., 2000, Li et al., 2023). Acute or chronic inflammation accelerates HSC differentiation and shifts metabolism toward oxidative respiration, increasing ROS exposure that drives DNA damage, senescence, and functional decline (Mistry et al., 2019, Shao et al., 2011, Bogeska et al., 2022). Progeny of senescent HSCs, along with other aging niche constituents, secrete inflammatory cytokines that further compromise HSC pool integrity and reduce clonal diversity over time (Bulati et al., 2011, Desdín-Micó et al., 2020, Valletta et al., 2020). The progressive accumulation of intrinsic and extrinsic damage ultimately drives the characteristic features of aged haematopoiesis, reduced regenerative capacity, lineage bias, impaired immunity, the elevated risk of myeloid disease and the sudden age-related loss of HSC clonal diversity.

Throughout most of adult life, haematopoiesis is maintained by a polyclonal HSC pool, with many individual stem cells contributing relatively evenly to mature lineages (Mitchell et al., 2022). However, aging is accompanied by a pronounced decline in HSC clonal diversity, with a few expanded clones increasingly dominating haematopoietic output. This shift, which typically occurs between 65–75 years of age, can result in 30–60% of blood production being derived from as few as 12–18 independently expanded HSC clones (Mitchell et al., 2022). These dominant clones gain a Darwinian fitness advantage within the bone marrow niche, enabling them to outcompete neighbouring HSCs and contribute disproportionately to peripheral blood, a phenomenon known as clonal haematopoiesis (CH) (Jaiswal and Ebert, 2019). CH is driven by clonal expansion of PZMs in the HSC compartment. Ageing HSCs steadily accumulate

somatic mutations over the lifespan (Mitchell et al., 2022). Estimates suggest that HSCs acquire ~17 SNPs and 0.7 INDELs per cell per year postnatally, accompanied by gradual telomere attrition of ~30 bp per year (Mitchell et al., 2022). Exonic mutations occur less frequently, at roughly 1 per HSC per decade (Welch et al., 2012). While most mutations arise stochastically, HSCs are also exposed to exogenous and endogenous genotoxic stressors, adding further DNA damage that, if unrepaired, is maintained as somatic mutations (Nohmi, 2018). The cumulative effect is a genetically heterogeneous HSC population, producing mature blood and immune cells with variable fitness within the BM niche.

Clonal expansion in HSCs arises from selective advantages conferred by driver mutations, which may enhance proliferative capacity, apoptosis evasion, or, most critically, self-renewal potential (Challen and Goodell, 2020). Cell characteristics that confer proliferative advantage are usually associated with malignancy. Mutations that confer such advantages are frequently located in genes also implicated in haematologic malignancies, encompassing both myeloid and lymphoid lineages (Niroula et al., 2021). The most recurrent CH driver mutations occur in epigenetic regulators *DNMT3A*, *TET2* and *ASXL1*, followed by splicing factors *SF3B1* and *SRSF2*, and apoptosis regulators *TP53* and *PPM1D* (Kessler et al., 2022, Bick et al., 2020b). Expanded studies in larger and more diverse cohorts have revealed a growing list of genes implicated in CH, with differential enrichment across myeloid and lymphoid lineages (Niroula et al., 2021).

The fitness of an HSC clone is influenced by the specific driver mutation, its genomic context, and evolving evolutionary pressures within the BM niche. Clones harbouring *DNMT3A* or *TP53* mutations expand slowly (~5% per year), whereas *TET2*, *ASXL1*, *PPM1D*, and *SF3B1* mutations confer faster expansion (~10% per year), and *SRSF2* mutations show the highest growth rates (~15–20% per year) (Fabre et al., 2022). These rates are dynamic across the lifespan, with *DNMT3A* clones expanding more rapidly in younger individuals (~15% per year) than in the elderly (~6.2%), while *TET2* clones expand consistently, and *SRSF2* clones only acquire selective advantage late in life (Fabre et al., 2022, Watson et al., 2020). These nonlinear dynamics underscore the importance of changing intrinsic and extrinsic pressures in clonal selection.

While natural selection drives most clonal expansions, stochastic genetic drift may occasionally eliminate or expand a clone (Ní Leathlobhair and Lenski, 2022). However, recent studies have shown that selection is the most dominant factor driving the clonal

expansion of HSCs (Watson et al., 2020, Mitchell et al., 2022). Moreover, germline variants influence clonal expansion. For example, *TERT* variants modulate telomere maintenance and predispose HSCs to acquiring additional driver mutations, illustrating a synergistic interaction between germline and somatic genomes consistent with Knudson's two-hit hypothesis (Bao et al., 2020, Ní Leathlobhair and Lenski, 2022, Knudson, 1971). Lastly, extrinsic exposures to the BN niche, including infection, chemotherapy, and smoking, further modulate clonal expansion rates by altering selective pressures over time (Dharan et al., 2021, Coombs et al., 2017, Kessler et al., 2022). Driver mutations may arise decades before clonal expansion is detectable, often established before age 40, with expansion delayed until later in life (Mitchell et al., 2022). This delay has been hypothesized to reflect a "selective shadow," where the force of natural selection wanes as reproductive potential declines, allowing deleterious or fitness-altering mutations to expand unchecked in late life (Kirkwood and Melov, 2011, Mitchell et al., 2022). Consequently, CH is strongly age-associated, and low-level clonal expansions are ubiquitous in individuals over 70 (Mitchell et al., 2022).

It is important to note that most expanded clones do not progress to malignancy but increase the risk of myeloid (HR = 7.0) or lymphoid (HR = 1.1) cancers in a mutation-dependent manner (Niroula et al., 2021). Approximately 60–80% of driver mutations occur outside known malignancy-associated genes (Jaiswal and Ebert, 2019). Non-malignant clonal expansions driven by mutations in cancer-associated genes are defined as clonal haematopoiesis of indeterminate potential (CHIP) to distinguish them from overt malignancy and reflect the associated uncertainty (Steensma et al., 2015). By definition, CHIP requires a VAF ≥2% and is limited to SNPs and INDELs (Genovese et al., 2014). Malignancy risk is modulated by the fraction of cells harbouring the mutation, the growth rate of the clone, the number of distinct drivers, and the function of affected genes, with TP53 and splicing factor mutations conferring higher risk (Goldman et al., 2023, Miller and Steensma, 2020).

Haematological malignancy is not the only endpoint of CHIP. HSCs and their progeny are free to move throughout the body and immune cells influence almost every organ (Wright et al., 2001). Therefore, there is the potential of mutations in HSCs to impact the health status of these organs as well as impacting immune response and inflammation. Driver mutations in *DNMT3A*, *TET2*, and *JAK2* are associated with elevated risk of cardiovascular disease, early-onset myocardial infarction, stroke, chronic pulmonary disease, lung cancer, and osteoporosis (Bhattacharya et al., 2022,

Jaiswal et al., 2017, Kessler et al., 2022, Miller et al., 2022, Kim et al., 2021). The health consequences of CHIP appear driver-specific, reflecting how somatic mutations in single genes can propagate functional changes across tissues. This raises questions about the effects of more complex mosaic events, such as large structural variants, on CH dynamics and systemic health.

While the aetiology of CH has predominantly focused on small variants in the form of CHIP, MCAs represent a distinct and highly impactful form of clonal expansion in the haematopoietic system. MCAs typically span megabase scales and therefore have far larger genomic footprints than CHIP drivers, with correspondingly greater potential to disrupt gene dosage, unmask recessive alleles through CN-LOH, or alter 3D chromosome topology (Liu et al., 2021). Large cohort SNP-array and WGS MCA screens have identified thousands of autosomal and sex-chromosome MCAs across hundreds of thousands of people, demonstrating that MCAs are a common, age-associated driver of CH (Liu et al., 2024, Loh et al., 2020, Terao et al., 2020, Zekavat et al., 2021, Jakubek et al., 2023a, Jakubek et al., 2025). Autosomal MCAs (aut-MCA) are typically distinguished from mosaic loss of Y chromosome (MLOY) in men and mosaic loss of X chromosome (MLOX) in women, which show marked sex-specific prevalence, and are linked to distinct systemic health outcomes (Jakubek et al., 2023b, Jakubek et al., 2025). Because of these differences in frequency, fitness, and clinical associations, MLOY and MLOX are usually analysed separately from aut-MCAs. Interestingly, aut-MCAs have been recurrently detected at specific loci, suggesting non-random patterns of mutagenesis and selection across the genome (Jakubek et al., 2023a, Jakubek et al., 2025, Watson and Blundell, 2023). Autosomal-MCAs are rarely observed in healthy individuals under 40 years of age (<1%), but their frequency rises to 2-5% between ages 40 to 60 and increases exponentially thereafter, reaching 8-10% in those over 60 and 10-20% beyond 80 (Jakubek et al., 2023b). In contrast, sex-chromosome losses are far more common, with MLOY in approximately 40% of men and MLOX in over 25% of women over 80 years of age (Jakubek et al., 2023b, Jakubek et al., 2025). While age remains a significant determinant of MCA prevalence, the distribution of events is not explained by aging alone.

Ancestry, sex, germline predisposition, and extrinsic exposures all shape MCA occurrence and clonal trajectories, underscoring the multifactorial basis of MCA-driven CH. Large population studies have shown that individuals of European ancestry show higher rates of aut-MCAs compared with African ancestry (AA) and East Asian groups,

whereas and Hispanic American individuals display elevated frequencies of chromosome X MCAs (Jakubek et al., 2023b, Jakubek et al., 2025). Males are at significantly greater risk of any aut-MCA occurrence, and half of all MCA types demonstrate significant sex-specific differences in mutation rates or fitness effects, with differences in mutation rates reaching up to 12-fold for 15q gains in men, while fitness effects generally vary more modestly (Watson and Blundell, 2023). This suggests a significant role of sex-hormones such as oestrogens and androgens in MCA mutagenesis or selection. Germline genetic predisposition further contributes through both cis- and trans-acting variants as identified through multiple MCA GWAS (Loh et al., 2018, Loh et al., 2020). Rare mutations in DNA repair genes *ATM*, *NBN*, and *MRE11* promote CN-LOH of damaged alleles, whereas *MPL* variants bias duplication of wild-type alleles, and common variants in telomere maintenance genes *TERT*, and *TERC* and immune function loci in *SP140* increase genome-wide MCA risk (Loh et al., 2020, Jakubek et al., 2023b). Environmental exposures such as smoking, chemotherapeutic therapy, and exposure to genotoxins have also been associated with increased risk of MCA driven CH (Jakubek et al., 2023a, Hubbard et al., 2023).

Recent quantitative analyses showed that although MCAs occur at mutation rates roughly ten-fold lower than CHIP drivers, the fittest MCAs confer selective advantages comparable to the fittest SNPs in CH (Watson and Blundell, 2023). The clonal dynamics of MCA in CH reflect strong positive selection rather than neutral drift, with recurrent events clustering at loci where dosage changes or CN-LOH unmask classical CH driver genes (Watson and Blundell, 2023). Much like CHIP-associated variants, these events frequently disrupt tumour suppressors or enhance oncogenic signalling, providing a mechanistic link between MCA-driven expansion and malignant transformation (Loh et al., 2018, Loh et al., 2020). Overall autosomal MCAs are associated with approximately 10-fold increased risk of haematological malignancy (Loh et al., 2018, Loh et al., 2020). Autosomal MCA carriers face substantially elevated haematological malignancy risk, with lymphoid cancers showing odds ratios of ~3, increasing to 3.78 for high cell fraction events, and reaching 5.64 for lymphoid-specific MCAs (Jakubek et al., 2023b, Niroula et al., 2021). Myeloid malignancies demonstrate even stronger associations, with odds ratios of 5.42 for aut- MCAs overall and 7.77 for high cell fraction events (Jakubek et al., 2023b, Niroula et al., 2021). Sex chromosome alterations show more modest effects with MLOX in women conferring 3.85-fold increased haematological malignancy risk, while MLOY shows weaker associations

(1.5-fold) with myeloid malignancies (Zekavat et al., 2021). Autosomal MCAs show substantial heterogeneity, with risks varying according to chromosomal location and the nature of the copy-number change. Even events affecting the same arm or copy-number state can have divergent effects, depending on which genomic elements are altered. For instance, CN-LOH at 9p, 11q, or 14q and deletions of 7q are strongly linked to myeloid neoplasms, whereas chromosome 12 gains and 13q deletions preferentially associate with lymphoid malignancies, particularly chronic lymphocytic leukaemia (Niroula et al., 2021). While most MCAs do not progress to overt malignancies, they confer a substantially increased future locus- and malignancy-specific risk, particularly when expanded to >10% cell fraction.

Mosaic chromosomal alterations have also been associated with several diverse non-haematological disease outcomes and overall mortality which has been extensively reviewed elsewhere (Hubbard et al., 2023, Evans and Walsh, 2022). Autosomal MCAs have been associated with increased risk of solid-malignancy such as bladder cancer incident infection including septicaemia and pneumonia, as well as modest risks for metabolic conditions such as type 2 diabetes mellitus (T2DM) (Song et al., 2024, Lin et al., 2021, Zekavat et al., 2021). Mosaic loss of Y chromosome exhibits a paradoxical risk pattern, with protective associations with key metabolic disorders including T2DM, primary hypertension obesity, and gout, while simultaneously increasing cardiovascular disease (CVD) mortality risk in high cell fraction carriers (Lin et al., 2021, Sano et al., 2022). Mosaic loss of X chromosome in women is primarily associated with infectious disease susceptibility, with elevated risks for pneumonia and other respiratory infections (Lin et al., 2021). Granular analyses have revealed locus-specific associations between aut-MCAs and disease outcomes, including 13q CN-LOH with chronic kidney disease, 9p CN-LOH and 13q loss with increased CVD risk, and 13q loss or 15q gains with increased infection susceptibility (Zhao et al., 2025). These distinct phenotypic patterns underscore the diverse mechanisms by which MCAs influence disease and establish them as multisystem biomarkers of age-related pathology, with clinical relevance extending well beyond haematological malignancy. Importantly, most associations have been identified in cohorts of predominantly European ancestry, with limited data available for other populations and a particular paucity of studies in individuals of African ancestry, underscoring a critical gap in understanding ancestry-specific MCA aetiology and health outcomes.

1.6 Genetic mosaicism in Africans

In the past decade, technical advances have eased entry into somatic mutation research. Specialised tools have been developed that account for the intricacies of its detection (Dou et al., 2020, Loh et al., 2020). Access to large biobank array and sequencing data has facilitated large GM epidemiological studies, especially in CH (Loh et al., 2020). Advances in single-cell sequencing and long-read sequencing have expanded the study of somatic mutation across tissues (Coorens et al., 2024, Martincorena et al., 2018, Wang et al., 2021b). Interest is at an all-time high, with rapid methodological innovations driving discovery (Yu et al., 2024). Although these advances have reshaped our understanding of genetic mosaicism across all fields, studies in continental African populations remain sparse. This gap limits insight into population-specific patterns, disease associations, and the broader impact of genetic mosaicism to health in Africa. The following section reviews the literature on genetic mosaicism in Africans, beginning with its role in rare diseases and DDs and then in clonal haematopoiesis.

1.6.1 Rare disorder mosaicism in Africa

Finding relevant papers from either African based institutions or that report genetic mosaicism in African ancestry cases is challenging. A search term was generated to identify papers describing “mosaicism” in “Africans” in a rare disorders context from the PubMed database (https://pubmed.ncbi.nlm.nih.gov/) and the Scopus database (https://www.scopus.com/). The search was expanded to include somatic mutation, gonadosomatic, germline mosaicism, chromosomal mosaicism, mosaic aneuploidy, and mosaic as terms. The term then added the names of 54 African countries, African ancestry, and African regions not limited to sub-Saharan Africa to capture regional requirements. Finally, a list of terms was added to filter out papers from other fields in which “mosaic” is used to describe some form of non-related process. A final total of 90 papers that describe clinical mosaicism cases in African ancestry people, from cases identified in African countries, or by African institution affiliated authors was generated. The list of publications can be found in the Appendix D. No restrictions were put on the year of publications, which ranged from 1967 to 2023. The earliest report of GM in an African ancestry case comes from South Africa, which described a case of mosaic Klinefelter syndrome in 1967 (Segal et al., 1967).

Most publications describing genetic mosaicism in continental Africans originated from Northern African countries, with only a handful from Sub-Saharan African (SSA) countries outside of South Africa (Appendix D). These reports were largely case studies of GM identified during routine clinical diagnostics, typically following a phenotype-to-genotype approach. The molecular methods employed were predominantly cytogenetic, with occasional use of targeted single-variant PCR or Sanger sequencing. Only two studies reported next-generation sequencing data from larger clinical cohorts, both involving African ancestry cases in the United States (Posey et al., 2016, Odgis et al., 2023). Only one of those publications explicitly set out to screen for GM in rare disease cases (Odgis et al., 2023). This is in stark comparison to the large screenings for GM contributions to DDs, ASD, and congenital malformations that have been described in European ancestry cases and cohorts (Cook et al., 2021, Dou et al., 2017, Freed and Pevsner, 2016, Gambin et al., 2020, Sherman et al., 2021, Wright et al., 2019, Hsieh et al., 2020, King et al., 2015, King et al., 2017, Eberhardt et al., 2023). To date, no genotype-first studies designed to systematically screen for GM in large continental African rare disease cohorts have been published (Appendix D).

While this was not a comprehensive systematic review, such as one adhering to the STROBE guidelines and incorporating multiple databases, it nonetheless highlights a striking gap in knowledge of clinically relevant GM in African ancestry populations (Von Elm et al., 2007). Over the past 60 years, only 90 publications have reported cases of GM in African ancestry individuals or GM cases from African institutions (Appendix D). In contrast, applying the same search terms to European ancestry studies from Europe and North America resulted in 367 PubMed papers and 197 Scopus papers. African genomics is commonly underrepresented in the literature, to the detriment of the global community (Rotimi et al., 2017, Wonkam et al., 2022). A similar disparity was identified in a comprehensive review of African publications in developmental disorders, with only 23% of the 287 publications coming from researchers outside of South Africa, Egypt, Tunisia, and Marocco (Baine-Savanhu et al., 2023). Several explanations for this disparity in DD research have been proposed, which could apply to GM in Africans.

Firstly, even in environments with adequate resources and trained personnel, developmental disorders are notoriously difficult to diagnose due to heterogenous phenotypes, variable expressivity, incomplete penetrance, and pleiotropic or epistatic genetic aetiologies (Lumaka et al., 2022, Lumaka et al., 2017). This is particularly

relevant for African ancestry DDs, where many phenotypes remain poorly characterized and often present atypically compared to European ancestry cases in which they were initially discovered (Lumaka et al., 2022, Lumaka et al., 2017). While this is a shared challenge across the globe, the paucity of reported African cases in comparison to European cases amplifies the problem in Africans. This is further complicated in mosaic cases, which often present with an attenuated phenotype correlated with the proportion of and types of affected tissues (Takagi et al., 2012, Lubala et al., 2018). In African populations, phenotypes may deviate from Eurocentric dysmorphology references. In a scoping review of African ancestry DDs, 53 publications documented atypical cases that expanded known phenotypes or identified novel variants in established DD genes (Baine-Savanhu et al., 2023).

Identification of mosaic attenuations in African DD cases is rare. For example, a Central African patient with mosaic Fragile X syndrome presented with a phenotype so atypical that a facial recognition algorithm trained on DD facial dysmorphia could not aid in diagnosis (Lubala et al., 2018). Despite Fragile X syndrome being a well-characterized monogenic intellectual disability for over 40 years, molecular confirmation in African ancestry patients has been reported only in South Africa, Ghana, and Cameroon cases (Peprah et al., 2010, Essop and Krause, 2013, Chiurazzi et al., 1996). The Central African case, identified as potentially mosaic by an experienced clinical team, prompted targeted molecular testing, confirming size-mosaic Fragile X syndrome (Lubala et al., 2018). This case extended the known Fragile X phenotype and represented the first molecularly confirmed diagnosis in Central Africa (Lubala et al., 2018). Even with the significant underreporting of rare disease incidence, and the lack of reliable epidemiological data in Africa, it is extremely unlikely that there are no other cases in a region that large and that populous (Omotoso et al., 2022, Malherbe et al., 2021). Rather, as the authors point out, it is likely an underreported condition due to known barriers to medical genetics in Africa (Kamp et al., 2021, Lubala et al., 2018). These barriers are pervasive across continental Africa and include a scarcity of medical genetics services, limited funding, logistical constraints, low public awareness, and cultural factors such as fear of discrimination (Kamp et al., 2021, Omotoso et al., 2022, Lubala et al., 2018, Lumaka et al., 2022). The key issue underlying these challenges is the scarcity of resources, especially for rare disease, on a continent where infectious disease is pervasive and the focus of most healthcare policies (Lumaka et al., 2022).

Sub-Saharan African (SSA) nations spend an average of 0.44% of their gross domestic product (GDP) on research and development and, to date, no SSA country has spent more than 1% GDP (current global average = 2.21%) (World Bank, 2024a). Despite being classified as an upper-middle-income country, South Africa currently spends only 0.60% of its GDP on research and development which is on a downward trend from its peak of 0.83% in 2018 (current upper-middle-income average = 2.06%) (World Bank, 2024a, World Bank, 2024c). In terms of healthcare, SSA spends 5.11% GDP, while South Africa spends 8.27% (upper-middle-income = 5.82%). This is across all sectors, with health sciences making up only a small portion of allocated funding. Most impactful African genomics research relies on some form of international agency grant funding, and the lack of public infrastructure means most medical genetics services are tied to academic institutions (Kamp et al., 2021, Krause et al., 2018). There is limited funding for the training of medical geneticists and genetic councillors and yet, at least in South Africa, the rate of graduating continues to increase (Gomes et al., 2024). Training opportunities for medical geneticists and genetic counsellors remain limited, and employment shortages drive qualified personnel out of Africa, reducing capacity to detect rare diseases and mosaicism phenotypically (Gomes et al., 2024). In South Africa, there are currently 13 registered medical geneticists and 31 genetic councillors, which is estimated to cover only 10% and 5% the national needs (Gomes et al., 2024). Of these professionals, only 1 medical geneticist and 3 genetic councillors practice outside the greater Johannesburg and Cape Town regions in the Gauteng and Western Cape provinces (Gomes et al., 2024). Rural populations are particularly underserved, resulting in under identification and under-referral of heterogeneous DD, rare disease, and GM cases (Baine-Savanhu et al., 2023, Gomes et al., 2024, Kamp et al., 2021, Krause et al., 2018). This underreporting is a significant contributor to the observed paucity in the literature.

Infrastructure limitations further impede African research. Many SSA countries lack the capacity for somatic mutation or sub-chromosomal MCA screening, which requires validated microarrays, expensive NGS platforms, and trained personnel (Sun et al., 2015, Srivastava et al., 2019, Waggoner et al., 2018, Savatt and Myers, 2021). Most publications in the review relied on cytogenetics, Sanger sequencing, or targeted PCR, paralleling findings from the African DD review, where only 21,2% reported microarray or NGS use (Appendix D) (Baine-Savanhu et al., 2023). While techniques like karyotyping, Sanger sequencing, MLPA, Southern blot, and PCR can detect

mosaicism they are confirmatory and phenotype-dependent, limiting large-scale or genome-wide discovery. (Verhoef et al., 1999, van Veghel-Plandsoen et al., 2011, Lubala et al., 2018, Essaddam et al., 2023). Thus, these techniques are not suitable for screening for GM in African rare disease at scale.

We have known about GM in African since at least 1967, regularly detecting mosaic cases in rare disease through mostly incidental findings (Appendix D) (Segal et al., 1967). The paucity of literature is not due to a lack of scientific interest in Africa but rather is likely a reflection of limited access to large, clinically characterized cohorts and insufficient capacity for molecular GM diagnosis. Studies of this scale require multidisciplinary teams, extensive funding, and specialized infrastructure, resources that have historically been low in Africa (Dou et al., 2018, Savatt and Myers, 2021, Thorpe et al., 2020, Kamp et al., 2021, Krause et al., 2018). African studies need funding and support at many levels and can be extremely successful if properly resourced. The success of large initiatives such as DDD-UK demonstrates the potential of well-supported studies to elucidate GM at scale (Wright et al., 2023). The unfortunate result is the observed paucity of data on African mosaicism in rare diseases literature. We do not know what the true extent of genetic mosaicism in African rare disease is.

The good news is that initiatives such as H3Africa have started to address this inequity through the pooling of resources across the continent, providing access to bioinformatics infrastructure, establishing training networks, facilitating the establishment of larger cohorts, and developing appropriate Afrocentric tools (Mulder et al., 2018). DDD-Africa, one of its major outputs, is a multidisciplinary study aimed at determining the genetic aetiology of DDs in continental Africans from South Africa and the Democratic Republic of the Congo (DRC). Leveraging trio-WES, partnership with DDD-UK, and an experienced African medical genetics team, the study enables detection of post-zygotic mosaicism in probands and parental gonadosomatic mosaicism. This design provides the necessary resolution and scale to uncover GM contributions to DDs in Africans, simultaneously addressing the underrepresentation of African DD and GM data in the literature. (DDD-Africa, https://h3africa.org/index.php/ddd-africa/). The recruitment strategy and methodology of DDD-Africa is detailed in Chapter 2.

1.6.2 Clonal haematopoiesis in Africa

Prior to 2020, large-scale characterizations of GM in African ancestry populations were extremely limited. Most major genetic epidemiological studies of somatic mosaicism, including clonal haematopoiesis, were conducted in European ancestry cohorts, with African ancestry participants included only as small sub-cohorts (Machiela et al., 2015). The only study reporting MCA prevalence in African ancestry individuals was from a meta-analysis of approximately 120,000 participants, predominantly of European ancestry, which included a small group of African American and Afro-Caribbean participants (Machiela et al., 2015). In this study the authors reported that genotype-inferred African ancestry participants, had significantly lower odds of MCA occurrence than their European ancestry counterparts (OR = 0.28) (Machiela et al., 2015). However, the study had significant limitations. The SNP arrays used were optimized for European populations and may not have captured the full spectrum of African genomic variation, while the MCA-calling algorithm relied on heterozygous sites to detect B-allele frequency imbalances (González et al., 2011). Given the higher heterozygosity of African genomes, these methodological constraints likely contributed to under-detection of MCAs, suggesting that the reported lower prevalence may not reflect the true frequency in African ancestry populations. This culminated in the rationale for and design of Chapter 3.

Since 2020, the representation of African ancestry participants in studies of CHIP and MCA-driven CH has increased substantially, driven by the release of larger, more diverse cohorts (Table 1.2). Notably, the Trans-Omics for Precision Medicine (TOPMed; ~51,000 African ancestry samples) and the Million Veterans Program (MVP; ~180,000 African ancestry samples) include significant African ancestry cohorts (Taliun et al., 2021, Gaziano et al., 2016). Advances in sequencing technologies and the development of specialized pipelines for calling somatic mutations in single-sample, non-matched tissues have facilitated large-scale CHIP analyses in these diverse biobank cohorts (Loftfield et al., 2018, Francis et al., 2024, Liu et al., 2024, Jakubek et al., 2025, Bhattacharya et al., 2024). Improvements in MCA-calling algorithms applied to widely available SNP-array data originally designed for GWAS have similarly enabled the characterization of MCAs in thousands of African ancestry participants (Loftfield et al., 2018, Francis et al., 2024, Liu et al., 2024, Jakubek et al., 2025, Bhattacharya et al., 2024, Jakubek et al., 2023b, Zhou et al., 2023). However, only a few studies to date have included samples from continental African cohorts

(Bhattacharya et al., 2024, Zhou et al., 2023). MCAs in inferred African ancestry samples have generally been analysed in comparison to European ancestry distribution s(Jakubek et al., 2023b, Machiela et al., 2015). Most studies continue to report lower prevalence of MCA-driven CH in African ancestry participants (Jakubek et al., 2023b, Machiela et al., 2015). The results from these studies can be found in Table 1.2 and are summarised here.

In TOPMed, no differences in overall CHIP prevalence were observed between European and African ancestry participants after adjusting for age and smoking, although European ancestry participants showed higher risk than Hispanic or Asian ancestry individuals (Bick et al., 2020b). They did, however, identify an African specific *TET2* variation rs79901204 that increased the risk of CHIP status in carriers by 2.4-fold ($p = 4.0 \text{ X } 10^{-09}$) through disruption of the *TET2* enhancer in HSCs which decreased *TET2* expression and increased self-renewal (Bick et al., 2020b). Analyses in the UK Biobank and Geisinger Community Health Initiative (GHS) cohorts reported higher CHIP rates in European ancestry than African ancestry participants, but these differences were not significant after multivariate adjustment, reflecting a potential sample size skew (Kessler et al., 2022). Similarly, in the All of Us (AoU) cohort, CHIP was associated with age, smoking, and Hispanic ethnicity, but not African ancestry, highlighting the continued need for diverse cohorts (Vlasschaert et al., 2023). Earlier studies linking environmental exposures to CHIP enrichment did not find significant differences in CHIP prevalence between African and European ancestry participants (Bolton et al., 2020).

Table 1.2: Clonal haematopoeisis studies with African ancestry participants.

Study	ΔOdds	Incidence (%)	OR	P-value	N_{AA}	Cohort
CHIP						
Bick *et al.* (2020)	AA = EA	-	0.96	$4.2x10^{-01}$	31261	TOPMed
Bolton *et al.* (2020)	AA = EA	AA=27.00/ EA=31.00	0.90	$5.3x10^{-02}$	1491	MSK-IMPACT
Pincez *et al.* (2021)	SCD ↑ nSCD	-	13.50	5.3×10^{-05}	1398/6211	SCD/nSCD
Kessler *et al.* (2022)	AA = EA	AA=2.48/ EA=5.98	-	ns	15463	UKB
	AA = EA	AA=2.25/ EA=7.84	-	ns	7984	GHS
Vlasschaert *et al.* (2023)	AA = EA	-	1.01	ns	21035	All of Us
Bhattacharya *et al.* (2024)	SSA ↓ NA	-	0.57	$1.9x10^{-05}$	1878	REPRIEVE
MLOY						
Lotfield *et al.* (2018)	AA ↓ EA	AA=0.40/ EA=1.80	-	$3.0x10^{-03}$	3285	UKB
Cheng *et al.* (2024)	AA ↓ EA	-	0.34	$1.0x10^{-03}$	576	ILCCO
Francis *et al.* (2024)	AA ↓ EA	AA=14.10/ EA=26.40	-	-	99103	MVP
Jakubek et al. (2025)	AA ↓ EA	AA=5.80/ EA=15.43	0.69	$1.6x10^{-07}$	7653	TOPMed
MLOX						
Jakubek *et al.* (2023)	AA ↑ EA	AA=15.94/EA=13.67	1.67	$2.5x10^{-33}$	20217	TOPMed
Cheng *et al.* (2024)	AA = EA	-	0.75	ns	576	ILCCO
A-MCA						
Machiela *et al.* (2015)	AA ↓ EA	-	0.23	$3.0x10^{-03}$	14588	Meta-analysis
Gao *et al.* (2021)	EA ↑ AA	AA=0.71/EA=1.19	1.50	$3.3x10^{-02}$	2088	MSK-IMPACT
Jakubek et al. (2023)	AA ↓ EA	AA=3.64/EA=5.30	0.76	$1.2x10^{-05}$	20132	TOPMed
Cheng et al. (2024)	AA = EA	-	0.92	$2.3x10^{-01}$	576	ILCCO
Zhou *et al.* (2023)	SSA ↑ EA	SSA=14.20/ EA=6.20	-	-	674	Ghanaian men
	BL ↑ BL-f	BL=8.40/ BL-f=3.40	2.80	5.70×10^{-11}	931/3822	BL / BL-f

ΔOdds = lower or higher odds of occurrence in ancestry A vs B, with B as reference; OR = odds ratio; A-MCA = autosomal mosaic chromosomal alteration; MLOY = mosaic loss of Y chromosome; MLOX = mosaic loss of X chromosome; CHIP = clonal haematopoiesis of indeterminant potential; N = African ancestry count in cohort; UKB = UK Biobank; MVP = Million Veterans Programme; MSK-IMPACT = Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets; GHS = Geisinger MyCode Community Health Initiative; TOPMed = Trans Omics for Presession Medicine; REPRIEVE = Randomized Trial to Prevent Vascular Events in HIV; ILCCO - International Lung Cancer Consortium; BL = Burkett's Lymphoma children; BL-fr = Burkett's Lymphoma free children; AA = African ancestry; EA = European Ancestry; SSA = sub-Saharan African; NA = North American; SCD = sickle cell disease affected AA; nSCD = non-sickle cell AA controls; ns = reported not significant but does not report p-value.

Only one study has reported a significant difference in CHIP prevalence between African and European ancestry participants (Table 1.2). In a cohort of people living with HIV, CHIP odds were significantly lower among participants from SSA than among those from higher-income North American countries (OR = 0.57 [0.40–0.81], p = 0.0019) (OR = 0.57 [0.4-0.81], p = 0.0019). Interestingly, CHIP was also associated with greater HIV severity, although ancestry modification of this association was not reported (Bhattacharya et al., 2024). This study is one of only two that have included participants from continental Africa. The other, a case–control study, demonstrated that African ancestry individuals with sickle-cell disease were more likely to develop CHIP than unaffected African ancestry controls (Pincez et al., 2021). Overall, a binary CHIP-status outcome does not appear to be significantly different between African and European ancestry participants (Table 1.2).

In contrast risk of MLOY has consistently been reported at lower prevalence in African ancestry men (Table 1.2). In both the UKB and the MVP cohorts, MLOY was detected less frequently in African ancestry than European ancestry participants (Francis et al., 2024, Loftfield et al., 2018). In MVP, the largest ancestry-stratified analysis of MLOY to date with 99,301 African ancestry and 400,970 European ancestry men, MLOY was detected in 14.1% and 26.4% of participants, respectively. An unadjusted analysis indicated European ancestry men were approximately twice as likely to have MLOY (OR = 2.19, $p < 0.0001$) than African ancestry men, with similar effects across age strata (Francis et al., 2024). The MVP study also identified 46 SNP loci which were associated with MLOY risk in African ancestry participants, including 5 SNPs in the *GGT5*, *MPL*, *NKX2-3*, *ETV6*, and *BLCAP* genes that were reported as African ancestry specific (Francis et al., 2024). This highlighted ancestry-related genetic contributions to MLOY risk.

Other large cohorts from TOPMed and the International Lung Cancer Consortium (ILLCO) studies confirmed significantly lower MLOY prevalence in African ancestry men after adjustment for age (Cheng et al., 2023, Jakubek et al., 2025). Methodological considerations, such as higher heterozygosity rates in African genomes leading to greater call loss when down-sampling sites for MoChA-based detection (17% African ancestry versus 3% European ancestry), did not alter the direction of effect, and admixture analyses showed European ancestry proportion was positively correlated with MLOY in African Americans (Jakubek et al., 2025). They identified two autosomal loci, rs2296311 in *TCL1A* and rs2376992 in *BCL2L1* that were associated with lower odds of MLOY in TOPMed, and that are found at higher allele frequencies in African versus European ancestry participants (Jakubek et al., 2025). Together, these findings suggest both shared and ancestry-specific germline loci may shape MLOY risk.

Mosaic loss of the X chromosome shows a different pattern. In TOPMed, African ancestry women had significantly higher odds of chrX MCAs than their European counterparts (OR = 1.67, $p = 2.5 \times 10^{-33}$), with consistent associations for both chrX losses (OR = 1.59) and gains (OR = 2.06) (Jakubek et al., 2023b). These associations persisted after random down-sampling of heterozygous sites, though African ancestry women lost 26% of calls compared to 7% in Europeans. Lower MCA prevalence was observed in admixed women with reduced African ancestry, and one associated SNP was more frequent in Africans, indicating possible ancestry-linked germline influences

(Jakubek et al., 2023b). By contrast, the much smaller ILCCO cohort did not identify ancestry-related differences in MLOX (Cheng et al., 2023).

Autosomal MCAs odds were more varied in African ancestry participants (Table 1.2). One of the early meta-analysis of approximately 120,000 participants reported significantly lower odds of MCA occurrence in African ancestry inferred participants compared with European ancestry participants (OR = 0.23, p = 0.003) (Machiela et al., 2015), However, this study was limited by the use of SNP arrays that do not fully capture African variation, thus reducing resolution, and a European ancestry skewed cohort (Machiela et al., 2015). Subsequent sequencing-based studies also reported similar trends of lower MCA prevalence in African Americans (Table1.2). One such study detected MCAs from whole blood of cancer patients from targeted deep sequencing data generated for a cancer risk study (Gao et al., 2021). European ancestry participants had higher odds of having MCAs with cell fractions above 10% than African ancestry participants (OR = 1.50, p = 0.033) (Gao et al., 2021). No significant difference in autosomal MCA rate was found between African ancestry and European ancestry participants in the ILCCO cohort, and this lack of significance was repeated when stratified by MCA type and chromosome arm (Cheng et al., 2023).

The strongest evidence comes from TOPMed, where autosomal MCAs were again less common in African ancestry participants (OR = 0.76, p = 1.2 X 10^{-05}), despite accounting for probe coverage (Jakubek et al., 2023b). Heterozygosity rates were once again down sampled to match the lowest median rates in Southeast Asians, as done for MLOY and MLOX. The result was African ancestry participants losing 23% of all autosomal MCA calls, versus just 4% in European ancestry participants, with the authors reporting significance and direction of odds remaining the same both prior and after down sampling (Jakubek et al., 2023b). After down sampling African ancestry participants were reported to have significantly lower risk of autosomal MCA incidence than European ancestry participants (OR = 0.76, p = 1.2x10^{-05}) (Jakubek et al., 2023b). Stratified analyses revealed specific differences in MCA incidence by ancestry, with African ancestry participants having higher incidence of 8q gains, and lower incidence of 12q gains, 14q CN-LOH, 20q losses, and 13q CN-LOH than European ancestry participants (Jakubek et al., 2023b).

A subsequent GWAS of autosomal MCAs in *cis* germline loci revealed six deleterious coding, promoter, and splice SNPs in the *ATM* gene that were significantly associated with 11q CN-LOH in Africans (Jakubek et al., 2023b). These SNPs had ORs ranging

13.94-29.11 and were common in Africans (MAF = 5.52-7.73%), but rare in Europeans (MAF <0.1%). Allelic ratio analysis revealed that the 11q CN-LOH replaces wild-type *ATM* alleles with the deleterious allele, disrupting *ATM* function and leading to 11q CN-LOH selection. A single rare *MPL* SNP rs6086 was associated with 1p CN-LOH (OR = 533.64) in Africans (MAF = 1.21%), with 1p CN-LOH replacing the deleterious wild-type with the alternate allele (Jakubek et al., 2023b). Aggregate rare variant burden analysis revealed further associations with *ATM* and 11q CN-LOH, and *MPL* and 1p CN-LOH. Additional rare variant burdens were reported in *DCPS*, *ADAM17*, *PPP1R16B,* and *TET2* with any autosomal MCA, as well as in *OR4C16* with MLOX. For the commonly observed MCAs in African ancestry participants, CHIP mutation co-occurrence was reported for 1q CN-LOH and *ZBTB33* CHIP, 5q loss and *ASXL2* CHIP, 8q gain and *DNMT3A, CUX1, YLPM1,* or *U2AF1* CHIP, 9p CN-LOH with *TET2*, *ASXL1*, *JAK2*, and *U2AF1* CHIP, and *KRAS* with 12p and 12q gains (Jakubek et al., 2023b). Suggesting that somatic mutations in these genes predispose African ancestry participants to MCAs or vice versa, which could be used as markers of future haematological malignancy risk (Jakubek et al., 2023b).

The additional stratification of autosomal MCAs into more granular structures reinforced evidence of significant heterogeneity underlying the broader association of autosomal MCA status with genetic ancestry (Jakubek et al., 2023b). Certain loci were enriched for distinct autosomal MCA classes in either European or African ancestry participants (Jakubek et al., 2023b). The observed increased risk of aggregated autosomal MCAs in European ancestry participants was reinforced by the results showing enrichment of more unique MCA loci as well (Jakubek et al., 2023b). Without the additional stratification of MCAs, the heterogeneity underlying the aggregate associations would be missed, and the results may be interpreted as European ancestry participants having increased risk of all MCA classes. This is clearly not the case. The non-random autosomal MCA distributions across ancestries point to both shared and ancestry specific germline variant predispositions or environmental exposure risk factors which impact the rates of MCA formation and selection across the genome. In both cases the information expands our understanding of MCAs in CH and is more useful than broader categorical associations, which are prone to aggregation bias(Gelman and Auerbach, 2016).

The results from TOPMed, UKB, and MVP are robust, and are suggestive of underlying factors which increase the risk of aggregate MCA occurrence in European ancestry

participants versus non-European ancestry participants (Francis et al., 2024, Jakubek et al., 2025, Jakubek et al., 2023b, Kessler et al., 2022). Some of the strongest evidence stemmed from the proportion of European or African genetic ancestry correlating with MCA risk alleles associations in admixed individuals (Francis et al., 2024, Jakubek et al., 2025, Jakubek et al., 2023b, Kessler et al., 2022). However, the results from these studies are only reflective of MCA rates and/or distributions in African ancestry people or African diaspora living in the environmental conditions of Northern America or Europe (Jakubek et al., 2023b, Francis et al., 2024). These associations cannot be generalised to continental African populations that live in diverse environments and that have substantial genomic diversity (Figure 1.10) (Choudhury et al., 2020, Wonkam et al., 2022, Agyemang et al., 2016, Marzà-Florensa et al., 2021, Fortes-Lima et al., 2024).

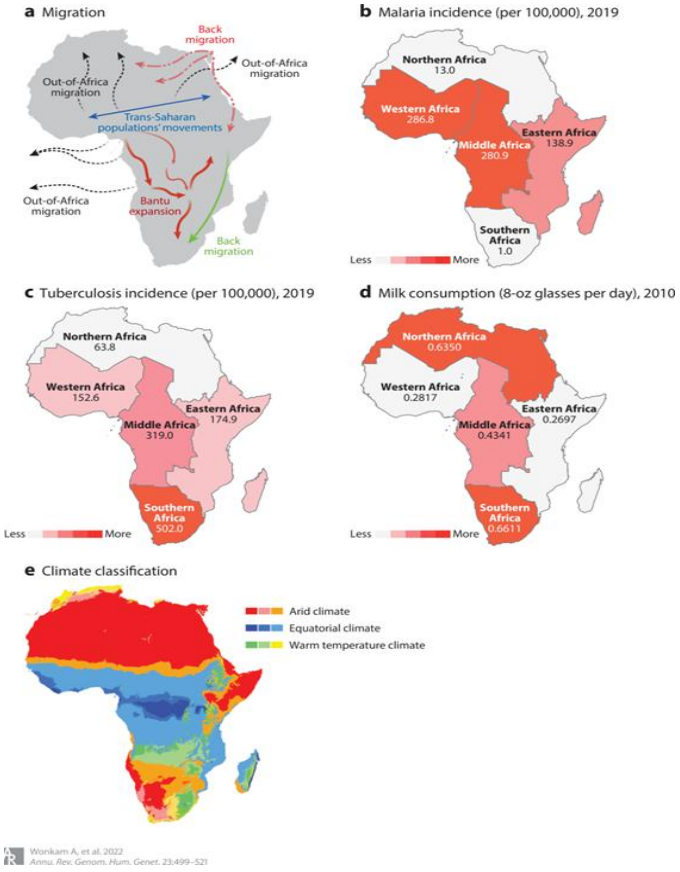


Figure 1.10: Environmental factors that shape genomic diversity within Africa.

Several intersecting factors have shaped genomic diversity within African populations across human history. (a) Migratory events within and out-of-Africa have led (b and c) exposure to diverse pathogen burdens across Africa impacts adaptive selection of African genomes. (d) significantly different socioeconomic status with a wide array of diets and exposure to rural and urban. (e) the African climate and ecosystem vary significantly. Figure from (Wonkam et al., 2022) under Creative Commons Attribution 4.0 International License.

It might be worthwhile, at this point, to discuss the inclusion of ancestry as a factor in CH studies. Variation has been observed between subpopulation groups for many human traits and disorders, including CH (Bick et al., 2020b, Jakubek et al., 2023b). Variability in the observed rates and distributions of CH, underlying somatic mutations, clonal trajectories, and fitness effects have been reported across ages, between sexes, and now between broadly defined genetic ancestry groups (Bick et al., 2020b, Jakubek et al., 2023b). The categorization of cohorts into age, sex, ethnicity/ancestry, or other groups is common in association analysis to capture aggregate variation in the outcome that may be explained by underlying biological mechanisms common to that grouping. The underlying mechanism does not necessarily need to be fully understood. If the association with the aggregate grouping is robust, then it needs to be accounted for so as not to confound additional independent variables of interest. The broader categories act as proxies of the actual underlying mechanisms and should not necessarily be considered inherent modifiers in themselves.

The inclusion of such categorizations in a study should have clear rationale, as they describe heterogenous human traits often derived from political and historical definitions (Gombault et al., 2023). For example, ethnicity is a multifaceted social construct encompassing shared geographical, historical, and social traits, such as country of origin, genetic ancestry, language, religion, and traditions (Gombault et al., 2023). Some of these aspects of ethnicity would clearly have no logical impact on CH risk with a biologically sound mechanism of action. Additionally, the definitions for racial/ethnic categorizations differ across the globe, change over time, and often do not sufficiently account for complexity in genetic admixture (Gombault et al., 2023). In the context of CH, ancestry is often a broad proxy categorization or measure of shared genetic similarities within individuals, which are derived from common ancestor population groups and should be reflective of alleles or haplotypes that were fixed across time and space as these groups migrated, mixed, and settled through the forces of drift, bottlenecks, and selection (Mathieson and Scally, 2020). In the absence of an empirical estimation of genetic ancestry, from methods such as principal component analysis, self-reported ancestry is often used as the proxy of these shared alleles. However, ancestry categorisation is also reflective of the current geographical location of the participants in cross-sectional analysis, which is sometimes an overlooked property. The selective pressures which modify CH risk are not identical for individuals, even with significant shared genetic ancestry, in every environment. In much of the CH

literature, ancestry grouping has been used as a proxy for shared germline alleles and not for the later. Little effort has been made to further define ancestry categorisations in a geographical context, or to acknowledge the bias that may arise from this.

European ancestry participants from North America may share more alleles with European ancestry participants from Iceland than with individuals of African descent residing in North America. Even though the environmental pressures they face can differ markedly from those in Iceland. It is not correct to assume that CH mutation rates and distributions from individuals with shared genetic ancestry from one environmental context will apply to those in a different environmental context. Thus, aggregate ancestry categorisations should be interpreted in context of region of recruitment, serving as further proxies for environmental modifiers of CH mutation rate and fitness effects. This is especially important in African ancestry populations, which are known to have the greatest genetic diversity. As a hypothetical example, a CH risk allele identified in Ghanaians residing in Accra may be shared to a high degree with African ancestry groups in the UK, but could be totally absent from Kenyans residing in Nairobi (Fortes-Lima et al., 2024, Rotimi et al., 2017). Additionally, that same risk allele may be more significant in the UK environment than in Ghana, where other risk alleles may be more important in modifying CH risk (Marzà-Florensa et al., 2021, Agyemang et al., 2016). In the above example, the aggregate African ancestry grouping may obscure important subpopulation-specific modifiers of CH risk.

Modern human populations have lived longer on the African continent than anywhere else (Campbell et al., 2014). Long-established continental African populations have genomes that have been shaped by migratory events across diverse ecological zones, promoting extensive admixture and gene flow over thousands of years (Figure 1.10) (Campbell et al., 2014, Fortes-Lima et al., 2024). Important migrations, including the Bantu-speaking agriculturalist expansions from Western and Central Africa into Eastern and Southern Africa, the migration of Nilo-Saharan speakers from Sudan to Lake Chad, Kenya, and Tanzania, and subsequent out-of-Africa dispersals with reintroductions of Middle Eastern, European, and Southeast Asian lineages, have all occurred under selective pressures in varied ecosystems, climates, and pathogen exposures (Campbell et al., 2014, Choudhury et al., 2020, Fortes-Lima et al., 2024, Wonkam et al., 2022). This has ultimately contributed to unparalleled genetic diversity in continental Africans. There are almost as many defined genetic ancestries within African populations as there are in the rest of the world combined, with African

populations having smaller and more diverse haplotype blocks and a lower incidence of runs of homozygosity (Rotimi et al., 2017, Auton et al., 2015). The extent to which this diversity extends to germline MCA and CH risk modifier alleles is not yet known. Recent analysis has suggested that most pathogenic mutations in DDR pathways arose after the latest out-of-Africa migration, with relatively few being shared between African and non-African populations (He et al., 2024). Thus, the landscape of germline CH risk alleles may be just as diverse in continental Africans. Additionally, the application of recently identified alleles in African ancestry participants from Northern America might not apply to continental African populations, especially from non-Western African populations (Francis et al., 2024, Jakubek et al., 2025, Jakubek et al., 2023b).

Environmental pressures such as infection burden, genotoxic exposure, sociodemographic factors, and human behaviour traits have all been shown to significantly impact MCA driven CH risk (Demanelis et al., 2023, Dharan et al., 2021, Jakubek et al., 2023a, Zekavat et al., 2021, Young et al., 2024). The diversity of African genomes has been shaped by the diversity of environmental selective pressures across very long periods of time (Figure 1.10) (Rotimi et al., 2017, Wonkam et al., 2022). Many continental African populations are in an epidemiological transition, brought upon by rapid urbanisation. However, a large proportion of Africans also still live in more rural agrarian communities. Rural communities make up 70% of Kenyans, 67% of Burkinabe, 41% of Ghanaians, and 31% of South Africans (World Bank, 2024b). Within African regions, the environmental exposures of its rural and urban populations can also be significantly different. A significant contributor to CH risk is air pollution. Urban populations are exposed to air pollutants from industrial activities, vehicle exhaust, and waste disposal, whereas rural populations generally experience lower overall air pollution but may have greater exposure to biomass burning from smaller fires (Health Effects Institute, 2022). The burden of air pollution varies across African countries. For example, Ghanaians face high levels of ambient particulate pollution ($PM_{2.5}$), primarily due to Saharan desert dust, while South Africans are exposed to $PM_{2.5}$ mainly from industrial emissions and vehicle exhaust (Health Effects Institute, 2022). In contrast, Kenyans experience lower ambient $PM_{2.5}$ levels but significantly higher exposure to household air pollutants from burning biomass or indoor cooking (Health Effects Institute, 2022). Particulate air pollution has been significantly associated with increased risk of MLOY, while the impact on other CH is

not as clear (Wong et al., 2018, Guan et al., 2024). Thus, CH risk through air quality likely differs across the continent and is likely very different from air pollution exposures in other global regions (Health Effects Institute, 2022).

Genotoxic exposure often occurs as an occupational hazard. Rural agricultural workers in Africa are often exposed to pesticides and herbicides and poorly purified water sources, while urban workers in Africa are often exposed to industrial carcinogens, heavy metals in the mining industry, or water that is contaminated with industrial waste such as arsenic (Fasinu and Orisakwe, 2013). Many occupational genotoxins have been associated with CH, including the broad-spectrum herbicide glyphosate, which is commonly used in agriculture and is associated with increased risk of MLOY expansions in farmers (Chang Vicky et al., 2023). Water contaminated with arsenic has also been associated with increased MLOY risk (Demanelis et al., 2023). Additionally, exposure to polycyclic aromatic hydrocarbons, found in fossil fuel emissions, has been associated with increased risk of MLOY in male coke-oven workers (Liu et al., 2020b). While no other specific links with MCAs driven CH and occupational carcinogens have been reported, known carcinogenic compounds such as heavy metals, cadmium, chromium, pyrethroids, organophosphate insecticides, organochlorines, and many others are likely to impact MCA formation through direct DNA damage inducing DSBs and dysregulation of DDR pathways (Tabrez et al., 2014). Thus, industry and agriculture practices, that may impact CH risk, also differ across the continent.

The African continent's diverse climates shape both biodiversity and agricultural practices, leading to a wide variety of diets among African populations (Figure 1.10) (Niohuru, 2023). Urban areas offer a greater abundance and diversity of food options compared to rural areas, but they also have a higher prevalence of highly processed foods. Urbanising populations in Africa have been shown to consume more pro-inflammatory diets, characterized by higher intakes of processed sugars, and unhealthy fats, leading to increased immune system activation, which may impact CH selection (Pourmontaseri and Khanmohammadi, 2024, Ramaboli et al., 2024). In contrast, rural populations typically adhere to more traditional diets that are richer in whole grain cereals such as millet and sorghum, fruits, vegetables, and legumes, which have been shown to lower levels of inflammatory cytokines such as interleukin-1 (IL-1) and interleukin-6 (IL-6) (Malongane et al., 2024). Dietary intake across Africa varies considerably. Populations in Ghana typically consume diets rich in fish and other seafood, while those in Kenya increasingly rely on maize supplemented with traditional

staples like beans and plantains (Niohuru, 2023). In contrast, rapid urbanization in South Africa has led to a shift towards more pro-inflammatory diets which are characterized by higher intakes of fats, sugars, and salts, coupled with lower consumption of proteins, fruits, and vegetables (Niohuru, 2023). Undernourishment is still prevalent in many countries across SSA, especially in poorer rural settings, where it primarily impacts younger people (Niohuru, 2023). Undernourishment impairs immune system response, gut health, and increases oxidative stress, all of which contribute to chronic inflammation (Stumpf et al., 2023). Deficiencies in key nutrients, such as proteins, vitamins, and minerals, can trigger inflammatory pathways and increasing the production of pro-inflammatory cytokines (Stumpf et al., 2023). The extent to which diet impacts CH remains unclear and needs to be studied further. Diet may be a key modulator of CH risk in both rural and urban African settings, especially in combination with other lifestyle factors such levels of physical activity, alcohol consumption, and tobacco smoking.

The impact of tobacco smoking on CH risk is well established (Young et al., 2024). Smoking is strongly linked to an increased risk of all forms of CH, including CHIP, autosomal MCAs, MLOX, and MLOY, with current smokers exhibiting the highest risk across all categories (Young et al., 2024). In contrast, the relationship between alcohol consumption and CH remains less clear and appears to be modified by sex, with heavy-drinking women showing a significantly higher risk for MLOX, whereas no such association has been observed in men and MLOY (Young et al., 2024). Smoking prevalence varies across African countries and regions, though men consistently report higher smoking rates than women in all surveyed populations (Niohuru, 2023). South Africa has the highest smoking prevalence, with approximately 34% of men and 6.5% of women smoking (World Bank, 2020a). In Kenya, smoking rates are lower, with 19.5% of men and 2.7% of women smoking, while in Burkina Faso, the prevalence is 22.1% in men and 6.4% in women. Ghana reports the lowest smoking rates, with only 0.3% of women and 6.6% of men smoking (World Bank, 2020a). However, substantial regional differences exist within countries. For instance, in Ghana, smoking prevalence is significantly higher in the remote northeastern region, reaching 31.2% (Niohuru, 2023). Similar trends are observed for alcohol consumption, with men generally consuming more than women. Per capita alcohol intake varies widely across SSA, with significantly lower consumption in countries with a higher proportion of Islamic practitioners. Among the surveyed countries, Burkina Faso reports the highest per

capita alcohol consumption at 11.27 litres, followed by South Africa at 7.13 litres. In contrast, intake is lower in Ghana at 4.26 litres and Kenya at 2.26 litres (World Bank, 2020b). Although the evidence is less conclusive, heavy alcohol consumption may modulate inflammatory pathways and alter clonal dynamics in a sex-specific manner, potentially impacting the prevalence and expansion of MCAs and CH (Evans and Walsh, 2022).

The relationship between environmental exposures and lifestyle factors that do not directly induce DNA lesions or disrupt DDR pathways with CH is often mediated through inflammation (Avagyan and Zon, 2023). CH is associated with elevated pro-inflammatory cytokines, which serve as key drivers of clonal expansions (Bick et al., 2020a). Exogenous environmental exposures, including air pollutants, pro-inflammatory or nutrient-poor diets, tobacco smoking, and excessive alcohol consumption, can induce oxidative stress and chronic systemic inflammation, thereby increasing DNA damage and mutation rates in HSCs (Avagyan and Zon, 2023, Evans and Walsh, 2022, Jakubek et al., 2025, Jakubek et al., 2023b, Malongane et al., 2024, Niroula et al., 2021, Uddin et al., 2024, Young et al., 2024, Bick et al., 2020a). This pro-inflammatory environment, compounded by the inherent aging process that impairs DNA repair and immune surveillance, creates a selective environment that favours the expansion of clones with mutations impacting genes such as *DNMT3A*, *TET2*, and *PPM1D* (Bick et al., 2020a, Zhang et al., 2019). The diverse environmental exposures, dietary patterns, and aging trajectories across the African continent may contribute to region-specific inflammation profiles, which, in turn, shape the selection of specific CH driver mutations. These processes occur within the broader context of regionally distinct infection burdens, further influencing clonal dynamics and selection pressures.

In African regions with high burdens of chronic infections such as tuberculosis (TB), malaria, and Human Immunodeficiency Virus (HIV), persistent inflammatory responses may further exacerbate these processes. Sustained immune activation and elevated cytokine/chemokine production can disrupt HSC homeostasis, promoting the development and expansion of CH clones by altering hematopoietic differentiation and selective pressures within the bone marrow microenvironment (Caiado et al., 2021). The relationship between chronic infections and CH is an emerging area of research, with increasing interest in how infection induced inflammation may shape CH selection and, conversely, how CH induced immune system dysregulation might influence infection susceptibility (Zekavat et al., 2021, Caiado et al., 2021). HIV burden varies

markedly across SSA, with the highest incidence, prevalence, and death rates recorded in Southern African populations (Niohuru, 2023). Eight Southern African countries have estimated HIV prevalences greater than 10% of the adult population between the ages of 15 to 49 (World Bank, 2022). HIV prevalence in South Africa is 17.8% of the adult population, with women being particularly impacted. In comparison, HIV prevalence is much lower in Eastern and Western African nations, with Kenya reporting 3.7%, Ghana 1.7%, and Burkina Faso 0.6% among the adult population (World Bank, 2022). This substantial regional disparity suggests that Southern African populations likely experience significantly higher rates of chronic HIV modulated inflammation, which has been shown to increase CHIP but not MCA risk (Han et al., 2024, Lin et al., 2023). People with advanced or untreated HIV are at increased risk of opportunistic infections, which can trigger immune reconstitution inflammatory syndrome (Hsu et al., 2018). Recent findings suggest a twofold increased risk of CHIP in individuals with HIV, with CH risk being linked to HIV disease severity (Han et al., 2024). Specifically, higher CD8+ T-cell counts and lower nadir CD4+ T-cell counts have been associated with an increased likelihood of CH in people with HIV (Rocco et al., 2024). Both HIV infection and CH have been shown to independently drive systemic inflammation, potentially creating a pathological feedback loop in which the presence of either CH or HIV exacerbates the risk of acquiring or accelerating the other (Avagyan and Zon, 2023, Grund et al., 2016, Han et al., 2024, Rocco et al., 2024). However, the directionality of this association remains unclear and warrants further investigation.

Direct associations between CH and *Mycobacterium tuberculosis* (*Mtb*) infection have not been extensively studied. However, research indicates that *Mtb* can infiltrate the bone marrow niche as early as 10 days post-infection, where it reprograms HSCs for extended periods, leading to suppressed myelopoiesis and impaired development of protective immunity via type I interferon signalling (Khan et al., 2020) Interestingly, this dysregulation of myelopoiesis may also mitigate the progression of cytopenia to more severe leukaemia phenotypes (Xiu et al., 2023). Consequently, the successful treatment of TB may also increase the odds of CH progressing to malignancy (Xiu et al., 2023). Additionally, *Mtb* infection induces a distinctive pro-inflammatory and oxidative stress environment, which is associated with disease severity and anaemia (Ashenafi et al., 2023, Caiado et al., 2021, Gil-Santana et al., 2019). Given that chronic inflammation can drive clonal expansion, the persistent inflammatory state induced by TB may contribute to the development or progression of CH, particularly in aging

individuals with other predisposing factors or co-infection with HIV, potentially selecting for CH-associated mutations that confer protection against myelopoiesis loss (Ashenafi et al., 2023, Caiado et al., 2021, Gil-Santana et al., 2019). The incidence of pro-inflammatory profiles induced by TB likely varies across SSA populations, corresponding to the wide variation in TB incidence rates per 100,000 people across the region (Figure 1.10) (Niohuru, 2023). Southern and Central African populations exhibit some of the highest TB incidence rates in SSA with South Africa reporting 427 cases per 100,000 individuals and the Central African Republic reaching 540 cases per 100,000 (World Health Organization, 2024a). In contrast, TB burden is comparatively lower in Western and Eastern Africa, with Ghana reporting 129 cases per 100,000, Burkina Faso 43, and Kenya 223 (World Health Organization, 2024a). Thus, the potential impact of TB and TB treatment on CH risk is likely to vary among continental African ancestry populations. However, this relationship has not yet been systematically investigated.

Malaria infection induces profound alterations in haematopoiesis that may also create conditions favourable for CH, with infection inducing highly proliferative states in HSCs in mouse models (Vainieri et al., 2016). While overall, malaria induces increased HSC proliferation, it also skews differentiation. Acute malaria infection depletes mesenchymal stromal CXCL12-abundant reticular (CAR) cells, leading to a reduction in CXCL12 chemokine and interleukin-7 cytokine signalling (Lee et al., 2024). This disruption shifts haematopoiesis toward increased myelopoiesis and erythropoiesis while depleting lymphopoiesis(Lee et al., 2024). Malaria severity has also been correlated with specific pro-inflammatory cytokine profiles, with IL-17A and IL-1β increasing significantly in severe cases (Obeng-Aboagye et al., 2023). Malaria incidence per 1,000 people at risk contrasts sharply with the burden of HIV and TB across the African continent, being lower in Southern Africa and substantially higher in Western and Eastern Africa (World Health Organization, 2024b). In South Africa, overall incidence is well under 1.0 per 1,000 people at risk, though regional variation exists, with the Vhembe region in Limpopo experiencing a higher incidence of 3.79 per 1,000 people at risk (Niohuru, 2023, World Health Organization, 2024b). In contrast, malaria incidence is markedly higher in Kenya, Ghana, and Burkina Faso, with reported rates of 69, 159, and 354 per 1000 people at risk respectively (World Health Organization, 2024b). While no direct association between malaria and CH has been reported, there is also very little data available.

The interplay between chronic infection burden and CH in Africa is likely region-specific, shaped by interactions with climate, diet, and lifestyle factors that influence inflammation-driven selection for CH across populations. The relationship between HIV, TB, malaria, and CH exemplifies how chronic infection can establish a persistent pro-inflammatory state that selects CH, illustrating the complex interplay between infection, inflammation, and PZMs that lead to clonal expansions (Avagyan and Zon, 2023, Grund et al., 2016, Han et al., 2024). Of course, there are many other infection burdens which may impact CH specifically through inflammation in Africa, with regional incidence dependant on climate, population density, diet, water and sanitation, access to medical care, and many other factors (Ashenafi et al., 2023, Niohuru, 2023). The diverse spectrum of infection burdens and environmental factors across Africa presents a unique opportunity to explore the interconnected relationship between environment, inflammation, infection, and CH risk.

Given the significantly different exposures in Africa compared to North America and Europe, along with the immense genetic diversity within African populations, it is not appropriate to assume that CH and MCA profiles observed in these regions would be directly applicable to continental African populations. Similarities in MCA distributions between European ancestry, African American, and continental African ancestry cohorts could help identify inherent risk factors influencing MCA rates across populations. Distributions that are similar between African American and continental African ancestry cohorts, but distinct from those observed in European ancestry populations, may reveal African ancestry-specific variables, such as shared germline modifiers of MCA risk. Conversely, differences between African American and continental African ancestry MCA distributions could point to environmental modifiers of MCA selection specific to each region or help fine-map germline loci that modify MCA risk in multi-ancestry cohorts. However, there is a lack of comprehensive data on CH in continental African ancestry populations, making it difficult to discern the similarities or differences between continental Africans and African ancestry populations not living within the African continent.

To date, only a single study has investigated MCAs in CH in small cohort of SSAs. In this analysis, MCAs were detected in 931 children with Burkett's Lymphoma (BL), and 3822 children without (BL-f) from Uganda, Kenya, Tanzania, and Malawi. This was then compared to 674 cancer-free Ghanian men and 2618 cancer-free European men (Zhou et al., 2023). Autosomal MCAs larger than 2Mb in size were more likely to be

detected in BL children than BL-free children after adjusting for age, sex, country, and *P. falciparum* infection history (8.40% vs 3.40%, OR = 2.80, $p = 5.70 \times 10^{-11}$) (Zhou et al., 2023). Chromosome X MCAs were also more likely to be detected in BL girls than BL-f girls (3.7% vs 1.7%, OR = 2.15, $p = 3.74x10^{-02}$), while chromosome Y MCAs did not differ between BL and BL-f boys (4.50% vs 4.70%, OR = 1.01, p = 0.91) (Zhou et al., 2023). Country of origin was not found to have a significant impact on MCA detection in SSA children. Autosomal MCAs were detected in 14.20% of the Ghanaian men and 6.80% of cancer-free European men, with African rates consistently higher across all age groups (Zhou et al., 2023).

The distribution of MCAs across the autosome differed between BL children, BL-free children and cancer-free Ghanaian male adults (Zhou et al., 2023). The most prevalent MCAs in BL children were gains on chromosomes 1q and 8, CN-LOH on chromosomes 1, 6, and 11, and losses on chromosomes 9 and 13. In BL-f children the most frequent MCAs were gains on chromosome 5, CN-LOHs on chromosomes 10, 14, and 16, and losses on chromosome 19. In Ghanaian men, the most frequently occurring MCAs were gains on chromosomes 5, 12, and 21 (Zhou et al., 2023). No data on germline variant associations with MCA rates were generated in this study, mostly likely due to the smaller sample sizes. Additionally, the autosomal MCA distributions of the European ancestry controls were not reported on, and no additional comparisons between cohorts were made. The authors note that the study served as a proof of principal, or pilot study, for characterising MCA driven CH in continental Africans (Zhou et al., 2023). The overall MCA incidence in continental cancer free Africans in this study was higher than the reported incidence of autosomal MCAs in African ancestry participants from TOPMed and MSK-IMPACT (Gao et al., 2021, Jakubek et al., 2023b, Zhou et al., 2023). Gains of chromosome 12 were reported as commonly occurring in both TOPMed and Ghanaian men and TOPMed reported low fraction gains of 21, which was also common in Ghanaian men. Apart from these two autosomal MCAs, the distributions and rates appear to be different between the cohorts (Jakubek et al., 2023b, Zhou et al., 2023). However, a directed comparison between the SSA and TOPMed studies is not advisable.

While both studies detected MCAs with MoChA, their respective methodology was very different. The TOPMed studies detected MCAs from WGS data, while the SSA study used high density SNP-array data (Jakubek *et al.*, 2023b; Zhou *et al.,* 2023). A comparison of MCAs detected from either WGS or SNP-array in the TOPMed studies

showed that while SNP-arrays detect many of the same MCAs as WGS, with high concordance, WGS data resulted in significantly more autosomal and chrX MCAs overall(Jakubek et al., 2023b). WGS generates more informative heterozygous sites than SNP-arrays but also has lower effective depth than SNP-arrays. Each technique is susceptible to platform specific errors and has different QC considerations. Each study also required different MCA parameter thresholds in post-processing. For example, the study in SSAs only included MCAs >2 Mb in size, while the TOPMed studies required MCA calls to have a minimum of 2000 informative markers after random down sampling of heterozygous SNPs (Jakubek *et al.*, 2023b; Zhou *et al.,* 2023). More confidence in the MCAs calls, came at the cost of smaller MCAs being excluded in both studies. Additionally, the effective sample sizes, and confounder distributions such as age and sex of African ancestry participants, differed between the studies (Jakubek *et al.*, 2023b; Zhou *et al.,* 2023). These technical factors such as methodology, data sources, cohort makeup, quality control, and analysis approaches were different across all the studies in Tabe 1.2. Without effective data harmonization, it is extremely challenging to untangle variance in somatic mutation rates and distributions introduced by technical differences from biologically significant results.

Overall, with the addition of newer studies, there is a trend of lower frequencies or lower odds of autosomal MCAs and MLOY, and higher odds of MLOX in African ancestry cohorts when compared to European ancestry cohorts (Table 1.2). However, there is no full consensus across these studies, as differences in technical limitations, methods, and quality control processes may influence the outcomes. While most studies account for known CH risk factors, our understanding of CH remains incomplete. As such, there may be unidentified risk-modifying factors that impact mutation rates and/or fitness effects, which could influence the likelihood of observing these mutations.

A final consideration is the level of stratification performed in each study. Some studies utilize a simple binary classification of overall or aggregated CH or CHIP status, followed by association analysis adjusted for confounding factors. However, further stratification of CH status based on mutation class or consequence, as well as grouping by broader shared loci such as chromosomal arm, may significantly alter these associations. Insufficient stratification or variable annotation can lead to aggregation bias or mask heterogeneity (Gelman and Auerbach, 2016). For example, grouping MCAs by chromosome losses may indicate a higher likelihood of losses on

chromosome 2 in men. Further stratification by chromosomal arm could reveal that the association is limited to the 2p arm, which may only be a hotspot of losses with shared breakpoints near the terminal region. This stratification could reveal that terminal losses are more prevalent in men, thus driving the observed association at the arm and chromosome level. Without this detailed annotation, a misleading conclusion could be drawn that all 2p losses are enriched in men. However, there are limitations to further stratification in the analysis of rare mutational events, even in large cohorts. Increased multiple testing raises the risk of false positives, while smaller group sizes reduce statistical power and introduce sparse data bias, potentially inflating association signals (Gosho et al., 2023). Given the volume of data generated in these studies, it is understandable that a balance must be struck between aggregation and sparse data biases in the analysis and publication process.

An insightful example of this effect was published recently on CH and haematological malignancy risk (Niroula et al., 2021). Earlier publications reported associations of aggregate autosomal MCA with lymphoid malignancies, while CHIP status was primarily associated with myeloid malignancies (Desai et al., 2018, Loh et al., 2020). Niroula and colleagues further stratified aggregated CHIP and autosomal MCA status into more granular groupings, identifying those that were linked to either lymphoid or myeloid malignancy. This stratification enabled the accurate prediction of malignancy risk with much greater precision (Niroula et al., 2021). This reduced aggregation biases, emphasizing that not all drivers of CH carry the same risk. Thus, when interpreting the associations reported in studies summarized in Table 1.2, it is crucial to consider the variability introduced by methodology. Given the skewed ratio of European ancestry to African ancestry participants in many cohorts, it is advisable to evaluate each study independently until appropriate data harmonization is achieved, and replication studies from continental African ancestry cohorts in diverse geographical regions are incorporated into the literature.

1.7 Thesis rationale

The role of GM in human aging, health, and disease is becoming increasingly clear as advances in technology have allowed for its detection in increasingly larger cohorts. However, these characterisations have overwhelmingly focused on European ancestry populations, with little inclusion of more diverse cohorts from across the globe. This is especially true for continental African ancestry cohorts, for which there is a severe paucity of data. The paucity is prevalent in all GM research contexts, spanning GM across the human lifespan, from early rising PZMs which contribute to rare childhood disorders to somatic mutation accumulation in normal aging tissues. We do not know what the full extent and impact of GM in continental African ancestry populations is. This has important implications for African rare diseases and the genetic epidemiology of aging related phenotypes, both of which have important implications for Africans at the individual and public health levels. As more continental African ancestry genomics studies are facilitated through collaborative initiatives such as H3Africa, it is essential that the role of GM is investigated by African genomics researchers, who have a unique insight into African specific challenges and are motivated to address them.

Large clinical genomics projects in Africa, such as DDD-Africa, are shedding light on the genetic ethology of DDs in African populations. Similarly, large genetic epidemiological studies like AWI-Gen are exploring the genetic contributions to cardiometabolic diseases across diverse African populations. While GM has been found to play a significant role in these contexts in other populations, the extent and impact of GM in continental African ancestry populations remain unclear. More specifically, we lack knowledge regarding the burden and contribution of GM to the aetiology of DDs in Africans, as well as the extent and distribution of MCAs in age-related CH. From small variants to MCAs, and from the mechanisms of their formation to their fitness effects, the paucity of data on African GM is ubiquitous.

Most of what has been reported on GM in African ancestry people has come from African American and African diaspora cohorts from European and North American studies. While this data is deepening our understanding of GM, we do not know the extent to which it can be generalised to continental African ancestry cohorts. African ancestry is not a monolithic classification. Continental African ancestry populations are genetically diverse, shaped by a wide range of ecosystems and evolutionary pressures across the continent since the dawn of modern human history. Given that GM is influenced by both germline risk alleles and environmental selection pressures, its

incidence and distribution are likely just as diverse across different continental African populations. However, very few studies have included participants from continental African populations, largely due to challenges in recruiting and accessing appropriate cohorts. Those that have often fail to use tools that accurately account for the vast genetic diversity of these populations. Without sufficient African ancestry data, our global understanding of GM remains incomplete in both clinical and epidemiological contexts, to the detriment of all. This gap in research must be addressed. African GM research cannot be sidelined or used solely for comparisons and replications by international research groups, without respecting the priorities and mandates of African genomics. Thus, protocols for characterizing GM in African ancestry datasets need to be established and tested.

1.7.1 Research questions

A central research question was generated to try to address the lack of African GM data:

Can the scarcity of continental African ancestry data be overcome by deploying advanced bioinformatics tools to detect GM in existing continental African ancestry datasets?

Several additional questions further shaped the aims and objectives:

I) Which types of African datasets are suitable for detecting GM in both clinical and epidemiological contexts?
II) Are there any existing datasets which can be used for GM detection?
III) Which classes of GM can be detected in these datasets?
IV) What additional considerations must be taken into account when detecting GM in these datasets?
V) What are the appropriate tools for detecting GM in these datasets? Which are suited for trio-WES data, and which are appropriate for SNP-array data?
VI) Can these tools be run locally using existing computational infrastructure?
VII) Can GM detection strategies be incorporated into existing African genomics research analysis pipelines?
VIII) What are the limitations of detecting GM from each of these datasets with these strategies?

IX) Should African genomics research groups incorporate GM pipelines into their analyses?

X) What recommendations can be made to improve GM detection strategies in the future?

1.7.2 Hypothesis

The central hypothesis of this thesis is that state-of-the-art bioinformatics tools can be successfully integrated into existing local analysis pipelines to detect and characterize GM in appropriate continental African ancestry datasets. Mosaic small variants can be identified in trio-WES data, and MCAs can be characterized in SNP-array data.

1.7.3 Aim and objectives

Therefore, the aim of this thesis was to identify appropriate detection strategies to explore the extent and impact of GM in large continental African ancestry trio-WES and SNP-array datasets.

To achieve this aim and address the research questions, the following two core objectives were established:

1) To implement a bioinformatics pipeline to screen for pathogenic mosaic small variants that may contribute to DDs in South African trio-WES data from the DDD-Africa study.

2) To implement a bioinformatics pipeline for the exploratory characterisation of autosomal MCAs in diverse continental African populations using H3Africa SNP-array data from the AWI-Gen study.

The first objective is presented in Chapter 2, and the second objective is presented in Chapter 3.

Chapter 2: Screening for pathogenic genetic mosaicism in the Deciphering Developmental Disorders in Africa study

2.1 Introduction

Genetic mosaicism is a well-established contributor to developmental disorders (Moog et al., 2020, Wright et al., 2019). However, its clinical impact remains significantly underestimated due to systematic underdiagnosis and profound research inequities across global populations (Lumaka et al., 2022, Baine-Savanhu et al., 2023, Geiger et al., 2024). Pathogenic post-zygotic mutations arising during embryogenesis can manifest as somatic mosaicism in affected probands or gonadosomatic mosaicism in asymptomatic parents, fundamentally altering recurrence risk counselling and therapeutic management strategies (Bernkopf et al., 2023, Geiger et al., 2024). Large-scale characterizations of DDs in European cohorts indicate that mosaic diagnoses account for approximately 1-3% of cases (Wright et al., 2023). However, the prevalence and spectrum of genetic mosaicism in developmental pathology remain largely unexplored in African populations, particularly in the context of substantial clinical heterogeneity and resource-limited settings.

African DD molecular epidemiology is characterized by a profound paucity of data, largely due to the lack of large, ancestrally diverse cohorts (Baine-Savanhu et al., 2023). As many as 63% of African countries have no published DD data, while within South Africa 66% of published studies have focused on predominantly European ancestry patients (Baine-Savanhu et al., 2023). The contribution of PZMs to DDs in continental African populations is even less well characterized, with no published large-scale studies to date. Technical challenges in PZM detection further compound these disparities. Many pathogenic PZMs present at low variant allele fractions and are frequently filtered out by standard variant-calling pipelines designed to remove sequencing artifacts (Huang and Lee, 2022). As a result, clinically relevant mosaic variants may be systematically missed unless dedicated detection strategies are deployed. Trio-based whole exome sequencing is a powerful solution to these challenges, enabling high-confidence identification of both somatic mosaicism in

probands and gonadosomatic mosaicism in parents through family-aware variant filtering strategies. The DDD- UK study demonstrated that trio analysis increases diagnostic yield 4.7-fold compared to singleton sequencing, and specialized mosaic detection tools for trio-WES, such as EM-Mosaic, can recover clinically relevant PZMs that would otherwise be discarded at negligible additional cost (Wright et al., 2023, Hsieh et al., 2020).

The analysis in this chapter forms part of the DDD-Africa initiative and leverages trio-WES to screen for pathogenic mosaic small variants in South African trios and evaluate their clinical impact on DDs. This approach is particularly critical in African populations, where diagnostic pipelines and allele frequency databases are often Eurocentric, limiting variant interpretation. By combining a custom variant filtering strategy with robust statistical analysis and the EM-Mosaic tool, this chapter screens for candidate post-zygotic SNPs and INDELs in affected probands and candidate gonadosomatic variants in parents from 64 South African trios. Identifying causative DNMs and PZMs in African DD trios will not only improve local diagnostic yields but also contribute globally by enhancing variant interpretation across diverse populations (Choudhury et al., 2020, Baine-Savanhu et al., 2023). For affected families, a molecular diagnosis would shorten years-long diagnostic odysseys and inform precise, lifelong clinical care (Blesson and Cohen, 2020, Carmichael et al., 2015).

2.2 Methodology

2.2.1 Study design and ethics

A cross-sectional variant prioritisation analysis was performed on a sub-set of 64 South African WES-trios as part of the ongoing DDD-Africa study (DDD-Africa, https://h3africa.org/index.php/ddd-africa/). Ethics approval for this study was granted by the University of the Witwatersrand Human Research Ethics Committee (HREC) under clearance certificate number M2011144 (Appendix A). Ethics approval for DDD-Africa was granted by the University of the Witwatersrand HREC under clearance certificate number M180678 (Appendix B).

2.2.2 The DDD-Africa cohort

The DDD-Africa study aimed to recruit approximately 500 trios with severe undiagnosed DDs from South African and the DRC (DDD-Africa, https://h3africa.org/index.php/ddd-africa/). Affected participants underwent detailed clinical phenotyping by registered medical geneticists, to facilitate clinical interpretation of candidate variants. WES data was generated for multiple parallel variant calling strategies to uncover pathogenic variants in DD associated genes. At the time of writing, 64 South African trio-WES from the first batch were available for the analysis in this chapter. Of the 64 trios, 62 probands were from unrelated families, and 2 probands were indicated as dizygotic twins.

The following is a summary of the research protocol, that was carried out members of DDD-UK and DDD-Africa research team, which produced the 64 trio-WES files used in this analysis. Participants from the DRC were not included in this study. The South African participants were recruited from three sites in Johannesburg, South Africa. These were the Chris Hani Baragwaneth Academic Hospital, the Charlotte Maxeke Academic Hospital, and the Rahima Moosa Mother and Child Hospital, Johannesburg. Participants were included in the DDD-Africa if they met defined inclusion criteria without overlap with defined exclusion criteria (Appendix E). Parents were required to give informed consent for participation. Relevant medical history and sociodemographic data was voluntarily captured through questionnaires. Participants were assessed by a registered clinical geneticist. Clinical information was then

captured and curated, and phenotypes were linked to appropriate Human Phenotype Ontology (HPO) terms (Robinson et al., 2008). Participant clinical data and documentation was then uploaded to the REDCap electronic data capture platform hosted at the University of the Witwatersrand, and the Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources (DECIPHER) and made available for this study (Harris et al., 2019, Firth et al., 2009).

Samples for DNA extraction and WES were taken from two tissue sources. Whole-blood samples (5-10ml) were drawn from the primary trio and other any other relevant or affected family members. A second sample for affected probands was obtained from saliva swabs. Whole-blood and saliva swap samples were stored on site at the Division of Human Genetics, University of the Witwatersrand, Johannesburg following established in-house protocols. The DDD-Africa study used the whole-blood samples as the primary DNA source for WES and parallel variant calling analysis. Saliva swab sample WES data was not available for this analysis. Registered lab technicians extracted whole-blood sample DNA for each participant following a verified in-house adaptation of the established salting-out protocol (Miller et al., 1988). DNA samples were stored on-site, where they were quantified and processed for WES following recommended practices (Li and Durbin, 2009).

2.2.3 DDD-Africa whole exome sequencing

Genomic data for the DDD-Africa study was generated in partnership with the DDD-UK study at the Welcome Sanger Institute in the United Kingdom (WSI). Processed DNA samples were sent to the WSI for WES. WES data was generated for each participant using Index Sequence Capture with the Twist exome library capture kit (Twist Bioscience, USA), followed by 150bp paired-end sequencing on the Illumina NovaSeq 6000 platform (Illumina, USA). A mean depth of coverage of 45X was achieved across the cohort. Primary bioinformatics analysis of the raw unmapped WES read data was performed at the WSI to produce multi-sample VCF files and CRAM files with linked identifiers. Pre-processing and variant calling procedures for germline short-calling discovery was performed following the Genome Analysis Toolkit (GATK) best practices pipeline (McKenna et al., 2010, Van der Auwera et al., 2013). Importantly, the variant calling pipeline was not setup by DDD-Africa with somatic variant calling as a primary objective and was intended for larger scope multiple parallel

analyses. As such this protocol was not as sensitive to somatic variants as a dedicated somatic caller such as Mutect2 would have been (Cibulskis et al., 2013). However, it was determined that an exploratory screening strategy could still be designed for future refinement and deployment in the full cohort; thus, achieving the objective of this chapter.

Pre-processing for germline variant discovery involved mapping of short-read sequences to the GRCh38 reference genome with the Burrows-Wheeler aligner backtrack algorithm, producing initial raw BAM files (Li and Durbin, 2009). BAM file improvement including duplication marking, INDEL realignment, and base quality score recalibration, was performed using GATK and SAMtools (Li et al., 2009, McKenna et al., 2010). The GATK HaplotypeCaller was used to call SNPs and INDELs in per sample in full DDD-Africa dataset using a multi-sample calling mode (GVCF) to improve sensitivity of variant detection (McKenna et al., 2010). The HaplotypeCaller sample-ploidy setting was set to 2.0 in this step, which is the default diploid setting that tells the variant calling algorithm to exclude heterozygous calls at the extreme ends of VAF deviations from an expected 0.5 as artificial errors. This limitation was known, and it was expected that the variant calling would be less sensitive to mosaic PZMs going forward. Single sample GVCFs were then merged with GATK GenomicsDBImport, and multi-sample joint calling was done with GATK GenotypeGVCFs. The raw variant call set was filtered for machine artifacts using GATK Variant Quality Score Recalibration (VQSR) and ApplyVQSR following best practices (Van der Auwera et al., 2013). VQSR truth sensitivity thresholds were set to 99.7% for SNPs and 99.5% for INDELs, and variants over these scores were labelled as PASS for the filter.

The resulting data was transferred back to the WITS high-performance cluster, where it was further processed by the DDD-Africa bioinformatics team. The SNP VCF and INDEL VCF were concatenated into a single multi-sample VCF with BCFtools merge. Filtered variants were annotated with Ensembl Variant Effect Predictor (VEP) plugins to include per site transcript predictions, functional consequence predictions, splice site annotations, ClinVar predictions, selected *in silico* tool pathogenicity predictions from the dbNSFP database, and population frequency data (McLaren et al., 2016, Liu et al., 2020a). Trio-VCF files were extracted from the main multi-sample VCF using a trio pedigree file and BCFtools +split. The BCFtools trio-dnm2 plugin was deployed to annotate trio-VCFs with formatted statistics that were used by DDD-Africa to call likely DNMs (Danecek et al., 2021). The trio-dnm2 plugin annotates each variant site shared

between the proband, mother, and father, with a VAF and assigns a negative log-scaled quality score (DNM score) for each variant that can be used to screen for putative DNMs. The 64 fully annotated trio-VCF files produced by this extensive bioinformatics pipeline and accompanying CRAM files were made available by DDD-Africa for variant screening and prioritization in this analysis.

2.2.4 Candidate variant prioritization strategy

The scope of this analysis was to leverage the existing DDD-Africa trio-WES data to screen for candidate mosaic small variants in affected DD probands that would be missed by standard DNM variant prioritization strategies. And to screen for gonadosomatic small variants in either parent, which would appear as inherited germline small variants in DD affected probands. To achieve this, two separate variant prioritization strategies were implemented which incorporated two statistical filtering methods to identify evidence of mosaicism from small variant VAFs. Each strategy has three components:

- **Variant quality control and initial filtering**
- **Statistical determination of mosaicism**
- **Variant prioritization and curation**

The two prioritization strategies were: i) a proband level DNM and PZM variant filtering strategy to identify a wide range of *de novo* variants and capture mosaic candidates, and ii) a gonadosomatic parental variant prioritization strategy. These were designed specifically for this analysis by adapting and combining elements from two publications which identified PZMs from trio-WES data DNM pipelines (Hsieh et al., 2020, Wright et al., 2019). The EM-Mosaic tool, which was specially designed to detect mosaic SNPs from WES, and a high sensitivity variant prioritization strategy employed by DDD-UK on WES-trios with similar read depth (Hsieh et al., 2020, Wright et al., 2019). Additional quality control parameters were evaluated at each level to account for potential mosaicism, and to support downstream statistical analysis. Variant filtering was performed using BCFtools (Danecek et al., 2021). A comprehensive workflow for both strategies can be seen in Figure 2.1 and Figure 2.2.

2.2.4.1 Candidate variant quality control and filtering strategy

Variant level statistics generated by GATK VQSR were used to exclude likely false positive variant calls following GATK best practices (Van der Auwera et al., 2013). SNPs and INDELs were hard filtered to pass their respective VQSR filters. Additional SNP and INDEL specific hard filters for strand bias were applied to account for the high number of variants passing VQSR with evidence of strand bias observed in downstream validation. SNPs were required to have VQSR StrandOddsRatios (SOR) annotations of less than 3, while INDELs were required to have FisherStrand (FS) annotations of greater than 200. These cutoffs were chosen following GATK hard filtering recommendations and further scrutiny was applied to variants near these cutoffs (Van der Auwera et al., 2013).

Additional hard filters were applied to reduce mapping and alignment errors (Treangen and Salzberg, 2012). Variants falling in known segmental duplication (SegDup) regions were excluded as they are known to map to multiple parts of the genome and can lead to alignment artifacts (Numanagić et al., 2018). Variants in low complexity repeat regions (LCR) were excluded as they are known to create ambiguity in read mapping and induce realignment errors (Li, 2014). Known SegDup and LCR regions were downloaded in BED file format, using the Table Browser tool hosted on the USCS Genome Browser, from the genomicSuperDups and Repeat Masker tracks respectively (Karolchik et al., 2004, Kent et al., 2002). Variants in these regions were then excluded using BCFtools. Additionally, variants in the *MUC* and *HLA* loci were also excluded as they are known to be highly polymorphic, have complex architecture, and are prone to structural variations which can confound accurate read mapping leading to an enrichment of false positive mosaic calls. A custom script from the EM-Mosaic tool was run to remove variant clusters, which were defined as two or more neighbouring variants located within 10bp of each other. This removed additional erroneous calls that were likely sequencing artefacts, misaligned read mapping, and WES library preparation errors (Hsieh et al., 2020). Finally, a maximum minor allele frequency (MAF) cutoff of less than 0.001, from the VEP annotation POPMAX_AF<0.001, was applied to restrict candidate variants to very rare disease causing across all populations within the gnomAD v2.1.1 database (Karczewski et al., 2020, Liu et al., 2020a).

i. Candidate proband variant quality control and filtering strategy

To identify high probability DNMs in affected probands, trio-VCFs were hard filtered to include only heterozygous variants in probands with no alternate read (ALT) support in both parents. All variants were required to have a minimum DP of 10 across the trio. A high sensitivity proband VAF threshold of 0.10≤VAF≤0.90 was set to capture candidate PZMs that would otherwise be missed by a standard filter of 0.30≤VAF≤0.70. VAFs were limited to >0.10 as it has been shown that calls below this threshold cannot be distinguished from known WES platform errors at average depths of 50X (Wright et al., 2019). Proband DNMs were required to have at least 5 ALT reads supporting a call, thereby enriching for higher probability PZMs, considering the expected rate of false positives introduced by WES library preparation and sequencing (Hsieh et al., 2020). To further reduce false positives, variants were required to have a maximum DDD-Africa cohort wide allele count (AC) of one in unrelated trios, and two in the sibling pair. A similar approach in singleton sequencing cohorts is achieved with cohort specific panel of normals filtering (Dou et al., 2018). A limitation to this approach is that potential reoccurring constitutive DNMs, which have been reported in larger DD cohorts, were excluded in this analysis (Wilfert et al., 2017). Finally, SNPs and INDELs that were called in intronic regions with no predicted impact on splicing were excluded.

ii. Candidate parental low- and high-level gonadosomatic quality control and variant filtering strategy

A second filtering strategy was deployed to identify constitutive heterozygous SNPs and INDELs in affected probands that were inherited from low-level and high-level parental gonadosomatic mosaicism. Trio-VCFs were hard filtered for high probability germline heterozygous variants with by requiring 0.27≤VAF≤0.73 in probands with at least 5 REF and 5 ALT reads (DP = 10). This VAF range was selected based recommendations from the DDD-UK study, in which variants with VAF>0.27 were more likely to be validated as germline DNMs (Wright et al., 2019). For gonadosomatic mosaicism in the mother, proband heterozygous variants were required to have a minimum of 2 ALT reads in the mother and 0 ALT read support in the father. For gonadosomatic mosaicism in the father, the reverse was applied. For low-level parental mosaics, no minimum DP or VAF threshold was required for the parent under

investigation, while a minimum DP of 10 reads was required for the second parent and the proband. For high-level parental PZM, a minimum DP of 10 was required across the trio with VAF>10 in the affected parent. The ALT read threshold was chosen at the recommendation of the DDD-UK mosaicism study, which reported that candidate low-level parental mosaic variants with 1 ALT read support were validated as true only 5% of the time, while those with 2 or more ALT reads were validated 42% of the time (Wright et al., 2019). Variants were required to have minimum DDD-Africa cohort-wide observed AC of 2 and a minimum AC of 3 for the twin sibling pair with a common corresponding parent. Finally, SNPs and INDELs that were intronic with no predicted impact on splicing were removed.

2.2.4.2 Statistical determination of mosaicism evidence

As true PZMs are present in only a fraction of the total cells tested in WES. their presence causes allelic imbalances that can be detected through VAF deviations in from an expected theoretical binomial distribution (Dou et al., 2018). Statistically significant deviations from an expected VAF outcome of 0.5 for putative heterozygous DNMs can be determined using a two-tailed binomial test. The DDD-UK study screened for PZMs using an unadjusted $p<0.0001$ from the two-tailed binomial test (sensitivity = 69%, specificity = 84%) with the number of observed ALT reads supporting evidence of mosaicism considered as a function of the total read depth (Wright et al., 2019). In WES datasets the variability in observed VAFs is much higher than expected for a binomial distribution (Hsieh et al., 2020). The variance in VAFs exceeds the mean VAF for that cohort and individual observed VAFs are spread out more than what you would expect at the total read depth under a binomial distribution (Hsieh et al., 2020). The EM-Mosaic computational tool accounts for this overdispersion using a beta-binomial distribution correction based on the observable VAFs from detected DNMs. It employs an expectation-maximization algorithm to detect mosaic SNPs based on VAF and total read depth and assigns a per-site posterior odds of a variant being mosaic compared to germline (Hsieh et al., 2020).

Thus, two statistical methods were employed to detect deviations of VAF from the expected distributions. Candidate mosaic SNPs in trio probands were screened using the EM-Mosaic computational tool (Hsieh et al., 2020). As a quality control step, EM-Mosaic calculated expected DNMs per sample controlling for cohort size and a

genome wide mutation rate, and samples with more DNMs than expected were excluded (Hsieh et al., 2020). Additionally, it deploys an FDR-based minimum ALT read threshold cutoff for the total read depth supporting a DNM call. This cutoff was calculated taking expected DNM mutation rates and sequencing error rates into consideration, and controls for FDR of 10% (Hsieh et al., 2020). The minimum FDR ALT read threshold removes additional variants that are likely to be sequencing errors. One major disadvantage to EM-Mosaic is that it can only be used to detect mosaic SNPs, as the authors report that the downstream parameter estimates for INDELs are unstable (Hsieh et al., 2020). It also cannot be used to detect parental gonadosomatic mosaicism without grandparent WES data. Thus, a standard two-tailed binomial test p-value was used to screen for candidate mosaic INDELS in probands, and parental gonadosomatic SNPs and INDELS.

iii. Screening for candidate proband PZMs

Candidate mosaic SNP and INDELs were screened for evidence of mosaicism in affected probands through two separate statistical approaches centred on VAF and total read depth. High quality *de novo* SNPs were run through the EM-Mosaic pipeline to screen for evidence of mosaicism (Hsieh et al., 2020). All candidate SNPs passed the minimum FDR-based ALT threshold QC set by EM-Mosaic, as they had at least 5 ALT reads supporting the call. Of the probands run through the tool, five were determined to be outlier samples with excessive *de novo* SNP candidate counts (Table 2.1). These five samples were excluded from further analysis proband PZM variant prioritization. The remaining SNPs were stratified into likely mosaic and likely germline *de novo* by EM-Mosaic by their posterior odds scores. Candidate mosaic SNPs were required to have a posterior odds ≥10 to be considered likely mosaic, to restrict the FDR to 9.1% (Hsieh et al., 2020). SNPs with posterior odds <10 were considered germline *de novo*. Both candidate mosaic and germline *de novo* SNPs were carried forward to clinical curation.

Table 2.1: Probands with excessive DNM counts flagged by EM-Mosaic.

Proband ID	Total *de novo* SNPs (cohort %)	Total *de novo* INDELs (cohort (%)
11	80 (13.45)	3 (6.93)
29	100 (16.81)	4 (9.30)
33	76 (12.77)	2 (4.65)
70	89 (14.96)	1 (2.33)
96	68 (11.43)	4 (9.30)
Total (% of cohort)	**413 (69.41)**	**14 (32.56)**
Mean DNMs [SD]	**82.60 [11.02]**	**6.51 [2.71]**

SD = standard deviation, DNM = *de novo* mutation

Candidate INDELs could not be run through the EM-Mosaic pipeline due to INDEL downstream parameter instability hampering overdispersion correction (Hsieh et al., 2020). The EM-Mosaic QC statistics were built on prior probabilities for mosaic SNPs that cannot be carried over for INDELs. Therefore, INDELs were tested separately from SNPs using the approach published by DDD-UK (Wright et al., 2019). The same 5 outlier samples that were excluded from the mosaic SNP analysis were excluded from the INDEL analysis for consistency (Table 2.1). To detect candidate mosaic INDELs in affected probands, a two-tailed binomial test p-value was calculated for each INDEL VAF centred around an expected 0.5. Binomial p-values were FDR adjusted using the Benjamini-Hochberg method, and an FDR<0.1 cutoff was chosen to account for 10% false positives due to multiple testing. INDELs with FDR adjusted binomial p-values<0.1 were considered significant, and likely mosaic candidates. INDELs with FDR adjusted p≥0.1 were considered germline *de novo*. Following DDD-UK recommendations, INDELs with unadjusted binomial p-values <0.0001 were considered high confidence mosaics (Wright et al., 2019). Both candidate mosaic and confirmed *de novo* INDELs were carried forward to clinical curation.

iv. Screening for low- and high-level parental gonadosomatic PZMs

No binomial test was applied to candidate low-level parental VAFs. Variants inherited from mothers and fathers were analysed separately. The five samples that were identified as outliers in the proband PZM pipeline were not excluded in this pipeline (Table 2.1). Variants with insufficient evidence of low-level parental gonadosomatic mosaicism were excluded from further analysis and clinical curation.

To detect candidate high-level parental mosaicism, parental VAF deviations from an expected 0.5 were detected by calculating a binomial test p-value for each candidate

SNP or INDEL. Binomial p-values were FDR adjusted using the Benjamini-Hochberg method, and an FDR<0.1 cutoff was chosen to account for 10% false positives due to multiple testing. SNP and INDELs with parental FDR adjusted binomial p-values <0.1 were considered significant, and likely high-level parental PZM candidates. Following DDD-UK recommendations, variants with unadjusted binomial p-values <0.0001 were considered high confidence parental mosaics (Wright et al., 2019). SNPs and INDELs were analysed separately. Variants inherited from mothers and fathers were also analysed separately. The five samples that were identified as outliers in the proband PZM pipeline were not excluded in this pipeline (Table 2.1).

This study overlaps with a small variant analysis previously undertaken by members of the DDD-Africa research study. The DDD-Africa analysis used a parallel pipeline to detect pathogenic germline DNMs in the same 64 WES trios as part of a separate PhD. Identification of previously detected germline DNMs with this pipeline served as positive validation for the DNM results presented in this thesis. Previously identified DNMs that were considered constitutional DNMs in parallel analysis with new evidence of mosaicism in this analysis were flagged for future validation to confirm VAF. A table of 16 variants identified in parallel analysis can be found in Appendix D.

2.2.4.3 Candidate variant interpretation and clinical curation

Two virtual gene panels were used to prioritize candidate variants. The first panel was genes known to be associated with monoallelic DDs in the curated Development Disorder Genotype - Phenotype Database (DDG2P) database (https://www.ebi.ac.uk/gene2phenotype/) (Thormann et al., 2019). A list of 1025 genes was extracted from the full set of known 2362 DDG2P genes reported in the database as of 20-12-2023 (https://www.ebi.ac.uk/gene2phenotype/) (Thormann et al., 2019). In the DDG2P database, “monoallelic” genes are that where pathogenic variation in a single allele is sufficient to cause disease, regardless of mechanism (loss-of-function, dominant-negative, or gain-of-function), and serves as a mode of inheritance descriptor (Thormann et al., 2019). Dominant refers more broadly to the clinical inheritance pattern observed in families. A secondary panel of candidate genes was selected to identify variants in genes that are linked DD cases but are not in the DDG2P database due to limited evidence or incomplete curation. Additionally, this filter was selected to identify potential novel dosage sensitive genes in the DDD-Africa cohort

that may not currently be associated with classical DD phenotypes in the DECIPHER database. This secondary gene panel consisted of non-DDG2P genes with significant evidence of loss-of-function (LOF) intolerance, that was defined from the estimated probability of loss-of-function intolerance score (pLI) >0.9 from the gnomAD V4.1.0 database. A total of 2365 non-DDG2P overlapping genes with annotated pLI >0.9 were extracted from the most up-to-date gnomADv4.1.0 database (Karczewski et al., 2020, Chen et al., 2024). LOF genes with functionally relevant candidate variants were furthered assessed for links to DD phenotypes in the literature. Heterozygous variants with evidence of mosaicism in DDG2P biallelic genes, associated with recessive DDs, were also identified. Manual curation of the affected trio-VCF was carried out to identify potential compound heterozygote candidates inherited from a parent to account for potential two-hit cases. However, as no recessive two-hit candidates were identified, these variants were not taken forward or reported here.

Candidate variants were then visualised with the Integrative Genomics Viewer (IGV) tool (v2.16.0) across all members of the trio for further quality control (Robinson et al., 2011). Sites were evaluated for evidence of alignment errors, base and read quality confirmation, strand bias, errors due to tail end proximity in supporting reads, and errors of soft clipping reads. Variants failing the IGV inspection were excluded. The resulting candidate variants in DDG2P or LOF-intolerant genes were clinically curated with Ensembl VEP canonical transcript annotations, genotype-to-phenotype associations, previous ClinVar classification, and a suite of calibrated computational *in silico* tool predictions (Pejaver et al., 2022, Liu et al., 2020a). Functional predictions were interpreted in line with the published and updated American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP) guidelines for sequenced variants (Richards et al., 2015). Additional guidance on interpretation and application of ACMG/AMP codes was obtained from the Association of for Clinical Genome Science (ACGS) Best Practice Guidelines for Variant Classification in Rare Disease 2024 framework, which aggregated results from recent ACMG/AMP working groups and variant curation expert panels into a unified and updated guidelines for clinical curation based on ACMG/AMP classification (Durkie et al., 2024).

The ACMG/AMP guidelines recommend determining direct support for functional prediction from multiple lines of evidence including population frequency data, established disease databases, computational *in silico* predictions, functional study

results, disease and gene specific case data, and segregation data (Richards et al., 2015). Each source is assigned a corresponding classification code that indicates evidence for variant pathogenicity or benignity. Classification codes are further stratified by strength of evidence as: supporting, moderate, strong, very strong or standalone (Richards et al., 2015). Classification codes can be combined and adjusted to enhance prediction of pathogenicity or benignity that have been adapted for specific diseases or genes (Richards et al., 2015, Tavtigian et al., 2018). The guidelines have since been updated to include a system that stratifies strength of evidence as a scale of additive points that fit a Bayesian formulation (Tavtigian et al., 2020). Codes applied to a variant are assigned a score on this scale, based on supporting evidence of pathogenicity or benignity, which are added together cumulatively. This cumulative score is proportionate to established posterior probability boundaries of pathogenicity prediction which was used to assign predicted variant functional consequence as ether benign, likely benign, uncertain significance (VUS), likely pathogenic, and pathogenic (Tavtigian et al., 2020). Rare variants that were precited to be LOF in candidate genes were considered high priority for clinical curation and validation recommendation. LOF variants included those that were predicted to be nonsense SNPs, frameshift INDELs, or variants impacting the canonical splice-sites within 2bp of an exon-intron junctions. These variants were assigned putative PVS1 codes following updated recommendations on LOF variant interpretation published by the ClinGen Sequence Variant Interpretation Working Group (Abou Tayoun et al., 2018). The updated recommendations were implemented by assessing each candidate LOF variant with the automated PVS1 strength classifier AutoPVS1 (http://autopvs1.genetics.bgi.com/) (Xiang et al., 2020).

All variants were required to be rare in initial the filtering (POPMAX_AF <0.001). A PM2-moderate code was applied for novel variants after confirmation of annotated MAF being absent from the gnomAD V4.1.0 database. PM2-moderate was also applied where the variant specific gnomAD V4.1.0 GroupMax filtering allele frequency was lower than a gene specific MAF threshold calculated from reported prevalence, penetrance, allelic and genetic heterogeneity estimates (Chen et al., 2024). Gene specific MAFs were calculated with data from OMIM and DECIPHER with the recommended online tool (https://cardiodb.org/allelefrequencyapp/) (Whiffin et al., 2017). The ACMG/AMP guidelines recommend downgrading PM2 from moderate to supporting to avoid double counting with PP3 when *in silico* tools incorporate

population allele frequencies into their models. However, a report by the ACGS determined that this approach would be overly conservative in cases where PP3 is also capped at supporting level (Durkie et al., 2024). They then showed that employing empirically calibrated *in silico* prediction thresholds to dynamically determine PP3 levels up to strong, only mitigated some of the impact (Durkie et al., 2024, Pejaver et al., 2022). They conclude by recommending keeping PM2 at moderate level, until a planned review and overhaul of scoring PM2 is completed, and utilizing calibrated PP3 thresholds from models that were not trained with population allele frequency (Durkie et al., 2024). This study followed this ACGS guideline for both PM2 and PP3.

Evidence of the pathogenicity of missense variants was assessed by *in silico* tool prediction and PP3/BP4 scores. Recent advances have converted *in silico* tool outputs into quantitative thresholds aligned with the ACMG/AMP Bayesian framework, allowing calibration of PP3 and BP4 evidence strengths (Tavtigian et al., 2020). Following these recommendations, a single predictor with empirically calibrated thresholds was employed to classify variants into benign-supporting, uncertain, pathogenic-supporting, pathogenic-moderate, or pathogenic-strong categories (Table 2.2) (Tavtigian et al., 2020). This approach was chosen for two reasons. First, it prevented double counting of population allele frequencies between PM2 and PP3, allowing PM2 to be retained at moderate strength (Pejaver et al., 2022). Second, it avoided biases from tools trained on allele frequency data from outdated population databases with limited African representation, which could confound predictions in the predominantly African ancestry DDD-Africa cohort (Bope et al., 2019).

Table 2.2: Calibrated *in silico score* thresholds selected for this study.

Tool	B.Sp.	VUS	P.Sp.	P. Md.	P. St.	Reference
BayesDelnoAF	≤ −0.18	-0.18 - 0.13	≥ 0.13	≥ 0.27	≥ 0.50	(Pejaver et al., 2022)
REVEL	≤ 0.290	0.290 - 0.644	≥ 0.644	≥ 0.773	≥ 0.932	(Pejaver et al., 2022)
CADD	≤ 22.7	22.7 - 25.3	≥ 25.3	≥ 28.1	-	(Pejaver et al., 2022)
SpliceAI Δ	< 0.1	0.1 - 0.2	≥ 0.2	≥ 0.5	≥ 0.8	(Logan et al., 2023)

B = benign; P = pathogenic; Sp. = supporting; Md. = moderate; St. = Strong; VUS = variant of uncertain significance.

The ensemble machine-learning algorithm BayesDel, that integrates predictions from nine tools into a naïve Bayesian framework trained on curated pathogenic and neutral variant datasets, was chosen as the primary *in silico* predictor for scoring missense SNPs with PP3 (Feng, 2017). It has two modes: BayesDeladdAF, which incorporates population allele frequency, and BayesDelnoAF, which excludes it. BayesDel has consistently been ranked among the top ensemble *in silico* tools in terms of accuracy, sensitivity, and specificity (Anderson and Lassmann, 2022, Borges et al., 2021, Tian et al., 2019). BayesDeladdAF was ranked the top *in silico* predictor of nonsense and missense SNP deleteriousness in abnormalities of metabolism, the immune system, and nervous system (Anderson and Lassmann, 2022). While not as accurate as the model that included allele frequencies, BayesDelnoAF still performed exceptionally well in these scenarios and was the top predictor out if the tools that do not incorporate allele frequencies (Anderson and Lassmann, 2022). Variants were annotated with BayesDel scores using Ensembl VEP dbNSFP v4 and interpreted using published calibration thresholds that map BayesDelnoAF outputs onto the spectrum of PP3/BP4 evidence strengths (Table. 2.3) (Pejaver et al., 2022, Liu et al., 2020a, McLaren et al., 2016).

Variants were also annotated with scores from REVEL and CADD (v1.6), two top meta-predicting tools (Ioannidis et al., 2016, Rentzsch et al., 2021). REVEL was included as a secondary predictor for comparison as it combines predictions from 13 algorithms in a random forest model without incorporating population allele frequency data (Ioannidis et al., 2016). The third tool CADD uses a machine-learning model that incorporates scores from a more than 60 annotations into a logistic regression model to predict deleteriousness (Kircher et al., 2014). The 1.6 version is significantly improved for determining the effects of splicing variants (Rentzsch et al., 2021). However, these additional scores were not used for PP3 scoring and were included for comparison purposes in cases of uncertainty (Table 2.2).

The ACMG/AMP PM4 code was used as evidence for in-frame INDELs predicted to disrupt protein length in evolutionary conserved non-repeat regions. A PM4-moderate code was applied in cases where INDELs impacted multiple amino acid. Scores from the *in-silico* tool MutationTaster2021 were generated for INDELs of interest as well, as further evidence of pathogenicity (Cannon et al., 2023, Durkie et al., 2024, Steinhaus et al., 2021). Benchmarking has shown that MutationTaster2021 predicted in-frame INDEL pathogenicity in DD cohorts with high sensitivity and specificity. However, it

could only be applied at a moderate strength to support PM4 evidence in line with recommended guidelines (Cannon et al., 2023, Durkie et al., 2024, Steinhaus et al., 2021). In-frame INDELs occurring in repetitive regions were not scored with PM4, whereas those predicted to disrupt splice sites were instead assessed with SpliceAI (Jaganathan et al., 2019). The impact of SNPs and INDELs on non-canonical splice sites was assessed with SpliceAI, a deep neural network *in silico* predictor of global splice site alteration trained on GENCODE isoform data (Jaganathan et al., 2019, Strauch et al., 2022). SpliceAI is the top performing *in silico* predictor of splice site disruptions and is particularly accurate for non-canonical splicing variants (Jang et al., 2022, Rowlands et al., 2021, Riepe et al., 2021). Candidate variants with predicted splicing impact scores Δ≥0.2 were considered PP3 at supporting level, while those with Δ≥0.5 and Δ≥0.8 were considered PP3 at moderate and levels, respectively (Table 2.2) (Logan et al., 2023). Variants affecting canonical splice sites were interpreted according to PVS1 criteria (Logan et al., 2023).

Evidence for a causal relationship between genotype and phenotype was evaluated using the ACMG/AMP codes PS2 and PM6, which assess *de novo* status in conjunction with confirmation of parentage, as well as specificity of phenotype explained. Because the dataset was derived from trio-WES, parentage of probands was directly confirmed, allowing application of PS2 to both DNMs and postzygotic mutations. Variants with evidence of gonadosomatic mosaicism were assigned a lower strength PM6 –moderate or supporting score. In these cases, the candidate variant was assumed to have arisen as a *de novo* PZM in the parents, which could not be confirmed due to our inability to test parental germline cells. The strength of PS2/PM6 evidence was further refined by considering the degree to which the variant accounted for the proband's reported phenotype from medical records. Variants associated with a high degree of overlapping phenotypes to those in the trios were assigned PS2-strong or PM6-moderate, whereas variants explaining only a subset of the reported clinical features had their evidence level reduced accordingly.

Phenotypic specificity was evaluated subjectively, based on established genotype–phenotype correlations for the gene in question as well as the individual's clinical presentation. This component of the analysis was captured under the PP4 criterion. In presumed cases of parental gonadosomatic mosaicism where the parent exhibited a consistent, overlapping phenotype with the proband, a PP1-supporting score was also applied, reflecting evidence of familial segregation from a single meiotic transmission

event. By contrast, PP1 was not applied when the putative mosaic parent was clinically unaffected, given the absence of segregation evidence.

Finally, additional functional evidence was integrated from multiple resources. The online Varsome platform was used to rapidly annotate features such as protein length alteration, mutational hotspots, critical protein domains, and gene constraint metrics (Kopanos et al., 2019). Findings were validated through assessing data on each candidate from DECIPHER, gnomAD, ClinVar, UniProt, and OMIM databases (Gudmundsson et al., 2022, Hamosh et al., 2005, Landrum et al., 2014, The UniProt, 2023, Firth et al., 2009). Currently, *in silico* tools are very poor predictors of gain-of-function (GOF) pathogenicity (Stein et al., 2023). In cases where the mechanism of pathogenicity was through GOF, the LoGoFunc *in silico* tool was used assess pathogenicity (Stein et al., 2023).

The functionality of affected genes was assessed through online queries to the UniProt, DECIPHER, ClinGen and OMIM databases (Hamosh et al., 2005, Rehm et al., 2015, The UniProt, 2023, Firth et al., 2009). Associations of impacted genes with DDs was determined through queries primarily to the DECIPHER database, followed by ClinGen curation reports, OMIM database reports, and case reports in the literature. Genes linked known disorders were prioritized through manual assessments of these reports and comparison to the affected participants DD presentation. Information on the mode of inheritance, reported variable expressivity, and penetrance of phenotypes in related disorders was noted. Trio phenotypic data was extracted from the DECIPHER and REDCap databases in the form of curated HPO terms, which were then manually compared to the reported phenotypes of linked gene disorders. If a plausible overlap of one or more phenotypes was found, the associated variant was shortlisted for deeper investigation. For the secondary set of variants in non-DDG2P LOF intolerant genes queries to OMIM were prioritised, followed by extraction of associated phenotypes in open access DECIPHER patients with variants in the same gene, and a brief review of new literature in which new links to phenotypes may have been made.

The combination of evidence from these approaches was used to assess if candidate variants explained DD pathology in part, in combination, or in full. The evidence was interpreted in the context of the candidate variants being standalone mosaic, germline *de novo*, inherited parental mosaic or in combination. In the cases where phenotypes were positively correlated with a candidate variant, a putative PS2 code was applied to confirmed DNM and PZM variants with strength proportionate to the phenotypes

explained. PS2 codes could not be assigned to inherited parental PZMs. In affected parents, segregation rules were applied. In unaffected parents the variant was treated as if it was *de novo* in the proband. The combined variant identification and curation strategies were used to generate a candidate list of putative somatic mosaic variants and DNMs in trio probands, and parental gonadosomatic variants inherited as constitutional heterozygous variants in trio probands that may be contribute to DD phenotypes. Variants were considered good candidates if the met the likely pathogenic or pathogenic criteria during curation. These variants were taken forward as candidates for future validation and clinical curation by the DDD-Africa multidisciplinary team. An overview of the variant prioritization strategies used in this thesis chapter is summarized in Figure 2.1.

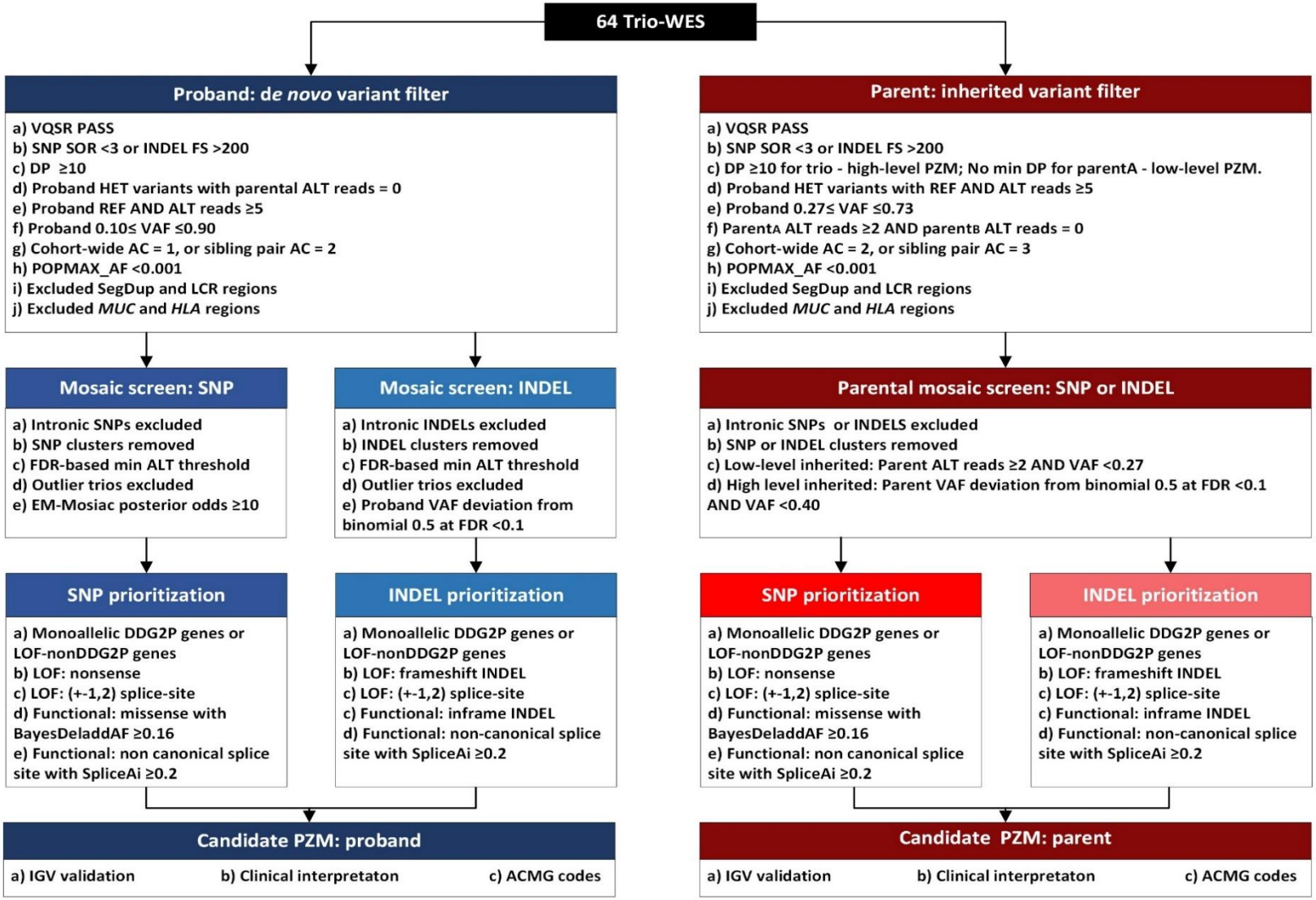


Figure 2.1: Workflow of variation prioritization strategies deployed to identify candidate mosaic variants from trio-WES in the DDD-Africa cohort.

2.3 Results

2.3.1 Study demographics

A summary of the 62 trios can be seen in Table 2.3. Of the 62 probands, 37 (59.68%) were males and 25 (40.32%) were females. The ages of affected probands ranged from 0-14 years at time of recruitment (Mean = 5.87, SD = 3.29). There was no significant difference in the distribution of ages of recruitment between male and female probands (p=0.32). The mothers of affected probands ranged in age from 17-40 years (Mean = 29.25, SD = 5.55) at the time of birth. The fathers of the affected probands ranged in age from 21-58 years (Mean = 34.18, SD = 6.52) at the time of birth. On average, the fathers were older than the mothers at the time of birth ($p = 1.32X10^{-05}$). Only three (4.84%) of the probands were born outside South Africa. A total of 58 (93.55%) probands self-reported as black or African ancestry, one proband was reported as multi ancestry (self-reported as Coloured in the Southern African context), one proband was reported as Indian ancestry, and two probands were reported as white or European ancestry.

Table 2.3: Summary of DDD-Africa demographics

Proband sex	**N (%)**
Male	37 (59.68)
Female	25 (40.32)
Mean Age (years)	**Mean (SD)**
Proband at recruitment	5.87 (3.29)
Mother at birth	29.25 (5.55)
Father at birth	34.18 (6.52)
Country of birth	**N (%)**
South Africa	57 (91.94)
Other Africa	3 (4.83)
Unknown	2 (3.23)
Self-reported proband ancestry	**N (%)**
Black	58 (93.55)
White	2 (3.23)
Coloured	1 (1.61)
Indian	1 (1.61)
Total	62 (100)

Probands presented with an array of diverse phenotypes at time of recruitment which are summarized in Table A. 2 in Appendix E. A significant proportion of the probands

(40.32%) had prior confirmed genetic diagnoses by parallel DDD-Africa analysis (Table 2.5). There was no significant difference in the proportion of male and female participants that had a prior confirmed diagnosis (p = 0.27). Additionally, the mean age of the probands at recruitment with no confirmed diagnoses was not significantly different from the mean age of those with prior diagnoses (p = 0.58). However, the mean ages of both the mothers (p = 0.04) and fathers (p = 0.01) of those with prior genetic diagnoses were significantly higher than those with unsolved cases.

Table 2.4: Probands with prior genetic diagnoses.

Prior diagnoses	**Yes (N = 25)**	**No (N = 37)**
Proband sex	**N (%)**	**N (%)**
Male	17 (45.95)	20 (54.05)
Female	8 (32.00)	17 (68.00)
Total	25 (40.32)	37 (59.68)
Age	**Mean (SD)**	**Mean (SD)**
Proband at recruitment	5.88 (2.80)	5.86 (3.63)
Mother at birth	31.08 (5.51)	27.98 (5.28)
Father at birth	36.17 (5.22)	32.86 (6.59)

N = count; SD = standard deviation

2.3.2 Proband variant prioritization results

The results of the proband variant prioritization are summarized in Figure 2.2. An initial QC filtering strategy was employed to detect rare coding *de novo* SNPs and INDELs in affected probands. This filtering strategy was expanded to sensitively include mosaic SNPs (m-SNPs) and INDELs (m-INDELS), while stringently controlling for sequencing artifacts. Five trios were excluded as they were determined to have an excess of DNMs by EM-Mosaic which was indicative of either contamination or sample handling errors (Table 2.1). The proband *de novo* variant filter resulted in 182 *de novo* SNP candidates in 53 probands, and 29 *de novo* INDEL candidates in 22 probands. Probands from families 36, 58, 84, 88, 102, and 116 had no detectable *de novo* SNP candidates passing initial QC. Additionally, probands from families 84, 88, 102, and 116 had no detectable *de novo* INDEL candidates passing QC. Thus, no candidate proband DNMs were detected in these four trios. A total of 40 probands had no detectable *de novo* INDELs passing QC. This equated to an initial detection rate of 3.43 *de novo* SNPs per proband and 1.32 *de novo* INDELs per proband.

1

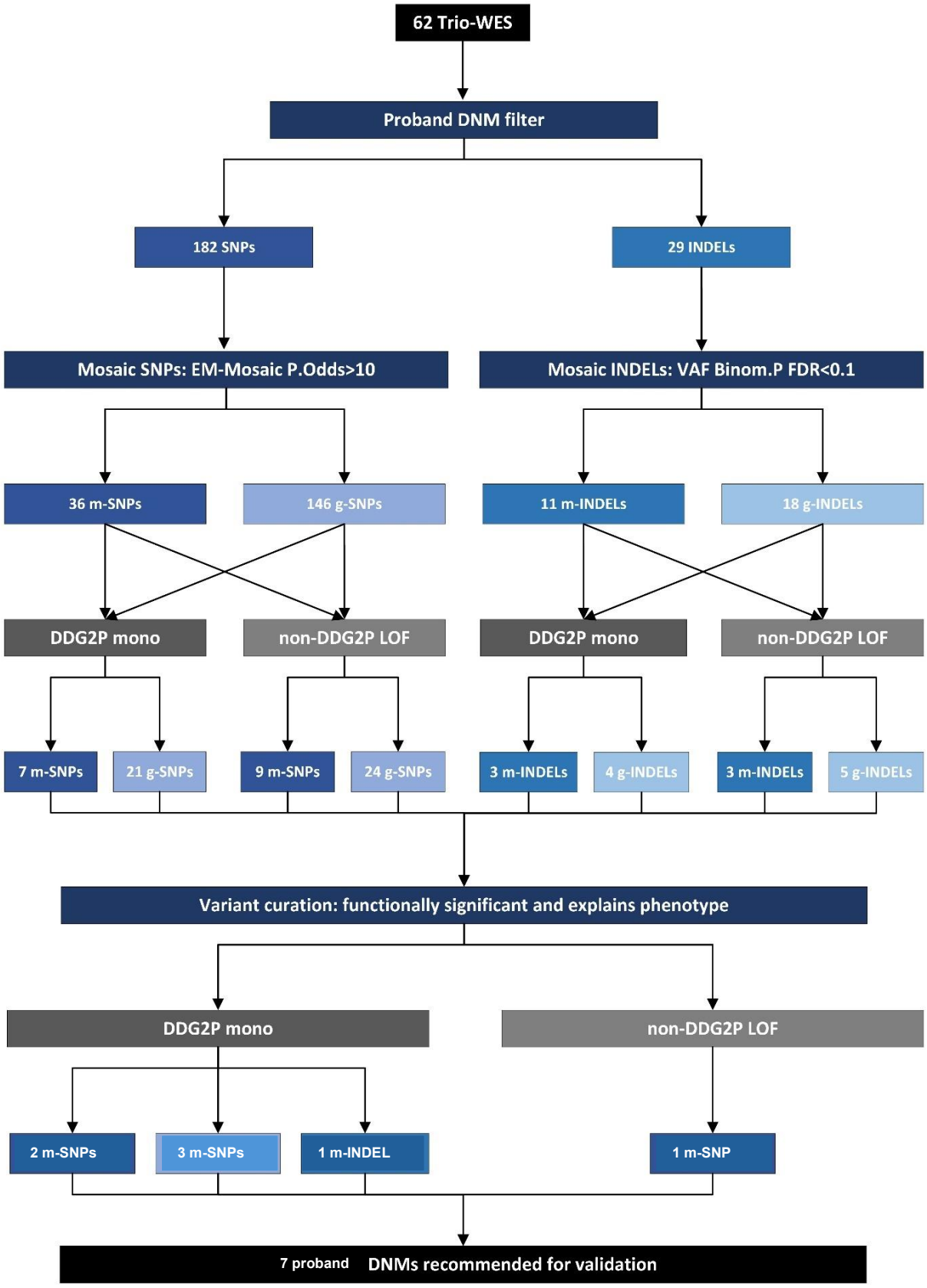


Figure 2.2: Results from the proband *de novo* variant prioritization pipeline for 62 DDD-Africa trios.

m- = mosaic, g- = germline, P.Odds = EM-Mosaic posterior odds, Binom.P = p value from binomial test, mono = monoallelic genes, LOF = loss-of-function intolerant genes.

The 182 candidate *de novo* SNPs were used as inputs for EM-Mosaic and assessed for evidence of mosaicism. All 182 candidate SNPs passed the FDR-based minimum ALT allele filter for total read depth (DP), which further strengthened evidence against being sequencing errors or other artefacts (Figure 2.3).

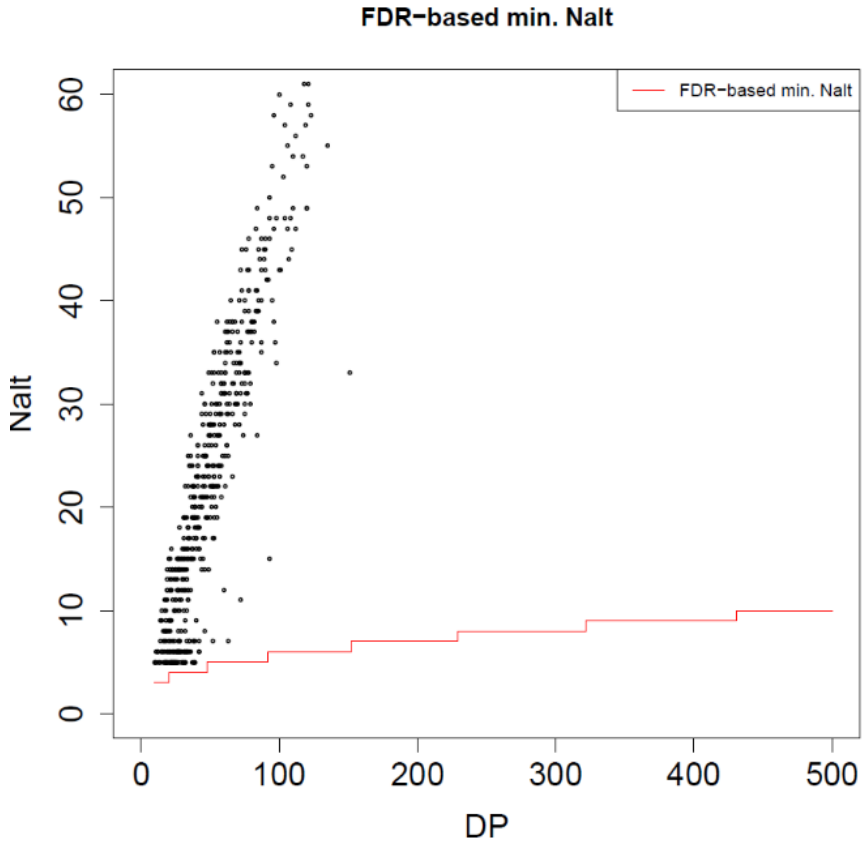


Figure 2.3: Candidate SNPs passing the EM-mosaic FDR-minimum alternate read depth threshold.

Alternate read support for each variant is plotted against the corresponding total read depth at that locus. The red line indicates the minimum alternate reads required to minimize artifact error false positives. DP = total read depth. Nalt = alternate read. FDR = false discovery rate

The VAF distribution of the 182 candidates indicated an initial mosaic variant fraction of 51.55% (Figure 2.4). The initially high estimate of mosaic variant fraction was likely inflated by germline DNMs whose VAF deviations, caused by technical artifacts, were incorrectly classified as mosaic under binomial assumptions, resulting in an overdispersion of VAF distribution (Figure 2.5).

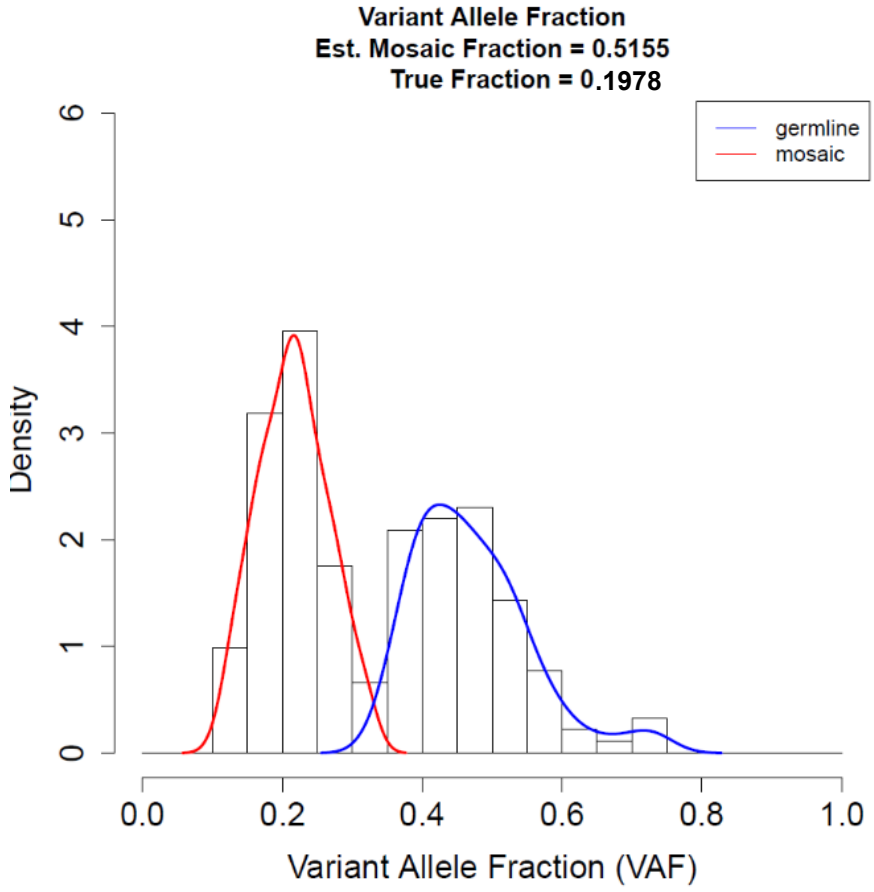


Figure 2.4: EM-Mosaic generated variant allele fraction density plots of likely mosaic and likely germline candidate SNPs.

The relative counts of SNPs in defined variant allele fraction bins. EM-mosaic estimates a prior probability from these distributions which is inputted into the EM algorithm.

Overdispersion in the data was addressed by EM-mosaic through modelling the VAF distribution using a beta-binomial framework. The degree of overdispersion (θ) was estimated directly from the observed variance in VAFs among candidate SNPs. As shown in Figure 2.5, the study-specific beta-binomial model provided a substantially better fit to the observed VAF variance across total depth compared to a standard binomial model. A theta value of 33.76 was derived from the candidate SNP data and used to correct for overdispersion. This corrected model was then employed by the expectation-maximization algorithm to iteratively update the posterior probability that each SNP was mosaic rather than germline *de novo*, using individual VAF observations deviating from the expected 0.5.

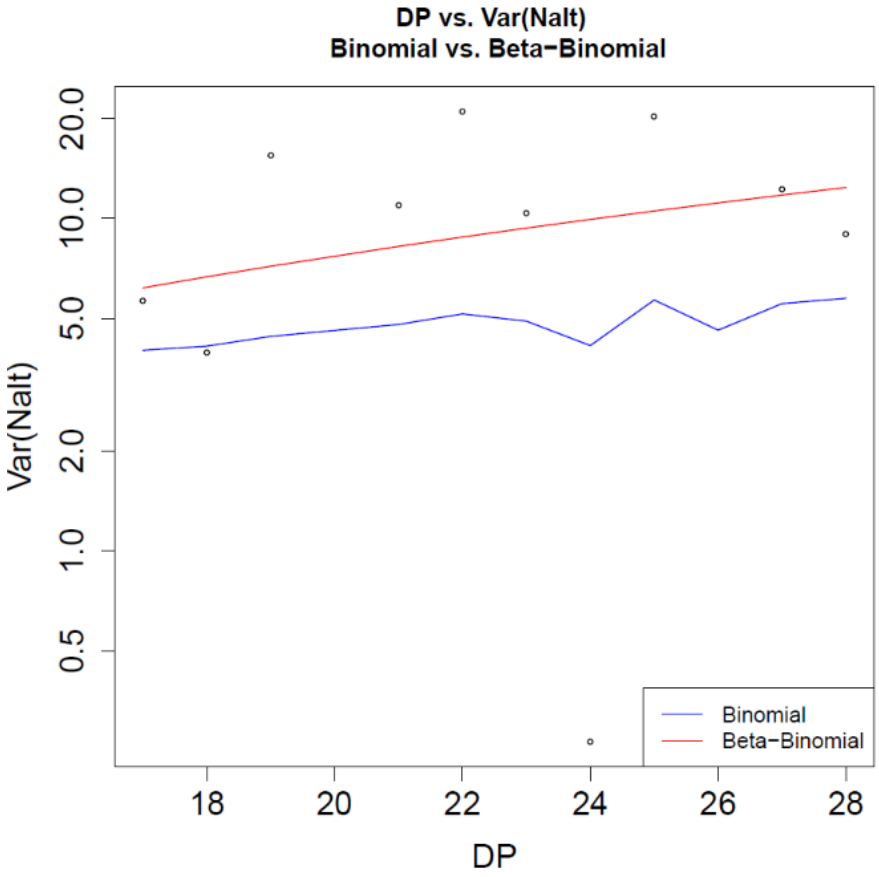


Figure 2.5: Overdispersion plot of calculated VAF variance across observed read depth values.

A higher variance in the VAF of observed VAFs of candidate SNPs was observed given a binomial model (blue). A beta-binomial model (red) was fitted to better account for this variance and was used to correct prior mosaic probabilities and overall mosaic fraction through the EM-Mosaic expectation-maximization algorithm. DP = read depth. Nalt = alternate read. Var = variance.

Using a stringent posterior odds threshold of >10, the EM-Mosaic expectation-maximization algorithm classified 36 (19.78%) of the candidate SNPs as likely mosaic and 146 as likely germline *de novo* SNPs (Figure 2.6). This refined the estimated proportion of m-SNPs to 19.78% under the beta-binomial framework (Figure 2.4). The high-confidence m-SNPs had an average VAF of 0.16 (SD = 0.02), while the germline *de novo* variants exhibited a higher mean VAF of 0.38 (SD = 0.13). Interestingly, among the 146 germline calls, 52 (35.62%) displayed posterior odds >1 and a lower average VAF of 0.24 (SD = 0.03), suggesting possible mosaicism (Figure 2.6A, lower left quadrant). Although these variants may represent true mosaics, they did not meet the high-confidence threshold due to limited read support. Each was further reviewed for borderline posterior odds (just below 10) and evaluated alongside their sequencing

depth, especially in cases with potential clinical relevance. Improved read depth at these loci may ultimately clarify their mosaic status.

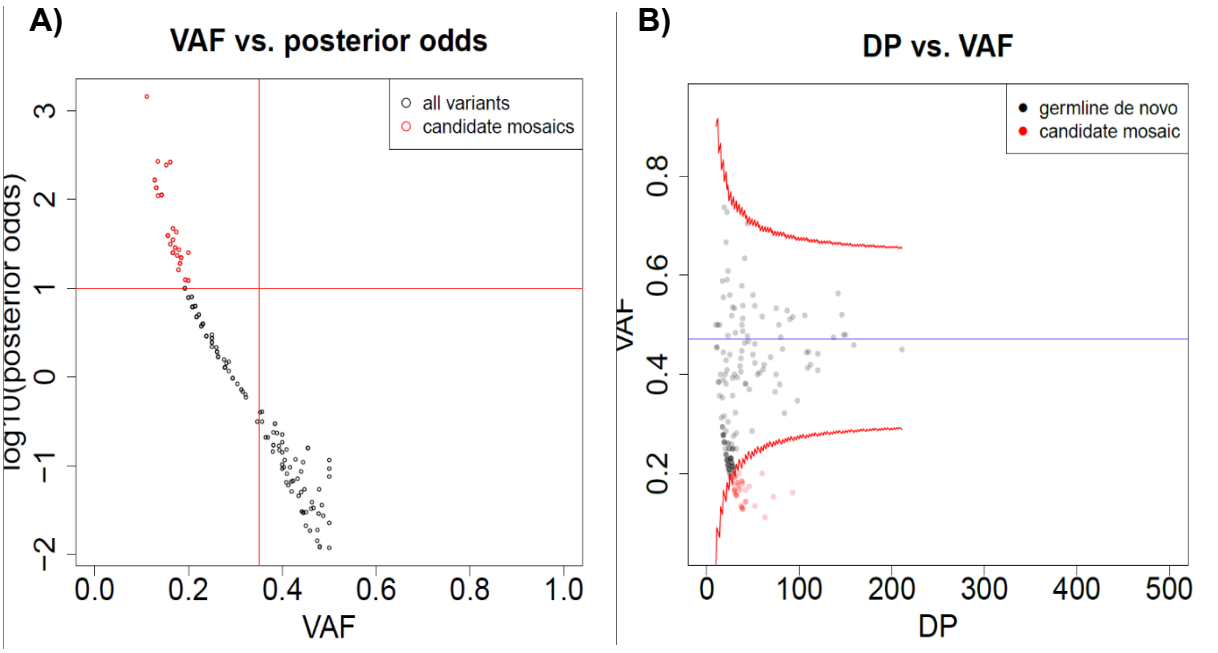


Figure 2.6: Candidate proband SNPs predicted to be mosaic versus germline *de novo* by EM-Mosaic with posterior odds >10.

A) Candidate SNP VAF with corresponding posterior odds of being mosaic versus germline under a beta-binomial model. SNPs predicted to be high confidence mosaic are red and germline de novo SNPs are in black. The red lines indicate the posterior odds cutoff of 10 on the x-axis and a VAF cutoff of 0.35 on the y-axis, B) The read depth versus VAF distribution of each candidate SNP. The blue line denotes a mean VAF of 0.49. The red lines denote a 95% confidence interval of the beta-binomial model. SNPs predicted to be high confidence mosaic are red and germline de novo SNPs are in black DP = read

The depth of sequencing coverage for each candidate SNP substantially influenced the ability to confidently identify evidence of mosaicism. At the average WES depth of ~45X in the DDD-Africa cohort, the estimated power to detect m-SNPs with a variant VAF of 0.2 and posterior odds >10 was only around 33% (Figure 2.7). This analytical approach had even less sensitivity for variants with a VAF of 0.1 and for distinguishing false positive *de novo* calls near a VAF of 0.27. EM-Mosaic modelling indicated that increasing sequencing depth would significantly improve detection power across all VAF levels (Figure 2.7). To achieve 95% power for detecting mosaic SNPs using this method, an average WES depth of over 300X would be required. As such, the actual burden of m-SNPs in the probands is likely underestimated, and the current variant set

may include an inflated number of false positive *de novo* SNPs due to the limited sequencing depth.

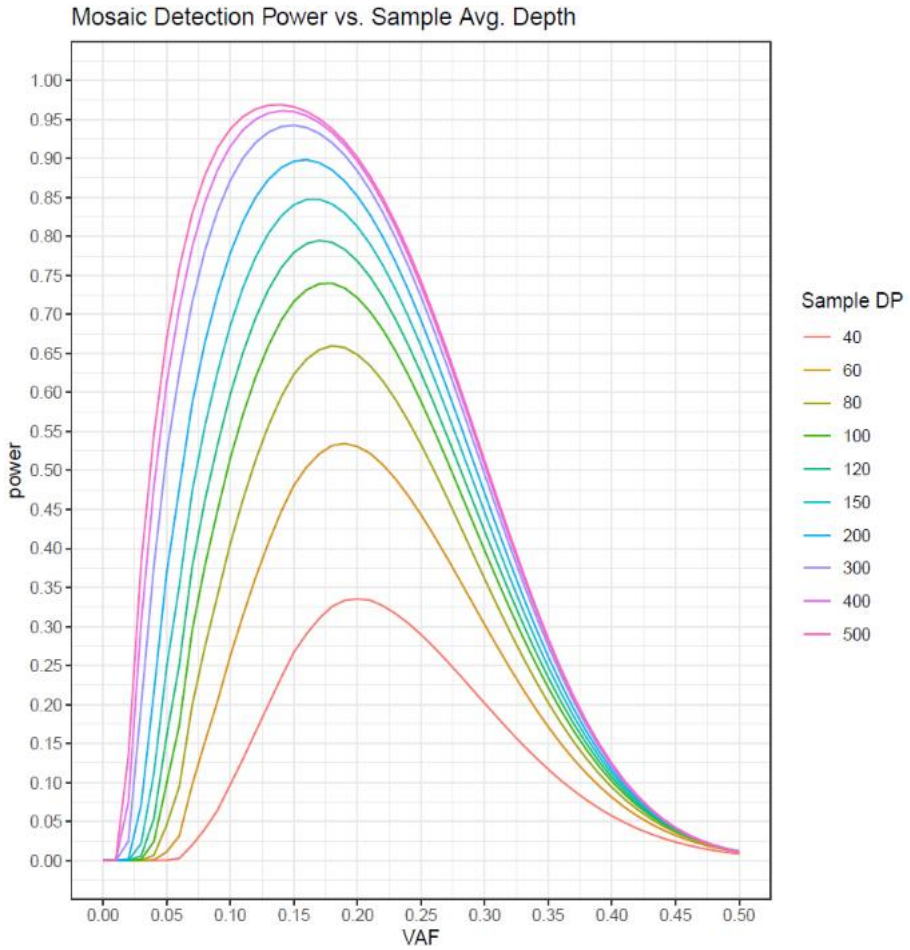


Figure 2.7: Mosaic detection power curves.

The estimated power to detect mosaic SNPs at various VAFs as a function of average read depth from this studies data. DP = read depth. VAF = variant allele fraction.

Evidence of mosaicism in candidate INDELs could not be assessed using EM-Mosaic. The VAF distributions of candidate INDELs were assessed for significant deviations from an expected heterozygous 0.5 with two-tailed binomial tests. Candidate INDELs with an FDR-adjusted p-value below 0.1 were considered to show suggestive evidence of mosaicism. A more stringent threshold of the unadjusted p-value<0.0001 was used to identify higher-confidence m-INDELs. Among the 29 candidate INDELs analysed, 11 (37.93%) met the FDR threshold, with a mean VAF of 0.16 (SD = 0.05). Eight of these met the more stringent unadjusted p-value threshold and were classified as higher-confidence m-INDELs. In contrast, the candidate INDELs presumed to be germline *de novo* exhibited a higher average VAF of 0.42 (SD = 0.07). Although a

greater proportion of candidate INDELs were flagged as mosaic compared to candidate SNPs (37.93% vs. 19.78%), the overall VAF distributions of these two variant types did not differ significantly (p = 0.65). After classifying SNPs and INDELs into likely mosaic or germline *de novo* categories, adjusted estimated mutation rates were calculated for the proband cohort. On average, each proband harboured 2.75 germline *de novo* SNPs and 0.82 germline *de novo* INDELs, along with 0.67 m-SNPs and 0.5 m-INDELs.

These combined candidate variants were then prioritized for future validation through clinical curation and variant interpretation. To focus on variants with the highest likelihood of pathogenicity, both mosaic and germline *de novo* candidates were filtered using two virtual gene panels, monoallelic DDG2P genes, and non-DDG2P LOF genes (Figure 2.2). This filtering approach identified 7 m-SNPs, 3 m-INDELs, 21 germline *de novo* SNPs, and 4 germline *de novo* INDELs in DDG2P monoallelic genes. An additional 9 m-SNPs, 3 m-INDELs, 24 germline *de novo* SNPs, and 5 germline *de novo* INDELs were identified in LOF-intolerant non-DDG2P genes. Variants not found within these two gene sets were excluded from further analyses.

2.3.3 Parental gonadosomatic variant prioritization results

Putative parental gonadosomatic mosaic variants that were transmitted to affected DDD-Africa probands as apparent germline mutations were identified using a specialized parental gonadosomatic prioritization pipeline (Figure 2.8). This pipeline was designed to detect variants consistent with low- or high-level gonadosomatic mosaicism, analysing maternal and paternal contributions independently. From the maternal samples, 5197 heterozygous SNPs were identified as inherited by 62 probands, and 167 heterozygous INDELs were inherited by 58 probands. Notably, no INDELs meeting QC thresholds were inherited from the mothers of probands 29, 87, 115, or 118. On the paternal side, 4,941 heterozygous SNPs were identified in 61 probands, alongside 181 heterozygous INDELs in 54 probands (Figure 2.8).

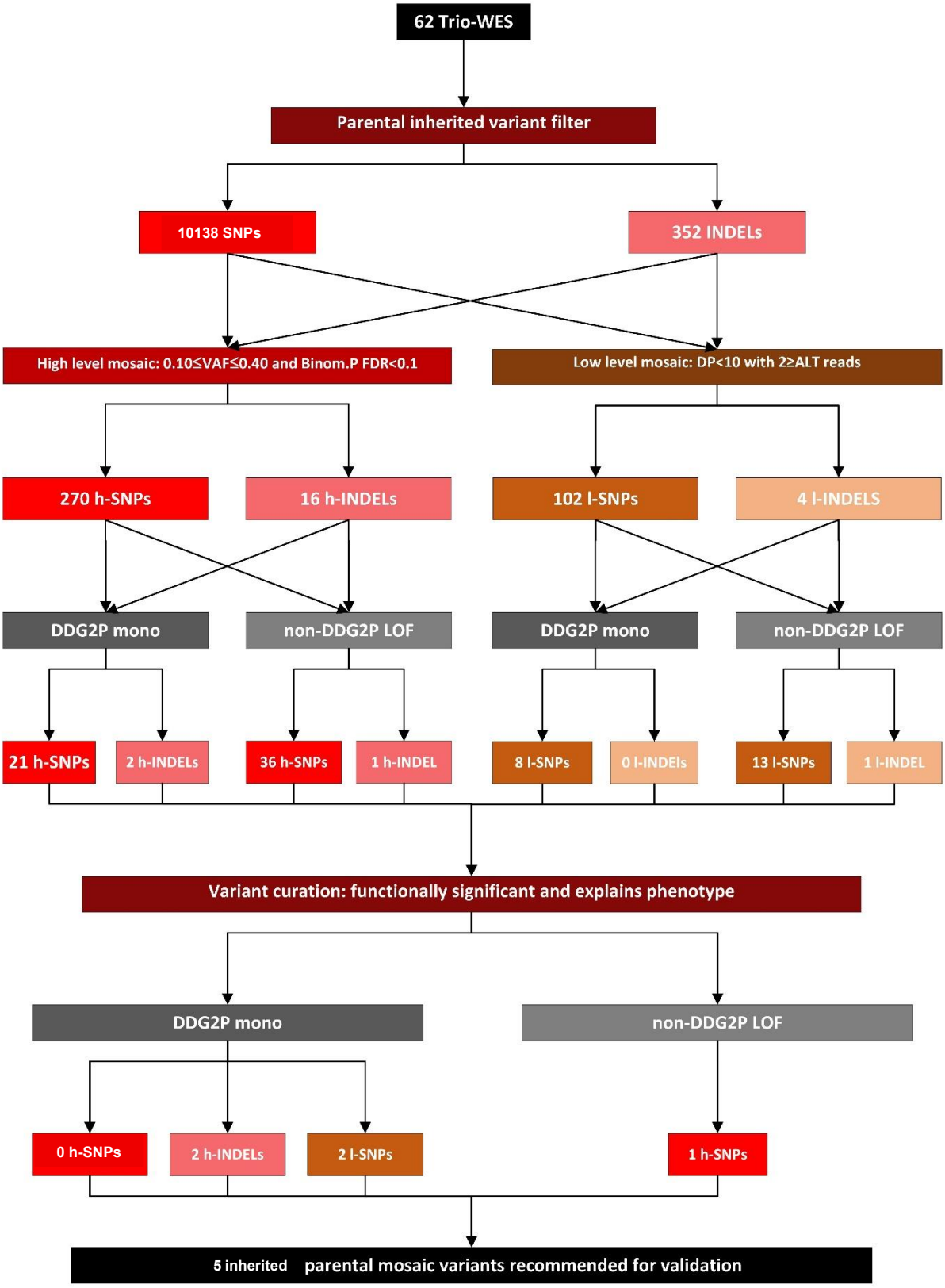


Figure 2.8: The parental gonadosomatic variant prioritization pipeline for 62 DDD-Africa trios outlining the results at each step.

h- = high-level mosaic, l- = low-level mosaic, Binom.P = p-value from binomial test, mono = monoallelic genes, LOF = loss-of-function intolerant genes, DP = read depth, VAF = variant allele fraction, ALT = alternate allele read support

No SNPs were found to be inherited from the fathers of proband 29, and only two SNPs were inherited from the fathers of probands 11 and 70, respectively. Furthermore, no INDELs were inherited from the fathers of probands 11, 29, or 70. These three trios had previously been excluded from the proband analysis pipeline due to an abnormally high number of DNMs, reinforcing the possibility of sampling errors. To identify putative gonadosomatic mosaicism in parents, two complementary variant prioritization strategies were employed (Figure 2.1). The EM-Mosaic algorithm could not be applied to parental samples because it assumes that all input variants are confirmed *de novo*, which could not be validated in the absence of grandparental data. Additionally, EM-Mosaic automatically excludes individuals with excessive counts of *de novo* variants, which would result in the exclusion of multiple parental samples. As an alternative, evidence for high-level parental mosaicism was assessed through two-tailed binomial testing of parental VAFs. Variants with FDR-adjusted binomial p-values <0.1 were considered indicative of potential mosaicism.

This analysis identified 270 SNPs and 16 INDELs, all of which were transmitted as heterozygous variants to affected probands. Among these, 156 candidate m-SNPs and 7 m-INDELs were identified in the mothers of probands 56 and 7, respectively. Likewise, 114 m-SNPs and 9 m-INDELs were detected in the fathers of probands 52 and 9, respectively (Figure 2.8). High-level m-SNP candidates had a mean VAF of 0.30 (SD = 0.05) in maternal samples and 0.29 (SD = 0.06) in paternal samples. High-level m-INDEL candidates showed a mean VAF of 0.29 (SD = 0.08) in mothers and 0.32 (SD = 0.06) in fathers. There were no statistically significant differences in the distribution of VAFs between maternal and paternal high-level m-SNPs ($p = 0.37$) or m-INDELs ($p = 0.63$). However, the median VAFs of candidate m-SNPs and m-INDELs detected in probands were significantly lower than those observed in either parent. Specifically, the median VAF of m-SNPs was lower than that of the mothers ($p = 0.0001$) and fathers ($p = 0.0001$), and the median VAF of m-INDELs was similarly lower compared to mothers ($p = 0.021$) and fathers ($p = 0.0005$). On average, parental gonadosomatic mosaic variants exhibited higher VAFs than the PZMs observed in probands.

To detect potential low-level parental gonadosomatic mosaicism, candidate SNPs and INDELs of interest with at least 2 ALT reads were evaluated. This approach yielded 102 candidate low-level SNPs and 4 candidate low-level INDELs across the cohort (Figure 2.8). These variants were retained with the understanding that they may be

enriched for false positives. Only low-level variants predicted to have functionally significant consequences in relevant genes were considered for downstream clinical curation. To prioritise putative pathogenic variants, parental candidates were filtered using the same two virtual gene panels (Figure 2.8). Within monoallelic DDG2P genes, 21 high-level m-SNPs, 2 high-level m-INDELs, and 8 low-level m-SNPs were identified. Within non-DDG2P LOF-intolerant genes, 36 high-level m-SNPs, 1 high-level m-INDEL, 13 low-level m-SNPs, and 1 low-level m-INDEL were identified. All other variants with evidence of mosaicism that fell outside these gene lists were excluded from clinical curation.

2.3.4 Candidate variants passing clinical curation and variant interpretation

Variants within the candidate gene lists were clinically curated according to the ACMG/AMP guidelines. Likely pathogenic and pathogenic curations represented the endpoint of the analysis and were considered significant candidates for recommended validation (Table 2.5). A total of seven candidate variants were identified through the proband variant prioritization pipeline, including two m-SNPs, three germline *de novo* SNPs, and one m-INDEL in monoallelic DDG2P genes. Additionally, a single m-SNP was identified in a non-DDG2P LOF-intolerant gene. Of these candidates, four were classified as likely pathogenic and three were classified as pathogenic (Table 2.5).

Five candidate variants were identified through the parental gonadosomatic variant prioritization pipeline (Table 2.5). These included one high-level m-SNP in a non-DDG2P LOF-intolerant gene and two high-level m-INDELs in monoallelic DDG2P genes. Two low-level candidate m-SNPs were also identified in monoallelic DDG2P genes. Among these variants, two high-level parental gonadosomatic variants were classified as pathogenic, one as likely pathogenic, and both low-level candidates as likely pathogenic (Table 2.5). Five variants were classified as variants of unknown significance (VUS) (Table A.4, Appendix G). In total, four proband PZMs, three proband germline DNMs, and five parental gonadosomatic candidate variants were recommended for future validation (Table 2.5). The clinical phenotypes of the affected probands in which these variants were identified is summarized in Table 2.6.

Table 2.5: A final list of candidate variants recomended for furthur validation.

ID	Variant	Consequence	Panel	BayesDel[$]	SpliceAI Δ	ACMG/AMP code	Class[^]	VAF_P	FDR	EM_{Post}	VAF_F	VAF_M
Proband PZM candidates												
1	*CNOT1*:c.2654G>T	missense	DDG2P	0.14	-	PS2, PM1, PM2, PP3	LP	0.17	$7.99x10^{-04}$	35.35	0.00	0.00
1	*CHM*:c.434C>A	stop gained	DDG2P	0.78	-	PVS1, PS2, PM2	P	0.20	-	-	0.00	0.00
21	*NSD2*:c.3160_3183del	in-frame deletion	DDG2P	-	-	PM2, PM4, PS2	LP	0.24	$4.21x10^{-03}$	-	0.00	0.00
66	*RFX7*:c.812-3C>A	splice region	LOF	-	AG - 0.54	PM2, PP3, PS2	LP	0.20	$1.58x10^{-02}$	7.84	0.00	0.00
Proband germline DNM candidates												
23	*KIF5C*:c.1288G>T	missense	DDG2P	0.30	-	PS2, PM2, PP3, PP2	LP	0.25	$8.97x10^{-02}$	2.20	0.00	0.00
41	*CDKL5*:c.978-14T>A	splice polypyrimidine tract	DDG2P	-	AL - 0.89	PS2, PM2, PP3	P	0.48	$9.54x10^{-01}$	0.03	0.00	0.00
130	*KMT2A*:c.3576A>C	missense	DDG2P	0.28	-	PS2, PM2, PP3, PM1, PP2	P	0.63	$2.07x10^{-01}$	0.00	0.00	0.00
High-level parental gonadosomatic candidates												
17	*PMP22*:c.448G>C	missense	LOF	0.58	-	PM1, PM5, PP5, PP3, PM6, PM2	P	0.41	$1.30x10^{-07}$	-	0.14	0.00
66	*EHMT1*:c.2073dup	frameshift	DDG2P	-	-	PVS1, PM2, PM6	P	0.56	$1.03x10^{-12}$	-	0.00	0.11
112	*NFE2L2:c.1688-1690del*	in-frame deletion	DDG2P	-	-	PM2, PM6, PM4	LP	0.64	$3.08x10^{-02}$	-	0.00	0.22
Low-level parental gonadosomatic candidates												
18	*MTOR*:c.5734G>T	missense	DDG2P	0.33	-	PM6, PM2, PP3, PS4	LP	0.25	-	-	<0.10*	0.00
53	*NIPBL:c.2584A>G*	missense	DDG2P	0.25	-	PM6, PM2, PP3, PP2, PP1	LP	0.5	-	-	<0.10*	0.00

ID –Family identifier; [$] BayesDel no allele fraction model; [^]Class – ACMG/AMP classification; FDR – adjusted binomial p-value for either proband or parental dependant on classification; [#] EM_{Post} – EM-mosaic posterior odds; LP – likely pathogenic; P – pathogenic; VAF_P – proband variant allele fraction; VAF_F – father variant allele fraction; VAF_M – mother variant allele fraction; AG – acceptor gain, AL – acceptor loss; DL - donor loss. *2 ALT read

Table 2.6: A summary of clinical phenotypes for DDD-Africa trios with potentially pathogenic candidate variants.

.ID	Sex	Proband phenotype	Parent phenotype	HGNC gene	Prior detection	HGVSc	Category	#OMIM
1	M	Severe global developmental delay; Delayed speech and language development; Broad-based gait; Dandy-Walker malformation; Microcephaly; Hypertonia; Cervical C3/C4 vertebral fusion; Clinodactyly of the 5th finger; Short stature; Myopia; Visual impairment; Cryptorchidism; Decreased body weight; Short stature; Dysmorphic facial features	No	*CNOT1*	No	NM_016284.5:c.2654G>T	PZM	Holoprosencephaly 12, with or without pancreatic agenesis (#618500) Vissers-Bodmer syndrome (#619033)
				CHM	No	NM_000390.4:c.434C>A	PZM	Choroideremia (#303100)
17	F	Moderate global developmental delay; Decreased body weight; Failure to thrive; Short stature; Severe left ventricular outflow tract obstruction; Atrioventricular canal defect; Malar flattening; High palate; Demyelinating peripheral neuropathy; Dysmorphic facial features	No	*PMP22*	No	NM_000304.4:c.448G>C	GON	Charcot-Marie-Tooth disease, type 1A/E (#118220) (#145900); Neuropathy, recurrent, with pressure palsies (#162500)
18	F	Severe global developmental delay; Delayed speech and language development; Microcephaly; Ventricular septal defect; Preauricular pit; Micrognathia; Craniosynostosis; Strabismus; Umbilical hernia; Bilateral hearing loss; Hypotonia; Dysmorphic facial features	No	*MTOR*	No	NM_004958.4:c.5734G>T	GON	Smith-Kingsmore syndrome (#616638)
21	M	Moderate global developmental delay; Delayed speech and language development; Decrease body weight; Failure to thrive; Hypospadias; Cryptorchidism; Clinodactyly of the 5th finger; Few cafe-au-lait spots; Ectopic kidney; Dysmorphic facial features	No	*NSD2*	Yes	NM_001042424.3:c.3160_3183del	PZM	Rauch-Steindl syndrome (#619695)
23	F	Mild global developmental delay; Ureterovesical junction obstruction; Thoracolumbar scoliosis; Lower limb asymmetry; Malar flattening; Short 5th finger; Dystrophic toenail; Dysmorphic facial features;	No	*KIF5C*	No	NM_004522.3:c.1288G>T	DNM	Cortical dysplasia, complex, with other brain malformations 2 (#615282)
41	F	Moderate global developmental delay; Delayed speech and language development; Hypopigmented skin patches; Preauricular skin tag; Infantile spasms; Hypotonia; Dysmorphic facial features	Type II diabetes mellitus in Mother	*CDKL5*	No	NM_001323289.2:c.978-14T>A	DNM	Developmental and epileptic encephalopathy 2 (#300672)
53	M	Moderate global developmental delay; Delayed speech and language development; Strabismus; Esodeviation; Micrognathia; 5th finger camptodactyly; Slit-like lateral ventricle; Spontaneous cerebrospinal fluid leak; Generalised hypotonia; Dysphagia; Poor suck; Brisk reflexes; Dysmorphic facial features	Hearing impairment in Father	*NIPBL*	No	NM_133433.4:c.2584A>G	GON	Cornelia de Lange syndrome 1 (# 122470)

66	F	Severe global developmental disorder; Absent speech; Microcephaly; Pectus excavatum; Clinodactyly of the 5th finger; Hypotonia; Hearing impairment; Joint hypermobility; Dysmorphic facial features	No	*RFX7*	No	NM_022841.7:c.812-3C>A	PZM	Intellectual developmental disorder, autosomal dominant 71, with behavioural abnormalities (# 620330)
66	F	Severe global developmental disorder; Absent speech; Microcephaly; Pectus excavatum; Clinodactyly of the 5th finger; Hypotonia; Hearing impairment; Joint hypermobility; Dysmorphic facial features	No	*EHMT1*	Yes	NM_024757.5:c.2073dup	GON	Kleefstra syndrome 1 (# 610253)
112	M	Moderate global developmental delay; Delayed speech and language development; Ileal atresia; Abnormal cardiac ventricle morphology; Butterfly vertebrae (possible); Short 3rd finger; Short 4th finger; Short 5th finger; Long 2nd finger; Cryptorchidism; Scrotal hypoplasia; Microphthalmia; Visual impairment; Abnormality of skin pigmentation; Facial asymmetry; Dysmorphic facial features	No	*NFE2L2*	No	NM_006164.5:c.1688_1690del	GON	Immunodeficiency, developmental delay, and hypohomocysteinemia (# 617744)
130	M	Profound global developmental delay; Absent speech; Autistic features; Bicoronal synostosis; Bilambdoid synostosis; Exotropia; Incisor macrodontia; Clinodactyly of the 5th finger; Hyperpigmentation of the skin; Dysmorphic facial features	No	*KMT2A*	No	NM_001197104.2:c.3576A>C	DNM	Wiedemann-Steiner syndrome (# 605130)

ID –Family identifier, M – male; F – female; HGNC - HUGO Gene Nomenclature Committee gene name; HGVSc - Human Genome Variation Society coding nomenclature; PZM – post-zygotic mutation in proband; DNM – germline *de novo* mutation in proband; GON – inherited heterozygote in proband with gonadosomatic mosaicism in parent; #OMIM - Online Mendelian Inheritance in Man MIM number

2.3.5 Clinical curation of selected candidate variants

The primary objective of this analysis was to prioritize the most plausible candidate variants for future validation. The *NSD2* and *EHMT1* variants, which were identified in parallel DDD-Africa analysis, both demonstrated evidence of mosaicism that was not reported prior (Table 2.6) (Table A.3, Appendix F). As a result, these variants were recommended for validation to confirm mosaic status. The IGV plots for variants classified as VUS are provided in Appendix H. The following results summarize the variant interpretation and curation presented in Table 2.5, utilizing the recorded data from Table 2.6.

2.3.5.1 Proband 1: *CNOT1*:c.2654G>T and *CHM*:c.434C>A

Two candidate m-SNPs from the proband variant prioritization pipeline were identified in proband 1. The first, a *CNOT1*:c.2654G>T missense SNP in exon 21, results in a p.Arg885Leu substitution in the Transcription Complex Subunit 1 (CNOT1) protein. CNOT1 is the central scaffolding component of the carbon catabolite repression 4-negative on TATA-less (CCR4-NOT) protein complex (Shirai et al., 2014). The CCR4-NOT complex plays a multifaceted role in mRNA stability regulation and protein ubiquitination (Shirai et al., 2014, Vissers et al., 2020). A highly conserved function of the CCR4–NOT complex is the regulation of global mRNA stability through deadenylation of the poly(A) tail, thereby modulating post-transcriptional gene expression (Stowell et al., 2016). The *CNOT1* gene is highly LOF-intolerant, with homozygous loss reported to be embryonically lethal (pLI = 1.0) (Kruszka *et al.,* 2019). *CNOT1* knockdown disrupts efficient CCR4-NOT complex assembly and impairs its regulatory activity, resulting in increased mRNA stability and global translation, which ultimately triggers endoplasmic reticulum stress–mediated cell death (Ito et al., 2011). Consistent with this essential role, CNOT1 has a missense Z-score of 9.16, indicating strong intolerance to missense variation and suggesting that such mutations are likely to exert functional effects contributing to disease.

CNOT1 is highly expressed during early embryogenesis and plays an essential role in forebrain development (Kruszka et al., 2019). It is listed in DDG2P and has been definitively linked to autosomal dominant holoprosencephaly 12 (OMIM: 618500) and Vissers-Bodmer syndrome (OMIM: 619033) by ClinGen expert panels (CCID:004485 and CCID:004484). Pathogenic variants in *CNOT1* are most often *de novo* and are

associated with a broad spectrum of developmental phenotypes, ranging from mild to severe GDD/ID and behavioural problems (Vissers et al., 2020). Affected individuals frequently present with structural brain anomalies such as corpus callosum malformations, white matter abnormalities, and holoprosencephaly, together with growth restriction, microcephaly, spinal defects, brachydactyly, clinodactyly, and vision impairment (OMIM:619033) (Vissers et al., 2020). Proband 1's reported phenotype showed substantial overlap with these symptoms (Table 2.6).

The *CNOT1*:c.2654G>T SNP showed significant evidence of mosaicism (EM-Mosaic posterior odds = 35.35, VAF = 0.17, ALT = 6, DP = 36) and no ALT read support in either parent (Table 2.5) (Figure 2.9). Accordingly, a PS2-strong code was applied for confirmed *de novo* status and parentage with significant phenotype overlap. A PP3-supporting code was applied based on a BayesDelnoAF score of 0.14 (CADD = 27; REVEL = 0.49). A PM1-supporting code was assigned as the variant lies in a UniProt region of interest in which two pathogenic variants have previously been reported (UniProt: A5YKK6-1). Finally, a PM2-moderate code was applied, as the variant was absent from gnomAD 4.1.0 in regions with good WES and WGS coverage. The cumulative ACMG/AMP evidence codes yielded a total score of eight, consistent with a classification of likely pathogenic.

The second candidate variant was *CHM*:c.434C>A, a nonsense SNP in exon 5 of the X-linked gene *CHM* (Table 2.5) (Figure 2.10). *CHM* s ubiquitously expressed and encodes component A of the Rab escort protein 1 (REP-1), a chaperone protein essential for intracellular membrane protein trafficking and photoreceptor homeostasis (Sarkar and Moosajee, 2022). REP-1 deficiency disrupts trafficking, resulting in degeneration of the photoreceptor layer through progressive cell death (Meschede et al., 2020). *CHM* is highly LOF-intolerant (pLI = 1.0) and has been definitively associated with monoallelic X-linked choroideremia in hemizygous males by the ClinGen Retina Gene Curation Expert Panel (OMIM:303100; CCID:004449). Choroideremia is characterised by progressive degeneration of the retinal pigment epithelium, photoreceptors, and choriocapillaris, with symptoms typically beginning as myopia and night blindness and advancing to central visual field constriction (Sarkar and Moosajee, 2022).



Figure 2.9: The *CNOT1*:c.2654G>T PZM candidate in Family 1.

A) Its genomic position in exon 21 of the CNOT1 protein impacting a domain of interest in relation to two previously reported pathogenic variants. Image adapted from the DECIPHER database (Firth et al., 2009). B) IGV plot of the coverage and mapped reads supporting the T change at chr16:58555488G>T in the proband (top), mother (middle), and father (bottom). Evidence of mosaicism is seen in proband 1 with 6 alternate reads at a read depth of 36. No evidence of sequencing artifacts, strand bias, read end bias, low variant call quality, or low mapping quality was observed.

Accordingly, the reported myopia in proband 1 overlapped with a known phenotype of *CHM* loss. The *CHM*:c.434C>A variant was confirmed *de novo* in proband 1 with evidence of mosaicism (VAF = 0.20, ALT = 7, DP = 35) and no supporting ALT reads in the either parent (Figure 2.10). As *CHM* is on the X chromosome in a male proband, VAF distribution does not follow a binomial model. This meant the VAF distribution could not be tested with EM-mosaic. Evidence of mosaicism was confirmed through manual examination of IGV plots and a Poisson test, which estimated the probability that the 7 ALT reads at the SNP locus arose from recurrent sequencing errors rather than a true variant (Figure 2.10). The test incorporated ALT read support, the Illumina NovaSeq 6000 error rate (0.01), and the estimated *CHM* nonsense mutation rate (0.000001893) from the gnomAD database.

The resulting p-value of 1.26×10^{-12} provided strong statistical evidence that the seven ALT reads were highly unlikely to be due to sequencing error, supporting the presence of a true mosaic variant. Thus, confirming *de novo* status for PS2 scoring. Choroideremia has a distinct clinical presentation that could not be fully ascertained from proband 1's records alone. Accordingly, PS2 was applied at a supporting level for the overlapping myopia phenotype. *In silico* tools predicted that the premature stop would trigger nonsense-mediated decay, and as loss-of-function is a known *CHM* disease mechanism, a PVS1-very strong code was applied. PM2-moderate was also applied, as the variant was absent from gnomAD 4.1.0 in regions with adequate sequencing coverage. The cumulative ACMG/AMP score of 11 classified the variant as pathogenic.

Taken together, the *CNOT1*:c.2654G>T and *CHM*:c.434C>A variants plausibly contributed to proband 1's reported phenotypes and showed strong evidence of being true mosaic SNPs. Both variants are therefore recommended for future validation.



Figure 2.10: The mosaic *CHM*:c.434C>A stopgain SNP detected in Family 1.

A) Its genomic position in exon 5 of the *CHM* gene leading to a premature stop codon and nonsense mediated decay. Image adapted from the DECIPHER database Firth et al. (2009). B) IGV plot of the coverage and mapped reads supporting the A change at chrX: 85963933C>A in the proband (top), mother (middle), and father (bottom). Evidence of mosaicism is seen in proband 1 with 7 alternate reads at a read depth of 35

2.3.5.2 Proband 17: *PMP22*:c.448G>C

A single candidate variant was identified in proband 17, with strong evidence of gonadosomatic mosaicism in his father (Table 2.5) (Figure 2.11). The *PMP22*:c.448G>C missense SNP in exon 5 results in a p.Gly150Arg substitution in peripheral myelin protein 22 (PMP22). *PMP22* is highly expressed in the plasma membranes of myelinating Schwann cells in the peripheral nervous system and is essential for the formation and maintenance of compact myelin sheaths (Snipes et al., 1992). Dysregulation of *PMP22* is associated with a spectrum of autosomal dominant neuropathies that vary with mutation type and gene dosage, including Charcot-Marie-Tooth disease Type 1A (OMIM:118220), Type 1E (OMIM:118300), Dejerine-Sottas disease (OMIM:145900), hereditary neuropathy with pressure palsies (OMIM:162500),

and Roussy-Lévy syndrome (OMIM:180800) (Hara et al., 2014). While PMP22 duplications are linked to Schwann cell apoptosis and peripheral nerve demyelination, loss-of-function is associated with milder hereditary neuropathy with pressure palsies (Hara et al., 2014). Small missense variants often lead to more severe phenotypes, likely due to accumulation of misfolded protein in the endoplasmic reticulum (Hara et al., 2014). *PMP22* has been definitively linked to Charcot-Marie-Tooth disease Type 1 and recurrent neuropathy with pressure palsies by the ClinGen Charcot-Marie-Tooth Disease Gene Curation Expert Panel (CCID:005837 and CCID:008314). Paediatric Charcot-Marie-Tooth Type 1 typically presents with hand and foot deformities, peripheral neuropathy-induced muscle atrophy, hyporeflexia, and, in severe cases, global developmental delay and hypo-myelinating neuropathy (OMIM:118220). Although uncommon, some patients have been reported with cardiac abnormalities, such as left ventricular hyper-trabeculation (Eltawansy et al., 2015).

Proband 17 presented with moderate GDD, normal speech development, and demyelinating peripheral neuropathy, consistent with a pathogenic *PMP22* variant (Table 2.6). While cardiac malformations are not commonly observed in *PMP22,* they have been reported in some Charcot-Marie-Tooth disease cases (Eltawansy et al., 2015). Neither parent exhibited overlapping phenotypes. *PMP22* was identified from the non-DDG2P LOF-intolerant gene panel (pLI = 0.99, missense Z-score = 0.70). The variant was germline heterozygous in proband 17 (VAF = 0.41, ALT = 24, DP = 58) and observed at low VAF in the father (VAF = 0.14, ALT = 10, REF = 71, FDR binomial p-value = 1.30×10^{-07}) supporting gonadosomatic mosaicism (Table 2.5, Figure 2.11). Accordingly, a PM6-moderate code was applied as the *de novo* status and parentage of proband 17's father could not be determined definitively, and it was determined that the phenotypic overlap was significant. Functional *in silico* prediction supported a PP3-strong code (BayesDelnoAF = 0.54, REVEL = 0.96, CADD = 27). The variant is present in a known mutation hotspot of a UniProt region of interest, in which 10 other pathogenic variants have been previously reported (UniProt: Q01453). This resulted in a classification of PM1-moderate, while PM2-moderate was applied as the variant was absent from gnomAD 4.1.0. Additionally, a PM5-strong code could be applied as three other pathogenic missense alleles reported at this position in the ClinVar database (*PMP22*:p.Glu150Cys, p.Glu150Asp, p.Glu150Val) (ClinVar: RCV004819225.1). Finally, a PS5-moderate code was applied as *PMP22*:p.Gly150Arg has been

previously classified as pathogenic in ClinVar (VCV000384327.6). The cumulative ACMG/AMP score of 15 classified the variant as pathogenic.

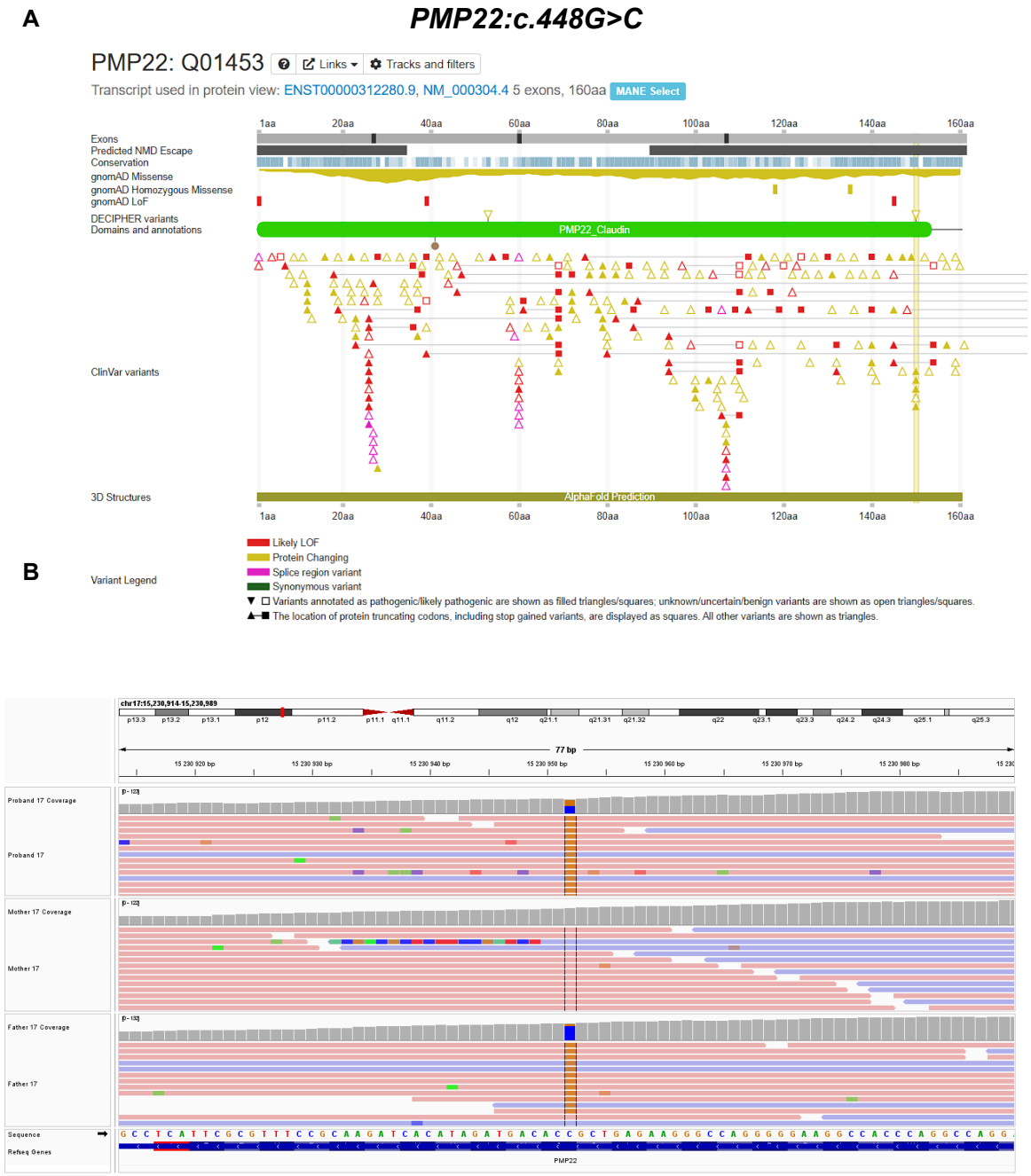


Figure 2.11: The *PMP22*:c.448G>C gonadosomatic SNP candidate detected in Family 17.

A) The genomic position (yellow line) in exon 5 of the *PMP22* leading to the pathogenic p.Glu150Arg amino acid change in a missense hotspot. Image adapted from the DECIPHER. B) IGV plot showing coverage and mapped reads supporting the G at chr17:15230952C>G in the reverse strand for the proband (top), mother (middle), and father (bottom). Strong evidence of gonadosomatic mosaicism is seen in father 17 with 10 alternate reads at a read depth of 71. No evidence for mosaicism in proband with 24 alternate reads at a depth 58. There is some strand asymmetry with a larger proportion of the call on the forward strand, there is no evidence of strand bias with an SOR=1.21 and a Fischer's strand Phred score of 5.12.

2.3.5.3 Proband 18: *MTOR:*c.5734G>T

In proband 18, a single germline heterozygous variant was identified, with some compelling evidence of low-level gonadosomatic mosaicism in his father (Table 2.5). The *MTOR*:c.5734G>T missense SNP results in a p.Asp1912Tyr substitution in the mechanistic target of rapamycin (mTOR), a serine/threonine kinase that nucleates mTOR complex 1 (mTORC1) and mTOR complex 2 (mTORC2) (Laplante and Sabatini, 2012). These complexes are central to signalling networks that respond to nutrients, energy status, oxidative stress, and growth factors, regulating protein synthesis, cell-cycle progression, metabolism, cytoskeleton organization, and cell survival (Laplante and Sabatini, 2012).

MTOR is critical during embryonic development, controlling neural stem cell proliferation, differentiation, and dendritic and axonal growth (Lipton and Sahin, 2014). Proper *MTOR* activity is essential for balancing these processes. Hyperactivation leads to excessive synaptic protein synthesis, increased neuronal size, and excitatory-inhibitory imbalance, while reduced activity is associated with hypomyelination, defective synapse formation, and decreased neuronal survival (Laplante and Sabatini, 2012; Lipton and Sahin, 2014).

MTOR is a highly dosage-sensitive monoallelic DDG2P gene (pLI = 1.0, pTriplo = 1.00, missense Z-score = 9.51), in which gain-of-function mutations are recognized as the definitive mechanism underlying overgrowth syndromes and cerebral malformations by the ClinGen Brain Malformations Gene Curation Expert Panel (CCID:005501). Pathogenic *MTOR* gain-of-function variants can result in either systemic Smith-Kingsmore syndrome (OMIM:616638) or somatic focal cortical dysplasia type II (OMIM:607341), depending on the location of the variant, the proportion of affected tissues, and impacted tissue. Clinical presentations include severe GDD/ID with speech delay, seizures, focal brain malformations or generalized white matter loss, macrocephaly, cryptorchidism, strabismus, cardiac malformations, hypotonia, dysmorphic features, and abnormal growth (OMIM:601231).

Thus, proband 18 exhibited a phenotype consistent with *MTOR* disorders, including macrocephaly and moderate GDD, although not all features were fully explained by this variant (Table 2.6). Additionally, neither parent displayed overlapping phenotypes. Curation of *MTOR*:c.5734G>T followed gene-specific ClinGen guidelines and the ACMG/AMP Variant Interpretation Guidelines Version 1.1.0, with adjustments to

account for GOF mechanism, mosaicism, and oncogenic potential. A PM6-moderate code was applied, reflecting potential inheritance from a low-level gonadosomatic mosaic father (proband VAF = 0.25, ALT = 6, DP = 24; father VAF<0.1, ALT = 2, DP = 21), and partial phenotype overlap (Table 2.5 and Table 2.6). Two putative ALT reads were seen in the father's BAM file in IGV (Figure 2.12). These reads were not present in the WES file of proband 18's father, potentially a result of the GATK-HC ploidy settings but were subsequently confirmed to be high-quality with no evidence of artifacts (Figure 2.12).

PM2-moderate was assigned due to the variant's absence from gnomAD, though it was noted that WES coverage across the region was reported to have failed gnomAD QC. A PP2-supporting code was applied as the *MTOR* missense Z-score of 9.51 exceeded an updated threshold of 3.09, and regional missense constraint was 5.35 in the surrounding genome. A PS4-supporting code was applied as an alternative variant at the same residue (*MTOR*:c.5734G>A) had been reported as a somatic GOF mutation in COSMIC (COSV63879685).

The LoGoFunc in silico predictor was queried to support PS3 scoring for the known GOF mechanism of *MTOR* (Stein et al., 2023). The *MTOR*:c.5734G>T variant was predicted to be GOF, with a score of 0.77. Additionally, the *MTOR*:c.5734G>A variant has been reported in the COSMIC database, where it was also predicted to be GOF. Thus, a PS3-strong code was applied as *in silico* and protein folding predictions for this conserved amino acid suggest a mechanistic impact. This candidate variant was classified likely pathogenic.



Figure 2.12: The *MTOR*:c.5734G>T gonadosomatic SNP candidate detected in Family 18.

A) The genomic position (black line) in exon 41 of the *MOTR* leading to the pathogenic p.D1912Y amino acid change in a missense hotspot. Image adapted from the DECIPHER database Firth et al. (2009). B) IGV plot showing coverage and mapped reads supporting the A at chr1:11128932C>A on the forward strand for the proband (top), mother (middle), and father (bottom). Some evidence of gonadosomatic mosaicism is seen in father 18 with 2 alternate reads at a read depth of 21. No evidence for mosaicism in proband with 6 ALT reads at a depth 24. No evidence of sequencing artifacts, strand bias, read end bias, low variant call quality, or low mapping quality was observed. The variant caller did not pick up the ALT reads in the father, likely due to ploidy settings removing extreme VAFs as sequencing error.

2.3.5.4 Proband 23: *KIF5C*:c.1288G>T

A single candidate germline *de novo* SNP was identified in proband 23 with no evidence of mosaicism across the trio (Table 2.5). The *KIF5C*:c.1288G>T missense SNP results in a p.Asp430Tyr substitution in exon 12 of KIF5C, a kinesin motor protein essential for neuronal development, axonal and dendritic function, and synaptic transport through microtubule-mediated trafficking of mRNAs, vesicles, mitochondria, and protein complexes (Banerjee et al., 2024, Hirokawa et al., 2010). The DDG2P *KIF5C* gene is highly LOF-intolerant (pLI = 1.0, missense Z-score = 3.63) and is moderately associated with autosomal dominant complex cortical dysplasia with other brain malformations 2, according to the ClinGen Brain Malformations Gene Curation Expert Panel (CCID:005233). However, since the published last curation, six additional pathogenic variants have been reported in *KIF5C* in the literature. Thus, pathogenic variants in *KIF5C* are rare, with only 12 cases being reported to date (Banerjee et al., 2024).

Reported phenotypes associated with pathogenic *KIF5C* variants include severe GDD/ID with absent speech, cortical abnormalities, corpus callosum anomalies, seizures, severe behavioural problems, stereotypical hand movements, foetal akinesia, microcephaly, and arthrogryposis (OMIM:615282) (Banerjee et al., 2024). In proband 23, the phenotype only partially overlapped with *KIF5C* associated features (Table 2.6). Proband 23 exhibited mild GDD with normal speech, and no seizures or brain malformations. Additionally, the scoliosis and cardiac malformations reported in proband 23 have not been previously associated with *KIF5C*. However, given the rarity of pathogenic variants in KIF5C, and their clustering in the microtubule binding domain, variants in other regions may manifest with atypical, milder presentations.

Consequently, a non-specific association between the GDD phenotype and *KIF5C* was considered as partially causal or contributing to proband 23's condition. Thus, PS2 was only applied at supporting level, with a confirmation *de novo* status through no ALT read support in either parent (Figure 2.13). Next, PM2-moderate was applied as the variant was not found in gnomAD. A PP3-moderate code was applied as the BayesDelnoAF prediction score was 0.30 (CADD = 31, REVEL = 0.629) (Table 2.5). Finally, a PP2-supporting score was applied as the *KIF5C* missense Z-score from gnomAD (3.63) was greater than the required 3.09. Thus, the cumulative ACMG/AMP score of six points resulted in a likely pathogenic classification.

A ***KIF5C*:c.1288G>T**



Figure 2.13: The *KIF5C*:c.1288G>T germline *de novo* missense SNP detected in proband 23.

A) Its genomic position in exon 12 of the *KIF5C* gene leading to a Asp430Tyr substitution. Image adapted from the DECIPHER database Firth et al. (2009). B) IGV plot of the coverage and mapped reads supporting the T change at chr2:148973506 in the proband (top), mother (middle), and father (bottom). Some evidence of mosaicism is seen in proband 23 with 5 alternate reads at a read depth of 20, however it was not deemed significant by EM-Mosaic. No evidence of sequencing artifacts, strand bias, read end bias, low variant call quality, or low mapping quality was observed.

2.3.5.5 Proband 41: *CDKL5*:c.978-14T>A

In proband 41, a candidate SNP was identified as germline *de novo*, with no evidence of mosaicism and zero ALT read support both parents (Table 2.5) (Figure 2.14). The *CDKL5*:c.978-14T>A variant is located 14 bp upstream of exon 12 in the cyclin-dependent kinase-like 5 (*CDKL5*) gene, which encodes a serine/threonine protein kinase involved in phosphorylation of substrate proteins critical for microtubule dynamics, neuronal morphology, and synaptic function (Van Bergen et al., 2022). *CDKL5* is widely expressed in neurons during brain development and regulates actin dynamics, synaptic vesicle recycling, dendritic growth, excitatory/inhibitory synaptic balance, and overall network formation (Van Bergen et al., 2022).

Loss of *CDKL5* function, which impairs dendritic branching and axonal morphology, has been definitively associated with infant-onset X-linked dominant *CDKL5* deficiency disorder by the ClinGen Rett/Angelman Expert Panel (CCID:004403) (Van Bergen *et al*., 2022). It is also known as developmental and epileptic encephalopathy 2 (OMIM:300672), and a Rett-like syndrome. However, it does not present with defining Rett features such as regression and intense gaze (Van Bergen et al., 2022). Although historically considered rare, pathogenic *CDKL5* variants have now been reported in the hundreds, establishing *CDKL5* deficiency as one of the more common causes of childhood epilepsy and severe neurodevelopmental delay (Van Bergen et al., 2022).

It is characterized by early-onset seizures (typically <3 months), often presenting as infantile spasms, tonic, tonic-clonic, or myoclonic seizures, alongside severe GDD/ID, speech delay, motor dysfunction, stereotypic hand movements, gastrointestinal issues, musculoskeletal abnormalities, microcephaly, hypotonia, dysmorphic facial features, and characteristic sleep disturbances (Van Bergen et al., 2022). Mosaic variants have been described in males, with phenotypic severity influenced by X-inactivation patterns in females (Van Bergen et al., 2022).

The reported *CDKL5* deficiency disorder symptoms overlapped significantly with proband 41's recorded phenotype (Tabel 2.6). Neither parent had any significant overlapping phenotype, with the mothers reported Type II diabetes mellitus not being considered relevant to this case. Variant classification followed the ClinGen Rett and Angelman-like Disorders Expert Panel Specifications to the ACMG/AMP guidelines for *CDKL5* (v4.0.0). The variant was confirmed germline *de novo* with no evidence of mosaicism in proband 41 (VAF = 0.48, ALT = 21, DP = 44, EM-mosaic posterior odds

= 0.03) and no read support in either parent (Figure 2.14). Thus, a PS2-strong code was applied in line with the *CDKL5* guidelines, in which pathogenic variants are mostly *de novo*. A PM2-moderate code was applied as the variant was novel, not being reported in the gnomAD database in a region with good coverage.

A PS3-strong code was applied for a SpliceAI acceptor loss prediction score of 0.89. The variant, located 14 bp from a canonical splice acceptor site in the polypyrimidine tract of exon 12, was predicted to create disrupt this site but introducing a cryptic acceptor 2 bp downstream. Additional annotation with SpliceVault, predicted exon 12 skipping as the most likely outcome (Dawes et al., 2023). Finally, two canonical splice acceptor site disruption variants *CDKL5*:c.978-2A>G and c.978-1G>A have recently been reported as likely pathogenic and pathogenic in ClinVar (RCV004724912.1 and RCV002249337.1, respectively). Thus, the *CDKL5*:c.978-14T>A SNP was classified as pathogenic with 10 ACMG/AMP points.

It is noted that a single reverse orientation read also carried the reference T allele, suggesting possible strand bias. However, the variant strand odds ratio (SOR = 0.13) was low, and no supporting reads were observed across other DDD-Africa samples, making sequencing error unlikely. Nevertheless, orthogonal validation remains essential to this out.



Figure 2.14: The *CDKL5*:c.978-14T>A *de novo* splice polypyrimidine tract SNP detected in proband 41.

A) Its genomic position in intron 11 of the *CDKL5* gene leading to disruption of the canonical splice acceptor gain site of exon 12. Image adapted from the DECIPHER database Firth et al. (2009). B) IGV plot of the coverage and mapped reads supporting the A change at chrX:18603888 in the proband (top), mother (middle), and father (bottom). No evidence of mosaicism is seen in proband 41 with 21 alternate reads at a read depth of 44. The region is only covered by forward reads at the intron exon boundary. A single reverse read as present in the proband, which carried the variant. This may indicate strand bias. However, SOR was 0.13. No evidence of sequencing artifacts, read end bias, low variant call quality, or low mapping quality was observed.

2.3.5.6 Proband 53: *NIPBL*:c.2584A>G

A single candidate variant was identified in proband 53, with some evidence of low-level gonadosomatic mosaicism in the father (Table 2.5) (Figure 2.15). The *NIPBL*:c.2584A>G missense SNP introduces a p.Lys862Glu amino acid substitution in exon 10 of the Nipped-B-like cohesion loading factor (NIPBL) protein. The NIPBL protein forms a heterodimeric cohesion loading complex that facilitates the recruitment of the cohesin complex onto DNA (Gao et al., 2019). It plays a crucial role in regulating cohesin activity and maintaining 3D genome organization, influencing chromatin remodelling, DNA repair, and transcriptional regulation at CTCF-binding sites (Gao et al., 2019). Mutations in *NIPBL* have been shown to dysregulate transcription and translation, induce DNA repair errors, and lead to genome instability (Garcia et al., 2021). *NIPBL* is a highly LOF-intolerant DDG2P gene (pLI =1.0, missense Z-score = 6.70) that has been definitively linked to autosomal dominant Cornelia de Lange syndrome 1 by the ClinGen Intellectual Disability and Autism Gene Curation Expert Panel (CCID:005644). Variants in NIPBL are most often *de novo* and are associated with more severe phenotype (Gruca-Stryjak et al., 2024).

Classic Cornelia de Lange syndrome presents with characteristic features, developing prenatally. This includes foetal growth restriction, limb deformities, hand abnormalities, cardiac malformations, microcephaly, micrognathia, congenital diaphragmatic hernia, renal anomalies, hypertonicity, GDD/ID with speech delay, and distinctive facial dysmorphism (OMIM: 122470) (Gruca-Stryjak et al., 2024). Mild to moderate sensorineural hearing loss has also been reported in up to 65% of patients, and mosaic pathogenic *NIPBL* variants have been identified in up to 20% of cases (Huisman et al., 2013, Gruca-Stryjak et al., 2024). Thus, clinical curation revealed significant overlap between proband 53's reported phenotype and Cornelia de Lange syndrome (Table 2.6). Importantly, the father of proband 53 was also recorded as having hearing impairment, which was identified as a relevant phenotypic overlap further supporting the mosaic diagnoses. However, it has also been shown that missense variants in NIPBL may present with milder phenotype (Gruca-Stryjak et al., 2024). Thus, it is possible that the variant is constitutionally heterozygous in the father. Unfortunately, the low sequencing coverage in the father (DP = 8, ALT = 4) made it impossible to determine the true VAF with high confidence. In fact, most WES pipelines would exclude this variant. The low-level mosaicism pipeline in this study only required 2 ALT read support in a parent for a clinically relevant variant in the proband for inclusion.

Thus, for the purpose of this analysis, it was assumed that the *NIPBL:*c.2584A>G missense SNP was gonadosomatic going forward, in line with the pipeline which led to its discovery. This did make curation a challenge, with conservative PM6-supporting code being assigned to account for inheritance and phenotypic uncertainty. The father's overlapping, but less severe hearing impairment phenotype, was evidence for the variant segregating as a single meiotic event within the family, justifying the scoring of PP1-supporting as well. The variant was present in gnomAD 4.1.0, at a WES derived MAF of $6.842x10^{-07}$, reflecting a total allele count of 1. However, as this frequency was less than the calculated maximum credible allele frequency for *NIPBL* and Cornelia De Lange Type 1 of $7.0X10^{-07}$, a PM2-moderate code could be applied. A PP3 code was applied at the supporting level for a BayesDelnoAF score of 0.25 (REVEL = 0.48, CADD = 22.3). Finally, a PP2 code was applied at a supporting level for a gnomAD missense Z-score was 6.70, suggesting significant missense intolerance in *NIPBL*, and missense mutations being a known mechanism of disease. The variant was reported as VUS on ClinVar (2stars) (VCV000450053.8), which was not sufficient as evidence for pathogenicity or benignity. The resulting ACMG/AMP score of six points lead to a classification of likely pathogenic.

Although the four ALT reads in the father suggest VAF constitutional heterozygosity, the low coverage makes definitive confirmation impossible. Multiple lines of evidence indicate that a gonadosomatic mosaic origin is at least equally, if not more, plausible. First, the variant is extremely rare in population databases (disease prevalence ~0.6 per 100,000), *NIPBL* is highly dosage-sensitive with an autosomal dominant mechanism of inheritance, and severe phenotypes usually manifest prenatally when *NIPBL* is disrupted (Gruca-Stryjak et al., 2024). Mild or late-onset presentations are uncommon, making it unlikely that a constitutive germline variant would remain largely silent for over 40 years in the father. This is consistent with prior reports showing that most pathogenic *NIPBL* variants arise *de novo* and are frequently mosaic (Gruca-Stryjak et al., 2024). Second, the father’s phenotype of isolated, milder sensorineural hearing loss aligns with the attenuated or tissue-restricted effects expected in mosaicism, which is in stark contrast to the proband’s broader syndrome consistent with full constitutional heterozygosity (Table 2.6). While variable expressivity is recognised in Cornelia de Lange syndrome, there is sparse evidence for true incomplete penetrance of *NIPBL* variants, further arguing against a lesser impactful germline heterozygote state (OMIM: 122470) (CCID:005644) (Gruca-Stryjak et al.,

2024). Third, the shallow sequencing depth at this locus reduces confidence in allele balance estimates, meaning the apparent VAF of 0.5 could reflect stochastic sampling rather than true constitutional heterozygosity. Taken together, a reasonable assumption is a gonadosomatic PZM in the father, present at sufficient frequency in germ cells to for high risk of inheritance, but at lower or variable levels in peripheral blood. However, a germline heterozygote with unusually mild expressivity cannot be excluded, as mosaic and germline cases often present with similar severity (Gruca-Stryjak et al., 2024). For this reason, PM6-supporting rather than PS2 was conservatively applied, reflecting this uncertainty and the strength of evidence for inheritance. This interpretation best explains the observed family phenotype, avoids a potential false positive *de novo* germline event with major counselling implications, and justifies prioritisation of this variant for targeted orthogonal validation. It noteworthy that variant is reported as VUS on ClinVar (VCV000450053.8), and that proband 53 also has a prior diagnosis for a confirmed *de novo* pathogenic *AUTS2* variant from parallel analysis (Appendix F). However, given the father's phenotype which is not explained by the *de novo AUTS2* variant, the *NIPBL* variant could plausibly contribute to the phenotype through epistatic interaction.

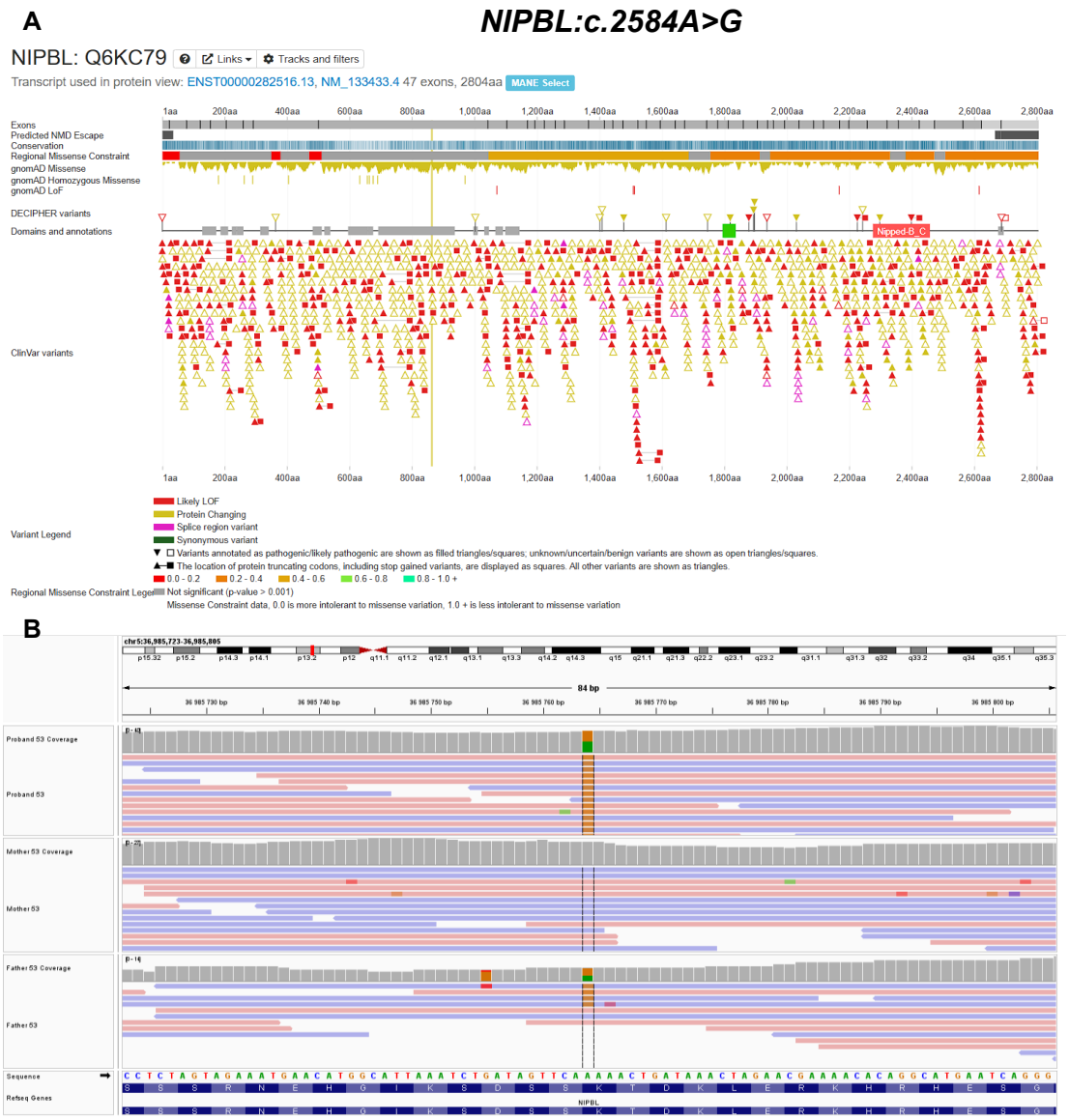


Figure 2.15: The *NIPBL*:c.2584A>G gonadosomatic missense SNP in Family 53

A) The genomic position (golden line) of the *NIBPL*:p.Lys862Glu amino acid change in a missense hotspot in exon 10. Image adapted from the DECIPHER database Firth et al. (2009). B) IGV plot showing coverage and mapped reads supporting the G at chr5:36985764 on the forward strand for the proband (top), mother (middle), and father (bottom). Some evidence of gonadosomatic mosaicism is seen in father 53 with 4 alternate reads at a read depth of 8. No evidence for mosaicism in proband53 with 7 ALT reads at a depth 14. No evidence of sequencing artifacts, strand bias, read end bias, low variant call quality, or low mapping quality was observed. Read depth was very low in this region especially for the father.

2.3.5.7 Proband 66: *RFX7*:c.812-3C>A

A candidate non-canonical splice site variant was identified in proband 66 with moderate evidence of mosaicism (Posterior Odds = 7.84, VAF = 0.20, ALT = 5, DP = 25) (Table 2.5) (Figure 2.16). Although the EM-mosaic posterior odds did not exceed the >10 threshold, post odds of 7.84 at this depth was considered sufficiently strong to

support mosaicism and warranted further investigation. The *RFX7*:c.812-3C>A variant was predicted by SpliceAI to create a cryptic splice acceptor site 65 bp upstream (AG = 0.54). The variant was identified through the non-DDG2P LOF panel (pLI = 1.0) as *RFX7* is not currently included in DDG2P. It encodes Regulatory Factor X 7 a highly conserved transcription factor with numerous gene targets and high expression in the brain (Sugiaman-Trapman et al., 2018). It has been shown to play an important role in cilia development during early embryogenesis and regulates central nervous system gene expression (Chu et al., 2010). Dysregulation of *RFX7* has very recently been associated with a spectrum of neurodevelopmental phenotypes, including GDD/ID with speech delay, hypotonia, abnormal brain morphology, ASD features with behavioural disturbances, hand and foot malformations, joint hyperlaxity, micrognathia, microcephaly, failure to thrive, and variable dysmorphic features (OMIM: 620330). Despite the relatively sparse literature, these findings indicate *RFX7* is functionally important during development. Apart from hearing impairment and pectus excavatum, proband 66's phenotype overlapped significantly with *RFX7* associated phenotypes (Table 2.6).

The *RFX7*:c.812-3C>A splice variant was confirmed *de novo* and mosaic with significant overlapping phenotype and no read support in either parent, resulting in a score of PS2-strong. The SpliceAI prediction of cryptic acceptor gain within exon 8 resulted in a PP3 classification at moderate level. Further annotation of the variant with SpliceVault, confirmed a predicted cryptic acceptor gain that results in aberrant splicing and loss of 62 bp of exon 8 (Dawes et al., 2023). Finally, a PM2-moderate score was applied as the variant present in gnomAD in a region with good overall sequencing coverage. The final ACMG/AMP score of eight led to a classification of likely pathogenic.

A pathogenic maternally inherited *EHMT1* variant was identified through parallel analyses (Appendix F). In this study, *EHMT1*:c.2073dup was shown to have strong evidence of gonadosomatic mosaicism in the mother (VAF = 0.11, ALT = 10, DP = 88, FDR binomial p-value = 1.03×10^{-12}, unadjusted binomial p-value = 3.33×10^{-14}) (Table 2.5). This variant also explains proband 66's phenotype. Thus, it was recommended for mosaicism confirmation, with significant implications for recurrence risk counselling. It is unclear whether proband 66's phenotype is primarily attributable to the mosaic *RFX7* splicing variant, the inherited *EHMT1*:c.2073dup variant or a potential interaction

between the two. Further functional validation and segregation testing are required to resolve the causal contribution of each variant.

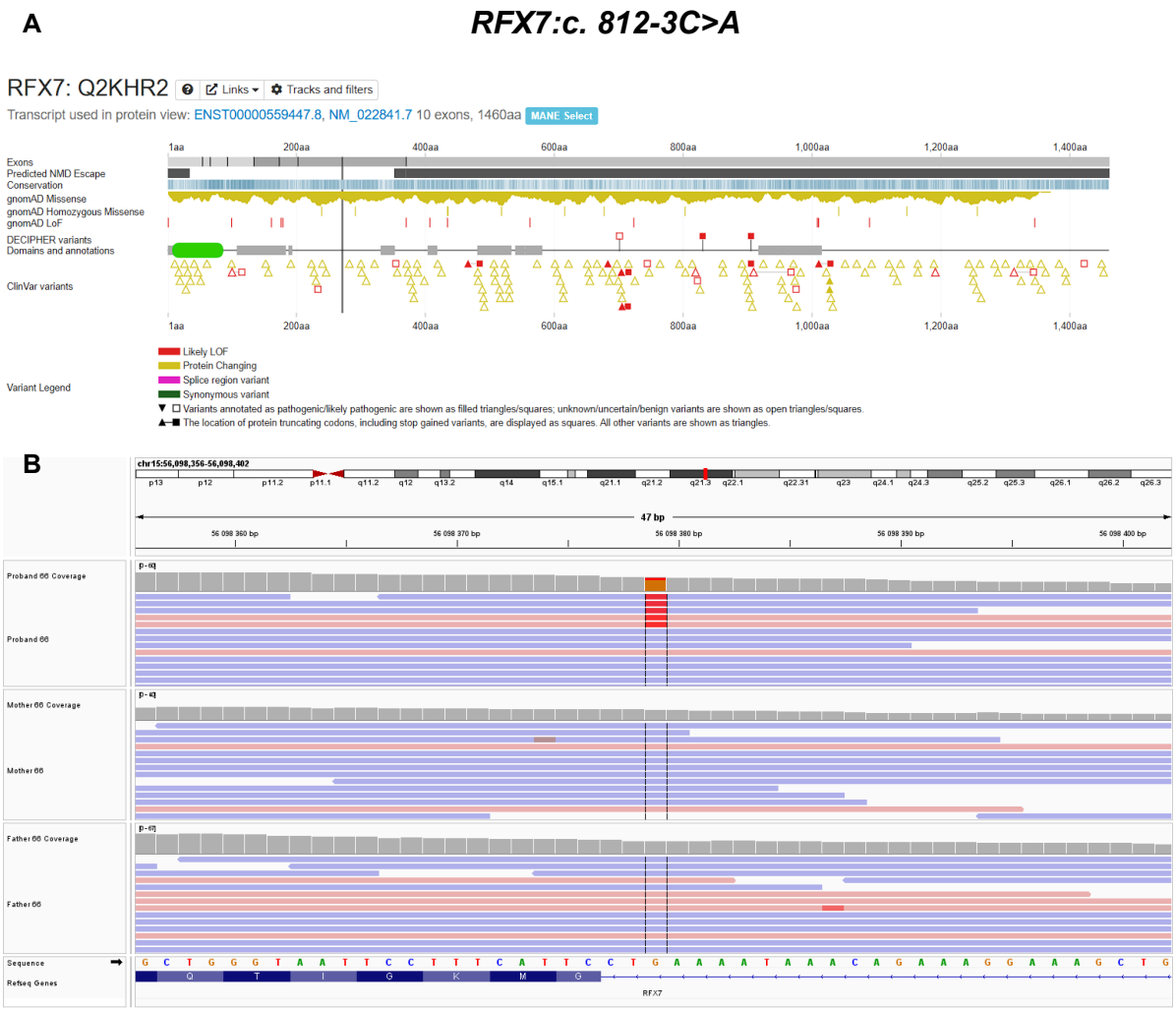


Figure 2.16: The *RFX7*:c812-3C>A mosaic splice altering SNP candidate in family 66.

A) The genomic position (black line) in exon 9 of the *RFX7* leading to an acceptor gain splice variant. Image adapted from the DECIPHER database Firth et al. (2009). B) IGV plot showing coverage and mapped reads on the forward strand for the proband (top), mother (middle), and father (bottom). Evidence of mosaicism is seen in Proband 66 with 5 alternate reads at a read depth of 25. No evidence of sequencing artifacts, strand bias, read end bias, low variant call quality, or low mapping quality was observed. Read depth was very low in this region especially for the father

2.3.5.8 Proband 112: *NFE2L2*:c.1688-1690del

In proband 112, the *NFE2L2*:c.1688-1690del variant was identified with strong evidence of maternal gonadosomatic inheritance (Tabel 2.5) (Figure 2.17). This in-frame deletion removes a single, highly conserved glutamic acid at position 536 of the Nuclear factor erythroid 2-related factor 2 (NRF2) transcription factor protein, which is encoded by the NRF2 like basic leucine zipper transcription factor 2 (*NFE2L2*) gene. NRF2 is a master regulator of the cellular stress response, orchestrating the expression of over 200 genes in response to oxidative stress, hypoxia, genotoxins, nutrient imbalance, and infection (Dai et al., 2020, Huppke et al., 2017). Normally NRF2 activity is suppressed. However, during embryogenesis it is strongly expressed in embryonic stem cells, where it is essential for survival, self-renewal, and differentiation through the regulation of pluripotency genes. NRF2 also links stem cell cycles to neuroectodermal development and, in later life, supports neuronal and mesenchymal stem cell survival, regeneration, and proliferation under stress (Dai et al., 2020, Huppke et al., 2017).

Gain-of-function missense variants in *NFE2L2* have been shown to disrupt NRF2's inactivation, resulting in aberrant upregulation of downstream stress-response genes. Pathogenic *NFE2L2* variants are associated with immunodeficiency, developmental delay, and hypohomocysteinemia, with affected individuals presenting with recurrent infections, GDD/ID with speech and learning difficulties, white matter abnormalities, and leukoencephalopathy (OMIM:617744) (Huppke et al., 2017). Additional phenotypic presentations include failure to thrive, short stature, muscular dystrophy and weakness, cardiac abnormalities, and abnormalities of the eyes (Huppke et al., 2017).

The clinical features of proband 112 showed significant overlap with published cases of *NFE2L2* dysregulation, including mild GDD with speech delay, failure to thrive, cardiac ventricular abnormalities, visual impairment, and pigmentation anomalies (Table 2.6). However, immunodeficiency and hypohomocysteinemia, which are hallmark features of previously reported cases, were not present (OMIM:617744). Proband 112's reported musculoskeletal abnormalities and facial dysmorphia have not been explicitly associated with *NFE2L2*, though DECIPHER records report hand and musculoskeletal anomalies in two unrelated individuals with pathogenic variants (https://www.deciphergenomics.org/gene/NFE2L2/phenotypes/). No relevant features were recorded in either parent (Table 2.6).

The *NFE2L2*:c.1688_1690del variant was confirmed germline heterozygous in proband 112 (VAF = 0.64, ALT = 9, DP = 14) and exhibited supporting evidence of gonadosomatic mosaicism in the mother (VAF = 0.22, ALT = 4, DP = 18, FDR-adjusted binomial p-value = 0.11). Although read depth was low in both proband and parents, and the FDR adjusted p-value did not reach the <0.1 significance threshold, the observed VAF of 0.22 was considered suggestive of mosaicism and sufficient to warrant future validation using an orthogonal method. On this basis, a PM6-moderate code was applied for assumed gonadosomatic *de novo* origin and overlapping phenotype. A PM2-moderate code was also applied due to the absence of the variant in gnomAD.

A PM4-moderate code was applied as this in-frame deletion is in a non-repetitive region and altered protein length. Further support for the maximum allowed level of PM4 as moderate was provided by MutationTaster2021 annotation, which classified the variant as deleterious. Additionally, the deleted glutamic acid is highly conserved across species. The reported mechanism of disease for *NFE2L2* is GOF. Although all four previously reported cases involved missense changes at residues 80 and 81, the c.1688_1690del variant lies just two residues from the boundary of the basic leucine zipper domain and upstream of predicted glycosylation sites (Figure 2.17). This suggests that deletion of Glu563 could plausibly alter domain conformation or interfere with glycosylation. While the specific mechanism by which this variant may act through GOF remains unclear, further investigation was recommended. A final ACMG/AMP score of six supported a classification of likely pathogenic.

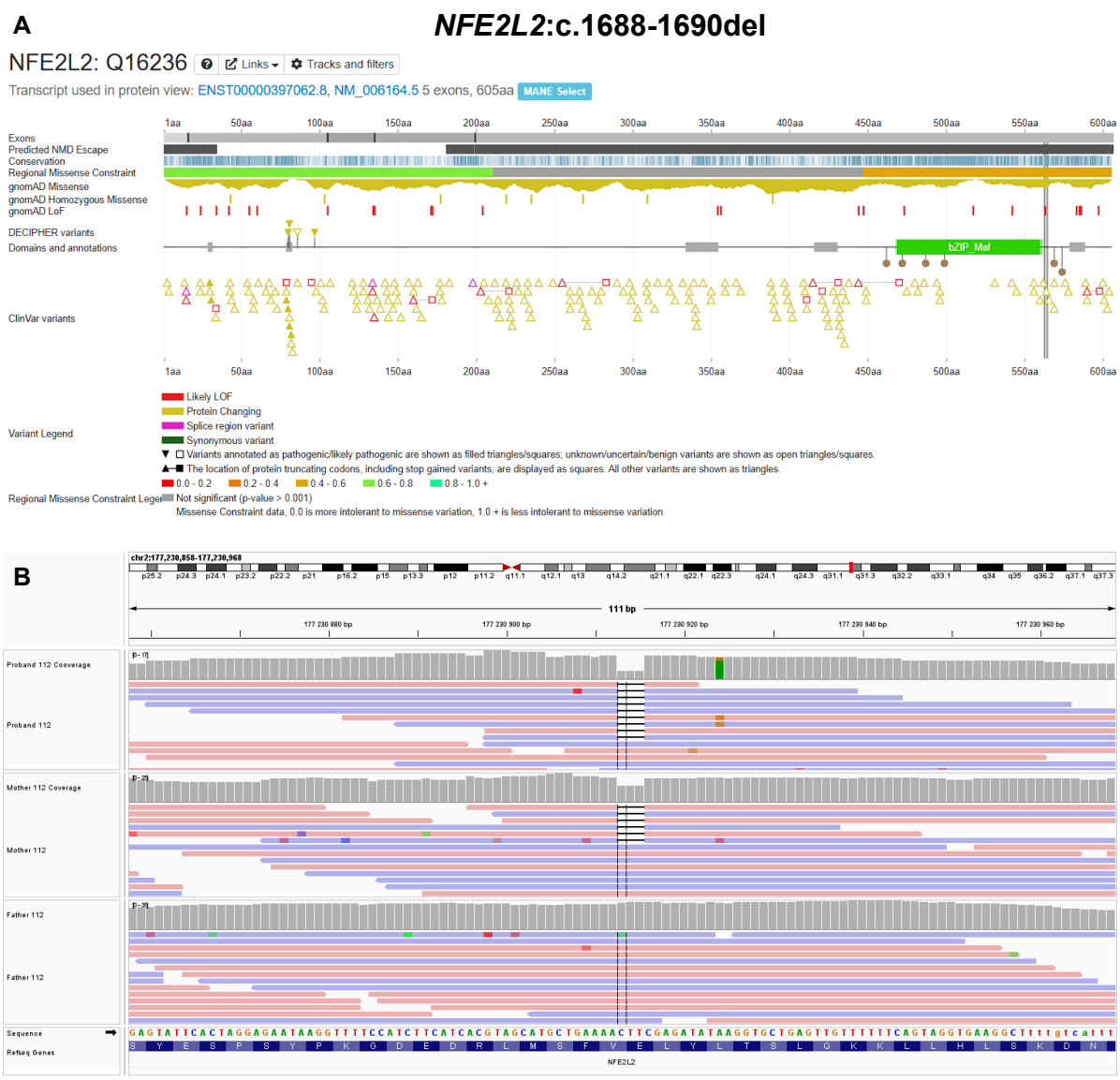


Figure 2.17: The *NFE2L2*:c.1688-1690del gonadosomatic in-frame INDEL in Family 112

A) The genomic position (golden line) of the *NFE2L2*:c.1688-1690del of Glu563 in exon 5 just downstream the bzip domain. Image adapted from the DECIPHER database Firth et al. (2009). B) IGV plot showing coverage and mapped reads supporting the deletion of CTT at chr2: 177230912 on the reverse strand for the proband (top), mother (middle), and father (bottom). Some evidence of gonadosomatic mosaicism is seen in mother 112 with 6 alternate reads at a read depth of 20. No evidence for mosaicism in proband53 with 9 ALT reads at a depth 14. No evidence of sequencing artifacts, strand bias, read end bias, low variant call quality, or low mapping quality was observed. Read depth was very low in this region.

2.3.5.9 Proband 130: *KMT2A*:c.3576A>C

A candidate missense SNP, *KMT2A*:c.3576A>C, which results in a p.Arg1192Ser substitution in exon 6 was identified in proband 130. *KMT2A* encodes lysine methyltransferase 2A, a ubiquitously expressed protein that catalyses the methylation of histone H3 at lysine 4 (H3K4), activating transcription through modulation of chromatin accessibility (Castiglioni et al., 2022). This activity is essential for early embryogenesis and haematopoiesis, with additional functions reported in the regulation of circadian rhythms (Castiglioni et al., 2022).

KMT2A is a highly LOF and missense intolerant DDG2P gene (pLI = 1.0, missense Z-score = 8.79) that has been definitively linked to autosomal dominant Wiedemann-Steiner syndrome by the ClinGen Syndromic Disorders Gene Curation Expert Panel (CCID:005249) (OMIM: 605130). The disease mechanism is typically LOF, but pathogenic missense variants have also been reported (CCID:005249). This rare syndrome is characterised by a spectrum of clinical findings, including mild to profound GDD/ID with speech delay, ASD features, behavioural dysregulation, agenesis or hypoplasia of the corpus callosum, musculoskeletal anomalies of the hands and feet, cardiac abnormalities, failure to thrive, hypotonia, and distinctive facial dysmorphia with abnormal dentition (OMIM: 605130) (Castiglioni et al., 2022). Dysregulation of *KMT2A* has also been reported in other chromatinopathies, including Coffin-Siris syndrome, Kabuki syndrome, Cornelia De Lange syndrome, and Rubinstein-Taybi syndrome (Castiglioni et al., 2022). Thus, proband 130's phenotype showed significant and specific overlap with these reported features (Table 2.6).

The *KMT2A*:c.3576A>C missense SNP was confirmed germline *de novo* in proband 130, with no evidence of mosaicism and no read support in either parent (VAF = 0.63, ALT = 26, DP = 41, EM-Mosaic posterior odds = 0) (Table 2.5) (Figure 2.18). Thus, a code of PS2 at the moderate level was applied. Additionally, a PM2-moderate code was applied as the variant was absent from the gnomAD database. The BayesDelnoAF prediction score for pathogenicity was 0.28 (CADD = 23.4, REVEL = 0.86) resulting in the application of a PP3-moderate code. A PP2-suporting code was applied as pathogenic missense mutations have been previously reported in *KMT2A* (missense Z-score of 8.79). Lastly, a PM1-moderate code was applied as Arg1192Ser is predicted to disrupt zinc and DNA-binding affinity of the CXXC-type zinc finger domain (UniProt: Q03164-3). The final ACMG/AMP score was 11, supporting a pathogenic classification.



Figure 2.18: The *KMT2A*:c.3576A>C germline *de novo* missense SNP detected in proband 130.

A) Its genomic position in exon 6 of the *KMT2A* gene leading to a Arg1192Ser substitution in a UniProt domain of interest. Image adapted from the DECIPHER database Firth et al. (2009). B) IGV plot of the coverage and mapped reads supporting the C change at chr11: 118480180 in the proband (top), mother (middle), and father (bottom). No evidence of mosaicism is seen in proband 130 with 26 alternate reads at a read depth of 41. No evidence of sequencing artifacts, strand bias, read end bias, low variant call quality, or low mapping quality was observed.

2.3.5.10 Variants of uncertain significance

Five of the candidate variants that passed both variant prioritization and quality control were classified as variants of uncertain significance (VUS) (Appendix H). Three of these occurred in *RORB*, *POLR3B*, and *OTX2*, which are known DDG2P genes. These cases showed partial phenotypic concordance with reported features but ultimately lacked sufficient evidence to support a likely pathogenic classification.

A novel *RORB*:c.601G>T missense SNP appeared to partially account for the GDD/ID observed in proband 17, particularly in the context of the *PMP22*:c.448G>C SNP (Table 2.6). However, *RORB* has been specifically associated with susceptibility to idiopathic epilepsy, and seizures were not reported in proband 17 (OMIM: 618357). Furthermore, the variant did not meet the criteria for PP3 even at a supporting level, although both CADD (27.1) and REVEL (0.66) suggested pathogenicity. In the absence of stronger functional evidence or a more specific phenotypic association, the variant remained classified as a VUS. Nevertheless, it may be contributing to proband 17's phenotype through epistatic interactions.

A second candidate variant, *POLR3B*:c.2714-16A>G, which was predicted to affect splicing (SpliceAI AG = 0.65) was identified in proband 58. *POLR3B* is associated with demyelinating Charcot–Marie–Tooth disease type 1I (OMIM: 619742). The variant was also exceptionally rare, being reported only five times in gnomAD v4.1.0. Inheritance was consistent with suspected gonadosomatic mosaicism in the unaffected father. While these findings suggested the potential for a likely pathogenic classification, proband 58 had a prior confirmed diagnosis of *MFSD8*-related neuronal ceroid lipofuscinosis through parallel analysis. This diagnosis also explained the presence of similarly affected siblings. Thus, the *POLR3B* variant could not be confidently classified as likely pathogenic, despite its apparent functional impact. The third case involved the *OTX2*:c.421C>T SNP in proband 75, which showed strong phenotypic concordance and good evidence for pathogenicity. However, the variant has been reported with an allele count of 5 in gnomAD, which is sufficient for a BS2-supporting code in autosomal dominant disorders. Thus, it could not be classified past VUS.

Two additional VUS variants were identified in *MATN3* and *TSC2*, both of which were predicted to be functionally impactful but occurred in genes with highly specific disease associations that did not sufficiently overlap (Appendix H). The *MATN3* variant, associated with autosomal dominant multiple epiphyseal dysplasia, was identified in

proband 46, who was reported to have signs of undergrowth. While this phenotype overlapped, the variant did not account for additional features of proband 46, including GDD with delayed speech, cerebral atrophy, and dysmorphic features. The variant may nevertheless be partially contributing to the phenotype, particularly as it was confirmed to be postzygotic mosaic. However, it was determined to be insufficient for a likely pathogenic classification. Proband 108 harboured a *TSC2* variant yet exhibited no features consistent with Tuberous Sclerosis Complex 2, which has a highly specific and recognisable phenotype.

Taken together, these five variants highlighted the interpretive challenges of variants of uncertain significance. They may partially explain observed phenotypes or suggest intriguing mechanisms, such as mosaic contribution, but without stronger support from functional, segregation, or population-level evidence, their pathogenicity remains inconclusive. Future clinical curations with emergent evidence may allow for reclassification.

2.3.6 Diagnostic yield potential in a best-case scenario

Of the 62 trios included in the final analysis, 24 already had confirmed molecular diagnoses through parallel DDD-Africa pipelines, giving a pre-screening diagnostic yield of 38.71%. Twenty-one of these were small variants, comprising 16 DNMs and five inherited variants (Appendix F). Thus, prior yields in this cohort were ~25.81% for DNMs and 8.06% for inherited small variants. The proband and parental variant prioritization pipelines presented in this chapter successfully captured all 16 previously confirmed DNMs and additionally identified one autosomal-dominant heterozygote missed by parallel analysis, demonstrating both the sensitivity of the approach and its viability as an alternative pipeline for trio-WES.

Beyond recapitulating the known diagnoses, this analysis added nine new candidate variants in probands 1, 17, 18, 23, 41, 53, 66, 112, and 130. Excluding proband 53, who already had a confirmed pathogenic diagnosis, eight were in previously undiagnosed probands. These comprised three putative germline DNMs, three inherited gonadosomatic variants, two mosaic PZMs, and one previously reported DNM reclassified as mosaic in this study. The probability that these candidates validate is difficult to predict, but insights from large-scale studies provide a benchmark.

Validation rates for mosaic candidates vary widely depending on detection method, mosaic variant characteristics, sequencing depth, VAF, and validation methodology (Wright et al., 2019). For proband SNPs, EM-Mosaic reported ~88% validation when posterior odds ≥10 were used with trio-WES at ~60× depth (Hsieh et al., 2020). However, validation rates for INDELs and inherited parental mosaics are more modest. The DDD-UK study reported sensitivity of 69% and specificity of 84% for child PZMs using binomial p-values <0.0001, with a positive predictive value (PPV) of 73% (Wright et al., 2019). For inherited parental mosaicism, only ~5% of candidates with a single ALT read validated as mosaics, compared with ~42% when ≥2 ALT reads were present. These results underscore that validation rates are strongly tied to VAF, which is itself dependant on read depth and related to binomial p-values (Wright et al., 2019). Additional biological and technical factors further influence rates of validation. Variants with low VAF (<10%) are challenging to confirm with Sanger sequencing which lacks the sensitivity to detect them, necessitating ultra-deep targeted NGS or digital droplet PCR (Wright et al., 2023). Tissue distribution adds another layer of complexity as mosaic variants confined to affected tissues may be undetectable in blood, while expanded CH in older individuals can confound interpretation of multi-tissue PZMs (Hsieh et al., 2020). Finally, validation rates depend on platform-specific error rates, and overall sequencing quality.

Given these considerations, realistic validation expectations for this study may range from 60-80% for the SNPs candidates, and from 40-60% for the inherited parental variants, dependant on VAF. DNMs are expected to validate at close to 100%. Thus, in a best-case scenario where all new candidates are confirmed, the yield across the 62 trios would rise to 51.61% (32/62), comprising 18 DNMs (29.03%), 3 proband PZMs (4.84%), and 8 inherited variants (12.90%). Even if fewer candidates validate, the ability to recapture all previously reported DNMs demonstrates the robustness of the strategy and highlights its potential to meaningfully increase diagnostic yield in African cohorts.

2.4 Discussion

The analyses described in this objective show that putative pathogenic PZMs could be screened for from standard trio-WES data in a resource limited African setting, with significant implications for DD affected families. This analysis leveraged the DDD-Africa trio study design to expand on conventional proband *de novo* small variant analysis to include putative PZM while accounting for likely false positives.

Two complementary variant prioritization strategies were successfully implemented within the DDD-small variant analysis to screen for evidence of post-zygotic mosaicism in affected DD probands and gonadosomatic mosaicism in their parents. The EM-Mosaic tool, applied through the proband variant prioritization strategy, successfully identified all previously confirmed germline DNMs from parallel DDD-Africa analysis, as well as three additional germline DNMs. By expanding the VAF thresholds beyond those used in conventional DNM prioritization pipelines as part of standard QC, this analysis effectively detected variants consistent with post-zygotic mosaicism, identifying three pathogenic or likely pathogenic small variant candidates in two probands. Additionally, one variant initially assumed to be a DNM in parallel analyses in *NSD2* demonstrated strong evidence of mosaicism, with important implications for recurrence risk assessment.

The use of trio-WES also enabled the screening of parental gonadosomatic mosaic variants, which are inherited as germline heterozygotes in affected probands. These variants would typically be missed by standard prioritization strategies employing stricter VAF filtering thresholds. Although the EM-Mosaic tool could not be incorporated into this aspect of the analysis, the approach nonetheless identified two high-level novel gonadosomatic candidates and two low-level candidates in four probands with no prior molecular diagnoses. Furthermore, this strategy detected evidence of gonadosomatic mosaicism in a previously identified *EHMT1* variant that had initially been assumed to be constitutionally inherited, with significant implications for recurrence risk assessment.

Overall, the expansion of DNM and inherited variant prioritization strategies to include variants with significant VAF deviations, combined with the implementation of filtering approaches and statistical tools to mitigate false positives from technical artifacts, improved the potential diagnostic yield of trio-WES in 62 DDD-Africa families. In total, twelve candidate variants were recommended for further validation using more

sensitive methods for detecting mosaicism. Novel candidate variants were identified in eight probands without prior diagnoses in DDD-Africa, and two previously reported variants were reclassified as having evidence of mosaicism, with important implications for recurrence risk counselling. The potential diagnostic yield increased from 37.81% to 51.62% across the 62 probands in a best-case scenario. These findings support the inclusion of mosaicism as a contributing factor to DD burden in trio-WES analyses.

2.4.1 Estimated DNM rates in DDD-Africa

Using this pipeline, 182 high confidence de novo SNPs and 29 high confidence de novo INDELs were identified in 53 probands and 22 probands respectively. This equated to 3.43 coding *de novo* SNPs and 1.32 coding *de novo* INDELs per DDD-Africa proband. The average mutational rates of SNPs and small INDELs across the genome are approximately $1.29X10^{-08}$ and $0.69X10^{-09}$ respectively, giving rise to approximately 40 de novo SNPs and 3 de novo INDELS genome wide per person per generation (Jónsson et al., 2017, Kloosterman et al., 2015). This translates to approximately 1 expected de novo variant in the protein coding region of the genome per person per generation, with coding INDELs being even rarer (Koboldt, 2020).

The coding de novo variant rate per proband in this analysis is 3-4 times higher than these estimates. Part of this discrepancy is the inclusion of PZMs to the calculation. However, even after stratifying into only the likely germline coding *de novo* variants (2.75 *de novo* SNPs and 0.82 germline *de novo* INDELs per proband) the rate of detection is still higher than previous estimates. It is possible that a significant proportion of these detected variants were false positives caused by sequencing artifacts that passed VQSR recalibration. Evidence for this hypothesis was supported by the number of variants that passed VQSR but had significant evidence of strand bias, in the form of elevated SOR, upon visual inspection in previous runs of this analysis. This led to the addition of hard strand bias filters during curation, a common and essential QC step when mosaic variants often incorporated into the analysis. However not all the 211 candidate *de novo* variants were visually inspected, especially those outside of candidate genes.

An intriguing hypothesis has been that local DNM rates differ across populations. African ancestry populations are known to harbour significantly higher genetic diversity than the European ancestry cohorts that make up most of the studies from which the

DNM rates have been calculated, with higher heterozygosity rates and smaller haplotype blocks across the genome (Choudhury et al., 2020). The “Heterozygote Instability” hypothesis states that gene conversion events at heterozygous sites during meiosis could increase local mutation rates (Amos, 2013). Substitution rates were investigated in the 1000 genomes project and were found to be consistently higher in African populations than non-Africans (Amos, 2013). However, a recent investigation into population specific DNM rates found no significant difference in DNM rates between populations even after accounting for parental age (Kessler et al., 2020). They also found no significant relationship between heterozygosity rate and DNM rate when accounting for parental age (Kessler *et al*., 2020). Finally, they found that DNM rates are not heritable and are more likely to be influenced by environmental factors (Kessler *et al.,* 2020). All together this suggests that the increased rate of DNMs seen in the DDD-Africa cohort it most likely due to technical factors, methodology, and cohort specific environmental factors that may influence DNM enrichment.

The largest factor that would impact DNM rate in this cohort is parental age. Constitutional proband DNMs originate as germline mosaic variations in a single parent, usually not being detectable in the parental somatic cells, and are transmitted to the offspring as constitutional heterozygote variants (Mohiuddin et al., 2022). Intuitively, a significant parental age effect at conception is seen in offspring with the number of DNMs detected increasing by approximately 1-3 per year for fathers and 0.24 DNMs per year for mothers (Acuna-Hidalgo et al., 2016). The sex disparity results from oogonia arresting in meiosis-I and only undergoing another round of DNA replication when maturing during meiosis-II, whereas spermatogonia stem cells are prolifically proliferative, generating ~1000 new cells per second (Fayomi and Orwig, 2018, Lee et al., 2023). Additionally, 80% of DNMs phase to the paternal haplotype (Sasani et al., 2019). The contribution of parental mosaicism to apparently constitutional DNMs has been significantly underestimated, with studies suggesting that 10–20% of reported single-event DNMs may originate from undetected parental germline or gonadosomatic mosaicism (Acuna-Hidalgo et al., 2016, Lee et al., 2023).

It was not possible to determine the parental origin of germline DNMs in the DDD-Africa probands in this analysis. This should be a future direction, as variants originating from the paternal haplotype could point to potential germline confined low-level mosaicism in the spermatogonia stem cells. These fathers could have their sperm cells genotyped at higher coverage, and in cases where low-level mosaicism is found

updated recurrence risk estimates could be given to affected families. Of particular interest would DNMs in DD associated genes. Previous investigations have found enrichment of damaging DNMs in DD cohorts (McRae et al., 2017). There is some weak preliminary evidence to suggest that this may be the case in the DDD-Africa cohort as well. Of the 211 DNMs, 59 (27.96%) were in known DDG2P genes. Another 41 (19.11%) were in non-DDG2P genes that are predicted to be high intolerant to LOF, in which additional DD associated genes are likely to be found in the future. The evidence becomes slightly stronger when considering only the most deleterious LOF SNPs and INDELs. Of the 39 predicted LOF variants, 14 (35.89%) were in known DDG2P genes.

A larger cohort of probands from full trios will be needed to fully investigate the DNM rate and enrichment in the DDD-Africa cohort, considering cohort specific mutation rates, genetic diversity in Africans, cohort specific selection pressures, parental age effects, and genetic drift. It will be very interesting to see these results as the higher genetic diversity seen across African populations and the lack of representative African samples in population matched control databases means that there is a significant chance for novel DD variant and gene discovery.

2.4.2 Technical considerations

This analysis used the EM-Mosaic tool to identify candidate *de novo* SNPs with significant evidence of arising post-zygotically in DD-affected probands. However, EM-Mosaic could not be applied to mosaic INDELs, as the authors note that the downstream beta-binomial model parameters for INDELs were unstable (Hsieh et al., 2020). Because EM-Mosaic estimates prior model parameters directly from the input data, the algorithm is highly sensitive to variation in the input set. When INDEL data are included, inaccurate overdispersion estimates are produced, which in turn lead to incorrect adjustment of mosaic SNP posterior odds (Hsieh et al., 2020). For this reason, and because INDELs have distinct mutational mechanisms and rates, INDELs were analysed separately from SNPs in the proband variant prioritisation strategy, although similar quality-control thresholds were applied to both to maintain consistency.

An additional advantage of EM-Mosaic is the incorporation of an excess DNM filter, and an FDR-based minimum ALT read support filter. The excess DNM filter identified

probands 11, 29, and 70 as having markedly elevated candidate DNM counts compared with the rest of the cohort (Table 2.2). Further investigation of inherited variant counts suggested that the paternal samples used in the trio-WES for these probands were likely incorrect, potentially due to sample handling errors or mislabelling. As contamination was already accounted for during variant-calling QC, sample misidentification remains the most plausible explanation. Another possible explanation is incorrect disclosure of paternity: if the reported fathers were not the biological fathers, paternally inherited variants would have been misclassified as DNMs. This hypothesis was supported by the observation that these probands had markedly fewer paternally inherited variants than the rest of the cohort, with no paternally inherited SNPs or INDELs passing QC in probands 11, 29, or 70, compared with cohort means of 14.10 and 0.50, respectively. As a result, these trios were excluded from the proband variant prioritisation strategy but were retained in the maternal gonadosomatic mosaicism pipeline.

The FDR-based minimum ALT read support filter further removes variants at total read depths that are more likely to be technical artefacts introduced by the Illumina NovaSeq 6000 platform. In comparison, the DDD-UK approach included all variants with ALT reads ≥2 (Wright et al., 2019). However, at low total read depths, in genomic regions that are difficult to sequence, and at the beginnings and ends of exons, it is difficult to distinguish sequencing errors from true mosaics. The DDD-UK strategy prioritised sensitivity over specificity, with confirmation achieved by ultra–high-depth sequencing of all candidates. In a resource-constrained setting, this approach is not always feasible. The FDR-based minimum ALT read support filter therefore adds a layer of statistical confidence, improving the specificity of mosaic calls.

A prior run of EM-Mosaic in this analysis, without applying a minimum ALT read support threshold for QC, resulted in all candidate DNMs with ≤3 ALT reads being excluded. None of these variants passed the filter at the total read depths achieved by WES on the Illumina NovaSeq 6000 platform. Including variants with ALT reads of 4 required substantial overdispersion correction and led to a significantly higher proportion of DNMs being classified as mosaic rather than germline. The initial mosaic fraction estimate by the EM algorithm exceeded 90% in this scenario. Variants with ALT = 4 at total read depths >10 tended to fall in a VAF range 0.20–0.30, which negatively affected EM-Mosaic's estimation of initial model parameters and inflated the initial mosaic fraction. These variants were therefore assumed to be enriched for false positives.

Restricting to variants with ALT ≥6 corrected this problem, resulting in VAF variance that fit the beta-binomial model more closely, less overdispersion correction, and downstream model estimates that aligned better with the literature. However, as EM-Mosaic WES in the original study was performed at higher coverage, this threshold caused a loss of sensitivity in the DDD-Africa data, leading to many true positive mosaic and germline variants being excluded. A compromise threshold of ALT ≥5 and REF ≥5 was therefore applied, balancing sensitivity and specificity. Variants passing this filter produced model estimates and downstream mosaic fractions more consistent with published reports. The trade-off was that true variants with fewer than 5 ALT reads in low-coverage regions that are a known issue in WES, were likely missed, as reflected by the estimated mosaic detection power of approximately 33% (Figure 2.7). Nevertheless, most sequencing artefacts were likely removed by the chosen approach, enriching the candidate list for true positives and enabling more confident recommendations for costly validation assays.

The estimated fraction of mosaic SNPs and INDELs in the DDD-Africa probands was 0.19 and 0.38, respectively, giving a combined fraction of 0.22 DNMs predicted to have arisen post-zygotically. In total, 47 PZMs were detected in 30 probands, equating to 1.57 PZMs per proband. Both SNP- and INDEL-based statistical approaches relied on detecting deviations of VAF from expectation, but only SNPs incorporated correction for overdispersion introduced by the WES protocol. This may partly explain the discrepancy between fractions. Attempts to incorporate INDELs into EM-Mosaic were abandoned, as their inclusion affected the mosaic status of SNPs in both directions, consistent with the tool being specifically designed to exclude INDELs (Hsieh et al., 2020). However, the detection of mosaic INDELs is a known limitation of all somatic variant callers presently available, especially at lower VAFs (Ha et al., 2023). Thus, even if another caller was used in this analysis, the INDELs would still need to be cautiously interpreted. More work is needed in the development of tools to overcome this limitation.

When using an FDR-adjusted binomial p-value <0.1 for SNPs, the estimated mosaic fraction was 0.47, more than double the EM-Mosaic estimate and higher than reported estimates of around 0.1 (Hsieh et al., 2020). The posterior odds cutoff of >10 chosen here is reported to correspond to an FDR of 9.1%, which may be considered too conservative (Hsieh et al., 2020). The EM-Mosaic authors predicted that around 61% of variants with posterior odds between 1 - 10 are mosaic (Hsieh et al., 2020). In this

thesis 52 candidate classified as germline by EM-Mosaic had posterior odds >1 and evidence of mosaicism. A substantial proportion of these are likely true mosaics, particularly given the limited power of detection at this depth. Indeed, 44 of these would also have passed the INDEL binomial test. Of the 11 INDEL PZM candidates, 8 were considered high confidence with binomial $p<0.0001$, as were 16 of 36 SNP PZM candidates. In the DDD-UK study, variants of this type validated as mosaic 73% of the time. Assuming a similar validation rate here, ~17 PZMs would be expected to validate, giving a mosaic fraction of 0.08 which is more comparable to estimates from other studies (Hsieh et al., 2020).

However, like for like comparisons between studies cannot be made without controlling for the variables which effect mosaic fraction estimates. These are the WES protocol from sample collection to variant calling, and the statistical approach used to detect evidence of mosaicism. This does also not take into consideration that both DNMs and PZMs may be enriched for in African DD cohorts. This cannot be determined from the data in this analysis for the above reasons. It must also be noted that the estimates from this analysis are from standard 45X WES data, with the VAFs of DNMs not being validated by higher depth sequencing. Additionally, EM-Mosaic power analysis revealed that the DDD-Africa WES depth was powered to detect only a third of mosaic SNPs in probands (Figure 2.7). A larger cohort of probands, with a larger number of DNM candidates from which to estimate model parameters and overdispersion would improve model estimates and fitting (Hsieh et al., 2020). In this context, the mosaic fractions and VAF distributions reported in this analysis are pre-validation estimates specific to only this cohort. To make fair comparisons to the estimates from the literature, the cohort size must be expanded to comparable numbers, and the methodology must be as similar as possible controlling for as many overdispersion factors as possible.

A preliminary analysis was performed to investigate factors that may have influenced the detection of PZMs in the DDD-Africa cohort. No significant relationships were found between proband sex (Coeff. = 0.052, $p = 0.85$), proband age at recruitment (Coeff. = -0.02, $p = 0.62$), paternal age at birth (Coeff. = -0.04, $p = 0.11$), or maternal age at birth (Coeff. = 0.057, $p = 0.06$) and the number of PZMs per proband with multivariate linear regression analysis ($R2 = 0.085$, $p = 0.29$). These findings are consistent with expectations for independent post-zygotic mutational events, and in line with previous reports showing no parental age effect on PZMs (Wright et al., 2019). While certain

loci may be more susceptible to post-zygotic mutagenesis, enrichment analyses were not conducted in this study. Identification of mutational signatures associated with platform specific errors could be a viable strategy to further reduce false positive calls. This could be achieved with the full DDD-Africa parental cohort to generate a platform, and cohort specific PON filter in place of relying on a sample size dependant allele count filter as was done in this analysis.

The VAF distributions of mosaic SNPs and INDELs were highly similar, with mean VAFs of 0.16 and a range of 0.11–0.27, consistent with the range most likely to validate in the DDD-UK study (Wright et al., 2019). Theoretically, these VAFs suggest mutations arising during the 2nd or 3rd embryonic cell divisions, affecting ~22–54% of cells at recruitment. However, embryogenesis is asymmetric and subject to tissue-specific selection pressures, which can drive clonal expansions or contractions (Abascal et al., 2021, Coorens et al., 2021, Li et al., 2021). In this study, VAFs showed no significant relationship with proband age (Coef. = -0.0025, p = 0.20) or sex (Coef. = 0.01, p = 0.34), likely reflecting the young age of the cohort, where large-scale clonal shifts in haematopoiesis are less common. Longitudinal data would be required to further assess these dynamics.

As DNA was sampled exclusively from blood, candidate variants were only detectable in the haematopoietic system. Since developmental disorders affect multiple organ systems, a PZM would need to be present in the relevant tissues at the time point in which it is important to be pathogenic. Each proband in this analysis, bar the two monozygotic twins from family 110, presented with wide ranges of neurological complications pointing to the central nervous system tissues being significantly impacted. For clinical curation of the candidate PZMs it had to be assumed that the candidate PZMs were present in other relevant tissues as well, at significant fractions, for the diseased phenotypes to manifest. Although confinement to blood cannot be excluded, two lines of evidence support systemic distributions. Firstly, there was no association between proband age and PZM VAFs. Secondly, the study design inherently excluded variants with VAF <0.10, meaning only high-level mosaicism (>20% cells) was captured. Given the high VAFs observed and the young age of the affected probands, these candidates are more likely to have arisen early in embryogenesis than as haematopoietic expansions. Following recent PZM classification guidelines, these variants would then fall into the moderate-to-severe tissue involvement categories (Martínez-Glez et al., 2020). The fractions of the most

impacted tissues in DDs, such as the central nervous system, the brain, the musculoskeletal system, and the heart could not be determined in this analysis without additional tissue sampling. While saliva swab DNA samples were attained for DDD-Africa probands, they had not been sequenced at the time of writing and, therefore, could not be used for comparison with blood-derived variants. Validation of candidates should be carried out on both blood and saliva samples in the future.

The sensitivity of PZM detection in this analysis was constrained by the sequencing protocol and variant calling framework employed by DDD-Africa. The project was designed to generate cost-effective, high-quality WES data to maximise diagnostic yield within a resource-limited context, with average sequencing depth typical of clinical WES. While sufficient for germline variant detection in trios, this depth was suboptimal for PZM discovery. Thresholds of VAF >0.10 and ≥5 ALT reads meant that reliable detection was restricted to regions with coverage of ~50X or greater, resulting in the likely under-calling of true variants at lower VAFs (Abascal et al., 2021, Hsieh et al., 2020, Wright et al., 2019). Detection power was further compromised by uneven coverage across exons, particularly at exon–intron boundaries, a known limitation of WES, which complicates the interpretability of candidates identified in these regions. More generally, sequencing strategy involves a trade-off between breadth and depth of coverage, with significant cost implications depending on the platform used. While the DDD-Africa protocol was appropriate for its diagnostic goals, higher-depth sequencing would substantially improve the sensitivity and accuracy of PZM calling. It would also improve VAF estimates, which is essential for recurrence risk estimation in gonadosomatic cases (Bernkopf et al., 2023).

Another important limitation arose from the use of GATK HaplotypeCaller for variant calling. GATK-HC, at the default ploidy of 2, is optimised for germline calling in diploid genomes and expects heterozygous variants to cluster around a VAF of ~0.5. Variants at the extremes of the VAF distribution are frequently filtered as artefacts, which is appropriate for germline analyses but results in loss of true somatic or mosaic variants. Consequently, many potential PZMs were likely never annotated in the VCF and were missed. Discrepancies between read counts in VCF files and those observed in IGV plots further compounded this issue. for example, the *NIPBL*:c.2584A>G variant in father 53 showed ALT = 3, DP = 6 in the VCF but 4/8 reads in IGV. At the modest sequencing depth available, such discrepancies had a disproportionate impact. Similar issues were evident in trio plots, where variants labelled as ALT = 0 in parents often

showed one or two ALT reads on manual inspection; these were usually sequencing errors, but their exclusion at the VCF level also prevented downstream filtering strategies from evaluating them.

Thus, sensitivity was limited not only by sequencing depth but also by the constraints of the variant caller itself. These shortcomings are well recognised for GATK-HC in somatic contexts, where increased ploidy settings (9-10 for 45X coverage) can improve sensitivity (Wang et al., 2021b, Ha et al., 2023). However, even with this adjustment, adjustments GATK-HC remains a poor somatic caller. Dedicated tools such as Mutect2 would have substantially improved performance but were beyond the scope of this study. This has been shown in clinical WES data like the data from this study (Sandran et al., 2024). Instead, this objective aimed to explore the feasibility of mosaic screening from data not specifically optimised for that purpose. Future work would benefit from higher, more uniform coverage, incorporation of a dedicated somatic-aware variant caller, larger cohorts to strengthen EM-Mosaic filtering and statistical power, and integration of mutational signature analyses to reduce artefactual noise.

2.4.3 Further validation of candidates is essential

The aim of the analysis in this Chapter was to sensitively screen candidate variants with evidence of mosaicism that could plausibly contribute to unexplained DDs in a resource-constrained setting, particularly for cases unresolved by parallel analyses. The candidate variants identified through these pipelines were not validated with appropriate orthogonal methods, as this was determined to be outside the scope of this study. The explicit goal was to determine if evidence of mosaicism could be identified in the DDD-Africa trio-WES data, which was not specifically designed to capture PZM, and then to prioritize the most significant mosaic candidates for downstream confirmation later by DDD-Africa. Ultimately reducing the potential cost of expensive validation protocols, which were not readily accessible in our setting. The screening strategy was set up to limit false positive candidate recommendations as much as possible. This was achieved by applying stringent filtering criteria and quality control specific to the DDD-Africa WES data, as well as requiring significant statistical evidence of mosaicism. Despite these efforts we do recognise that the candidate variants are most likely still enriched for false positive, which is inherent limitation of the WES data and the properties of mosaic variants. Sequencing artifacts are

introduced across the protocol, including in DNA extraction, DNA amplification, library preparation, and the actual sequencing (Dou et al., 2018). Thus, orthogonal validation through sensitive methods must be carried out. This limitation colours all conclusions.

One of the main considerations for not including validation in this study was the limited access to the platforms that would be sensitive enough to do so on hand in our research setting. That being Sanger sequencing, which offers limited validation applicability, as detection is generally restricted to variants with VAFs over 0.20 and requires quality primer design and high-sequencing fidelity (Martínez-Glez et al., 2020). Even under optimal conditions, the low-level peaks indicative of mosaic variants can be difficult to interpret. More sensitive approaches, such as ultra-deep targeted sequencing, ddPCR, or Multiple Independent Primer PCR Sequencing, provide superior validation power but require infrastructure and expertise that are not readily available in the resource constrained settings (Zemet et al., 2022). The recommendation for validation is to use a more sensitive platform than Sanger sequencing.

Another important consideration is tissue source. In cases of parental gonadosomatic mosaicism, accurate recurrence risk estimation depends on VAFs derived from gametes rather than peripheral blood (Bernkopf et al., 2023). Several studies have shown that blood-derived VAFs, particularly in men, correlate poorly with sperm VAFs, and lead leading to potential underestimation or overestimation of reproductive risk (Bernkopf et al., 2023, Yang et al., 2021). For this reason, future clinical studies should prioritise direct sequencing of paternal sperm samples whenever mosaicism is suspected. Similarly, in affected children, mosaic variants may be unevenly distributed across tissues due. Analysis risks missing low-level or tissue-restricted mosaics that could be clinically relevant, especially for neurodevelopmental disorders where brain-specific mosaicism may play a role. Thus, validation of candidate mosaics identified in this dataset should incorporate orthogonal testing from buccal swabs. If possible, additional sampling of sperm derived DNA in cases with suspected paternal gonadosomatic or germline mosaicism cases should be considered. This would improve sensitivity for detecting low-level mosaicism, enable more accurate assessment of tissue distribution, and provide better estimates of recurrence risk in suspected parental cases (Bernkopf et al., 2023).

2.4.4 Challenges in mosaic variant interpretation and curation

The clinical curation and variant interpretation of the candidate variants posed significant challenges. A key limitation was the absence of standardized clinical guidelines for evaluating genotype-phenotype associations in mosaic cases, where heterogeneity is considerable and is known to be attenuated by proportions of affected tissue in some cases (Waldvogel et al., 2024). A probabilistic framework that connects estimated VAFs, to phenotypic expression not been established to date. This uncertainty was compounded by additional challenges such as observable variable expressivity in developmental disorders phenotype, undefined or atypical genotype-phenotype correlations in African populations, uncharacterised epistatic interactions, pleiotropy, digenic inheritance, and the lack of direct sampling from affected tissues (Lumaka et al., 2022, Parenti et al., 2020).

Scoring ACMG/AMP codes to reflect phenotype associations in this context was inherently interpretive, with considerable potential for variability in future analyses. For the purposes of this screening, the objective was to prioritize plausible mosaic candidates rather than to provide final clinical curations. A deliberately cautious approach was adopted, assigning code strengths only where HPO terms indicated reasonable overlap between candidate variants and recorded phenotypes. These classifications should therefore be considered provisional. As ACMG/AMP guidelines continue to evolve, with expert panel-derived specifications for individual genes and conditions being updated, the curation of these variants may need to be revisited by the expert DDD-Africa clinical team to ensure accurate and clinically actionable interpretation post-validation.

In cases of potential gonadosomatic mosaicism, it was assumed that the proband inherited the variant from a parent carrying the variant in the germline at appreciable levels. This assumption could not be empirically tested, as germline derived DNA was not available, thus whole blood-derived VAFs were instead used as a proxy for systemic mosaicism. This approach rests on the rationale that variants fixed early in parental embryogenesis across both soma and germline can persist without impairing reproductive fitness (Sakumi, 2019). And that phenotype in asymptomatic parents may not be present due to tissue specific distributions, which were also not known. Whole blood-derived DNA is primarily derived from the mesoderm and may not fully capture PZMs in ectodermal tissues (Sakumi, 2019). Analysis of buccal-derived DNA, representing ectodermal lineages, would be needed to confirm systemic mosaicism

(Sakumi, 2019). While the design of this analysis relied on the theoretical validity of gonadosomatic mosaicism as a driver in parental cases, the variant prioritization strategy explicitly accounted for this uncertainty, with ACMG/AMP codes PM6/PS2 adjusted conservatively to reflect the limitations.

The paucity of African genomic data on DDs and in population databases further complicated curation. Many *in silico* prediction tools used to inform PP3 and BP4 rely on training datasets with sparse African allele representation, introducing biases that can misclassify variants in African cohorts (Bope et al., 2019). Thus, calibrated thresholds from predictors not trained on population allele frequencies were employed, thereby reducing dependency on skewed allele frequency priors. This strategy tried to account for the underrepresentation of African data in both population databases and predictive models. This paucity may also impact phenotypic curation, as databases such as ClinVar and the published literature contain very few clinically curated African cases for comparison, meaning that interpretation necessarily relies on limited non-African references.

Lastly, cases with multiple candidate variants, or those with prior confirmed diagnoses, introduced additional complexity to curation. It was often unclear which variant would contribute to a phenotype, given the potential for digenic contributions or epistatic interactions that are less recognised in DDs (Moreno-Ruiz et al., 2022). Known DD variable expressivity further compounded this uncertainty, as overlapping or atypical features could plausibly arise from more than one underlying genetic contributor. Although the prioritization in this study emphasized variants most likely to explain the phenotype in unsolved cases, plausible mosaic candidates in previously solved cases were also evaluated in full, acknowledging their potential to modify disease presentation or interact with constitutional variants. This interpretive challenge highlights both the limitations of current frameworks for variant curation and the opportunity for future studies to better integrate mosaicism into models of multilocus pathogenicity in rare disease.

2.4.5 Feasibility of trio-WES mosaic screening in African clinical settings

An important consideration is whether pipelines like the one described here can be translated from research to clinical use, particularly in resource-constrained African settings. This pipeline was specifically designed for cohort-based trio-WES screening in DDD-Africa, relying on EM-Mosaic for statistical modelling. It is not feasible for singletons, duos, or individual diagnostic cases and requires high-performance computing, labour-intensive variant prioritization, and specialized expertise that are generally unavailable in most African diagnostic laboratories. While DDD-Africa benefited from strategic partnerships and access to infrastructure, broader clinical implementation would demand substantial investment in sequencing platforms, computational resources, personnel training, and multidisciplinary capacity.

Routine clinical WES is optimized for germline variants, whereas mosaic detection requires higher sequencing depth, somatic-aware variant callers, statistical modelling, orthogonal validation, and expert interpretation. These requirements are compounded in African healthcare contexts by practical barriers such as incomplete documentation, limited medical record systems, and difficulty obtaining full trios. Even where local sequencing and bioinformatics capacity exist, this pipeline cannot be applied directly across different platforms without adaptation.

Given these limitations, the pipeline is best viewed as a research tool for prioritizing PZM candidates rather than a ready-to-use clinical workflow. Mosaic detection is highly sensitive to sequencing depth and quality. Clinically optimized approaches would include higher-depth sequencing (≥200-500X), targeted panels of known DDG2P genes for cost-effectiveness, somatic-aware callers such as Mutect2, and platform-specific panels of normals derived from local populations. Alternatively, long-read trio sequencing could further allow simultaneous detection of germline and somatic variants, including SNPs, INDELs, and structural variants. Overall, future clinical implementation will require careful adaptation, substantial resource investment, and integration with local genomic and clinical infrastructure to achieve reliable and interpretable mosaic variant detection.

Chapter 3: Characterising mosaic chromosomal alterations in an African cohort

3.1 Introduction

Mosaic chromosomal alterations are an important and distinct class of somatic variation with broad biological and clinical implications (Liu et al., 2021). When MCAs arise in haematopoietic stem cells, they can confer a selective advantage and drive clonal haematopoiesis, a phenomenon characterised by expanded clones contributing disproportionately to peripheral blood lineages (Loh et al., 2018, Terao et al., 2020).. Mosaic chromosomal alterations impact long stretches of DNA, altering the expression of hundreds of genes and disrupting 3D chromosomal topology (Liu et al., 2021). Selective advantage is gained when the expression of genes impacting HSC proliferative ability, self-renewal potential, or survival ability are altered in the background of the aging bone marrow niche (Jakubek et al., 2023a). Unsurprisingly, MCA driven CH has consistently been linked with an increased risk of haematological malignancy (Loh et al., 2020, Niroula et al., 2021, Terao et al., 2020).

The development of MoChA has allowed for the most accurate and sensitive detection of MCAs from high throughput genotyping data to date (Loh et al., 2018, Loh et al., 2020). Characterizations of autosomal MCAs in large biobank-scale cohorts have shown that MCA driven CH is a normal part of the aging process that does not always lead to malignancy (Loh et al., 2018, Terao et al., 2020). However, MCAs have also been increasingly associated with other adverse health outcomes including abnormal cell counts, immune deficiency, risk of diverse infection, and all-cause mortality (Dawoud et al., 2022, Lin et al., 2021, Loh et al., 2018, Terao et al., 2020, Zekavat et al., 2021).

These studies have transformed our understanding of MCA driven CH. However, an important limitation of the current literature is its lack of ancestry diversity, and most large MCA screens are heavily biased towards European-ancestry cohorts, with few studies that robustly sample continental African populations (Dawoud et al., 2022, Lin et al., 2021, Loh et al., 2018, Terao et al., 2020, Zekavat et al., 2021). The inclusion of African ancestry samples has been very limited, reflecting the paucity of African ancestry samples in large biobank-scale GWAS cohorts (Martin et al., 2019, Sollis et

al., 2023). Early evidence from small African-ancestry sub-cohorts suggests differing MCA risk of occurrence when compared with European samples, but methodological caveats and use of inferred ancestry from diaspora cohorts caution against overgeneralisation to continental populations (Jakubek et al., 2023b, Machiela et al., 2015). To date, there have been no studies characterizing MCAs in large continental African ancestry cohorts utilising SNP-arrays designed to capture African-specific genetic variations. It is likely that continental African ancestry populations have diverse genetic drivers of MCA selection, impacted by unique socio-economic and infection pressures, that have yet to be uncovered. Consequently, the complete landscape of MCA driven CH in African ancestry populations remains yet to be fully elucidated.

This chapter aimed to address this gap in the MCA knowledge base through an exploratory characterisation of autosomal MCAs in participants from the AWI-Gen cohort who had previously been genotyped H3Africa SNP-microarray. This analysis provided the first insights into the landscape of MCAs in continental African ancestry participants using appropriate genotyping and bioinformatic tools. These results significantly contribute to the global landscape of MCAs by adding African ancestry data to the literature. Results of this analysis will serve to identify regions of the African genome which are prone to MCA formation and will be used in future analyses to identify African-specific drivers of MCA selection that can be used as biomarkers of poor health outcomes in Africans.

3.2 Methodology

3.2.1 Study design and ethics

This was a retrospective analysis to detect MCAs in the cross-sectional AWI-Gen cohort from H3Africa SNP-array genotyping data using the MoChA tool. Ethics approval for this study was granted by the University of the Witwatersrand Human Research Ethics Committee (HREC) under clearance certificate number M2011144 (Appendix A).

3.2.2 Study cohort: AWI-Gen

AWI-Gen is a cross-sectional population study set up to investigate the genetic, environmental, and behavioral contributions to body composition and cardiometabolic disease on the African continent (Ramsay et al., 2016). The cohort is made up of 11,723 seemingly healthy male and female AA participants recruited from six sites located in Nanoro (Burkina Faso), Navrongo (Ghana), Nairobi (Kenya), Agincourt (South Africa), and Dikgale (South Africa), reflecting diversity from the Western, Eastern, and Southern African regions (Ramsay et al., 2016). Participants were randomly sampled with no recruitment criteria for infection history, disease history, behavioral history, body composition, or environmental exposures (Ali et al., 2018).

Data was available for 10,702 participants between the ages of 40 and 60 years recruited from all six sites, plus an additional 1021 participants over the age of 60 years from Agincourt (Ali et al., 2018). Extensive socio-demographic data was available for the cohort (Ali et al., 2018). Participant age, sex, region, and site of recruitment were extracted as variables of interest for this analysis as they have been shown to impact MCA driven CH in other cohorts (Loh et al., 2018, Terao et al., 2020). More details on the scope, methodology, and recruitment outcomes of AWI-Gen can be found in the relevant white papers (Ramsay et al., 2016, Ali et al., 2018).

3.2.3 Genotyping with the H3Africa SNP array

The full AWI-Gen cohort was previously genotyped on the Infinium H3Africa SNP array V1 (https://chipinfo.h3abionet.org/) at the Illumina® FastTrack™ Microarray services

(Illumina, San Diego, USA) as part of the AWI-Gen study. The H3Africa SNP array was developed by the H3Africa Consortium, in partnership with Illumina®, to address the need for a cost-effective African-centric array that accounted for the much larger genetic diversity and smaller haplotype blocks seen in African genomes (Mulder et al., 2018). The high-density array contains approximately 2,270,000 SNP markers. The base content of the array consists of approximately 1,700,000 SNP markers, common to African populations, that were selected from existing Illumina bead pools for their efficiency as tag-SNPs (Mulder et al., 2018). The array also contains approximately 700,000 additional SNPs selected to improve African genome coverage and downstream imputation accuracy (Mulder et al., 2018). The AWI-Gen H3Africa SNP array genotyping data is hosted on the H3ABioNet ZA-Wits-core high performance computing node and was made available for this analysis.

3.2.4 Detecting MCAs from SNP array data: MoChA

MoChA is a BCFtools extension designed to sensitively and accurately detect mosaic chromosomal alterations from SNP-array and high-coverage WGS data (https://github.com/freeseek/mocha/) The algorithm incorporates haplotype phase information from heterozygous SNPs into a Hidden Markov Model (HMM), allowing it to distinguish true mosaic events from technical artifacts (Loh et al., 2018, Loh et al., 2020). True MCA events alter the balance of maternal versus paternal genomic material in a cell population, causing a deviation in BAF values at heterozygous sites which are detected in computationally phased intensity data through the HMM (Loh et al., 2018, Loh et al., 2020). The full details of the statistical models underlying the MoChA tool can be found in its seminal papers (Loh et al., 2018, Loh et al., 2020). MoChA was selected for this study because it can detect smaller MCAs at lower cell fractions more accurately than other available methods (Loh et al., 2018, Loh et al., 2020).

3.2.5 Pipeline for detecting autosomal MCAs with MoChA

The AWI-Gen cohort was previously genotyped for various GWAS analyses (Choudhury et al., 2022). This generated a fully quality controlled multi-sample Binary Variant Call Format (BCF) file that was phased and imputed using the African Genome

Resources reference panel, currently the best reference panel for the phasing and imputation of African genomes (Sengupta et al., 2023a, Gurdasani et al., 2019). This BCF file was made available by AWI-Gen for this study; however, it did not contain LRR or BAF metrics for genotyped SNPs. These intensity values are not usually needed for downstream GWAS analysis but are essential to calling structural variations from SNP array data (Balagué-Dobón et al., 2022).

A new multi-sample BCF file with the required LRR and BAF metrics was created from the Illumina raw intensity files using the BCFtools plugin gtc2vcf (https://github.com/freeseek/gtc2vcf) following the authors recommendations. The AWI-Gen site and sample level quality control were applied to the new BCF file, which was subsequently used as the input for the MoChA pipeline. A NetFlow (https://www.nextflow.io/) workflow was generated to implement the gtc2vcf and MoChA pipelines on the H3ABioNet ZA-Wits-core high performance computing node. This workflow was set up with the help of Dr. Phelelani T. Mpangase (Sydney Brenner Institute for Molecular Bioscience, University of the Witwatersrand) and can be found at this GitHub repository (https://github.com/phelelani/nf-mochachoca). The workflow is summarized in Figure 3.1 and outlined in detail below.

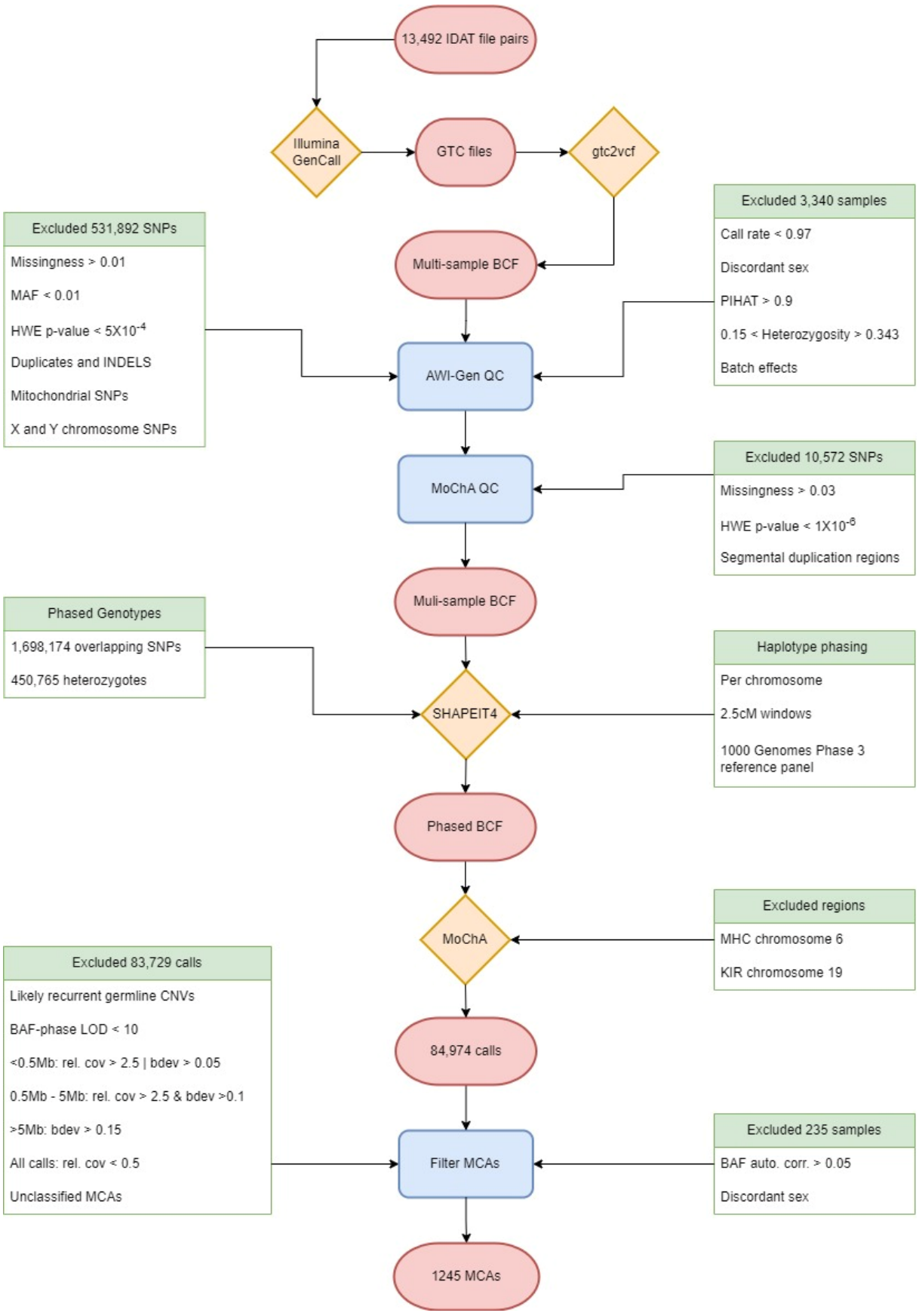


Figure 3.1: The MCA detection workflow for the AWI-Gen dataset.

Summary of the workflow implemented to detect MCAs using the MoChA pipeline. Red shapes indicated files. Blue shapes indicate processes. Yellow shapes indicate tools. Green boxes indicate specific parameters implemented in the workflow.

3.2.5.1 Converting Illumina H3Africa SNP array intensity data files for MoChA input.

The first step of the workflow implemented the gtc2vcf plugin pipeline to convert the raw H3Africa SNP array V1 intensity data (IDAT) files stored on the H3ABioNet ZA-Wits-core high performance computing node to a single multi-sample BCF file. The gtc2vcf pipeline implemented the Illumina Array Analysis Platform (IAAP) software (https://emea.support.illumina.com/array/array_software/illumina-array-analysis-platform.html) to call SNP genotypes from raw IDAT files using the Illumina GenCall algorithm. Illumina Genotype Call (GTC) files were generated for each participant from IDAT file pairs using a Bead Pool Manifest (BPM) file and a cluster (EGT) file. The manifest file contains SNP probe identities and genomic locations in relation to the GRCh37 reference genome, while the cluster file contains the expected intensity of each genotype cluster and was generated by the run at Illumina. The manifest and cluster files for the H3Africa SNP array V1 were made available by H3ABionet at https://chipinfo.h3abionet.org/downloads.

The original AWI-Gen genotyping was run in four independent batches containing overlapping samples to identify potential batch effects, resulting in 13,492 IDAT file pairs. All 13,492 IDAT file pairs were converted to GTC files using the gtc2vcf pipeline in a single run. The next step of the gtc2vcf pipeline converted the GTC files to a single multi-sample BCF file with the gtc2vcf plugin. All 13,492 files were called and converted to a single multi-sample BCF file in a single run to avoid additional batch effects. SNP probes were aligned to the GRCh37 version of the human genome reference. The GRCh37 reference genome was downloaded at (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/human_g1k_v37.fasta.gz) and prepared according to the gtc2vcf authors recommendations. In total 2,262,924 SNPs were called. A statistics file with per sample information on call rates and computationally inferred sex from sex chromosome probe intensity was generated for downstream quality control. The BCFtools annotation plugin was then used to create a minimal BCF file with only the required annotations for each SNP probe, including sample genotype, surrounding GC content and the LRR and BAF values.

3.2.5.2 Applying AWI-Gen GWAS quality control.

Site and sample level GWAS quality control (QC) was applied to the multi-sample BCF file, as required by MoChA prior to calling MCA events. Extensive site and sample level QC was previously performed using PLINKv1.9 on the H3ABioNet/H3Agwas pipeline (https://github.com/h3abionet/h3agwas) for downstream GWAS analysis (Chang et al., 2015). The AWI-Gen SNPs and samples passing this QC were deemed appropriate for this exploratory analysis. The AWI-Gen QC are outlined below.

SNP sites were excluded if they showed evidence of missingness (>0.01), which removed SNPs with low call rates across the cohort (Marees et al., 2018). Rare SNPs with minor allele frequencies (MAF) < 0.01 were excluded, as these SNPs are prone to genotyping errors and have low power to detect associations while still impacting multiple correction testing (Marees et al., 2018). Duplicate SNPs and INDELS were removed, along with SNPs that deviated from Hardy-Weinburg equilibrium (HWE) with p-values < $5X10^{-04}$, as these may be indications of genotyping error (Marees et al., 2018). Mitochondrial SNPs and sex chromosome SNPs were excluded to limit analysis to the diploid autosome (Sun et al., 2023). The exclusion of sex chromosome sites limited this exploratory analysis to autosomal MCAs. Future analysis, preceded by sex chromosome SNP QC, will be required to characterize mosaic loss of the X and Y chromosomes.

Samples with discordant sex information between recorded sex and genetically inferred sex were excluded. Genetic sex was inferred with PLINK using the F-statistic, a measure of the rates of heterozygosity/homozygosity on the X-chromosome (Chang et al., 2015, Marees et al., 2018). Individuals with F-statistics > 0.8 were inferred genetic males, while those with F-statistics < 0.2 were inferred genetic female (Marees et al., 2018). Samples with a hard mismatch between inferred genetic and reported sex, which may imply incorrect sample handling or labelling, were excluded (Marees et al., 2018). Samples with SNP call rates <97% were excluded. Participants with the PLINK calculated PIHAT score > 0.9 were considered as either biological replicates, sampling handling errors, or duplicates and were excluded. Samples with heterozygosity levels > 0.343, indicative of sample contamination, were excluded (Marees et al., 2018). Samples with heterozygosity levels < 0.15, indicative of inbreeding, were excluded (Marees et al., 2018). Finally, population stratification was assessed through principal component analysis based on a linkage disequilibrium

pruned set of SNPs. Additionally, genotype concordance of overlapping samples in the four batches sent for genotyping was used to rule out any batch effects.

AWI-Gen QC produced lists of 10,902 participant IDs and 1,731,032 SNP probe IDs that passed quality control for use in GWAS analyses. These lists were applied to filter the multi-sample BCF file in this study using BCFtools. During downstream analyses, it was discovered that an outdated participant file had been used to match AWI-Gen study IDs to IDAT file pair names during the sample filtering process. This outdated file did not include participants from batch 4. Consequently, the final set of participants included in the MoChA pipeline comprised only those from batches 1 to 3 who passed AWI-Gen QC, resulting in 10,152 participants and 1,731,032 SNPs taken forward for analysis.

3.2.5.3 Additional quality control as recommended by MoChA for redundancy.

An additional list of SNPs and samples to be excluded from the haplotype phasing and MCA-calling pipeline was generated following the recommendations of the MoChA authors. MoChA SNP QC filtered out SNPs with >3% missingness and those with HWE p-values < 1×10^{-6}. Although these thresholds were less stringent than those applied during AWI-Gen QC, the filtering was repeated as a precaution. SNPs located within segmental duplication regions with less than 2% divergence were also excluded. Segmental duplications are large, highly homologous genomic segments that are common drivers of structural variation and play important roles in genome evolution (Bailey et al., 2001). However, they can confound haplotype phasing and mosaic variant detection, and their exclusion is recommended by MoChA (Loh et al., 2018). A list of known segmental duplications was downloaded from the UCSC genome annotation database (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/genomicSuperDups.txt.gz) and annotated with minimal divergence intervals using BCFtools and MoChA preprocessing scripts. Using BCFtools, a BCF file was generated containing 10,527 additional SNPs to exclude from the phasing and MCA-calling pipelines. Sample-level QC confirmed that all participants had Illumina GenCall call rates >97%, and no additional samples were excluded, as all samples passing AWI-Gen QC met this threshold.

3.2.5.4 Statistical haplotype phasing the multisampling VCF with SHAPEIT4.

The core principle of MoChA for detecting MCAs is leveraging accurate long-range haplotype phase information to identify imbalances in maternal and paternal allelic fractions (Loh et al., 2018, Loh et al., 2020). Haplotype phasing estimates the sequences of alleles co-located on the same chromosome, allowing detection of allelic imbalance across the genome (Browning and Browning, 2011). Reference-based genotype phasing was performed on the filtered multi-sample BCF file using SHAPEIT4, as recommended by the MoChA authors (Delaneau et al., 2019). SHAPEIT4 allows for accurate haplotype phasing of large SNP-array data sets, allowing the incorporation of additional phasing information from population reference panels (Delaneau et al., 2019). For this analysis, the 1000 Genomes Phase 3 reference panel was used, containing 2,504 genomes, including 661 individuals of African ancestry, predominantly from West Africa (Auton et al., 2015).

Genotypes were extracted from the filtered multi-sample BCF file and split into 22 autosomal chromosome-specific BCF files for input into SHAPEIT4. Phasing was performed per chromosome using SHAPEIT4 v4.2.1 with default settings and a minimum phasing window of 2.5 cM, incorporating the 1000 Genomes Phase 3 reference panel. Chromosomal recombination rates were provided via GRCh37 genetic map files downloaded from the Broad Institute (https://data.broadinstitute.org/alkesgroup/Eagle/downloads/tables/genetic_map_hg19_withX.txt.gz). The GRCh37 version of the 1000 Genomes Phase 3 reference panel was downloaded from (ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/) and prepared for each chromosome with code provided by the MoChA pipeline. SHAPEIT4 reference-based phasing only considers SNP sites that overlap between the input cohort BCF and the reference panel. in this case, 1,698,174 SNPs overlapped, of which 450,765 were heterozygous. The resulting phased chromosome BCFs were concatenated into a single BCF file using BCFtools, and phased genotype annotations were incorporated back into the filtered multi-sample BCF file via BCFtools annotation. This final phased BCF served as input for MCA calling using MoChA. No imputation was performed, as it was unnecessary for MCA detection and germline SNP association analysis was outside the scope of this study.

3.2.5.5 Calling candidate autosomal MCAs with MoChA.

MCAs were called from all samples in the filtered and phased multi-sample BCF file with MoChA in a single run to ensure no additional batching. Call rates and computationally inferred sample sex were inputted from the statistics file generated by the gtc2vcf pipeline. Likely germline structural variants for downstream removal were flagged by MoChA through the addition of a BED file containing regions of the genome that contain recurring deletions and duplications. This file was generated from 1000 Genomes phase 3 data downloaded at (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/integrated_sv_map/ALL.wgs.mergedSV.v8.20130502.svs.genotypes.vcf.gz{,.tbi}) and prepared in accordance with the MoChA pipeline instructions. Additional options were included to flag the GRCh37 regions containing the ~6Mb Multiple Histone Complex region on chromosome 6 and the ~1Mb leukocyte receptor complex region on chromosome 19 following the MoChA pipeline recommendations. MCA calls within these regions were excluded by MoChA as they have been found to be enriched for false positives due to high sequence divergence between the HLA class I and II genes and the highly homologous Killer Cell Immunoglobulin-like Receptor genes (Loh et al., 2018).

The MCA detection run generated a genome statistics file with relevant run statistics for each sample in the BCF file and an MCA call set file with a list of 84,974 putative chromosomal alterations detected by MoChA. The total call set included both germline and false positive mosaic chromosomal alterations that required additional filtering and curation.

3.2.5.6 Post-processing and filtering the putative autosomal MCA call set.

Additional sample-level quality control was applied to the MoChA call set using the per-sample genome statistics file generated by the pipeline, following the authors' recommendations. Samples with phased BAF autocorrelation greater than 0.05 were excluded, as high autocorrelation across consecutive heterozygous sites can indicate DNA contamination or poor DNA quality. Samples showing discordance between the Illumina GenCall-inferred sex and self-reported sex at recruitment were also removed. This step resulted in the exclusion of 235 additional samples, yielding a final dataset of 9,917 participants.

To remove probable germline structural variations from the initial MoChA call set of 84,974 MCAs, a filtering strategy was implemented. Calls flagged by MoChA as overlapping likely recurrent germline structural variants, based on the 1000 Genomes Phase 3 common variation list, were removed. Putative calls with log10-odds (LOD) scores below 10 were excluded. Likely germline duplications were filtered using size-specific relative intensity and phased BAF deviation (bdev) thresholds as recommended by MoChA: MCAs <0.5 Mb with relative coverage >2.1 or bdev >0.05, MCAs 0.5–5 Mb with relative coverage >2.5 or bdev >0.1, and MCAs >5 Mb with bdev >0.15 were removed. Germline deletions were excluded by removing calls with relative coverage <0.5. No explicit size or cell fraction threshold was set for MCA exclusion. No limit on the number of informative heterozygous probe sites in a call was set. There were also many calls where the type could not be explicitly determined, due to the power of imbalance detection exceeding the power to distinguish between CN-LOHs and gains or losses (Loh et al., 2018). These were also removed. Following this post-processing, 83,729 putative MCAs were removed from the dataset.

3.2.6 Statistical analysis

Statistical analysis was performed using packages in the R V4.2.2 environment for statistical computing on the RStudio platform (Team, 2022a, Team, 2022b). Sample demographic statistics were extracted from AWI-Gen generated reports. These were combined with the sample statistics generated from the MoChA run used in downstream analysis.

3.2.6.1 Initial univariate analyses.

Variables were initially assessed for significant associations using univariate and bivariate analyses. Visualization and statistical testing were performed in R using the ggstatsplot and ggplot2 packages (Patil, 2021). The distribution of continuous variables including participant age, MCA length, and MCA cell fraction was assessed by visualization with histogram plots and with Shapiro-Wilks tests of normality. As none of the continuous variables were normally distributed (p-value < 0.0001), non-parametric statistical tests were used in downstream hypothesis testing. Spearman's rank correlation was used to measure significant correlation between continuous variables. The Chi-squared test of independence was used to test for association between

categorical variables. Where sample size was under 5, a Fischer's exact test was used instead. Fisher's exact tests were used to calculate odds ratios of categorical counts. Mann-Whitney U tests were used to test significant differences in the distributions of two categorical groups of continuous variables. The Kruskal-Wallis H tests were used to test for significant differences in the distributions of three or more categorical groups of continuous variables. P-values were adjusted for multiple comparisons with the Bonferroni - Holm and Benjamini-Hochberg false discovery rate procedures. Univariate binomial logistic regression was used to test relationships between categorical dependent variables and independent variables with a generalized linear model.

3.2.6.2 Multivariate logistic regression analysis for odds of autosomal MCA occurrence.

Following the identification of significant relationships between study variables and the occurrence of autosomal MCAs, multivariate analysis was conducted to investigate the combined effects of variables on MCA prevalence. Participants with at least one detected autosomal MCA were defined as positive cases versus those without any detected MCAs. Positive MCA cases were defined as the dependent variable. Participant age, sex, region of recruitment, site of recruitment, and sample batch were investigated as independent variables. Age, sex, and region of recruitment were assessed as fixed-effect variables in multivariate logistic regression models. Sample batch and site of recruitment were assessed as both fixed and random effects using the *lme4* R package.

Models were built stepwise by establishing significant impacts on MCA occurrence and investigating interactions with interaction terms and marginal effect plots, starting with null models. Model fit and diagnostic plots were assessed with the *performance* R package flowing the recommended workflow. Models were compared by assessing Akaike Information Criterion (AIC) scores, with the goal of reducing complexity and maximizing likelihood. Lower AIC was considered better. Chi2 tests of goodness of fit p-values < 0.05 were used to establish significant predictors and to compare significant model improvements. with lower values indicating better fits. Models were assessed for collinearity with variance inflation factor (VIF) scores. Variables were required to have VIF< 3. Model plots, marginal effect plots, and model summaries were generated

with the *sjPlot* and *jtools* R packages. Odds ratio heatmap plots were generated with *ggplot2* R package.

The impact of participant age and sex on MCA odds were assessed individually and then in combination with no interactions detected. A model of MCA ~ Age + Sex was taken forward. Region of recruitment was added to the model as a fixed effect, with no evidence of collinearity or interactions. The impact of batching effects was assessed by adding sample batch to the model. Batch was assessed as a fixed effect variable and as a random-effect variable in mixed models. Three primary models were assessed. A fixed-effect model of MCA~ Age + Sex + Region + Batch, a random intercept mixed effect model of MCA~ (1+Batch) + Age + Sex + Region, and a random slope mixed effect model of MCA ~ Age + Sex + Region + (Age + Sex + Region | Batch). The variance introduced by the random effect was assessed with the Intraclass Correlation Coefficient (ICC). ICC for batch in both cases was very small indicating no significant clustering of variables within batches. The simpler fixed effects model was taken forward.

Similarly, site of recruitment, which is nested within the region variable, was assessed as a random-effects variable a random intercept mixed effect model of MCA~ (1+Site) + Age + Sex + Batch + Region, and a random slope mixed effect model of MCA ~ Age + Sex + Region + (Site | Region). As the ICC was low in both cases, and no significant odds differences were observed between sites within a region in a fixed effects model of MCA ~ Age + Sex + Batch + Site, the site variable was dropped from the model to reduce complexity and improve fit. Addition of both site and region as fixed effects led to significant collinearity as expected. The addition of both Site and Batch as random effect variables did not improve model fit. The final model was MCA ~ Age + Sex + Region + Batch which captured the variable categories most associated with MCA occurrence odds. Region was the variable considered for comparisons in further analysis.

3.2.6.3 Visualization of autosome wide MCA distribution with Circos plots

Visualization of individual MCA calls was achieved through the plotting of Circos pileup plots (Krzywinski et al., 2009). Individual MCA calls were converted to BED file format and plotted in circular figures with the R package *circlize* (Gu et al., 2014). The proportion of the genome impacted by MCAs occurrence was assessed in 1Mb

windows. Each chromosome was split into 1Mb windows through a custom R script, generating a 1Mb window BED format file. The count of MCAs overlapping each 1Mb window was determined with BEDtools *intersect* (Quinlan and Hall, 2010). The proportion of MCAs in each window was calculated from the total MCA count per chromosome. Proportions were rescaled with the most common MCA in a chromosome always equaling 1. The average informative probe coverage across the autosome was assessed by extracting successfully phased heterozygous SNP positions passing QC from the phased multi-sample BCF file with BCFtools (Danecek et al., 2021). Extracted SNPs were outputted in BED file format, with positions reflecting all SNPs with at least one heterozygous genotype in the 9917 participants. This resulted in 418,595 informative SNPs used in the calling of phased BAF deviations across the genome by MoChA. The mean probe coverage per 1Mb genome window across the 9917 participants was determined by counting the SNPs overlapping each 1Mb window with BEDtools *intersect* (Quinlan and Hall, 2010). The mean SNP coverage, and MCA coverage was then plotted across the autosome.

3.2.6.4 Identifying plausible candidate genes underlying the selection of commonly occurring autosomal MCAs.

Two approaches were combined to comprehensively screen commonly deleted and duplicated regions of the autosome for candidate genes that could explain their increased prevalence in the AWI-Gen cohort. The first approach mirrored the method employed by the BBJ analysis, evaluating the impact of overlapping genes in an MCA by taking the total genes involved into account (Terao et al., 2020). MCAs were annotated with overlapping genes from the GrCH37 version of the National Center for Biotechnology Information (NCBI) RNA references sequences (RefSeq) curated database (O'Leary et al., 2016). Curated genomic coordinates were downloaded in BED file format from the UCSC Genome Browser Table Browser (https://genome.ucsc.edu/cgi-bin/hgTables/) RefSeq curated track aligned to GrCH37 (Kent et al., 2002). A BED file was generated for detected MCAs. BEDOPs *bedmap* was then used to annotate MCAs with overlapping genes regions (Neph et al., 2012). The number of genes overlapping each MCA was tallied. An adjusted score was given to each MCA which equaled 1/gene count. If a single gene was present in an MCA, it would be weighted as being the only contributor. Genes were then assigned numeric scores by adding all the count adjusted MCA scores of the MCAs that contained them.

A custom Python script was used to assign scores. Genes were considered candidate contributors if they overlapped with 3 or more MCAs in a chromosome and had a score of 0.4 or more. Genes were then manually curated and common regions were confirmed by inspecting pileup plots.

The second approach was adapted from the UKB study (Loh et al., 2018). Due to the comparatively smaller MCA count in this study, the first approach heavily penalized common regions with larger overlapping MCAs while inflating the scores of smaller MCAs. The number of genes in the annotation significantly impacted which regions were picked up by the algorithm. As a result, many significantly CH associated genes in CDRs were excluded. To account for this, larger 1Mb regions in common regions were investigated. The list of genes in the annotation was restricted to known pan-cancer drivers to limit the large number of genes found annotated in 1Mb windows. A candidate gene list was generated by combining published hematological malignancy driver genes, pan-cancer driver genes, and a list of literature review-based leukemia associated genes from the OncoVar (https://oncovar.org/) and dbLGL databases (http://soft.bioinfo-minzhao.org/lgl/) (Liu et al., 2018, Wang et al., 2021a).

Gene regions for the list of candidate genes were extracted from the GrCH37 NCBI RefSeq BED file. Additional annotations for the T-cell and B-cell V(D)J antigen receptor loci that were not present in the BED file were manually added. To account for length, MCAs were given a weighted score of 6Mb/MCA length. A maximum score of 1 was given to MCAs under 6Mb in length. The autosome was then divided into 1Mb windows in BED file format. BEDtools *intersect* was used to annotate each 1Mb window with a list of overlapping MCAs (Quinlan and Hall, 2010). The adjusted scores of overlapping MCAs were then added together for each 1Mb window. Windows with scores of 3 or more were defined as commonly impacted MCA regions. BEDOPS *bedmap* was used to annotate commonly impacted windows with overlapping candidate driver genes underlying the MCA calls. Only recurrent losses and gains were suitable for candidate driver analysis as they directly impact gene dosage. Losses and gains were analyzed separately. The mechanism behind CN-LOH selection involves unmasking rare homozygous alleles in driver genes, which requires GWAS level association and allele shift associations outside the scope of this exploratory analysis.

3.2.6.5 Determining statistically significant MCA distributions

To identify the recurrent MCA regions with enrichment between male and female participants, MCAs were stratified by type. The autosome was then divided into 1Mb windows by chromosome. Each discrete MCA was counted in the 1Mb windows in which they were detected. The total loss, gain, and CN-LOHs overlapping each 1Mb window was tallied separately for male and females. Fischer's exact tests were conducted on the proportion of MCAs in each 1Mb window between male and female participants to account for the reduced MCA counts introduced by stratification. Concurrent windows were manually curated and combined into larger window loci with evidence of MCA sex enrichment The process was repeated in a similar manner for participants from each region of recruitment. This analysis was univariate and dd not consider additional confounding factors. Significant regions were plotted as Circos pileup plots. Individual MCA calls were converted to BED file format and plotted in circular figures with the R package circlize (Gu et al., 2014). BEDtools *intersect* was used to annotate each significant MCA region with a gene region derived from oncogenic or tumor suppressive lists from earlier analysis (Quinlan and Hall, 2010)

3.2.6.6 Phase switch error analysis

Univariate analysis of phase switch errors was carried out similarly to the demographics analyses. The R packages ggstatsplot and ggplot2 were used for the visualization of relevant distributional plots and statistical hypothesis testing (Patil, 2021). Mann-Whitney U tests were used to test significant differences in the distributions of two categorical groups of continuous variables. The Kruskal-Wallis H tests were used to test significant differences in the distributions of three or more categorical groups of continuous variables. P-values were adjusted for multiple comparisons with the Bonferroni - Holm and Benjamini-Hochberg false discovery rate procedures.

Multivariate models of association were fitted to the independent variable of phase switch error counts adjusting for MCA size (Mb), MCA type (Loss/gain/CN-LOH), region of recruitment, and batch. Models were built stepwise by establishing significant impacts on MCA occurrence and investigating interactions with interaction terms and marginal effect plots, starting with null models. Model fit and diagnostic plots were assessed with the performance R package flowing the recommended workflow.

Models were compared by assessing Akaike Information Criterion (AIC) scores, with the goal of reducing complexity and maximizing likelihood. Lower AIC was considered better. Chi2 tests of goodness of fit p-values < 0.05 were used to establish significant predictors and to compare significant model improvements. with lower values indicating better fits. Models were assessed for collinearity with variance inflation factor (VIF) scores. Variables were required to have VIF< 3. Model plots, marginal effect plots, and model summaries were generated with the sjPlot and jtools R packages. Odds ratio heatmap plots were generated with ggplot2 R package.

An initial Poison model was tested with independent variables as fixed effects: Phase switch flip count ~ MCA size + Region + MCA type + Batch using a generalized linear model with the family set to Poison As the distribution of phase switch error flip counts was non-normally distributed when fitting a Poisson multivariate model which detected significant overdispersion., a negative binomial model was fitted instead using the R package *MASS*. Zero inflation models were additionally tested , with no evidence arising.

3.2.6.7 Determining MCA distribution differences between population cohorts

Data for the UKB and BBJ cohorts was taken directly from the relevant supplementary tables from each publication and was not changed in any way, with permission from Nature Springer (Loh et al., 2020, Terao et al., 2020). Univariate analysis was carried out as previously derived with the R packages *ggstatsplot* and *ggplot2* were used for the visualization of relevant distributional plots and statistical hypothesis testing (Patil, 2021). Mann-Whitney U tests were used to test significant differences in the distributions of two categorical groups of continuous variables. The Kruskal-Wallis H tests were used to test significant differences in the distributions of three or more categorical groups of continuous variables. P-values were adjusted for multiple comparisons with the Bonferroni - Holm and Benjamini-Hochberg false discovery rate procedures. To identify the differential distributions of MCA frequencies between studies, MCAs were stratified by type and chromosome into an MCA phenotype. Fischer's exact tests were conducted on the proportion of MCAs phenotypes between each study using the total number of detected MCAs in each study as reference.

3.3 Results

3.3.1 Study cohort demographic characteristics

The demographic distribution of the final study cohort is summarized in Table 3.1. The study cohort totaled 9,917 participants, of which 45.2% were male (Table 3.1). Participants ranged in age from 29 years to 81 years, with a total cohort median age of 51. The majority of study participants were recruited from sites in Southern Africa (N = 4,797), followed by West Africa (N = 3,209), and then East Africa (N = 1,911). The demographic makeup of participants varied across the three African regions (Table 3.1).

Table 3.1 Study cohort demographics stratified by African region and site of recruitment.

Study cohort	N (%)	Median age (range)	Male/ female (% male)
AWI-Gen	**9,917 (100)**	**51 (29-81)**	**4,482 / 5,435 (45.2)**
Southern Africa	**4,797 (48.4)**	**53 (29-81)**	**2,030 / 2,767 (42.3)**
Agincourt	2,177 (22.0)	57 (40-81)	909 / 1,268 (41.8)
Dikgale	1,183 (11.9)	52 (29-81)	351 / 832 (29.7)
Soweto	1,437 (14.5)	49 (39-61)	770 / 667 (53.6)
Eastern Africa	**1,911 (19.3)**	**48 (35-67)**	**880 / 1,031 (46.1)**
Nairobi	1,911 (19.3)	48 (35-67)	880 / 1,031 (46.1)
Western Africa	**3,209 (32.4)**	**50 (38-69)**	**1,572 / 1,637 (48.9)**
Nanoro	1,582 (16.0)	50 (38-69)	831 / 751 (52.5)
Navrongo	1,627 (16.4)	51 (40-61)	741 / 886 (45.5)

N = number of participants; % = percentage of cohort stratified by region or site

Differences in participant age and sex distributions, which have been shown to impact the risk of autosomal MCA occurrence, were investigated through stratification, univariate, and bivariate association analysis. Most of the cohort were between the ages of 40 and 60 years at the time of recruitment (89.2%) (Table 3.1). Only 10.3% of the cohort participants were over the age of 60 at the time of recruitment (Table 3.1). These elderly participants were of particular interest as they were considered the highest risk for MCA detection (Loh et al., 2018, Terao et al., 2020). Reflecting the AWI-Gen recruitment strategy, only 51 participants (0.5%) were under the age of 40 (Table 3.1, Figure 3.2A). The distribution of age was consistent between male and female

participants (Figure 3.2A), with no significant differences in the median age of males and females in the overall cohort (51 vs 51, p = 0.20) (Figure 3.2B).

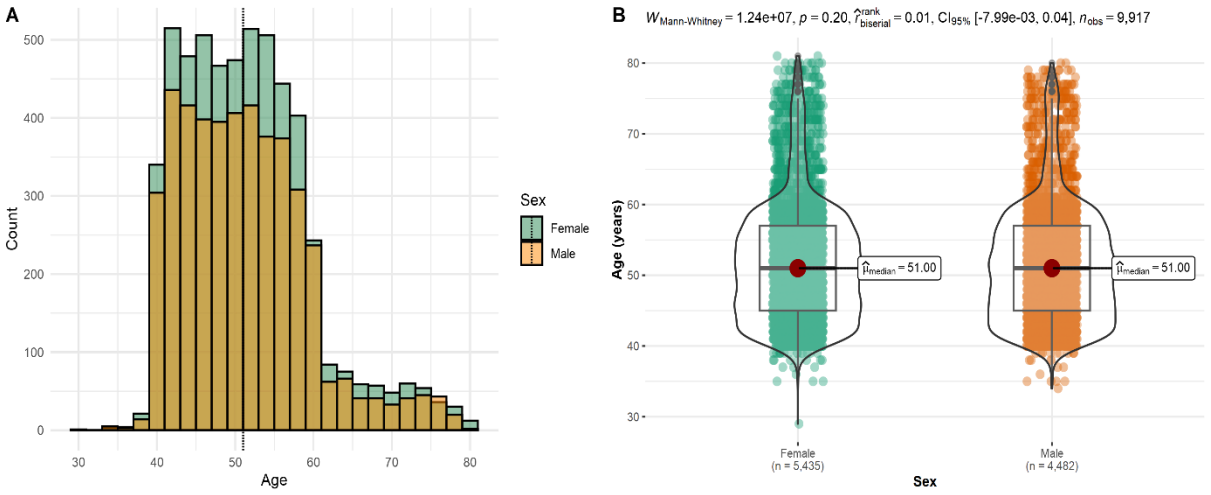


Figure 3.2: The age distributions of male and female AWI-Gen participants in this study.

A) histogram plot of the distribution of participant age stratified by sex. Median age is denoted by the dashed vertical line. B) Violin-box plot of participant age stratified by sex showing no significant differences in age distributions. Mann-Whitney U test statistics shown above plot.

The proportion of male and female participants from each of the three African regions differed significantly, with fewer male participants making up the Southern Africa cohort (42.3%) than either the Eastern African (46.1%) or Western African (48.9%) cohorts (p = 2.25 x 10^{-08}) (Tabe 3.1) (Figure 3.3A). The observed regional differences in participant sex proportions were driven by significant differences in participant sex proportions at each site of recruitment, with more female participants from each site apart from Soweto and Nanoro (p = 1.47 x 10^{-41}) (Table 3.1). Overall, none of the sites or regions were made up of equal proportions of male and female participants, which was accounted for in downstream analysis.

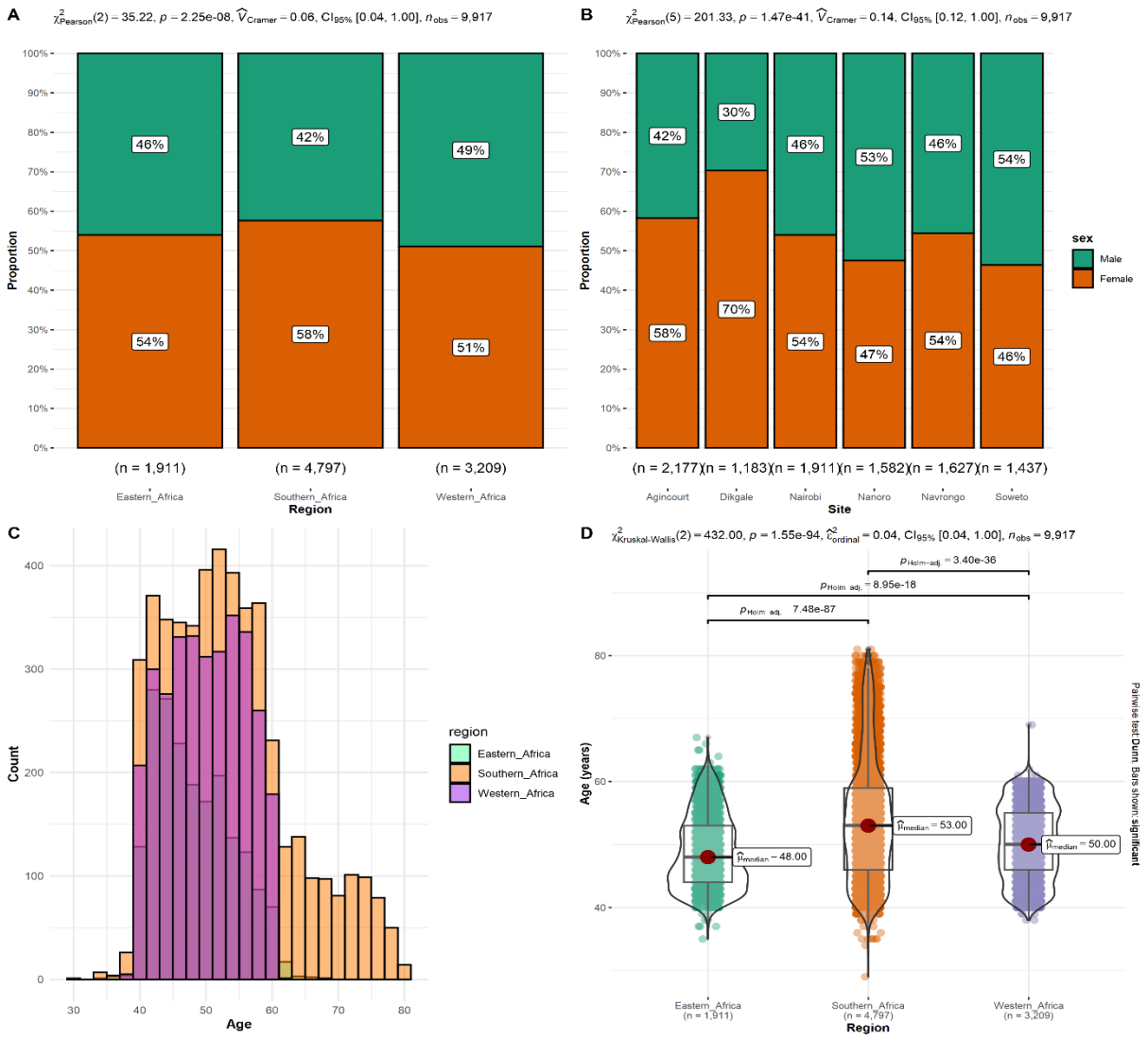


Figure 3.3: Univariate analysis results showing significant differences in the demographic makeup of participants from each region of recruitment.

A) the proportion of male and female participants differ significantly across the regions of recruitment. Chi Square test of independence statistics above plot show significant relationship between region of recruitment and participant sex. **B**) the proportion of male and female participants differ significantly across the sites of recruitment. Chi Square test of independence statistics above plot show significant relationship between site of recruitment and participant sex. **C**) a histogram plot of participant age groups across the three regions of recruitment showing 94.8% of participants over the age of 60 coming from Southern Africa. **D**) the distribution of age across the three regions of recruitment differs significantly. Kruskal-Wallis test statistics showing a significant relationship between region of recruitment and median age. Dunn's test p-values showing significant differences in the median age of each region compared.

The age distributions of participants also differed significantly across the three regions of recruitment (Table 3.1) (Figure 3.3). Of the 1,016 elderly participants over the age of 60, 963 (94.8%) were recruited in Southern Africa (Figure 3.3C). Comparatively, 48 (4.7%) of participants over the age of 60 were recruited in Eastern Africa, with only 5

(0.5%) coming from Western Africa (Figure 3.3C). Most of the study cohort were between the ages of 40 to 60 years at time of recruitment across all three regions, reflecting the recruitment strategy of the AWI-Gen study (Ali et al., 2018). The age distributions of participants from each African region were significantly different ($p = 1.55 \times 10^{-94}$) (Figure 3.3D). The median age of participants from Southern Africa was significantly higher than the median age of participants from both Eastern Africa (53 vs. 48 years, $p = 7.48 \times 10^{-87}$) and Western Africa (53 vs 50 years, $p = 3.4 \times 10^{-36}$) (Figure 3.3D). Additionally, the median age of participants from Eastern Africa was significantly lower than the median age of participants from Western Africa (48 vs 50 years, $p = 8.95 \times 10^{-18}$) (Figure 3.3D). The higher median age of Southern African participants was driven by the higher ages of participants recruited from the Agincourt and Dikgale sites, from which almost all the participants over the age of 60 (94.69%) were recruited. Overall, the demographic distributions of the study cohort were significantly different across the three regions of recruitment.

3.3.2 Autosomal MCAs were successfully detected in the AWI-Gen cohort

This study was able to successfully call autosomal MCAs with MoChA using the workflow outlined in Figure 3.1. A total of 1245 autosomal MCAs were detected in 998 of the 9917 AWI-Gen participants at a frequency of 10.06%. Autosomal MCAs were detected in participants ranging in age from 37 years to 80 years, with a median age of 53 years (Table 3.2). Most of the autosomal MCAs were detected in participants over the age of 50 years (62.01%). Of the 1245 autosomal MCAs, 646 (51.89%) were detected in male participants and 599 (48.11%) were detected in female participants (Table 3.2). MCA types were not detected at equal frequencies across the overall cohort. Most MCAs were CN-LOH events (42.01%), followed by losses (37.11%) and gains (20.88%) (Table 3.2). MoChA was able to call focal autosomal MCAs as small as 26.88Kb and mosaic aneuploidies as large as 243.19Mb (Table 3.2). MCAs over 5Mb in length made up the largest proportion of detected events (69.07%), while MCAs under 500Kb were the least frequently detected (11.89%). MoChA was able to detect autosomal MCAs at estimated cell fractions of 0.4%, though MCAs below an estimated cell fraction of 1% were the least frequently observed. Most MCAs were detected at cell fraction percentages between 1-10% (63.94%), with 237 (19.04%) MCAs being

classified as expanded MCA events at cell fraction percentages over 10% (Table 3.2). A summary of the overall characteristics of 1245 autosomal MCAs can be found in Table 3.2.

Table 3.2: Summary statistics of 1245 autosomal MCAs detected in the AWI-Gen cohort.

Sex	**N (%)**
Male	646 (51.89)
Female	599 (48.11)
Total	1245 (100.00)
Median age (years) (range)	53 (37-80)
Age group (years)	**N (%)**
≤ 50	473 (37.99)
> 50	772 (62.01)
Total	1245 (100.00)
MCA type	**N (%)**
Gain	260 (20.88)
Loss	462 (37.11)
CN-LOH	523 (42.01)
Total	1245 (100.00)
Median MCA length (Mb) (range)	27.80 (0.02 - 243.19)
MCA length group (Mb)	**N (%)**
< 0.5	148 (11.89)
0.5 - 5	237 (19.04)
> 5	860 (69.07)
Total	1245 (100.00)
Median MCA cell fraction (%) (range)	2.72 (0.40- 48.02)
MCA cell fraction group (%)	**N (%)**
< 1	212 (17.02)
1 – 10	796 (63.94)
≥ 10	237 (19.04)
Total	1245 (100.00)

N = count of MCAs; % = percentage of 1245 MCAs; MCA cell fractions estimated by MoChA were converted to percentages.

3.3.3 Participant demographic variables were significantly associated with the odds of autosomal MCA occurrence in the AWI-Gen cohort

The relationship between participant age, sex, and region of recruitment with the likelihood of autosomal MCAs occurrence was investigated in the AWI-Gen cohort to determine if the significant relationships reported in European and Asian cohorts are

true for Africans as well. Variables were first assessed independently to establish significant relationships between cohort demographics and the odds of MCA occurrence, which was defined as having detected at least one autosomal MCA in the participant. Variables were then assessed in combination to investigate the combined effects of significant relationships on MCA occurrence odds, and to establish any significant variable interactions. A multivariate logistic regression model was built by establishing stepwise significance of demographic factors on MCA odds. No significant relationship was found between sample batch and autosomal MCA occurrence in univariate analysis. However, batch was added to the multivariate analysis model to account for any uncategorized variance that may have been introduced by the different AWI-Gen genotyping runs. A summary of the demographic variables investigated for impact on the odds of MCA occurrence are seen in Table 3.3.

Table 3.3: A summary of the demographic makeup of 998 AWI-Gen participants with at least one detected autosomal MCA.

At least one autosomal MCA	Yes	No	Total
Median age (range)	53 (37-80)	51 (29-81)	51 (29-81)
Participant sex	**N (%)**	**N (%)**	**N (%)**
Female	477 (8.78)	4958 (91.22)	4482 (45.20)
Male	521 (11.62)	3961 (88.38)	5435 (54.80)
Region of recruitment	**N (%)**	**N (%)**	**N (%)**
Eastern Africa	153 (8.01)	1758 (91.99)	1911 (19.27)
Southern Africa	483 (10.07)	4314 (89.93)	4797 (48.37)
Western Africa	362 (11.28)	2847 (88.72)	3209 (32.36)
Site of recruitment	**N (%)**	**N (%)**	**N (%)**
Agincourt	256 (11.76)	1921 (88.24)	2177 (21.95)
Dikgale	109 (9.21)	1074 (90.79)	1183 (11.93)
Nairobi	153 (8.01)	1758 (91.99)	1911 (19.27)
Nanoro	172 (10.87)	1410 (89.13)	1582 (15.95)
Navrongo	190 (11.68)	1437 (88.32)	1627 (16.41)
Soweto	118 (8.21)	1319 (91.79)	1437 (14.49)
AWI-Gen batch	**N (%)**	**N (%)**	**N (%)**
Batch1	127 (9.44)	1218 (90.56)	1345 (13.56)
Batch2	476 (10.36)	4117 (89.64)	4593 (46.31)
Batch3	395 (9.93)	3584 (90.07)	3979 (40.12)
Total	**998 (10.06)**	**8919 (89.94)**	**9917 (100.00)**

N = number of participants; (%) = percentage of participants with or without at least one autosomal MCA by demographic variable.

3.3.3.1 The odds of autosomal MCA occurrence were significantly greater in male participants.

Of the 998 individuals who had at least one detectable autosomal MCA, 521 (52.2%) were male and 477 (47.8%) were female (Table 3.3) (Figure 3.4A). Computationally inferred participant sex was significantly associated with the MCA occurrence, with a greater proportion of male participants having at least one detectable autosomal MCA than female participants ($p = 2.71 \times 10^{-06}$) (Table 3.3) (Figure 3.4A). Univariate logistic regression of the MCA occurrence with sex as the explaining variable confirmed the association, with male participants having 1.37 times greater odds of having at least one detectable autosomal MCA than female participants (11.62% of males vs. 8.78% of females) (OR = 1.37, 95%CI [1.20-1.56], $p = 2.90 \times 10^{-6}$) (Table 3.3) (Figure3.4B). The sex bias in autosomal MCA prevalence that was observed across European and Asian populations was repeated in the AWI-Gen cohort when investigated in univariate analysis.

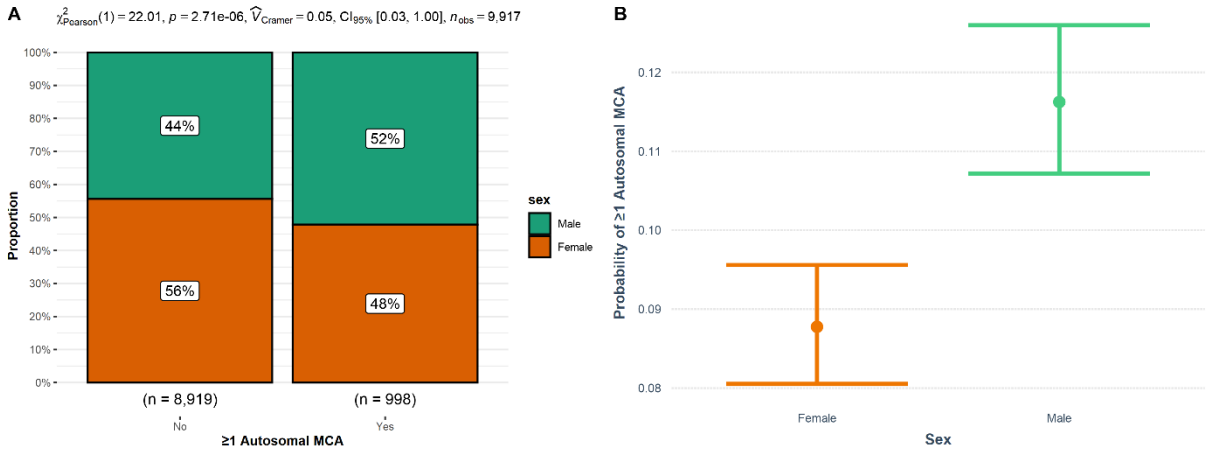


Figure 3.4: Univariate analysis showing the impact of participant sex on the odds of autosomal MCA occurrence in AWI-Gen.

A) the proportion of male participants with at least one detectable autosomal MCA was higher than the proportion of females. Pearson's Chi^2 test statistics are shown above the plot, showing a significant association between sex and MCA incidence. **B)** the marginal effect of sex on the likelihood of MCA occurrence showing that the likelihood of autosomal MCA occurrence was higher in males than in female following univariate logistic regression. Dots indicate mean probability scaled from 0 to 1. Error bars denote 95% CI.

3.3.3.2 The odds of autosomal MCA occurrence increased significantly with increasing participant age.

Increasing participant age was significantly associated with increased MCA prevalence in the AWI-Gen cohort (Figure 3.5). The prevalence of autosomal MCAs steadily increased from a combined 7.20% in participants under the age of 45 to a combined prevalence of 15.21% in participants aged 60-64 years (Table 3.5) (Figure 3.5A). The prevalence then decreased in participants between the ages of 65-69, followed by an exponential increase in participants over the age of 70 (Figure 3.5A). The oldest group of participants over the age of 75 years had the highest MCA occurrence in both males and females despite making up 1.95% of the overall cohort (Table 3.4) (Figure 3.5A).

Table 3.4: The number of male and female AWI-Gen participants with at least one autosomal MCA stratified by increasing age groups.

	≥1 autosomal MCA		*Total cohort*	
Age group	**N male (%)**	**N female (%)**	**N male (%)**	**N female (%)**
<45	77 (8.03)	71 (6.47)	958 (9.66)	1097 (11.06)
45-49	100 (9.87)	93 (7.51)	1013 (10.21)	1239 (12.49)
50-54	107 (10.73)	106 (8.73)	997 (10.05)	1214 (12.24)
55-59	130 (14.72)	112 (9.94)	883 (8.90)	1127 (11.36)
60-64	60 (18.13)	44 (12.46)	331 (3.34)	353 (3.56)
65-69	12 (10.34)	15 (9.09)	116 (1.17)	165 (1.66)
70-74	16 (17.58)	17 (12.14)	91 (0.92)	140 (1.41)
≥75	19 (20.43)	19 (19.00)	93 (0.94)	100 (1.95)
Total	**521 (11.62)**	**477 (8.78)**	**4482 (45.20)**	**5435 (58.40)**

N = number of participants in each age group; (%) - percentage of participants in each age group

The odds of a participant over the age of 70 having a detectable autosomal MCA was 1.86 times greater than those under the age of 70 (OR= 1.86, 95%CI [1.44-2.43], p = 1.49 $X10^{-05}$). Correspondingly, the median age of participants with at least one detectable autosomal MCA was significantly higher than in those in which no MCA was detected (53 vs. 51 years, p = 4.83 x 10^{-16}) (Figure 3.5C). The age effect was observed in both female and male participants, indicating that the relationship is independent of participant sex. This was further investigated using a multivariate logistic regression approach that added sex as another fixed effect variable (Table 3.5). The logistic regression model confirmed that both increasing age and participant sex were significantly associated with increased likelihood of observing at least one autosomal

MCA in the cohort (Table 3.5) (Figure 3.5D). The odds of MCA occurrence increased by 1.03 times per year when adjusting for the effect of participant sex (OR = 1.03, 95% CI [1.02-1.04], p = 1.19 x 10^{-15}). This equated to an estimated increased odds of at least one autosomal MCA occurrence by approximately 1.34 times per decade. The impact of sex remained significant when accounting for age, with male participants having 1.38 times higher odds than women (OR = 1.38, 95% CI [1.21-1.57], p = 1.59 x 10^{-09}) (Table 3.5).

Table 3.5: A logistic regression model showing the combined effect of participant age and sex on autosomal MCA occurrence odds.

Coefficients	Estimate	Std. Err.	Z-value	P-value
Intercept	-3.94	0.21	-18.72	<2.00 x 10^{-16}
Ref. Female				
Age	0.03	4.0 X 10^{-03}	7.95	1.19 x 10^{-15}
Sex (Male)	0.32	0.07	4.80	1.59 x 10^{-09}

Std. Err. = standard error of estimate; Ref = category varaible reference

There was no evidence of any significant interaction between participant age and sex in the model as shown by the overlapping 95% CI's in Figure 3.5D. Additional evidence was determined through introducing an interaction term for age and sex to the model, which was subsequently found to not be significant or improve the model (Age: Sex, p = 0.47). Thus, participant age and sex were confidently observed as independent predictors on the odds of autosomal MCA occurrence in the AWI-Gen cohort which is consistent with the literature. Visual inspection of the age plot in Figure 3.5A revealed a relatively linear relationship between increasing age and increasing autosomal MCA occurrence in participants up to the age of 50, after which MCA occurrence increased at an accelerated rate. However, adding an age^2 variable did not significantly improve on the linear model, and visual inspection of the model residuals confirmed a linear relationship between age and autosomal MCA occurrence in this data. Additionally, the low proportion of participants over the age of 60, especially from WA and EA, may have introduced error into the estimates of increasing age (Figure 3.3A and B).

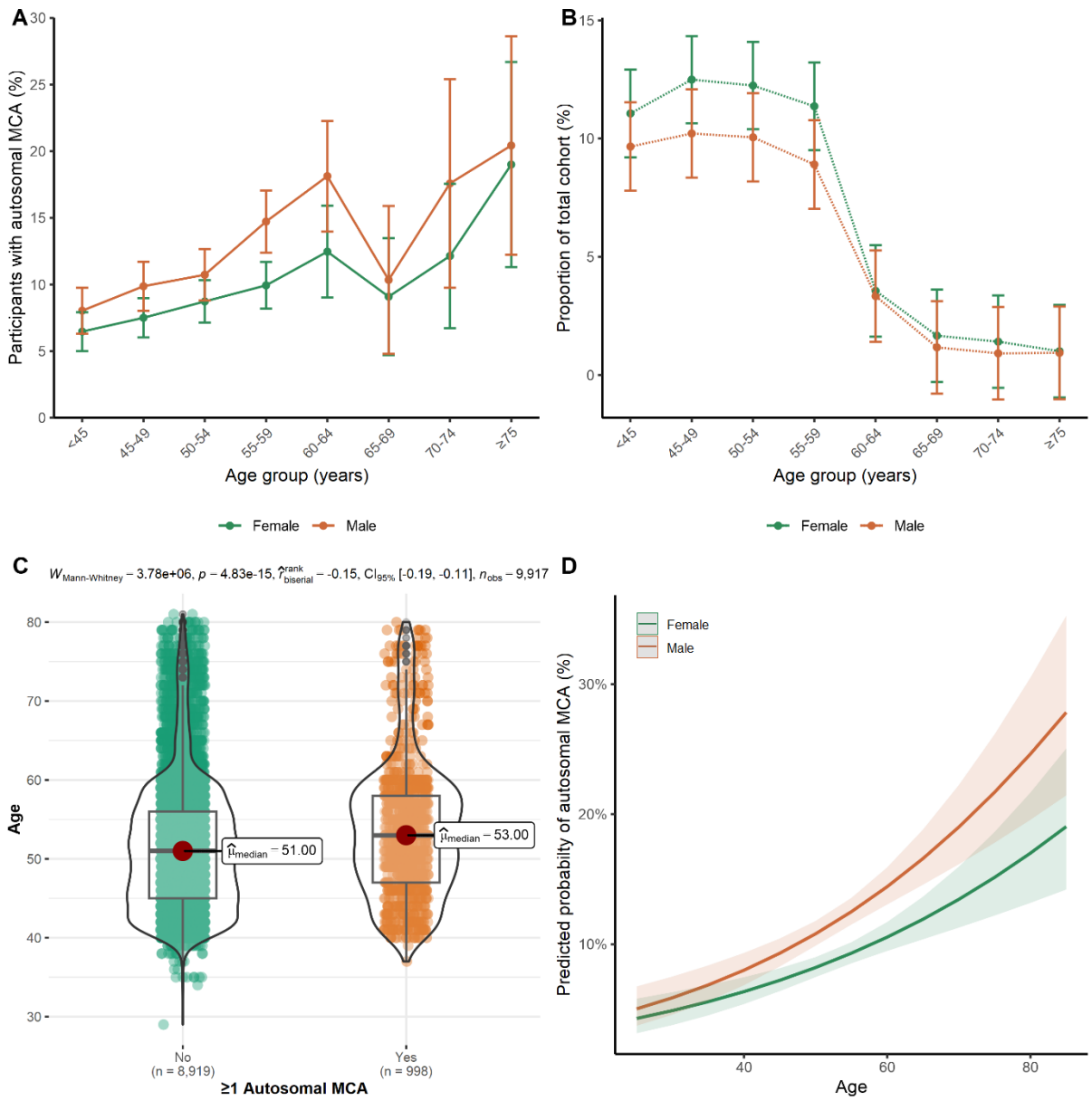


Figure 3.5: Increased age was significantly associated with greater odds of at least one autosomal MCA in both male and female AWI-Gen participants.

A) The percentage of male and female participants with at least one detectable autosomal MCA stratified by age group. The relative proportion of both male and female participants with at least one autosomal MCA increased with increasing age at recruitment. Error bars denote 95%CI of means. **B)** The total proportion of male and female participants recruited in each age group for the overall cohort of 9917 participants. Error bars denote 95%CI of means. **C)** the median age of participants with at least one autosomal MCA was significantly higher than the median age of participants with no detectable autosomal MCAs. Mann-Whitney U test statistics showing a significant relationship between age and MCA occurrence are shown above the violin-boxplot plot. **D)** the marginal effects of participant age and sex on the predicted likelihood of having at least one autosomal MCA following multivariate logistic regression with age and sex as independent variables. Shaded areas indicate 95%CI. Increasing participant age resulted in increased likelihood of having at least one autosomal MCA in both male and female participants. Overlapping 95%CI show that the relative strength of the age effect was not significantly different between male and female participants.

3.3.3.3 The odds of at least one autosomal MCA occurring differed significantly between participants from Southern Africa, Western Africa, and Easten Africa.

AWI-Gen participants were recruited from six independent study sites in Eastern, Southern, and Western African countries. The relationship between the region of participant recruitment variable and autosomal MCA occurrence was investigated in an initial univariate association analysis. The proportions of participants with at least one autosomal MCA were significantly associated with the region of recruitment ($p = 8.29 \times 10^{-04}$) (Table 3.3) (Figure 3.6A).

A follow-up univariate logistic regression showed that region of recruitment had a statistically significant impact on the autosomal MCA occurrence when compared to null model ($p = 7.0 \times 10^{-04}$). The odds of Southern African participants having at least one autosomal MCA was 1.29 times greater than participants from Eastern Africa (OR = 1.29, 95%CI [1.07-1.56], $p = 9.40 \times 10^{-03}$). The odds of at least one autosomal MCA were found to be even greater in participants from Western Africa when compared to Eastern Africans (OR = 1.46, 95%CI [1.20-1.78], $p = 1.0 \times 10^{-04}$) (Table 3.6) (Figure 3.6B). Western Africans had a marginally increased, but non-significant, likelihood of MCAs than Southern Africans, (OR = 1.14, 95%CI [0.98-1.31], p = 0.08) (Table 3.6) (Figure 3.6B).

Table 3.6: Univariate logistic regression models showing the impact of region of recruitment on the odds autosomal MCA incidence in AWI-Gen.

Coefficient	Estimate	Std. Err.	Z-value	P-value
Ref. Eastern Africa				
Intercept	-2.44	0.08	-28.97	$< 2.00 \times 10^{-16}$
Southern Africa	0.25	0.09	2.60	9.40×10^{-03}
Western Africa	0.40	0.10	3.75	2.00×10^{-04}
Ref. Southern Africa				
Intercept	-2.20	0.04	-45.64	$< 2.00 \times 10^{-16}$
Eastern Africa	-0.25	0.09	2.60	9.40×10^{-03}
Western Africa	0.13	0.07	1.73	0.08

Ref. = reference variable category; Std. Err. = standard error of estimate coefficient.

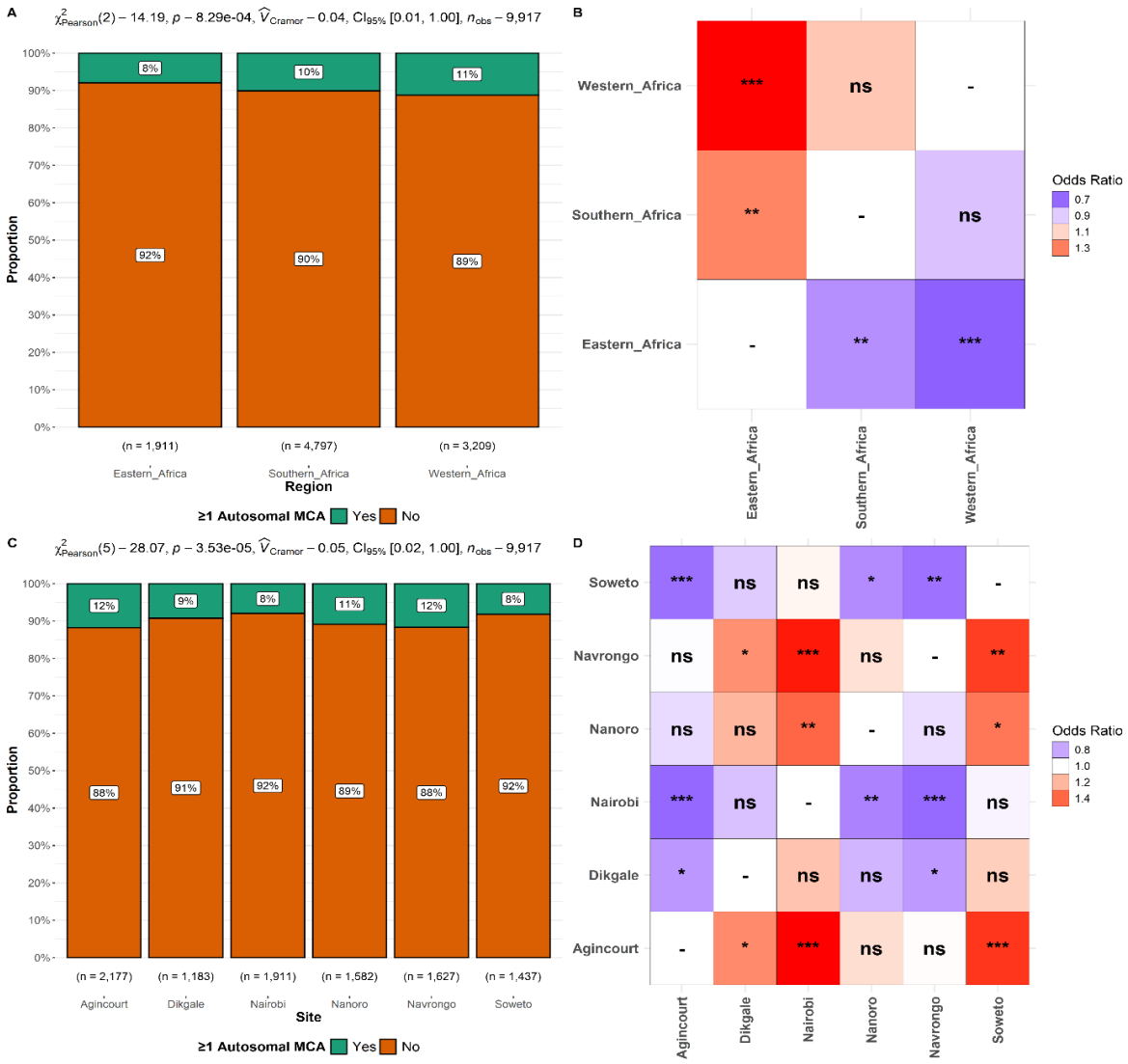


Figure 3.6: The proportion of participants with at least one autosomal MCA differed between both region of recruitment and study site in AWI-Gen.

The proportion of participants with at least one autosomal MCA was significantly associated with **A)** region of recruitment and **C)** site of recruitment. Pearson's Chi2 test statistics shown above the plots. The odds of at least one detectable MCA were significantly lower in **B)** Eastern Africans than in Western or Southern Africans and differed significantly by **D)** site of recruitment. Plots **B)** and **D)** show heatmaps representing the differences in the odds of autosomal MCA prevalence between region and sites of recruitment pairs. Comparisons are made from left to right on the x-axis and from top to bottom on the y-axis. Tile colours represent Fischer's Exact test odds ratio statistics with gradients reflecting the strength of the estimates. Tiles are annotated with FDR adjusted p-values. FDR<0.001 = "***", FDR <0.01 = "**", FDR <0.05 = "*", non-significant odds ratios = "ns"

These novel results from the initial exploratory analyses indicated the presence of underlying region variables that were differentially impacting the odds of MCA incidence in the AWI-Gen cohort. Participants were further stratified into their respective study recruitment sites to further investigate this association and to

determine if it holds true across sites within the same region. The significant association between proportions of participants with at least one autosomal MCA and region of participant recruitment was repeated when stratifying by study site ($p = 3.53 \times 10^{-05}$) (Figure 3.6C). Participants recruited from the Agincourt site in Southern Africa and the Western African sites of Nanoro and Navrongo had greater proportions of autosomal MCA events than participants from other sites (Figure 3.6C).

Univariate logistic regression once again showed that site of recruitment had a significant effect on the odds of detecting at least one autosomal MCA when compared to a null model ($p = 2.94 \times 10^{-05}$). Participants from the Agincourt site had significantly higher odds of having at least one detectable autosomal MCA than participants from Dikgale, Soweto, and Nairobi (Figure 3.6D). While participants from the Western African sites of Nanoro, and Navrongo had significantly higher odds of having a detectable autosomal MCA than those from all other sites except Agincourt and each other (Figure 3.6D). However, prior analysis showed that the age and sex distributions of participants from the various were significantly different which could confound univariate associations (Figure 3.3). Multivariate logistic regression analysis was employed to investigated combined effect of the participant age, sex, and region of recruitment on the odds of having at least one MCA event in Africans (Table 3.7). The SNP-array batch variable was included in the model to account for any potential technical biases introduced across the three separate genotyping runs.

A fixed effects logistic regression model was fitted to the data, with age, sex, batch, and region of recruitment as independent variables. This was significantly better at modeling the occurrence of at least one autosomal MCAs than a null model ($p < 2.20 \times 10^{-16}$). Batch was investigated as both a random-effect and fixed-effect variable. The fixed-effect model (AIC = 6387) was better fitting than both the mixed-effect random-intercept model (1 | Batch, ICC = 7.39×10^{-04}) (AIC = 6390, p = 0.04), and the random-slope model (Age + Sex + Region | Batch, ICC = 3.37×10^{-03}) (AIC = 6414). The random-effects and random-slope models captured very little variance in MCA occurrence that was not already accounted for by the simpler fixed-effects model. Thus, the fixed-effects model was selected to reduce model complexity and is summarized below in Table 3.7. The effects of participant age and sex remained independently associated with autosomal MCA incidence odds (Table 3.7) (Figure 3.7A). There was no evidence of interaction between combinations of region and age

(p=0.70), region and sex (p = 0.40), region and batch (0.27), or any other three- or four-way combinations of variables when iteratively adding interaction terms.

Table 3.7: A fixed-effects multivariate logistic regression of sex, age, region of recruitment, and sample batch on autosomal MCA occurrence odds.

Coefficients	Estimate	Std. Err.	Z-value	P-value
Ref. Eastern Africa, Female, Batch 2				
Intercept	-4.20	0.23	-18.55	$< 2.00 \times 10^{-16}$
Sex (Male)	0.31	0.07	4.55	5.32×10^{-06}
Age	0.03	4.17×10^{-03}	8.07	7.18×10^{-16}
Region (Southern Africa)	0.11	0.10	1.07	0.28
Region (Western Africa)	0.35	0.10	3.42	6.24×10^{-04}
Batch (Batch 1)	-0.09	0.11	-0.83	0.40
Batch (Batch 3)	-0.16	0.07	-2.17	0.03
Ref. Southern Africa, Female, Batch 2				
Intercept	-4.09	0.24	-17.12	$< 2.00 \times 10^{-16}$
Sex (Male)	0.31	0.07	4.55	5.32×10^{-06}
Age	0.03	4.17×10^{-03}	8.07	7.18×10^{-16}
Region (Eastern Africa)	-0.11	0.10	-1.07	0.28
Region (Western Africa)	0.24	0.08	3.14	1.68×10^{-03}
Batch (Batch 1)	-0.09	0.11	-0.83	0.40
Batch (Batch 3)	-0.16	0.07	-2.17	0.03

Ref. = reference variable category; Std. Err. = standard error of estimate coefficient.

The odds of male participants having at least one detected autosomal MCA remained 1.36 times greater than females when accounting for the marginal effects of age, batch, and region of recruitment (OR = 1.36, 95%CI [1.19-1.55], p = 5.32×10^{-06}) (Figure 3.7A). Increasing age also remained significantly associated with increased odds of autosomal MCA occurring with each increasing year increasing the odds by 1.03 times when accounting for the marginal effects of sex, batch, and region of recruitment (OR = 1.03, 95%CI [1.03-1.04], p = 7.18×10^{-18}) (Figure 3.7A).

Overall, the combined effect of batch was not a significant predictor of MCA occurrence (p = 0.09). However, the odds of samples from Batch 2 having at least one detectable MCA were 1.18 times greater than those from Batch 3 after accounting for the marginal effects of age, sex, and region (OR = 1.18, 95%CI [1.02-1.36], p = 0.03). No difference was seen between the other Batch pairs. After accounting for the marginal effects of sex, age, and batch, participants from Western Africa had a 1.42 times greater odds of autosomal MCA occurrence than participants from Eastern Africa (OR = 1.42, 95%CI [1.16-1.74], p = 6.24×10^{-04}) and a 1.27 times greater odds than participants from Southern Africa (OR = 1.27, 95%CI [1.09-1.48], p = 1.68×10^{-03}) (Table 3.7) (Figure

3.7). Southern Africans and Eastern Africans no longer had significantly different odds of autosomal MCA occurrence when accounting for differences in age and sex distributions (p = 0.28) (Table 3.7) (Figure 3.7B). Overall, region of recruitment was a significant and independent autosomal MCA odds (p = 8.0 x 10^{-04}). The significantly greater odds of MCA occurrence observed in Western Africans in the multivariate analysis once again supported evidence of regional specific factors on MCA formation or selection in Africa (Figure 3.7C).

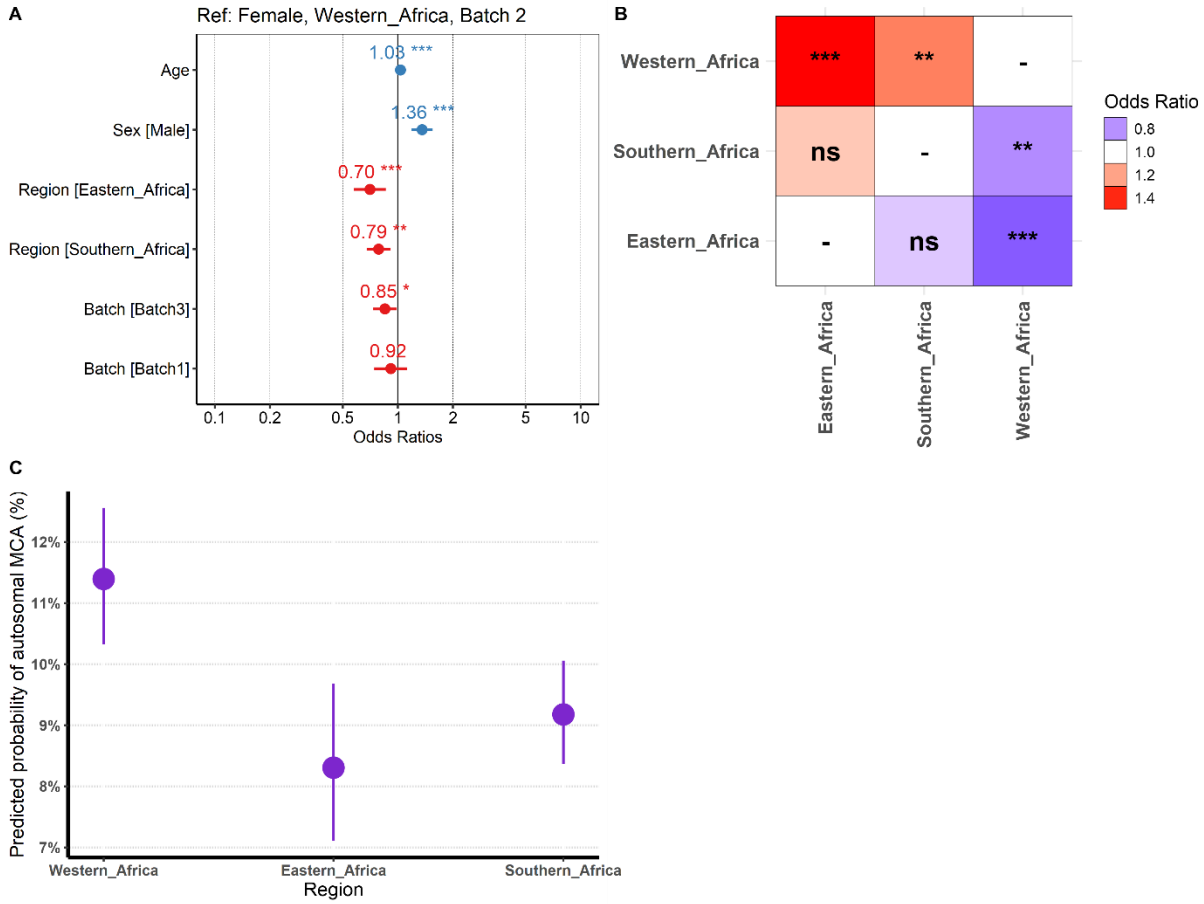


Figure 3.7: Multivariate logistic regression analysis showing the combined impact of region of recruitment, age, sex, and batch on differential odds of autosomal MCA occurance.

A) plot of multivariate logistic regression odds ratio estimates for autosomal MCA occurance predicted by sex, age, batch, and region of recruitment. Model reference variables (Ref.) are Western African females and batch 2. **B)** heatmap representing the differences in the odds of autosomal MCA occurance between regions when accounting for age, sex, and batch. Comparisons are made from left to right on the x-axis and from top to bottom on the y-axis. Tile colour represents multivariate logistic regression odds ratio statistics with gradients reflecting the strength of the estimates. **C)** the marginal effect of region on the likelihood of autosomal MCA occurance adjusted for age, sex and batch marginal effects. Dots indicate mean predicted probabilty. Error bars denote 95% CI. P-value symbols are p<0.001 = "***", p<0.01 = "**", p<0.05 = "*", p<0.1 = non-significant "ns".

Participants were stratified into lower-level site of recruitment variables and added to a logistic model with age, sex, and sample batch to determine if regional associations hold true. This model also served to further investigate the differences in the odds of autosomal MCA occurrence that was observed between the Southern African study sites in univariate analysis. The fixed-effect model of accounting for participant age, sex, batch, and site of recruitment was significantly better at predicting the occurrence of an autosomal MCA than a null model ($p < 2.2 \times 10^{-16}$). However, it did not improve on performance of the higher-level region of recruitment model and only added additional complexity to the analysis (AIC = 6392 vs 6387, p = 0.66). Once again it was shown that male participants had an approximately 1.37 times greater odds of having at least one autosomal MCA than females adjusting for the effects of age and site of recruitment (OR = 1.37 95%CI [1.20-1.56], $p = 3.17 \times 10^{-06}$). Increasing age remained significantly associated with increased odds of having at least one autosomal MCA, with each increasing year increasing the odds by 1.03 times adjusted for sex and study site effects (OR = 1.03, 95%CI [1.02-1.04], $p = 1.82 \times 10^{-12}$).

The marginal effect of site of recruitment on the odds MCA occurrence better reflected the observed marginal effects of the regions to which they belonged when accounting for age, sex, and batch (Figure 3.8). Participants from Nanoro had significantly greater odds of having at least one detectable MCA than participants from Soweto (OR = 1.40, 95%CI [1.09-1.79], $p = 8.44 \times 10^{-03}$) and Nairobi (OR = 1.41, 95% CI [1.12-1.79], $p = 4.06 \times 10^{-03}$) when accounting for age, sex, and batch effects. No significant difference in the odds of MCA occurrence was observed between Nanoro participants and those from Agincourt, Dikgale, or Navrongo once adjusting for age and sex (Figure 3.8B). Interestingly, participants from Navrongo had the highest overall odds of at least one autosomal MCA occurrence when accounting for age, sex, and sample batch effects (Figure 3.8). Their odds of MCA occurrence was significantly greater than participants from Soweto (OR = 1.41, 95% CI [1.10-1.80], $p = 5.87 \times 10^{-03}$), Nairobi (OR = 1.43, 95% CI [1.14-1.79], $p = 1.87 \times 10^{-03}$), Agincourt (OR = 1.21, 95% CI [0.97-1.50], p = 0.09), and Dikgale (OR = 1.25, 95% CI [0.97-1.61], p = 0.08 trending towards significance) (Figure 3.8A).

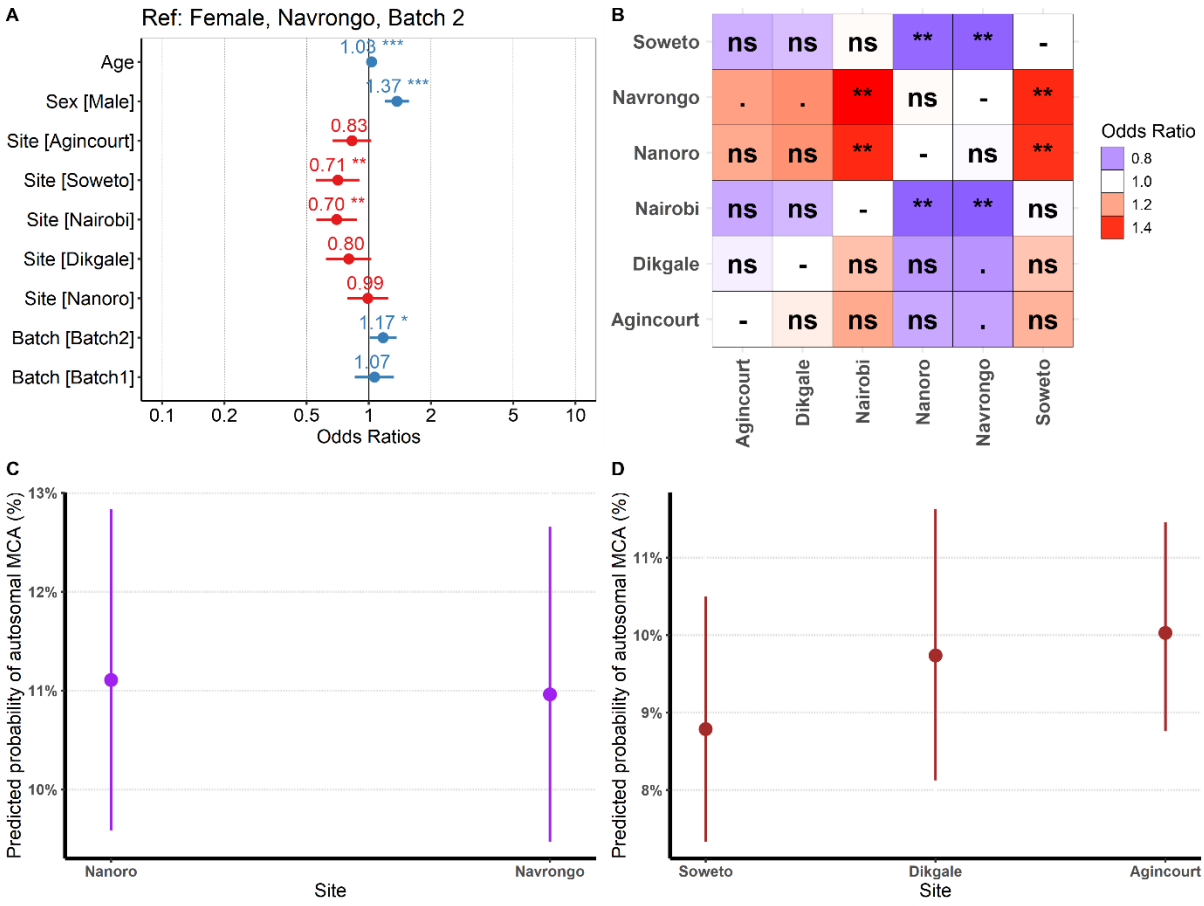


Figure 3.8: Multivariate regression anaylsis showed that particiapnts across study sites had differential odds of autosomal MCA occurance when adjusting for sex, age, and batch.

A) *A* plot of multivariate logistic regression odds ratio estimates for autosomal MCA incidence *accounting for* sex, age, batch, and site of recruitment. Model reference groups (Ref.) are females from the Navrongo site and Batch 2. Navrongo participants had higher odds of MCA ocurrence than pariticpants from Nairobi and Soweto, and higher odds than participants from Agincourt and Dikgale approaching signficance. Bars represent 95% CI of odds ratio estimates (dots). **B)** heatmap representing the differences in the odds of autosomal MCA prevalence between sites of recruitment when accounting for age, batch, and sex. Comparisons are made from left to right on the x-axis and from top to bottom on the y-axis. Tile colours represent multivariate logistic regression odds ratio statistics with gradients reflecting the strength of the estimates. **C)** the marginal effect of site on the probabilty of autosomal MCA incidence adjusted for age, batch, and sex in Western Africans. There was no significant difference in the likelihood of MCA ocurrence between sites within the Western African region. Dots indicate mean predicted probabilty. Error bars denote 95% CI. **D)** the marginal effect of site on the probabilty of autosomal MCA incidence adjusted for age, batch, and sex in Souther Africans. There was no significant difference in the likelihood of MCA ocurrence between sites within the Western African region. Dots indicate mean predicted probabilty. Error bars denote 95% CI. P-value symbols - p<0.001 = "***", p<0.01 = "**", p<0.05 = "*", p<0.1 = non-significant "ns". p-value approaching 0.05 = "."

Importantly, no significant differences in the odds of MCA occurrence were observed within the three Southern African sites or within the Western African sites once

observed differences of age and sex distributions were accounted for (Figure 3.8B). Addition of site of recruitment as a random effects variable nested within the region variable (Site | Region) with a mixed effect model did not improve on the model fit with age, sex, batch, and region as fixed effects (AIC = 6430 vs. AIC = 6388). This was confirmed by stratifying the cohort by region and running the model with age, sex, batch, and site as independent fixed effects (Figure 3.8C and D). The study site variable did not lead to increased predicted probability of MCA occurrence within the larger regions, with overlapping marginal effect probability intervals in both isolated Southern and Western African cohorts (Figure 3.8C and D). Higher levels of stratification reduced the effective sample sizes for testing when compared to grouped regional categories, which reduced statistical power. Consequently, the region of recruitment variable was not stratified by study site for the rest of the study analysis. Furthermore, it was determined that other observable study site effects between regions could be attributed to larger regional effects with more confidence.

Overall, the consensus from the multivariate logistic regression models was that participant age, sex, and the broader African region of recruitment all had significant and independent impacts on the odds of at least one autosomal MCA in the AWI-Gen cohort. Although robustly significant, the model's explanatory power was very weak, suggesting that that there were many other factors impacting the odds of MCA occurrence in the AWI-Gen cohort. These simple models were not tasked with accounting for mechanistic biases influencing MCA mutation rate risk or identifying the drivers of MCA selection and CH into account. This analysis did not explicitly set out to build a comprehensive model for predicting MCA prevalence in Africans but rather served to explore the demographic risk factors associated with MCA occurrence. Additional exploratory analysis was conducted to provide insight into these important factors which are the drivers of observed autosomal MCAs frequencies and distributions.

3.3.4 The sizes and cell fractions of autosomal losses, gains, and CN-LOHs differed significantly in the AWI-Gen cohort

A detailed investigation of the 1245 autosomal MCAs was carried out to comprehensively describe their physical characteristics and genomic distributions. MCAs were broadly categorized into losses, gains, or CN-LOHs based on their effect

on copy number. Further identification of complex MCAs and structural alterations with no impact on copy number was not possible with this data. A summary of the 1245 detected autosomal MCA characteristics is in Table 3.8.

Table 3.8: The characteristics of 1245 autosomal MCAs stratified by MCA type.

MCA type	Gains	Losses	CN-LOHs	Total
N (%)	260 (20.88)	462 (37.11)	523 (42.01)	1245 (100)
Male (%)	51.92	50.65	52.96	51.89
Median age	53.5 (40-79)	53 (40-80)	53 (40-79)	53 (37-89)
Median length (Mb)	33.84 (0.28-198.02)	1.25 (0.02-141.21)	42.14 (0.18-243.19)	27.80 (0.02-243.19)
Median CF %	2.28 (0.65-48.02)	8.56 (0.99-29.51)	1.24 (0.40-39.70)	2.72 (0.40-48.02)

Median (range). N = total count, Mb = mega base pairs, CF = cell fraction percentage

The most frequently detected autosomal MCAs in the AWI-Gen cohort were CN-LOH events (42.01%), followed by losses (37.11%), and gains (20.88%) (Table 3.8). No association was observed between overall proportions of losses, gains, and CN-LOHs detected and participant sex (Figure 3.9C) ($p = 0.77$). The size distributions of detected MCAs differed significantly between losses, gains, and CN-LOHs ($p = 7.42 \times 10^{-118}$) (Tabel 3.8) (Figure 3.9A). Losses were smaller on average, with significantly smaller median lengths than both gains ($p = 4.45 \times 10^{-67}$), and CN-LOHs ($p = 6.32 \times 10^{-102}$) (Figure 3.9A). Conversely, gains and CN-LOHs did not differ significantly in their size distributions ($p = 0.71$). Gains made up most of the detectable whole chromosome aneuploidies (84.62%), not counting acrocentric chromosomes 13,14,15,21,22 with no p-arm probe coverage.

Mosaic monosomy was very rare, with 2 cases in chromosomes 9 and 16 being detected in this cohort. Conversely, losses made up most interstitial focal events that spanned less than 1Mb in size (87.50%). Small interstitial CN-LOHs (5.60%) and gains (6.90%) under 1Mb were detected much less frequently. The distribution of the cell fraction estimates, that is the proportion of cells in which an MCA is estimated to occur, also differed significantly by MCA type ($p = 1.11 \times 10^{-117}$) (Table 3.8) (Figure 3.9B). Losses were detected at significantly higher median cell fractions than both gains ($p = 1.69 \times 10^{-29}$) and CN-LOHs ($p = 2.72 \times 10^{-118}$) (Figure 3.9B). CN-LOHs were found at significantly lower cell fractions than gains ($p = 2.08 \times 10^{-15}$). Losses made up 79.75% of expanded MCAs with cell fractions over 10%, while CN-LOHs made up 92.92% of

the MCAs that were detected at cell fractions less than 1%. Only a single loss was detected with an estimated cell fraction under 1%.

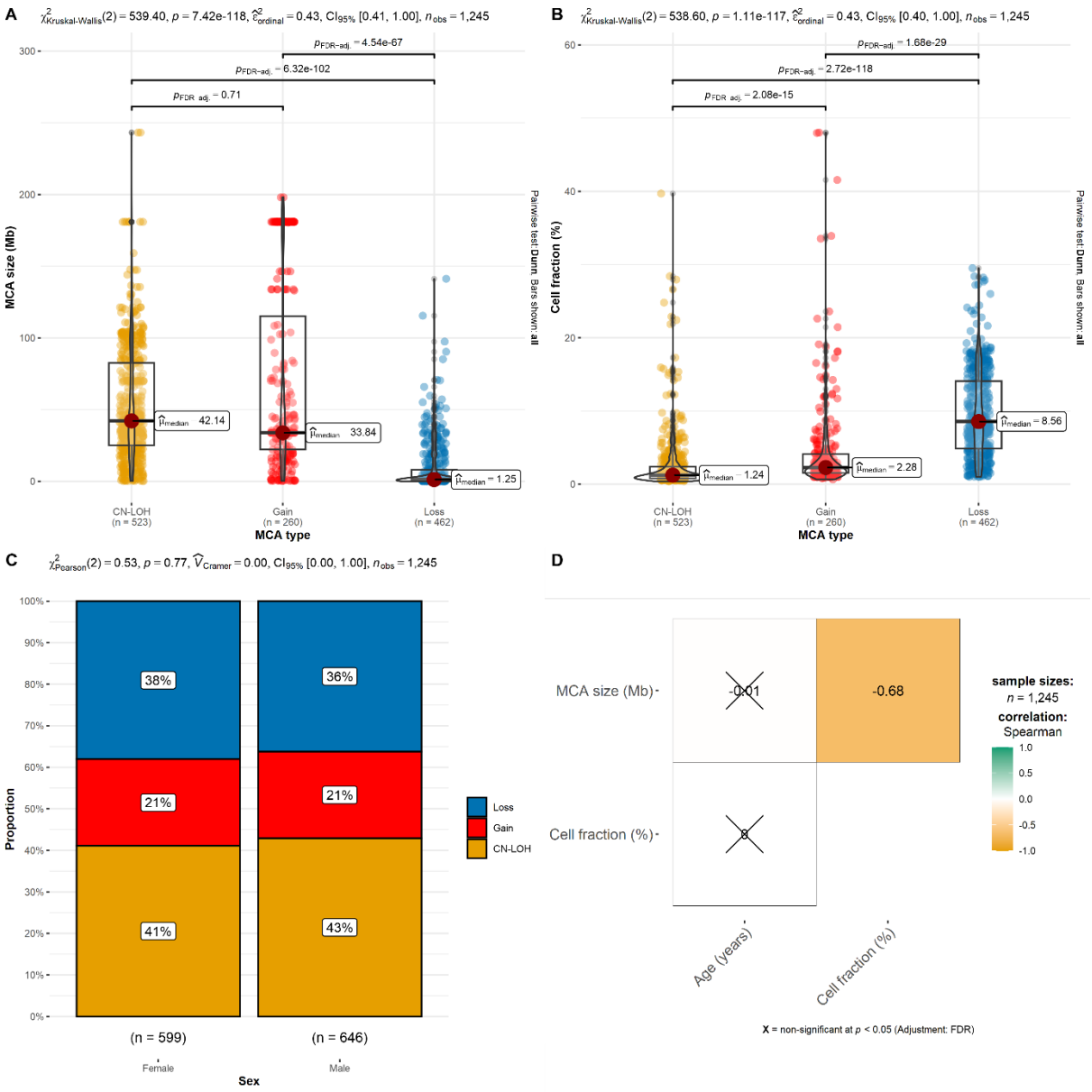


Figure 3.9 Univariate analysis showing CN-LOHs, gains, and losses differing in size and estimated cell fraction distributions.

A) Violin-boxplots of MCA size distributions stratified by MCA event type. Kruskal-Wallis test statistics are above, showing significant difference in size distributions. Losses were smaller on average than gain, and CN-LOHs **B)** Violin-boxplots of MCA cell fraction distributions stratified by MCA event type. Kruskal-Wallis test statistics are above, showing significant difference in cell fraction distributions. Losses were detected in higher cell fractions on average than gains or CN-LOHs**. C)** The proportion of MCA types in male and female participants has a similar distribution. Pearson's Chi^2 statistics are above. **D)** a correlation matrix of continuous MCA characteristics showing a strong negative correlation between MCA size and MCA cell fraction. On average, smaller MCAs were to be detected in higher fractions of cells, while larger MCAs were detected at lower fractions of cells. Participant age was not correlated with either MCA size or cell fraction.

No significant correlation was found between participant age and the MCA size or cell fractions (Figure 3.9D) in univariate analysis. However, a strong significant negative correlation was found between overall MCA size and the cell fraction (Rho = -0.69, p = 5.67 x 10^{-174}). Smaller MCAs were supported in higher cell fractions on average, while larger MCAs were supported in lower cell fractions on average. This may also be a result of the algorithm not detecting smaller MCAs at lower cell fractions. Though it is not possible to know which explanation is the contributing factor. The significant differential size and cell fraction distributions that were observed between MCA types suggested that the strengths or directionality of the size and cell fraction correlations may be different between losses, gains, and CN-LOHs as well. MCAs were stratified by type, and their individual sizes were plotted against their estimated cell fractions. Visual inspection of these plots revealed a nonlinear negative relationship between MCA size and cell fraction that was true for losses, gains, and CN-LOHs. A more detailed regression of size and cell fraction for each type was performed and can be seen in Table 3.9 and Figure 3.10.

Table 3.9: Parameter estimates for a power curve fitted regression of MCA event size predicted by MCA event cell fraction with the formula $Y=aX^b$.

MCA type	Estimate	Std. Err.	T-value	P-value
CN-LOH				
a	58.71	1.82	32.18	< 2.20 x 10^{-16}
b	-0.31	0.05	-5.90	6.63 x 10^{-09}
Gain				
a	112.65	8.50	13.25	< 2.20 x 10^{-16}
b	-0.63	0.11	-5.49	9.50 x 10^{-08}
Loss				
a	63.72	3.42	18.67	< 2.20 x 10^{-16}
b	-1.21	0.07	-17.79	< 2.20 x 10^{-16}

Std. Err. = standard error of estimate. a,b = estimated parameters for MCA type specific power curve regressions.

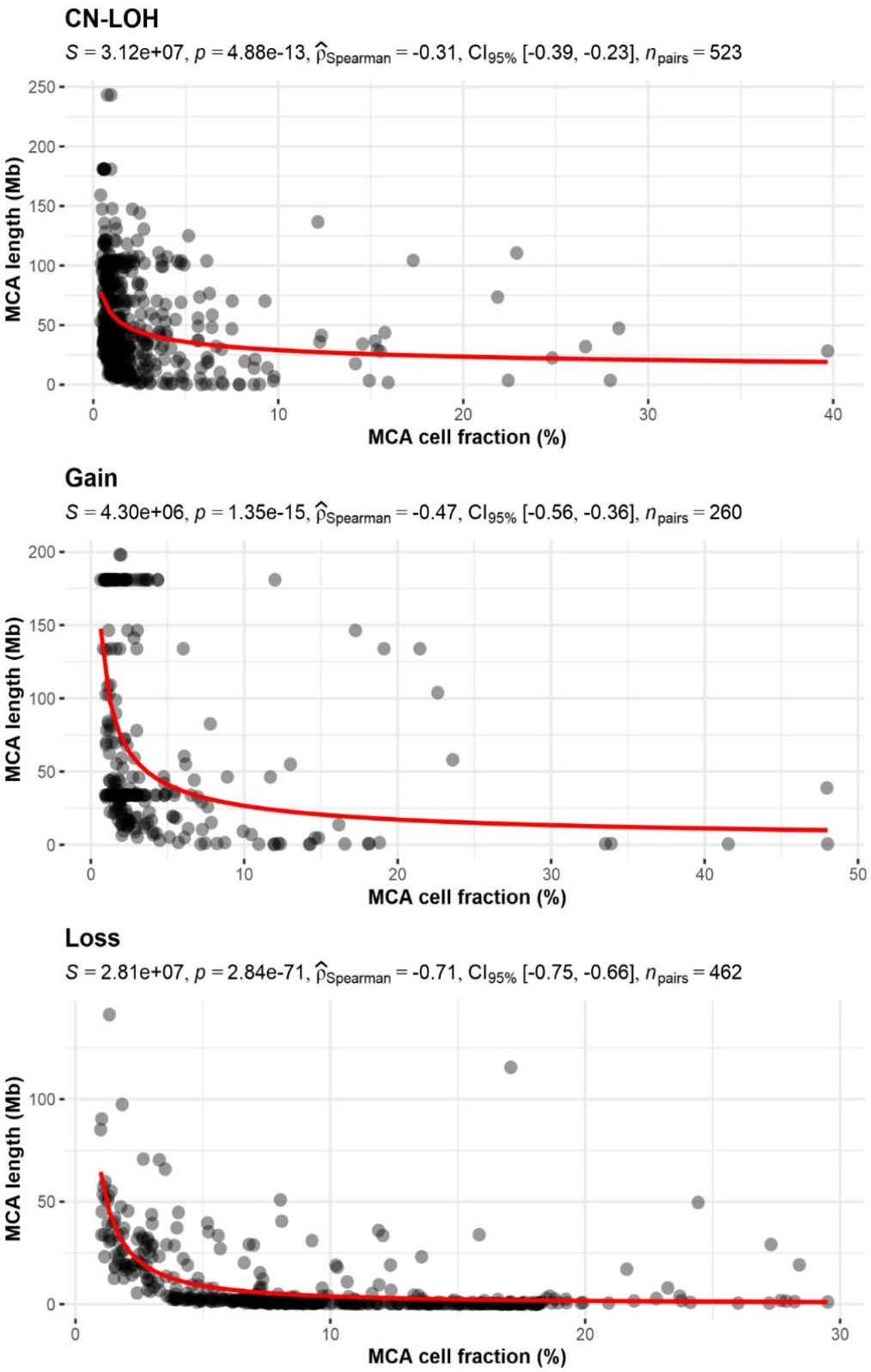


Figure 3.10: Scatterplots showing negative correlations between MCA size (Mb) and cell fraction stratified for losses, gains, and CN-LOHs.

Spearman's rank correlation statistics above. Power curve trendlines in are red.

The negative correlation between MCA size and its supported cell fraction remained significant when stratified by MCA type (Figure 3.10). The strength of that correlation differed by MCA type, with losses having the strongest negative correlation (Rho = -0.71, p = 2.84 x 10^{-71}) followed by moderate negative correlations in gain events (Rho = -0.47, p = 1.35 x 10^{-15}) and CN-LOH events (Rho = -0.31, p = 4.88 x 10^{-13}) (Figure 3.10). A non-linear negative power curve with the formula Y=aXb best predicted MCA event size by MCA cell fraction when the variables were regressed (Tabel 3.9) (Figure 3.10). This confirmed the observation that smaller MCAs were detected at higher cell fractions, with loss of genomic material having the strongest impact on clonal expansion overall. Larger MCAs encompass larger areas of the genome sequence features, modulating the impact of selection across its length.

3.3.5 The distribution of 1245 detected MCAs across 998 African autosomes in the AWI-Gen cohort

Autosomal MCAs were detected in all 22 chromosomes covered by the H3Africa -SNP array. However, the total number of MCA calls in each chromosome was different. Additionally, MCAs were unevenly distributed across the autosome, with relative proportions of losses, gains, and CN-LOHs detected in each chromosome differing significantly (Chi^2 = 482, p = 3.43 x 10^{-76}) (Figure 3.11). Chromosomes 1 (11.49%), 5 (8.17%), 14 (9.16%), and 21 (7.39%) had the highest overall MCA counts, while MCAs were detected least frequently on chromosomes 4 (1.85%), 10 (2.33%), 15, (1.61%), and 18 (1.45%) (Figure 3.11). This is in line with what has been reported in other MCA characterizations. However, the negative correlation between loss and gain frequencies within each chromosome that was previously reported in the UKB cohort was not repeated in the AWI-Gen cohort (Rho = 0.11, p = 0.63) (Loh et al., 2018). In this study, the proportion of CN-LOHs, gains, or losses appeared to be specific to each chromosome, which was suggestive of underlying mechanistic factors influencing relative mutation rates and selective forces shaping the differential frequency of MCAs in each chromosome (Figure 3.11).

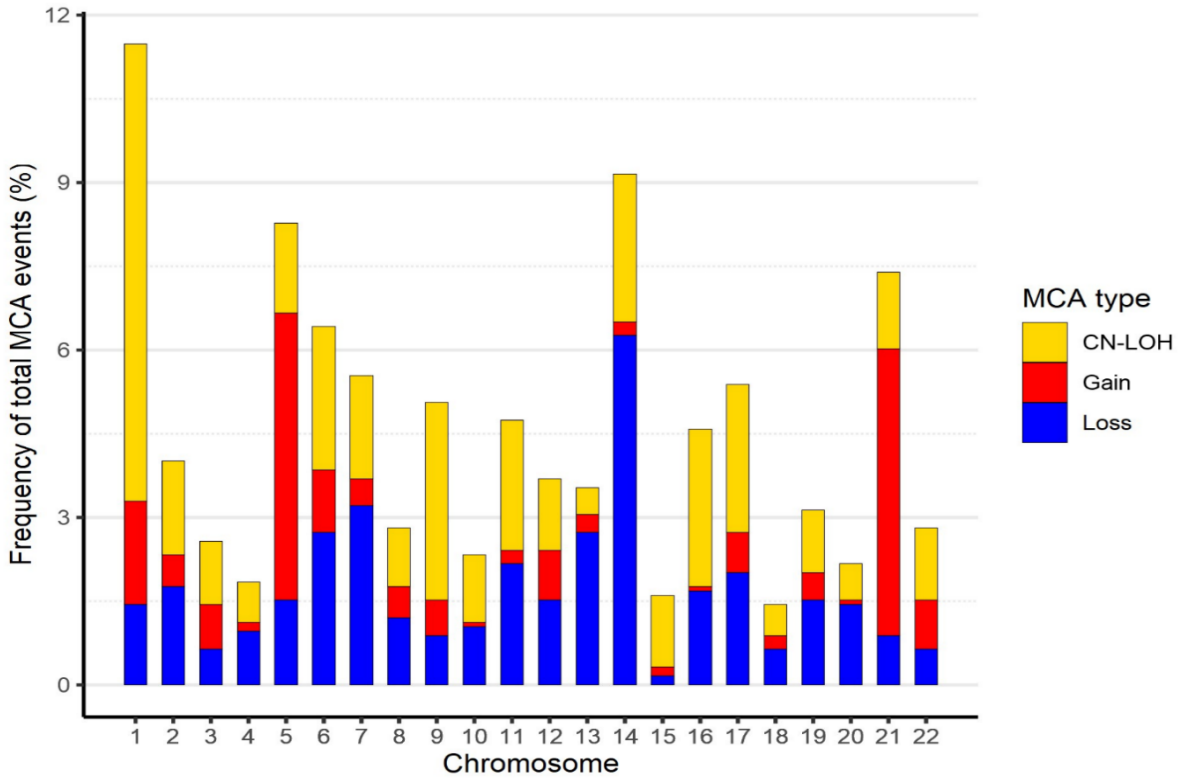


Figure 3.11: The proportion of MCAs detected in each chromosome in the AWI-Gen cohort

CN-LOHs were detected most frequently on chromosomes 1 (8.19%) and 9 (3.53%). Losses were detected most frequently on chromosome 14 (6.27%), followed by chromosomes 7 (3.21%), 6 (2.73%), and 13 (2.73%). Most gains were detected on chromosomes 1 (1.85%), 5 (5.14%). and 21 (5.14%) and were comparatively rarer in the other chromosomal regions (Figure 3.11). Striking differences in proportions of MCA types were seen across the autosome. For example, chromosome 13 MCAs were mostly losses, with few gains or CN-LOHs. Conversely, chromosome 15 MCAs were mostly CN-LOHs, with very few losses or gains (Figure 3.11). These differential proportions of MCAs suggested non-random distributions of MCAs across the autosome, which had previously been reported in MCA characterizations in other populations (Loh et al., 2018, Terao et al., 2020).

3.3.5.1 MCAs were non-randomly distributed across the autosome.

The chromosomal region spanned by each MCA was visualized through the generation of a Circos pileup plot, revealing non-random patterns of distribution across the autosome in more detail (Figure 3.12). Several MCAs appeared to be sporadically distributed across specific autosomal loci (Figure 3.12). These MCAs were generally smaller focal losses or gains in the larger chromosomes 2, 3, and 4 (Figure 3.12). However, many more MCA were observed to span common loci in each chromosome, with distinct regions frequently impacted by specific MCA types (Figure 3.12). Most of these MCAs were detected in different participants and appear to be recurrent. A large proportion of the commonly occurring gains detected in chromosomes 5 (75.00%), 12 (72.73%) and 21 (98.44%) were predicted to be mosaic aneuploidies that span the entire length of their respective chromosomes (Figure 3.12).

Most detected CN-LOHs extended to the terminal ends of the chromosomal arms (Figure 3.12). The larger median size of CN-LOHs that was observed in the MCA characterization could also clearly be seen. Overall, CN-LOHs were less frequently focal or interstitial than either losses or gains (Figure 3.12). Only 8.41% of all detected CN-LOHs did not span the terminal end of a chromosome arm. Conversely, small losses were frequently clustered across common recurrent interstitial loci, as well as in commonly deleted loci near the terminal ends of many chromosomal arms. These loss hotspots were seen in the arms of chromosomes 1p, 4p, 5p, 6q, 7p, 7q, 8p, 8q, 9q, 10q, 11p, 12q, 16q, 17p, 17q, 18q, 19q, 20q, and 21q (Figure 3.12). These terminal regions may be prone to telomere shortening in aging HSCs, resulting in increased incidence of DSB and losses (Muraki et al., 2012).

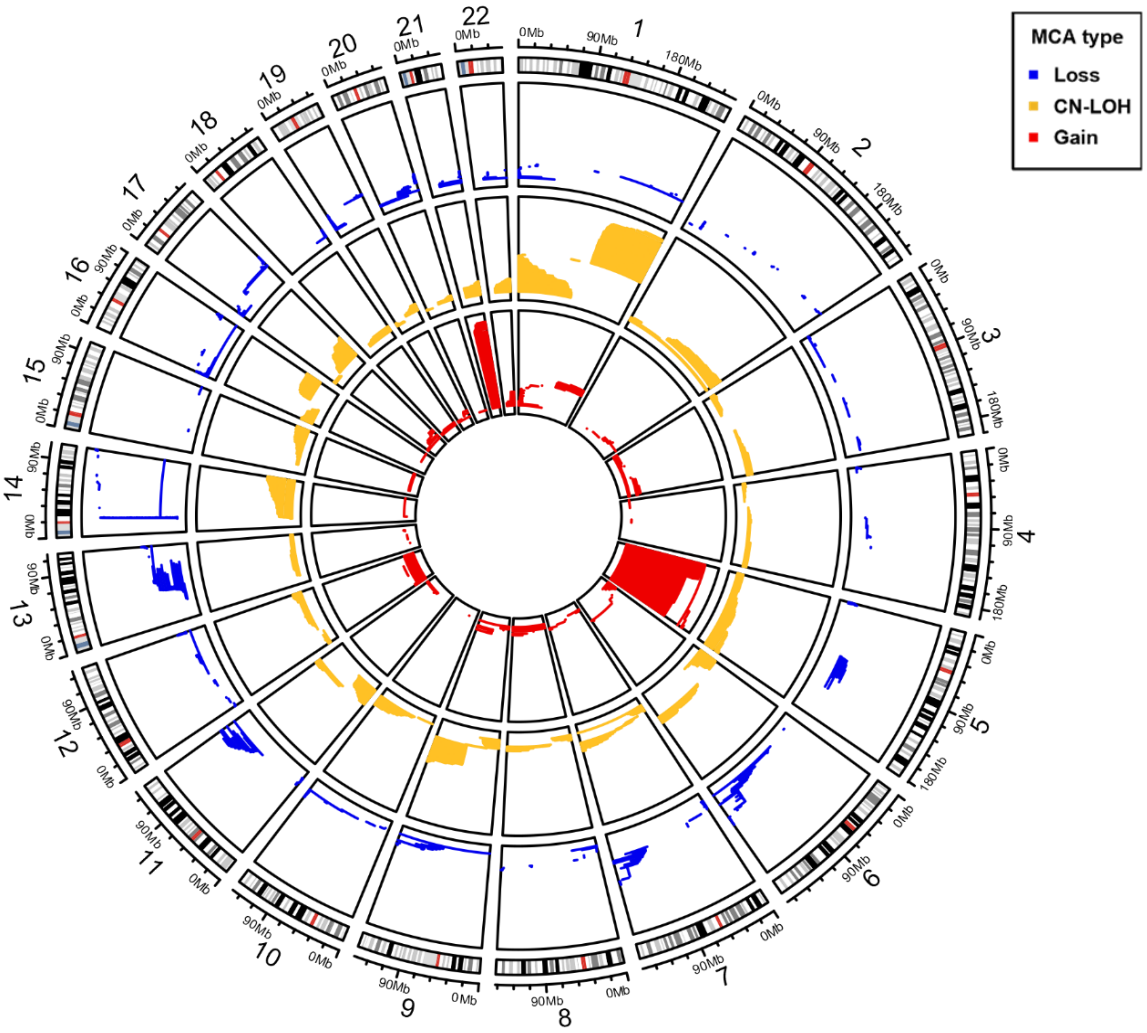


Figure 3.12: ***Autosomal distribution of*** **1245 MCA** ***loci*** **detected** ***in 998*** **AWI-Gen** ***participants*****.**

A circular pileup plot of the genomic locations of the 1245 autosomal MCAs detected in 998 participants. Each chromosome is numbered above with an accompanying ideogram and chromosome coordinates. Each line represents a discrete MCA event from their estimated start to end coordinates. Losses are plotted in blue in track 1, CN-LOHs in gold in track 2, and gains in red in track 3.

The distribution of MCAs is known to be influenced by loci specific mechanistic biases introduced by the physical properties of DNA sequence, in the context of 3D genome organizational architecture, and differential DDR pathways that influence DSB hotpots and non-random mutation rates, and selection biases modulated by driver mutations (Watson and Blundell, 2023). While the combination of differential loci specific mutation rates and fitness effects is most certainly the main driver of the observed MCA distributions in the AWI-Gen cohort as well, it may also be impacted by the coverage of the array probes, the quality of the array data, and the sensitivity of the MCA caller (Loh et al., 2018). The potential impact of probe bias in MCA detection was investigated

using the average informative probes per 1Mb genomic region. The informative probe coverage was plotted using the phased heterozygous probes underlying MCA calls in each of the 9917 participants and calculating the mean per 1Mb windows. This was compared to the proportion of each 1Mb window encompassed by losses, gains, or CN-LOHs to visualize recurrent MCA loci (Figure 3.13). The H3Africa array's high SNP probe density, coupled with the intentional design choice to include more efficient African tag SNPs, resulted in excellent mean genome wide probe coverage for this analysis. The coverage was not uniform, with mean hets/Mb fluctuating noticeably across adjacent 1Mb windows (Figure 3.13). This was expected as the array was designed for representative GWAS and not explicitly for MCA calling. However, very few windows outside of the centromeric and telomeric regions had less than 100 mean hets/Mb (Figure 3.13). The result was more heterozygotes for phasing, translating to more informative sites for MoChA to call smaller MCAs. Thus, there were sparce loci where calling small MCAs was limited by probe density.

Many of the loci with high proportions of small focal recurrent terminal losses can be seen in windows with lower-than-average informative probe coverage (Figure 3.13). Dips in the plotted mean hets/Mb line can be seen towards the terminal ends of all the covered chromosomal arms, coinciding with peaks in the proportion of loss events (Figure 3.13). A total of 114 terminal losses under 5Mb were detected in this analysis. Only 17 (14.91%) of these losses were smaller than 2Mb in length with a corresponding average of 47 (SD = 18) phased heterozygous SNPs underlying the call. The mean heterozygous SNPs informing the calls of all 114 focal deletions was higher, at 194 (Std. Dev. = 157). Thus, the coverage of phased heterozygous SNPs underlying these regions was determined to be sufficient for MoChA to make confident and statistically significant calls, with all MCAs having LOD scores over the recommended threshold of 10. It is likely that these terminal losses were likely real MCA calls. The calls that had too few data points to be confidently classified by MoChA were excluded from analysis during QC. However, caution was taken in their interpretation. The phenomenon of commonly deleted regions (CDR) has been observed across MCA studies, with population specific CDRs being reported in European and Asian cohorts (Loh et al., 2018, Terao et al., 2020). Recurrent MCA loci were identified in the AWI-Gen cohort and annotated with commonly impacted genes to uncover underlying driving their differential distribution across the African autosome.

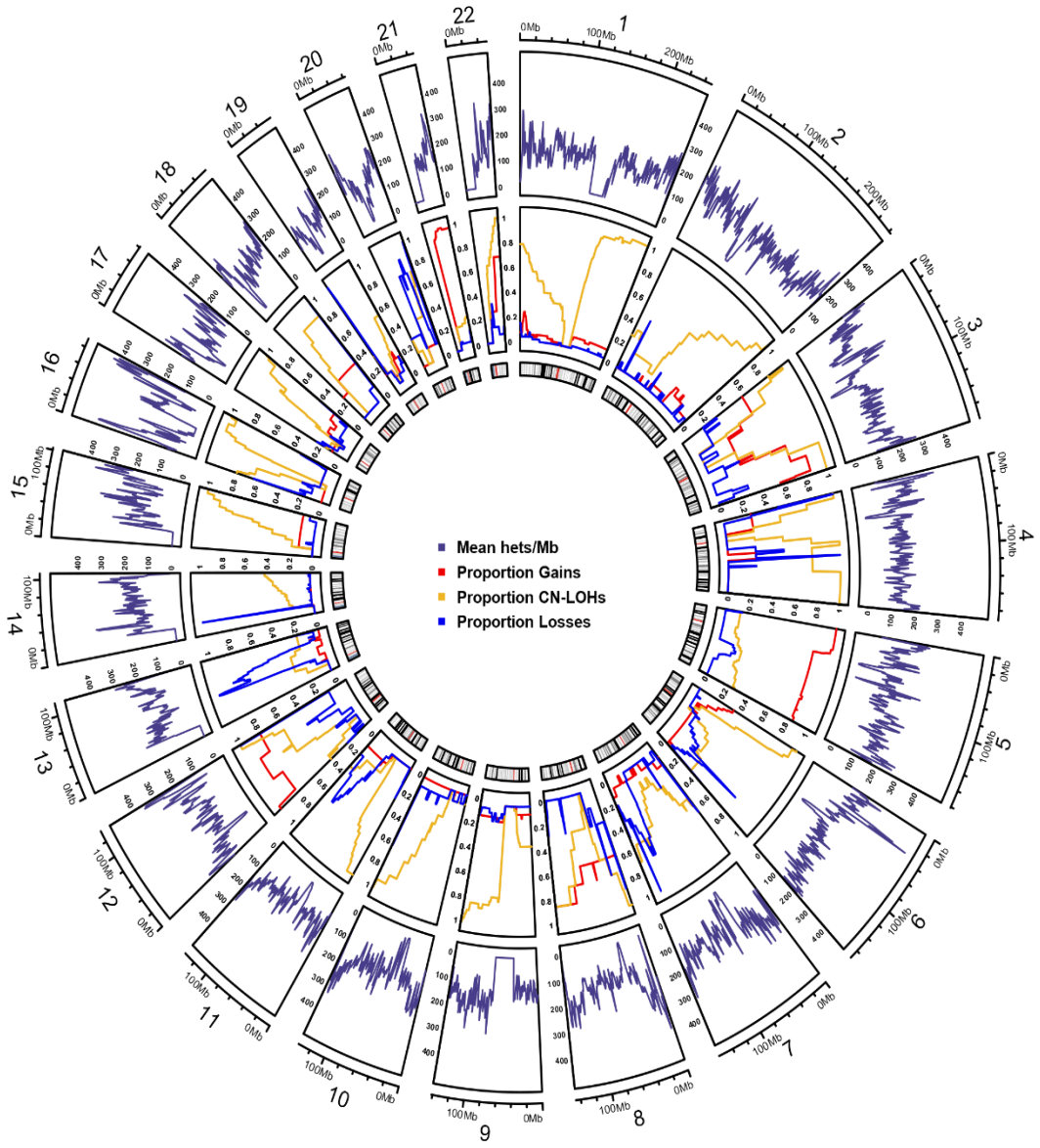


Figure 3.13: Circos plot showing the mean phased heterozygous SNP probe coverage per 1Mb window in the AWI-Gen cohort achieved with the H3Africa SNP array.

The mean phased heterozygous probes in each 1Mb window of the autosome was calculated compared to the proportion of autosomal MCA coverage by MCA type. Proportions are scaled to the total MCA count per chromosome.

3.3.5.2 Novel candidate drivers of common autosomal gain and loss loci.

While it was not possible to identify underlying mechanistic factors influencing MCA mutation rates with the SNP array data, it was possible to infer regions of the genome which may confer selective advantage to HSCs if altered by a loss or gain. This was achieved by identifying regions of the genome that were commonly deleted and commonly gained in the AWI-Gen cohort and then identifying potential driver genes that were impacted by these MCAs that could drive clonal expansions. Two approaches were combined to achieve this. The first approach was adapted from Terao *et al.* (2020) and involved annotating losses and gains with every gene encompassed by their breakpoints. Genes were given an MCA specific score of 1/total gene count underlying each MCA. This accounted for MCA length and the combinatorial impact of all disrupted genes on HSC function. Scores for each gene were tallied and those in 3 or more losses with scores over 0.4 were identified and manually curated as drivers.

A total of 49 candidate genes were impacted by 3 or more losses and had gene scores over the required 0.4 threshold (Table 3.10). Of these, 4 were V(D)J recombination loci, 6 were annotated as long intergenic non-coding RNAs (lincRNAs) locus, 5 were annotated as long-coding RNAs (lncRNAs), 6 were microRNAs (MIR), and 30 were canonical protein expressing genes (Table 3.10). No candidate drivers were identified for losses in chromosomes 1, 3, 9, 10, 11, 15, 16, 19, or 20. No candidate genes were identified in for gains, potentially due to the large size of most gains and rarity of focal loci penalizing individual gene scores. Thus, drivers could be lost in a sea of unimportant or uninvolved genes.

The drivers of AWI-Gen CDRs were compared to published CDRs in the UKB and BJJ cohorts (Loh et al., 2020, Terao et al., 2020). Drivers shared by all three cohorts were *DNMT3A*, *TET2*, *TNFAIP3*, *ARHGEF10*, *TRA*, and *TRD*. Additionally, *PERP* was common to the AWI-Gen cohort and the BBJ cohort. The rest of the regions were novel in the AWI-Gen cohort and were not identified as drivers in CDRs in either cohort. Of the commonly deleted loci in AWI-Gen with the first approach, *ARHGEF10*, *ASXL2*, *DNMT3A*, *IGL*, *TET2*, *TNFAIP3*, *TRA*, *TRB*, and *TRD* are known drivers of CH commonly mutated in CH. None of the other genes have been previously associated previously associated with CH in literature.

Table 3.10: Candidate genes driving the selection of commonly deleted loci scored by the gene approach.

Chromosome	Commonly deleted loci – gene scoring
2	*ASXL2, DNMT3A, MIR1301*
4	*PPA2, TET2*
5	*IRX4, LINC02116*
6	*DLL1, LINC01624, LINC02539, LINC02865, LOC154449, LOC285804, MIR4644, PERP, TNFAIP3, WAKMAR2*
7	*DNAJB6, FAM20C, LOC101927914, MIR11400, MIR153-2, PTPRN2, TRB, TARP*
8	*ARHGEF10, DMTN, FHIP2B*
12	*LINC02361, LOC101928416*
13	*CDC16, RASA3*
14	*TRA, TRD*
17	*METRNL, TBCD*
18	*ADNP2, CTDP1, NFATC1, PARD6G, TXNL4A*
21	*COL18A1, MIR6815*
22	*IGL, IGLL5, LINC01634, MICAL3, MIR650, VPREB1*

A second analysis was adapted from Loh *et al*. (2018) to identify additional candidate drivers that may have been missed by the strict penalization of larger MCAs inherent to gene centric approach. The genome was divided into 1Mb windows. Each MCA was given a score defined by 6Mb/MCA(Mb), with all MCAs under 6Mb being scored 1. While this also accounted for MCA size and relative gene density, it was much less stringent than the gene centric approach.

Common 1Mb genomic windows were identified by generating window scores for each MCA overlap. Windows with scores of 3 or more from loss overlaps were annotated with predicted tumor suppressor genes in pan-cancers and leukemias (Table 3.11). A total of 92 genomic windows were identified with weighted scores of 3 or more, encompassing 54 commonly deleted driver genes predicted to be tumor suppressors in the combined databases (Table 3.11). No commonly deleted windows were identified in chromosomes 3, 10, 15, or 18.

Table 3.11: Candidate genes driving the selection of commonly deleted loci scored by the window approach.

Chromosome	Commonly deleted loci – window scoring
1	*TNFRSF14*
2	*ASXL2, BCL2L11, DNMT3A*
4	*TET2*
5	*SDHA, TERT*
6	*CITED2, ECT2L, IFNGR1, TNFAIP3*
7	*CASP2, EZH2, POT1*
8	*ARHGEF10, DOK2*
9	*NOTCH1*
11	*ATM, CADM1, CBL, CDKN1C, H19, MIR210, MIR34B, MIR34C, ZBTB16*
12	*CHFR, POLE*
13	*DLEU1, DLEU2, KCNRG, MIR15A, MIR16-1, OLFM4, RB1*
14	*IGH, TRA, TRD*
16	*AXIN1, CBFA2T3, FANCA*
17	*ALOX15, DPH1, GPS2, MIR22, MIR132, NF1, SUZ12, TP53*
19	*STK11, TCF3*
20	*PTK6, SS18L1*
22	*BCR*

The known CH driver genes *DNMT3A*, *TET2*, *TNFAIP3*, *ARHGEF10*, *TRA*, and *TRD* were identified by both approaches, suggesting a significant role of these genes in recurrent losses (Table 3.11). The *POT1* gene identified as a candidate driver across the AWI-Gen, UKB, and BBJ, and *IFNGR1* was identified in AWI-Gen and BJJ only. Of the identified candidates associated with hematological malignancies only *ASXL2, ATM, BCR, CBFA2T3, CBL, DNMT3A, EZH2, FANCA, GPS2, IGH, NF1, NOTCH1, POT1, RB1, SUZ12, TET2, TNFAIP3, TNFRSF14, TP53, TRA, TRD, TERT, ZBTB16* had been identified as known canonical CH drivers through MCAs or in CHIP cases.

The other loci in Table 3.11 have not specifically been identified as drivers of CH, but their roles in malignancies suggests plausible mechanisms of dosage sensitive selection in CH. Whether or not the genes in Tables 3.10 and 3.11 outside of known CH genes are expressed in HSCs or their progeny or are significantly impacting the selection of loss events warrants further investigation. Additional larger genomic regions known to be associated with CH which were detected in AWI-Gen are del(5q), del(13q14), and del(20q). The location of these loci across the autosome are shown in Figure 3.14.

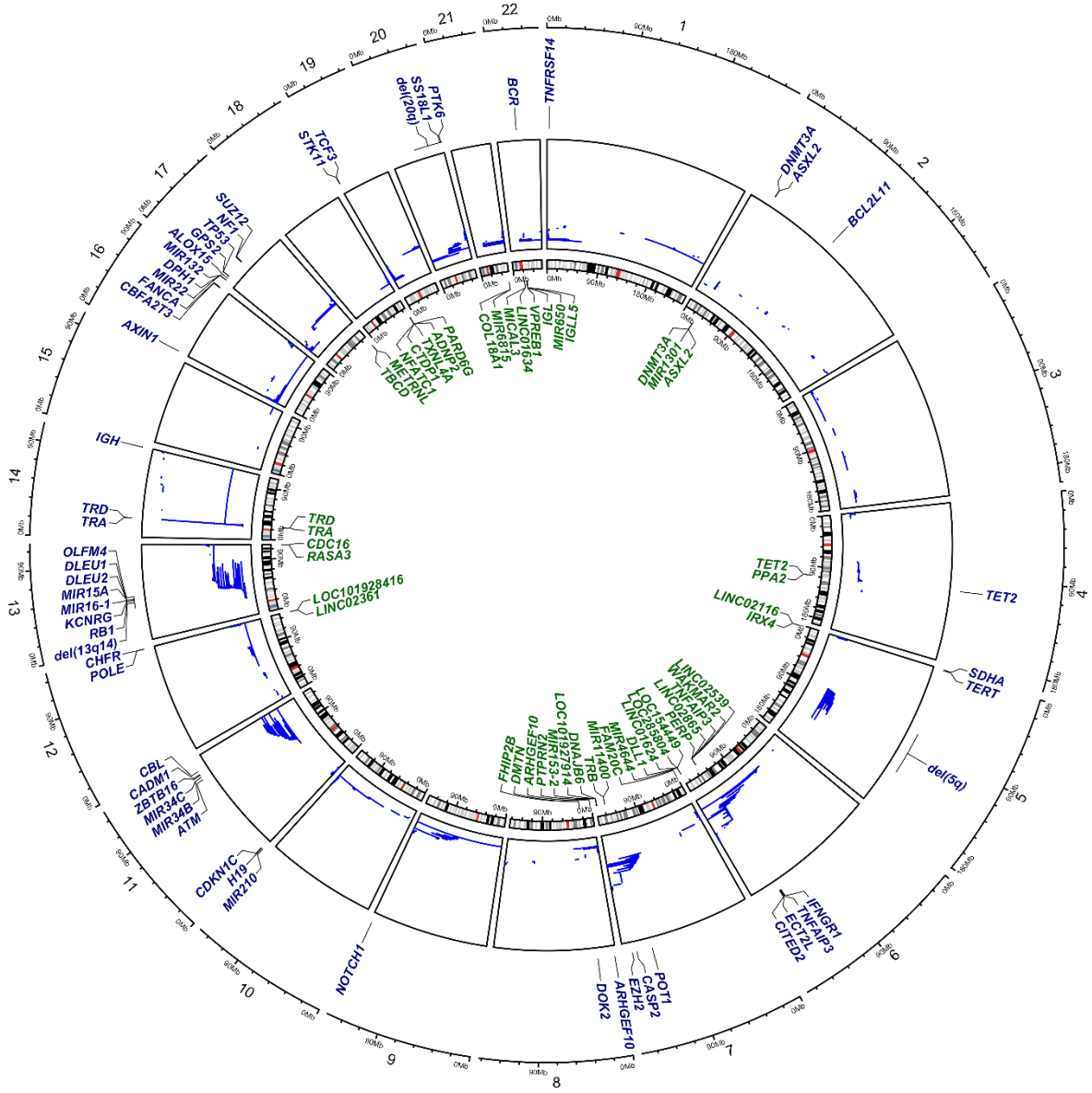


Figure 3.14: Recurrent focal loss loci in the AWI-Gen cohort commonly impacting candidate driver genes.

Deletion events across the autosome are plotted as blue segments in track 2. Each line is a discrete loss. Commonly deleted loci identified through scoring genes in Table 3.10 are shown on the inner track and are coloured green. Commonly deleted known pan-cancer and leukaemia associated driver loci are shown in the outer track and are labelled blue.

Far fewer focal duplication windows were identified with window scores over 3. In these windows, 20 genes that were reported as pan-cancer or leukemia associated oncogenes were identified in 4 chromosomes in the AWI-Gen cohort (Table 3.12). Of the identified candidate driver genes in recurrently duplicated loci, only *TERT* was shared between focal gains and losses suggesting a complex dual role in MCA

formation and selection in Africans. Additionally, *ERG, U2AF1, RUNX1, MTOR*, and *TERT* had all previously been reported as known CH drivers. No comparisons were made with UKB and BBJ as they only investigated CDRs in their studies. Making this one of the first studies to do so.

Table 3.12: Candidate pan-cancer and leukaemia associated driver genes recurrently duplicated in common loci scored by 1 Mb windows in the AWI-Gen cohort.

Chromosome	Commonly duplicated loci
1	*EPHA2, MIR34A, MTOR*
5	*CTNND2, IL7R, PRDM9, SKP2, TERT*
21	*ERG, ETS2, MIR155, MIR99AHG, OLIG2, RUNX1, TMPRSS2, U2AF1*
22	*APOBEC3B, MRTFA, PDGFB, PIM3*

It is likely that a great many loci that impact the formation of both loss and gain events in the AWI-Gen cohort were not identified in the above analysis. A combination of changes to the expression of all the genes underlying these MCAs, should they be expressed in mature blood and immune cell lineages, is likely to drive their expansion. However, the above analysis does provide initial insight into the commonly occurring focal regions in African populations and identifies novel associations with protein coding and regulatory loci that may contribute to MCA formation and selection not only in Africans, but also potentially in all aging populations. The non-random distribution of MCAs across the AWI-Gen cohort could be explained, in part, by the underlying genomic architecture features at these common loci being susceptible to MCA mutation in combination with changes in the expression of these genes conferring selective advantage. These locations would be acted upon by unique selection pressures. The demographic factors that were identified as significant in the regression analysis may play a role in that selection or serve as proxies for exogenous selection pressures acting on the AWI-Gen cohort. The impact of these factors on the non-random MCA distribution was investigated further.

3.3.5.3 Differential autosomal MCA distributions were associated with participant sex and age.

The overall proportions of detected MCAs stratified by the broad categories of loss, gain, or CN-LOH did not significantly differ between male and female participants (Figure 3.9C). However, specific differences in MCA distributions between male and female participants and by participant age were apparent when stratifying by MCA type and position relative to chromosomal arm (Figure 3.15). These associations were screened for in MCA phenotypes with cohort wide counts of 4 or more were grouped into MCA phenotypes to investigate the potential relationship between age, sex, and specific MCA occurrence in the AWI-Gen cohort (Figure 3.15). A table of the full 82 MCA phenotypes can be found in Appendix I.

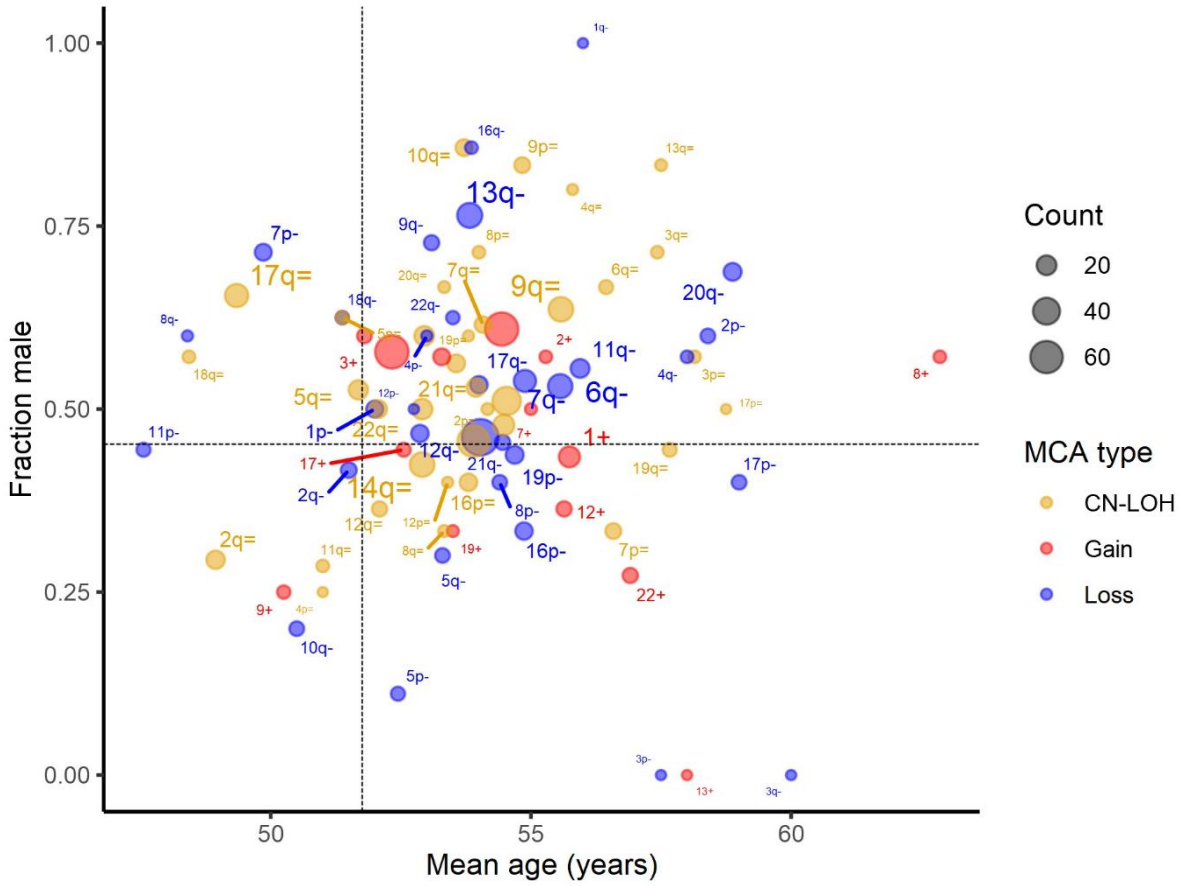


Figure 3.15: Bubblegraph showng autosomal MCA phenotype proportions by mean age and fraction male sex distributions.

MCAs were stratified by chromosome arm and type into MCA phenotypes. Bubbles represent discrete MCA phenotypes with counts of 4 or more plotted by the fraction of males and the mean age of the participants they were detected in. Bubble size reflects the MCA phenotype count. Bubble colour represents MCA type. q = q -arm, p = p-arm; - represents a loss, = represents a CN-LOH, + represents a gain. Dotted lines indicate study means for age and fraction of males.

The enrichment of detected MCAs in the elderly observed in the regression was reflected in the phenotype stratification analysis. Most stratified MCA phenotypes had male fractions upwards of the cohort mean male fraction of 0.45 and mean ages greater than the cohort mean age of 51 years (Figure 3.19). However, a wide range of sex and age distributions was observed across the stratified MCA phenotypes with some detected primarily in females or in younger participants (Figure 3.15). This suggested potential for sex and age-related bias of specific MCA phenotypes which did not follow the overall trend towards males and the elderly. To identify the recurrent MCA regions with enrichment, MCA phenotypes were stratified into 1Mb windows in which they were detected across the autosome. Fischer's exact tests were conducted on the proportion of MCAs in each 1Mb window between male and female participants to account for the reduced MCA counts introduced by stratification. Concurrent windows were manually curated and combined into larger window loci with evidence of MCA sex enrichment (Table 3.13).

Table 3.13: Autosomal window loci with evidence of differential distributions between 521 male and 477 female participants with at least one MCA.

MCA	Start	End	Count (M/F)	OR [95% CI]	P-value	Mean CF (SD)
Loss						
5p-	1	2022782	9 (1/8)	0.11 [0.00-0.85]	**0.02**	9.41 (4.80)
7q-	157082936	159138663	7 (1/6)	0.15 [0.00-1.25]	0.06	12.03 (8.52)
9q-	71033043	105114367	8 (7/1)	6.47 [0.84-292.38]	0.07	9.08 (8.46)
10q-	133933713	135534747	8 (1/7)	0.13 [0.00-1.01]	**0.03**	7.18 (2.90)
13q-	50323749	52448017	21 (15/6)	2.99 [1.04-10.51]	**0.03**	4.49 (4.06)
13q-	113326671	115169878	5 (5/0)	Inf [0.84-Inf]	0.06	8.20 (4.75)
14q-	105955668	107349540	4 (0/4)	0.00 [0.00-1.38]	**0.05**	17.04 (11.37)
Gain						
6p+	1	6978633	8 (7/1)	6.47 [0.83-292.38]	0.07	3.69 (2.94)
CN-LOH						
2q=	142359467	243199373	17 (5/12)	0.38 [0.10-1.16]	0.08	2.13 (2.77)
9p=	1	26576086	10 (9/1)	9.30 [1.31-404.36]	**0.01**	10.83 (13.64)
10q=	89534244	135534747	13 (11/2)	5.12 [1.11-47.73]	**0.02**	1.69 (1.58)

q = q -arm, p = p-arm; - represents a loss, = represents a CN-LOH, + represents a gain; Start/End of window base pair coordinates; M = male; F = female; OR = Fischer's Exact odds ratio with male reference; CI = confidence interval; P-value from Fischer's Exact test not adjusted for multiple tests; CF = cell fraction (%); SD = standard deviation of mean CF. Inf = infinity as Fischer's Exact cannot give exact OR estimates with counts of 0.

This analysis identified 11 autosomal window loci with evidence of differential distributions in either male or female participants suggestive of sex specific selection pressures in the AWI-Gen cohort (Table 3.13) (Figure 3.16). Males had significantly greater odds of having large losses on 13:50323749 – 52448017bp (OR = 2.99 [1.04-10.51], p = 0.03), CN-LOH on 9:1 – 26576086bp (OR= 9.30 [1.31-404.36], p = 0.01), and CN-LOH 10: 89534244 – 135534747bp (OR = 5.12 [1.11-47.73], p = 0.02) (Table 3.13) (Figure 3.16A). They also had greater odds of having losses which trended towards significance on 9:71033043 – 105114367bp (OR = 6.47 [0.84-292.38], p = 0.07), small terminal losses on the end of 13:113326671-115169878bp (OR = Inf [0.84-Inf], p= 0.06), as well as large gains on 6:1 – 6978633bp (OR = 6.47 [0.83-292.38], p = 0.07) (Table 3.13) (Figure 3.16B). Conversely, regions were identified where males had significantly lower odds of MCA presence than females (Table 3.13) (Figure 3.16). These included small terminal losses on 5:1- 2022782bp (OR = 0.11 [0.00-0.85], p = 0.02), small terminal losses on 10:133933713 – 135534747bp (OR = 0.13 [0.00-1.01], p = 0.03), and small focal losses of the *IGH* region on 14: 105955668 – 107349540bp (OR = 0.00 [0.00-1.38], p = 0.05) (Table 3.13) (Figure 3.16A).

Men had lower odds of small terminal losses that trended towards significance on 7:157082936 – 159138663bp (OR = 0.15 [0.00-1.25], p = 0.06), as well as large CN-LOH on 2:142359467 – 243199373bp (OR = 0.38 [0.10-1.16], p = 0.08) (Table 3.13) (Figure 3.16B). None of these apparent sex biases remained significant at FDR<0.1 after adjustment for multiple testing due to the substantial number of tests run by the per 1Mb window methodology. The lower MCA counts induced by this level of this stratification combined with multiple testing requirements for 2875 1Mb autosomal windows may have confounded testing. Only the individual combined window post curation Fischer's Exact p-values were reported in Table 3.13 as good evidence for association between specific genomic loci and participant sex that future validation or replication. Plausible driver genes from the generated candidate driver gene list were annotated across the MCA regions which showed sex bias (Figure 3.16).

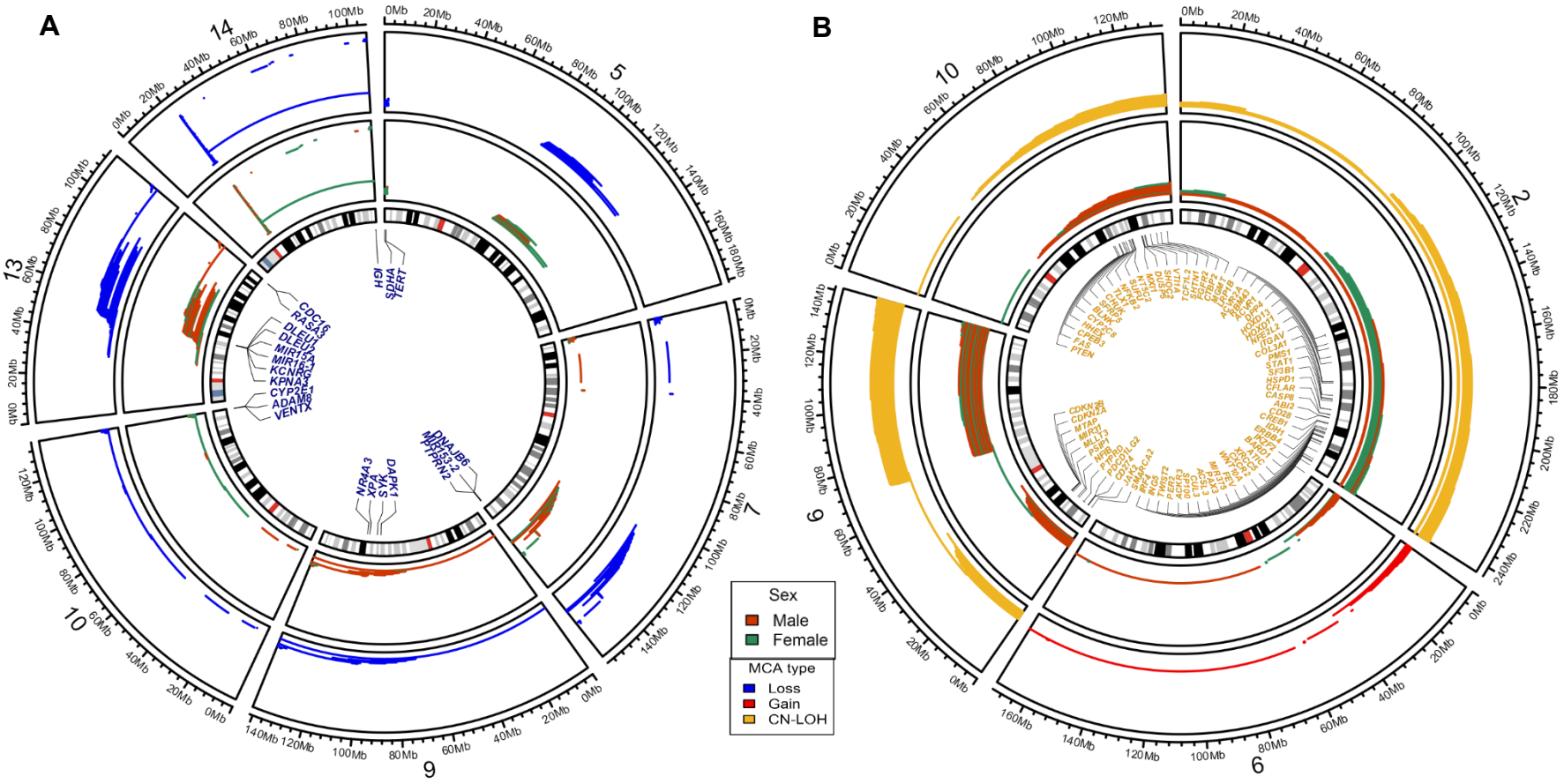


Figure 3.16: Autosomal MCA loci with evidence of sex biased selection in the AWI-Gen cohort.

A) Circular pileup plot of regions of the autosome which showed evidence for increased odds of losses (blue) in either males (orange) or females (orange). **B)** Circular pileup plot of regions of the autosome which showed evidence for increased odds of gains (red) or CN-LOHs (gold) in either males (orange) or females (orange). Chromosomes are numbered above. Each stacked coloured line represents a discrete MCA event in the outer track. All MCAs on each chromosome are shown for context and comparison of significant and non-significant regions. The second track shows the same MCAs as in track 1 coloured by the sex of the participant in which it was detected. Candidate diver genes encompassed by the region are annotated on a chromosomal ideogram in the inner track.

Most of the 82 autosomal MCA phenotypes (74.39%) with four or more counts were detected in participants over the median age of the cohort (51 years) reflecting the well-established age enrichment (Figure 3.14). Wilcoxon signed-rank one-sample tests were conducted to detect skewing of MCA class age distributions from an overall cohort median of 51 years. Despite being detected more often on average in the elderly, most MCA classes did not deviate significantly from the median cohort age. There were several MCA classes that were significantly enriched in the elderly in unadjusted analysis. These included gains in chromosomes 8 (median = 67 years, IQR [16.5], p = 0.04), 12 (median = 55 [3], p = 0.01), and 21 (median = 55 [8.5], p = 9.51×10^{-04}). Additional significance was observed in CN-LOH events in 1q (median = 55 [8], p = 0.02), 7p (median = 56 [8.25], p = 0.02), and 9q (median = 56 [12], p = 0.02). Finally, participants with losses in 6q (median = 56.5 [8.25], p = 2.94×10^{-03}), 7q (median = 56 [9.25], p = 0.04), 8p (median = 55 [6], p = 0.04), 14q (median = 54 [10.75], p = 0.02), and 20q (median = 57.5 [8], p = 0.01) were also observed to have age distributions higher than the median of the full cohort. Once p-values were adjusted for multiple comparisons, only gains in chromosome 21 remained significant at an FDR = 0.08. Losses in 6q, 14q, 20q were all at FDR = 0.12, potentially trending towards significance with a less stringent acceptance of false discovery.

Therefore, a cautious approach to the interpretation of apparent sex and age biases of specific MCA classes across the genome is warranted. Overall, there was compelling evidence that autosomal loci were enriched for specific MCAs in males and females or were more likely to be seen in the elderly. Increasing age and participant sex appeared to confer unique selection pressures on specific regions of the genome that were also copy-number dependent. The significant increase in the odds of MCA occurrence observed in the regression analysis may be driven by these unique selection pressures at these loci. Conversely, in regions of the autosome with similar observed MCA distributions across all ages and sexes, these factors may not have been as important in their selection. It was, therefore, hypothesized that participants from the three broad regions of recruitment were likely to experience additional selection pressures, unique to the regional environments, which may have contributed to the significant increased odds of MCA occurrence observed in WA participants in the regression analysis. The impact of participant region of recruitment on the observed distribution of detected autosomal MCAs enrichment was investigated at the stratified MCA level.

3.3.5.4 Autosomal MCA distributions differed across regions of recruitment.

Several analyses provided convincing evidence for differential distribution of MCAs at specific autosomal loci based on the region of participant recruitment. Autosomal loci were enriched for MCAs that were specific to Eastern, Western, and Southern African participants. The proportion of losses, gains, and CN-LOHs that made up the regional totals of detected MCAs was significantly different (Chi^2 = 11.52 p = 0.02). The frequencies of losses, gains, and CN-LOHs stratified by chromosome were plotted and compared between participants from SA vs EA, SA vs WA, and WA vs EA (Figure 3.17). Spearman's rank correlation analysis revealed strong significant correlations between CN-LOHs across all three regions after adjusting for multiple comparisons (Figure 3.17).

The strongest correlation between CN-LOH frequencies was observed between SA and WA (Rho = 0.85). A strong significant correlation was observed for loss frequencies between SA and EA participants (Rho = 0.84), while a moderate significant correlation was observed for loss frequencies between SA and WA (Rho = 0.57). The frequencies of losses per chromosome were also moderately correlated between EA and WA (Rho = 0.44) and trending towards significance. The biggest overall difference in MCA distributions across the autosome were observed for gains. The frequencies of detected gains were very strongly correlated between SA and EA (Rho = 0.94), while no significant correlations were observed for gains between WA and any of the regions (Figure 3.17).

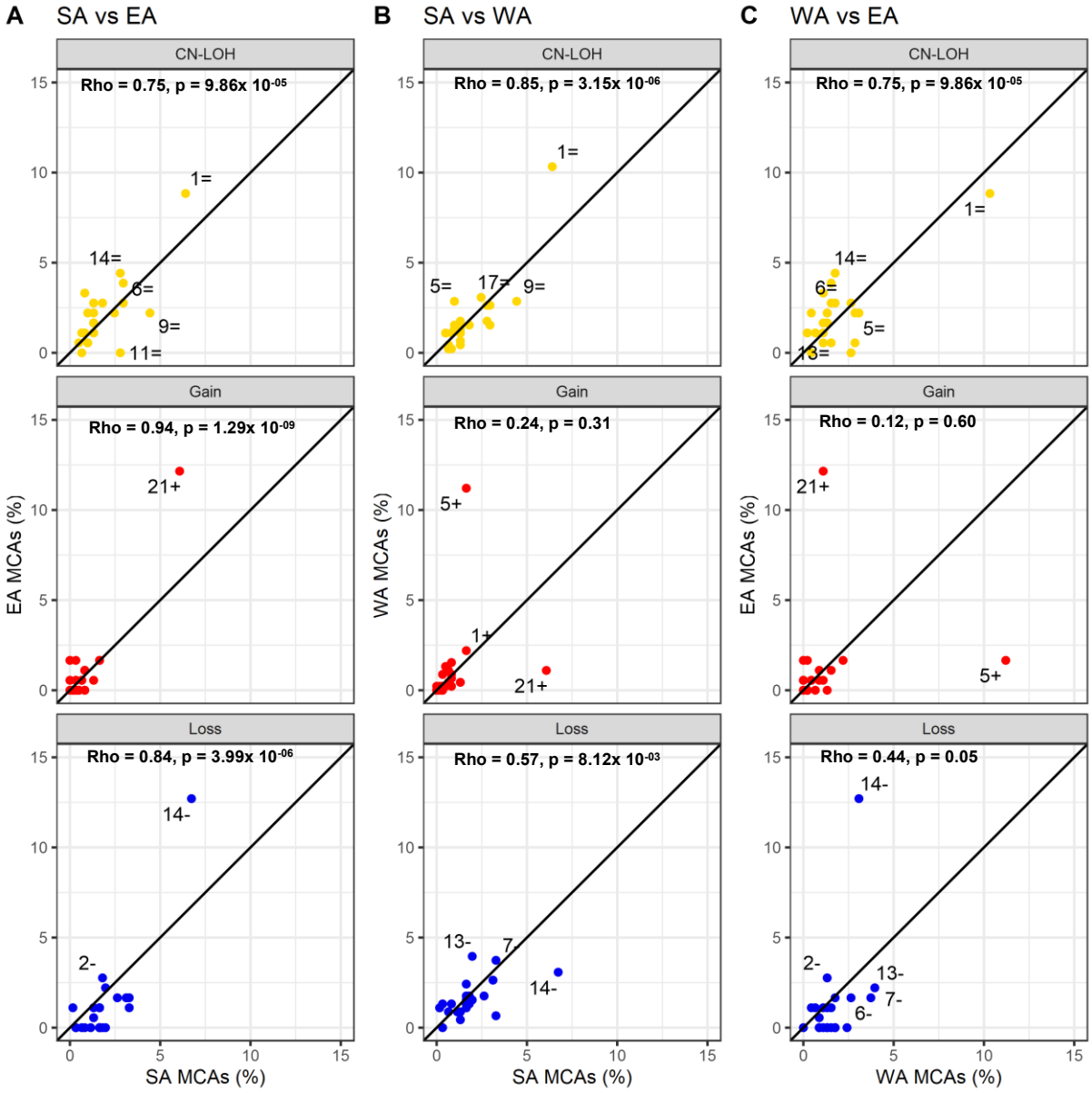


Figure 3.17: Autosomal MCA phenotype frequency correlations by region of recruitment.

The frequencies of autosomal MCAs as a function of the total MCA events detected in each region were compared between the three regions of recruitment by MCA type. Individual points represent each chromosome. Spearman's rank correlation statistics for each comparison by type and region are shown in each plot. q = q-arm, p = p-arm; - represents a loss, = represents a CN-LOH, + represents a gain. EA = Eastern Africa, SA = Southern Africa, WA = Western Africa.

The correlation analysis showed that MCAs broadly categorized by type and chromosome were similarly distributed between participants from all three regions. These regions were identified as common MCA regions in African populations, where MCA formation was not being significantly impacted by the exogenous selection pressures or germline variation differences that may exist between the three regions.

Significantly different frequencies could potentially indicate regions subjected to unique selection pressures. However, these overall MCA frequency comparisons by chromosomes did not reveal specific loci in which the regional differences could be accruing. MCAs were once again stratified by chromosomal arm and MCA type, and then by 1Mb autosomal windows in a manner done by participant sex. The counts of each MCA category were then evaluated for regional enrichment with Fischer's exact tests within the 998 participants with at least one MCA, within each 1Mb window. This was used to identify more specific manually curated autosomal loci within each MCA category that showed evidence of region biased selection (Table 3.14) (Figures 3.18-3.20).

Following manual curation and combination of the Fischer's exact tests run on each 1Mb window across the autosome, 16 windowed loci showed evidence of differential distribution between either EA vs SA, SA vs WA, or WA vs EA participants (Table 3.14). Small focal losses on chromosome 2: 25000001-27000000bp, encompassing the established *DNMT3A* CH driver gene, were significantly more likely to be found in SA vs WAs (OR = Inf [1.08-Inf], p = 0.02) with 7 of the 8 losses observed in SA participants and none being observed in WAs (Table 3.14) (Figure 3.18). The loss of *DNMT3A* has been reported as one of the most prolific CH drivers across many MCA and CHIP cohorts. Thus, there may be factors impacting its selection in these studied cohorts as well as the SA participants that are not shared by the EA and WA participants.

The loss event hotspots on the q-arm of chromosome 7 clustered together as a single phenotype were not significantly differentially distributed. However, upon closer inspection two distinct regional specific loss loci were uncovered (Figure 3.18A). Large loss events on chromosome 7:109000001-140000000bp were significantly more likely to be detected in WA vs SA participants (OR = 12.26 [1.69 - 538.36], p = 2.85 x 10^{-03}) (Table 3.14) (Figure 3.18). An additional recurrent small focal loss event loci was identified on chromosome 7:142035845-142513256bp which only spans the T-cell receptor beta (TRB) locus that had increased odds of being from SA vs WA participants (8 vs 1) trending towards significance (OR = 6.07 [0.81 – 270.20], p = 0.09) (Figure 3.18).

Table 3.14: Autosomal windows with evidence of differential MCA distributions between 483 SA, 362 WA, and 153 EA participants.

MCA	Start	End	Count (WA/EA/SA)	OR [95% CI]	P-value	OR [95% CI]	P-value	OR [95% CI]	P-value
Loss				*WA vs. EA*		*SA vs. WA*		*SA vs. EA*	
2p-	25000001	27000000	8 (0/1/7)	0.00 [0.00-16.67]	0.30	Inf [1.08-Inf]	**0.02**	2.25 [0.29-102.00]	0.69
7q-	109000001	140000000	11 (9/1/1)	3.87 [0.53-170.85]	0.29	0.08 [0.00-0.59]	**2.85 x10^{-03}**	0.32 [0.00-25.07]	0.43
11p-	1	2000000	6 (5/0/1)	Inf [0.39-Inf]	0.33	0.15 [0.00-1.33]	0.09	Inf [0.01-Inf]	1.00
11q-	111000001	124000000	14 (2/2/10)	0.42 [0.03-5.91]	0.59	3.78 [0.80-35.72]	0.08	1.61 [0.34-15.23]	0.74
13q-	45000001	84000000	20 (15/4/7)	1.61 [0.50-6.77]	0.61	0.34 [0.12-0.90]	**0.02**	0.55 [0.14-2.59]	0.31
14q-	22000001	23000000	69 (7/23/39)	0.11 [0.04-0.28]	**6.68 x10^{-08}**	4.43 [1.93-11.88]	**5.24x10^{-05}**	0.50 [0.28-0.91]	**0.02**
17p-	4000001	8000000	7 (0/0/7)	0.00 [0.00-Inf]	1.00	Inf [1.09-Inf]	**0.02**	Inf [0.46-Inf]	0.21
17q-	80000001	81000000	10 (1/0/9)	Inf [0.01-Inf]	1.00	6.81 [0.94-299.43]	**0.05**	Inf [0.63-Inf]	0.12
18q-	77000001	78000000	8 (6/0/2)	Inf [0.50-Inf]	0.19	0.25 [0.02-1.39]	0.08	Inf [0.06-Inf]	1.00
Gain									
5p+	1	47000000	60 (50/2/8)	12.07 [3.10-103.77]	**1.63 x10^{-06}**	0.10 [0.04-0.23]	**2.20 x10^{-12}**	1.28 [0.25-12.49]	1.00
5+	47000001	180000000	54 (50/2/2)	12.07 [3.10-103.77]	**1.63 x10^{-06}**	0.03 [0.00-0.10]	**<2.20 x10^{-16}**	0.31 [0.02-4.37]	0.25
6p+	1	25000000	8 (6/1/1)	2.59 [0.31-119.89]	0.68	0.12 [0.00-1.02]	**0.05**	0.32 [0.00-25.07]	0.43
21+	14000001	48000000	64 (5/22/37)	0.08 [0.02-0.24]	**1.52 x10^{-08}**	5.89 [2.27-19.39]	**1.72 x10^{-05}**	0.50 [0.27-0.92]	**0.02**
CN-LOH									
1q=	144000001	249000000	55 (29/9/17)	1.41 [0.63-3.47]	0.46	0.42 [0.21-0.80]	**0.01**	0.59 [0.24-1.53]	0.24
5q=	121000001	180000000	16 (12/1/3)	5.26 [0.77-226.61]	0.12	0.18 [0.03-0.68]	**0.01**	0.96 [0.08-50.51]	1.00
11p=	1	1000000	22 (9/0/13)	Inf [0.85-Inf]	0.06	1.08 [0.42-2.90]	1.00	Inf [0.98-Inf]	**0.05**

q = q -arm, p = p-arm; - represents a loss, = represents a CN-LOH, + represents a gain; Start/End of window base pair coordinates; WA = Western Africa, EA = Eastern Africa, SA = Southern Africa; OR = Fischer's Exact odds ratio; CI = confidence interval; P-value from Fischer's Exact test not adjusted for multiple tests; CF = cell fraction (%); SD = standard deviation of mean CF. Inf = infinity as Fischer's Exact cannot give exact OR estimates with counts of 0.

Losses on chromosome 11 trended towards significant differential odds from either WA or SA participants based on the arm loci (Table 3.14). Small terminal focal losses on the p-arm of chromosome 11:1-2000000bp were more likely to be detected in WA vs SA participants (OR = 6.74 [0.75 – 319.44], p = 0.09). Conversely, the large losses on the q-arm of chromosome 11:111000001-124000000bp were more likely to be found in SA vs WA participants (OR = 3.78 [0.80-35.72], p = 0.08) (Figure 3.18). Detected CN-LOH events on both arms of chromosome 11 were evenly distributed among participants from SA and WA sites, however, none were detected in participants from EA (Table 3.14) (Figure 3.20). No CN-LOHs on the chromosome 11:1-1000000bp were seen in EA participants resulting in lower odds that trended towards significance in both EA vs WA (OR = 0.00 [0.00 – 1.19], p = 0.06) and EA vs SA participants (OR = 0.00 [0.00 – 1.02], p = 0.05).

Interestingly, the same losses encompassing 13q14 that showed evidence of higher odds of detection in men were also more likely to be from WA vs SA participants (OR = 2.94 [1.11-8.61], p = 0.02) suggesting independent sex and regional factors playing a role in 13q loss selection or mutagenesis. The most observed loss event in the AWI-Gen cohort were small focal losses of the *TRA* region on chromosome 14: 22000001-23000000bp which made up 5.54% of all detected MCAs in the cohort. This recurrent localized loss had significantly lower odds of being seen in WA participants than both SA (OR = 0.22 [0.08-0.52], p = 5.24 x 10^{-05}) and EA participants (OR = 0.11 [0.04-0.28], p = 6.68 x10^{-08}) (Table 3.14) (Figure 3.18). Accounting for the lower overall count of EA participants in the cohort, this loss event was most likely to be seen in EA participants with significantly lower odds of SA vs EA (OR = 0.50 [0.28-0.91], p = 0.02). This suggested that there were specific regional factors shared between SA and EA participants which impacted the selection of this focal *TRA* deletion more significantly than in WA, which did not affect the selection of other 14q losses or large 14q CN-LOHs (Figure 3.18) (Figure 3.20).

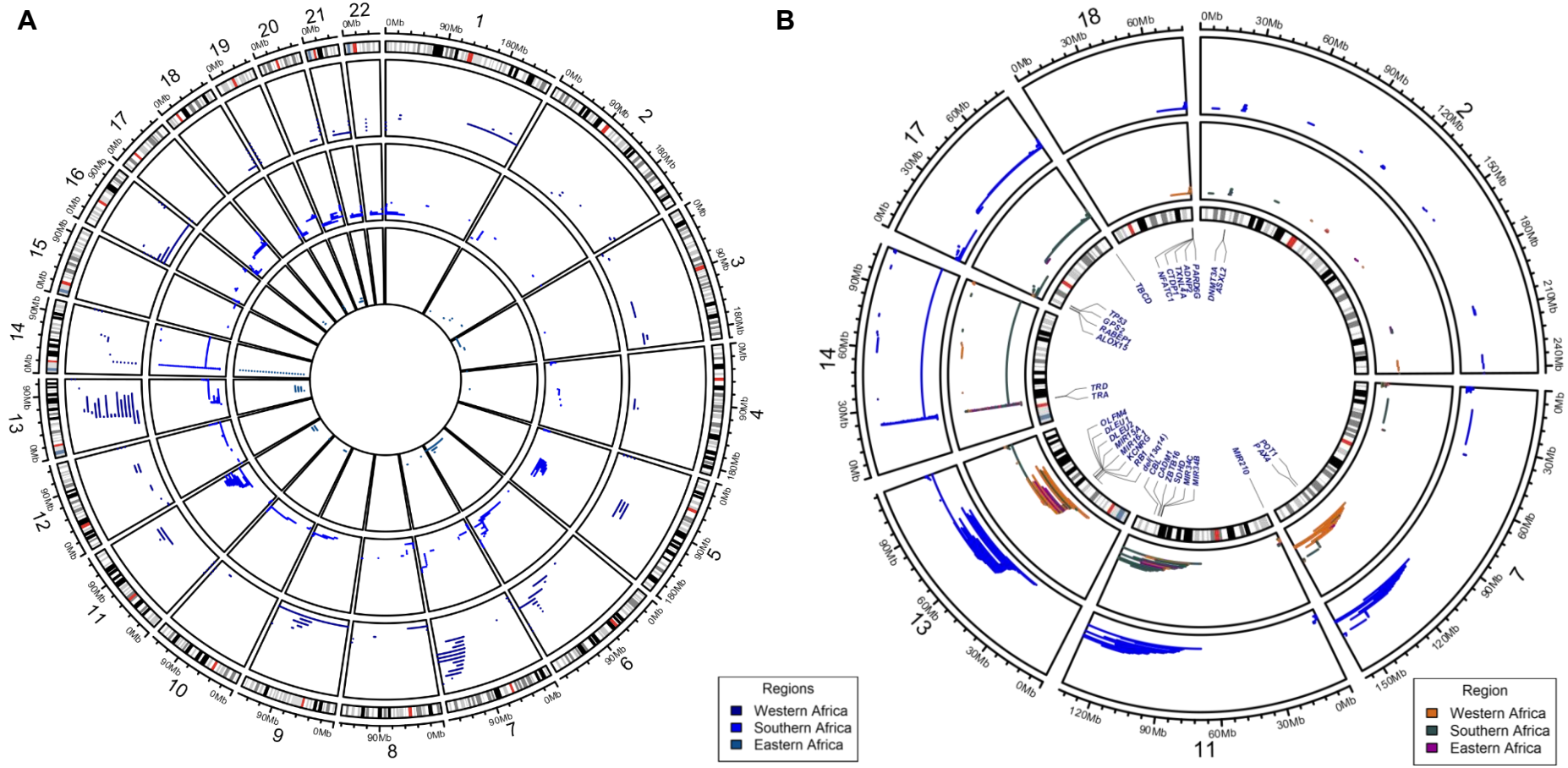


Figure 3.18: Autosomal loss loci with evidence of region of recruitment biased selection in the AWI-Gen cohort.

A) Circular pileup plot of all detected loss events detected in participants of each region of recruitment. Chromosomes with coordinates are numbered above accompanied by ideograms. All WA loss events are in track 1, all SA loss events are in track 2, and all EA loss events are in track 3. **B)** Circular pileup plot of loss loci with evidence for differential distributions by region of recruitment. Chromosomes with coordinates are numbered above. Each stacked blue line represents a discrete loss from start to end in track 1. Track 2 represents the same losses as in track 1 coloured by the region of recruitment of the participant in which it was detected. Candidate diver genes encompassed by the region are annotated on a chromosomal ideogram in track 3.

Loss events in chromosome 17 were mostly seen in participants from Southern African sites. Small focal losses on the p-arm at 17: 4000001-8000000bp showed significantly greater odds of being from SA vs WA participants (OR = Inf [1.09-Inf], p = 0.02) and were not detected in any WA or EA participants (Table 3.14) (Figure 3.18). Interestingly, all CN-LOHs detected on the p-arm of chromosome 17 were from SA participants as well, although this did not reach statistical significance (Figure 3.20). Additionally, small focal terminal losses on the q-arm of 17: 80000001-81000000bp had greater odds of being from SA vs WA participants that trended towards significance (OR = 6.81 [0.94-299.43], p = 0.05). The 17q losses include a newly identified novel CDR region which encompasses the *METRNL* and *TBCD* genes that may be specific to SA participants. Another novel African focal CDR on the terminal q-arm of chromosome 18: 77000001-78000000bp showed some evidence of greater odds in WA vs SA participants (OR = 4.04 [0.72-41.25], p = 0.08) while not being found in any EA participants (Table 3.14) (Figure 3.18). This CDR encompassed 5 plausible candidate genes not yet identified as CH drivers that may be driving selection specifically in WA participants.

The two most frequent gain events in the AWI-Gen cohort were detected on chromosomes 5 and 21, with each making up 5.15% of all MCAs (Figure 3.11). Both showed strong evidence of differences in their distributions based on region of participant recruitment that may be suggestive of regional specific selection pressures (Figure 3.19). Most of the gain events detected on chromosome 5 (75%) spanned the entire length of the chromosome (Figure 3.19). These somatic aneuploidies were detected almost entirely in WA participants, with only 1 out of the 48 being detected in an EA participant and 0 being detected in SA participants (Figure 3.19). When accounting for the number of WA, SA, and EA participants with at least one MCA, aneuploid gains of the chromosome 5 had significantly greater odds of being observed in WA vs SA participants (OR = Inf [18.55-Inf], $p < 2.20 \times 10^{-16}$) and WA vs EA participants (OR = 22.61 [3.79-916.81], $p = 5.89 \times 10^{-07}$). While much rarer than in WA, gains were detected in SA participants that encompassed the p-arm of chromosome 5 (Figure 3.19).

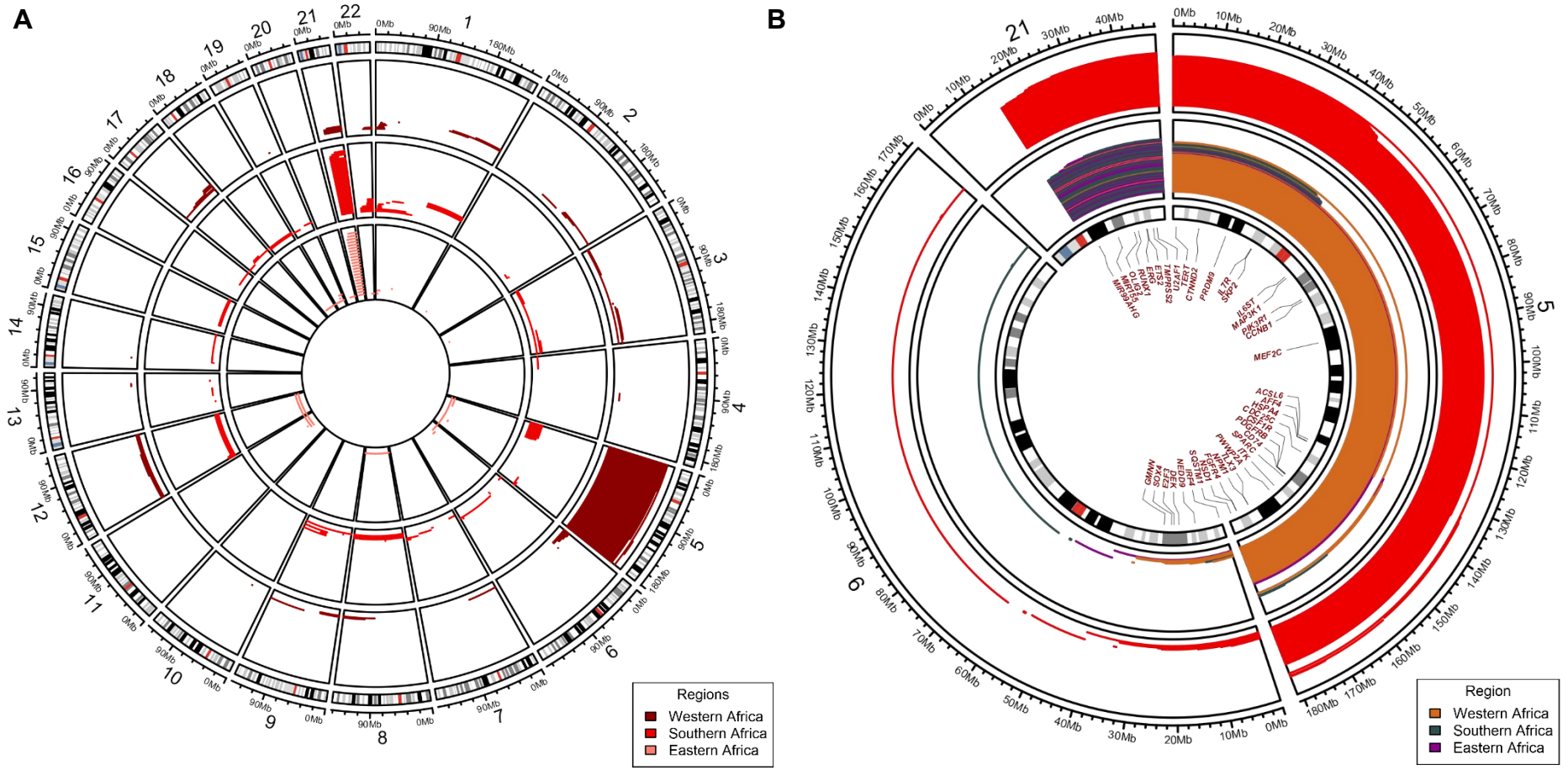


Figure 3.19: Autosomal gain loci with evidence of region of recruitment biased selection in the AWI-Gen cohort.

A) Circular pileup plot of all detected gain events detected in participants of each region of recruitment. Chromosomes with coordinates are numbered above accompanied by ideograms. All WA gain events are in track 1, all SA gain events are in track 2, and all EA gain events are in track 3. **B)** Circular pileup plot of gain event with evidence for differential distributions by region of recruitment. Chromosomes and coordinates are numbered above. Each stacked red line represents a discrete gain from start to end in track 1. Track 2 represents the same gains as in track 1 coloured by the region of recruitment of the participant in which it was detected. Candidate diver genes encompassed by the region are annotated on a chromosomal ideogram in track 3.

Several large gains were detected in SA participants on the p-arm of chromosome 5:1-47000000bp (Figure 3.19). These gains were formed through a different mechanism to the whole chromosome aneuploidies. When excluding the aneuploidies encompassing the same region, 8 of 10 were detected in SA participants. This resulted in greater odds of chromosome 5:5:1-47000000bp gains in SA vs WA participants (OR = 6.07 [0.81-270.20], p = 0.09) that trended towards significance. No significant differences in distributions were observed in SA vs EA participants. When considering the total coverage of gains in the chromosome 5:1-47000000bp region, WA participants had significantly greater odds than both SA (OR = 9.49 [4.39-23.50], p = 2.00 X 10^{-12}), and EA participants (OR = 12.07 [3.10-103.77], p=1.63 x 10^{-06}).

Overall, chromosome 5 gains were significantly enriched for formation and selection in WA participants. Interestingly, CN-LOHs of the q-arm of chromosome 5:121000001-180000000bp also showed significant enrichment in WA vs SA participants (OR = 5.48 [1.46-30.47], p = 0.01) (Table 3.14) (Figure 3.20). Taken together, these results suggested a WA population specific haplotype in the q-arm of chromosome 5 which may be positive selected for when activated or duplicated. The specificity of that haplotype remains to be identified in future analysis. Several candidate driver genes were identified in the region which, when duplicated, may confer selective advantage (Figure 3.19B).

Gains on chromosome 21 showed very strong evidence for regional specific differences in distribution. All but 1 gain extended to the terminal end of the q-arm, while only 6 did not extend to the centromere. The remaining 57 gains of the whole q-arm region were assumed to be an aneuploidy of the whole chromosome 21 region; however, this could not be verified as the p-arm is not covered by the array (Figure 3.19). Chromosome 21 gains were detected in participants from all three regions but were significantly less common among WA participants (Table 3.14) (Figure 3.19). Significantly increased odds of occurrence in SA vs WA participants (OR = 5.89 [2.27-19.39], p =1.72 x10^{-05}) and EA vs WA participants (OR = 11.93 [4.29-41.14], p = 1.52 x 10^{-08}) were observed. Additionally, EA participants had significantly greater odds of gains than SA participants (OR = 2.02 [1.09-3.66], p = 0.02) when accounting for relative EA vs SA cohort size.

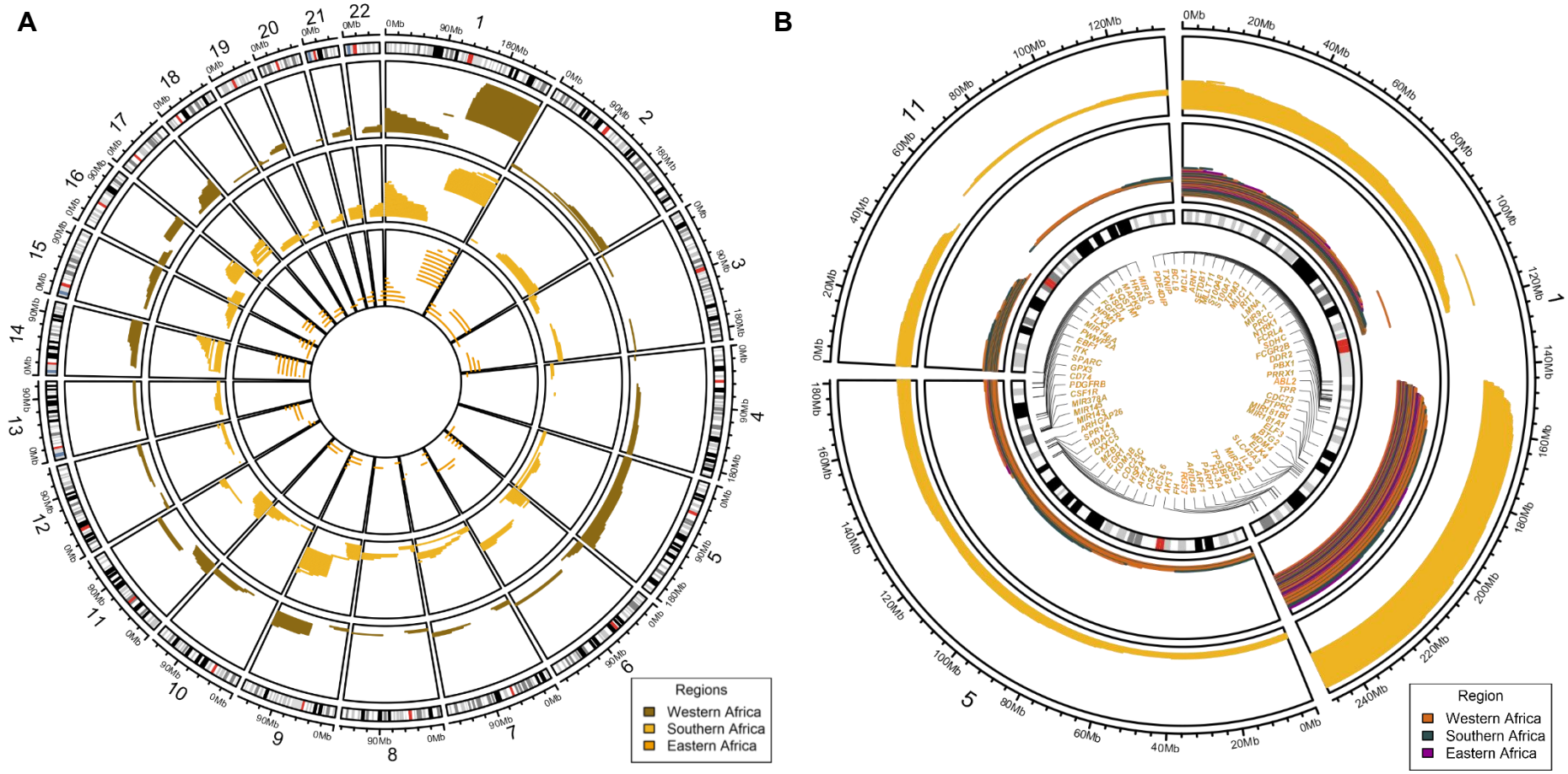


Figure 3.20: Autosomal CN-LOH loci with evidence of region of recruitment biased selection in the AWI-Gen cohort.

A) Circular pileup plot of all detected CN-LOH events detected in participants of each region of recruitment. Chromosomes with coordinates are numbered above, accompanied by ideograms. All WA CN-LOH events are in track 1, all SA CN-LOH events are in track 2, and all EA CN-LOH events are in track 3. **B)** Circular pileup plot of CN-LOH event with evidence for differential distributions by region of recruitment. Chromosomes and coordinates are numbered above. Each stacked gold line represents a discrete CN-LOH from start to end in track 1. Track 2 represents the same CN-LOHs as in track 1 coloured by the region of recruitment of the participant in which it was detected. Candidate diver genes encompassed by the region are annotated on a chromosomal ideogram in track 3.

Chromosome 6 gains extending to the terminal end of the p-arm were identified as another locus showing differential distributions across both sex and region (Table 3.13) (Table 3.14) (Figure 3.19). Participants from WA had significantly greater odds of having gains on 6:1-25000001bp than SA participants (OR = 8.11 [0.98-373.75], p = 0.05). While only 1 of these was detected in EA participants, the EA cohort was comparably smaller. These gains formed more commonly and/or confer stronger selective advantage to WA males than females as well (Tabel 3.13). Several oncogenes were identified from the candidate list which may play a role in this process (Figure 3.19B). Lastly, large CN-LOHs extending from the terminal end of the chromosome 1 q-arm showed considerable evidence of regional differential distribution. While these were detected in participants from all regions, they showed evidence for enrichment in WA vs SA participants with significantly greater odds of detection (OR = 2.39 [1.24-4.71], p = 0.01).

Taken together, these results demonstrate significant variation in MCA distributions at the regional level within the AWI-Gen cohort. The most robust signals were observed for recurrent losses at the *TRA* region on chromosome 14 and for gains on chromosomes 5 and 21 (Table 3.14). Whole-chromosome 5 gains were strongly enriched in WA participants, whereas *TRA* losses and chromosome 21 gains were enriched in SA and EA participants. In addition, several recurrent MCA hotspots appeared to be population-specific, suggesting either underlying germline variation predisposing to their expansion or differential selective pressures acting on the genes and regulatory elements within these regions. While candidate driver genes were identified for each locus, it remains possible that additional, uncharacterized drivers also contribute to these population-specific distributions.

3.3.6 Phase switch errors in autosomal MCAs differed by region of recruitment

Accurate long-range haplotype phasing is essential for the calling of larger MCAs at low cell fractions by MoChA (Loh et al., 2018, Loh et al., 2020). The accuracy of the haplotype phasing achieved in this analysis could not be directly assessed from the SHAPEIT4 output. MoChA was able to detect phase flips or phase switch errors in putative MCA calls which were reflective of phasing accuracy, with more flips in the

underlying call equating to more phasing errors across its length. The size and population makeup of both the cohort and reference panel are important for statistical phasing accuracy (Sengupta et al., 2023a). As the size of the sub-cohorts from each African region were different, and the representation of samples from each region in the 1000 genomes phase 3 reference panel was not equal, it was hypothesized that phasing accuracy in participants from each region may have differed significantly. Thus, the factors impacting phase flip counts of MCAs were investigated as it may have impacted the differential distribution of MCAs observed across the three regions of recruitment. Factors associated with phase flip count distributions were identified through univariate analysis followed by multivariable negative binomial regression analysis (Figure 3.21).

Univariate analysis showed no significant differences in phase flip distributions between male and female participants ($p = 0.88$) (Figure 3.21A). However, significant differences in phase flip counts were observed across MCA types ($p = 4.67 \times 10^{-03}$) and the three regions of recruitment ($p = 4.38 \times 10^{-13}$) (Figure 3.21B-C). On average, losses had significantly lower phase flip counts than both gains and CN-LOHs while no significant difference was observed between gains and CN-LOHs (Figure 3.21B). MCAs in Southern African participants had lower median phase flip counts than those from both Eastern and Western African participants (Figure 3.21C). While not significant, there was evidence that the genotyping Batch may have impacted phase flip count distributions. Additionally, a significant moderate correlation was observed between phase flip counts and the size of the MCA call ($Rho = 0.31$, $p < 2.20 \times 10^{-16}$), but not between any other continuous variables. Thus, MCA size, MCA type, and region of recruitment showed good evidence for association with underlying phase switch error counts. These factors were taken forward to a multivariable model.

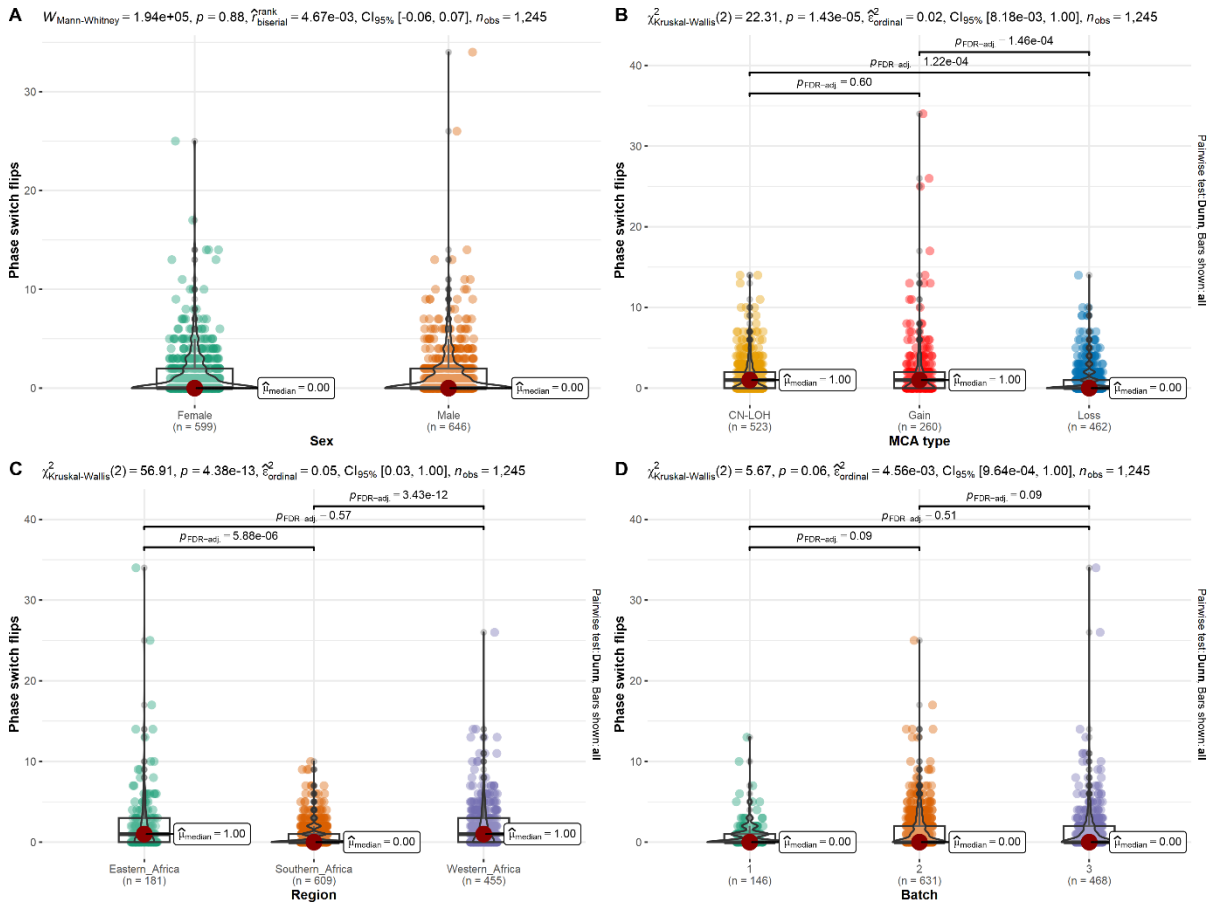


Figure 3.21: Univariate analysis of factors associated with phase flip counts in the AWI-Gen cohort.

A) Mann-Whitney U test statistics showing no significant relationship between sex and phase flip counts. B) Kruskal-Wallis test showing significant association between MCA type and phase flip counts. C) Kruskal-Wallis test showing significant association between region of recruitment and phase flip counts. D) Kruskal-Wallis test showing association between genotyping batch and phase flip counts trending towards significance. Test statistics are shown above violin plots.

The distribution of phase flip counts was non-normal, with substantial overdispersion observed when fitting a Poisson model (Mean phase flips = 1.36, Variance = 6.59, dispersion ratio = 3.09, $p < 0.001$). To account for this, a negative binomial regression model fitted to phase flip counts, using participant region of recruitment, MCA type, MCA size, and batch as independent variables. This model provided a significantly better fit than both the null model ($p < 2.20 \times 10^{-16}$) and a Poisson model with the same predictors (AIC = 3687.5 vs 4559.7, LRT $p < 2.20 \times 10^{-16}$). No evidence of zero inflation was observed, with 673 MCAs having no phase flips compared to 661 expected (ratio = 0.98). Importantly, the regression model was not intended as a definitive predictor of phase flip errors, as several relevant factors were not included in the dataset. Instead, the analysis was performed to test whether variables identified in the univariate

analysis remained significantly associated with phase switch errors after adjusting for one another. A summary of the regression results is provided in Table 3.15.

Table 3.15: Fixed-effects multivariate negative binomial regression of MCA size (Mb), MCA type, region of recruitment, and batch on MCA phase flip counts.

Coefficients	Estimate	Std. Err.	Z-value	P-value
Ref. Eastern Africa, CN-LOH, Batch 1				
Intercept	-0.39	0.18	-2.11	0.04
MCA size (Mb)	0.01	0.00	10.49	$< 2.00 \times 10^{-16}$
MCA type (Loss)	0.34	0.12	2.90	3.69×10^{-03}
MCA type (Gain)	-0.13	0.12	-1.06	0.29
Region (Southern Africa)	-0.78	0.13	-5.89	3.80×10^{-09}
Region (Western Africa)	-0.30	0.13	-2.21	0.03
Batch (Batch 2)	0.59	0.15	3.85	1.18×10^{-04}
Batch (Batch 3)	0.33	0.16	2.05	0.04
Ref. Southern Africa, Loss, Batch 2				
Intercept	-0.23	0.10	-2.42	0.02
MCA size (Mb)	0.01	0.00	10.49	$< 2.0 \times 10^{-16}$
MCA type (CN-LOH)	-0.34	0.12	-2.90	3.69×10^{-03}
MCA type (Gain)	-0.47	0.15	-3.21	1.34×10^{-03}
Region (Eastern Africa)	0.78	0.13	5.89	3.80×10^{-09}
Region (Western Africa)	0.48	0.10	4.73	2.29×10^{-06}
Batch (Batch 1)	-0.59	0.15	-3.85	1.18×10^{-04}
Batch (Batch 3)	-0.27	0.10	-2.69	0.01

Ref. = reference variable category; Std. Err. = standard error of estimate coefficient. Mb = mega bases

The size of MCAs, their type, participant region of recruitment, and the genotyping Batch were all significantly associated with the number of phase flips/switch errors in the negative binomial regression model (Table 3.15) (Figure 3.22). Increasing MCA size remained significantly associated with slightly increased phase flip counts after accounting for the marginal effects of MCA type, region of recruitment, and Batch (Table 3.15) (Figure 3.22). For every increase in 1 Mb the expected count of phase flips increased by 1.01 times (Incidence Rate Ratio (IRR) = 1.01 [1.01-1.01], $p < 2.0 \times 10^{-16}$). The impact of MCA type on phase flip counts remained significant, but the direction of the impact of Losses versus Gains and CN-LOH changed when accounting for size, region, and Batch (Tabe 3.15). The expected number of phase flips for losses was 1.40 times greater than CN-LOHs (IRR = 1.40 [1.10-1.78], $p = 3.69 \times 10^{-03}$) and 1.59 times greater than for gains (IRR = 1.59 [1.18 – 2.15], $p = 1.34 \times 10^{-03}$). There was no significant difference in the expected phase flip counts for gains versus CN-LOH when adjusted for size, region, and batch (Tabe 3.15). Region of recruitment remained significantly associated with phase flip counts after accounting for the marginal effects of the other variables (Tabel 3.15) (Figure 3.22). Participants from

Eastern Africa had 2.17 times the expected counts of phase flips than Southern Africans (IRR = 2.17 [1.67-2.84], p = 3.80 x 10^{-09}) and 1.34 times the expected counts than Western Africans (IRR = 1.34 [1.04-1.75], p = 0.03). Southern Africans had the lowest counts of phase flips with 0.62 times lower expected counts than Western Africans (IRR = 0.62 [0.50-0.76], p = 2.29 x 10^{-06}). Overall, the batch that the participant was genotyped in also was significantly associated with phase flip counts with Batch 2 containing the most phase flips, followed by 3 and then 1 (Table 3.15). (Figure 3.22).

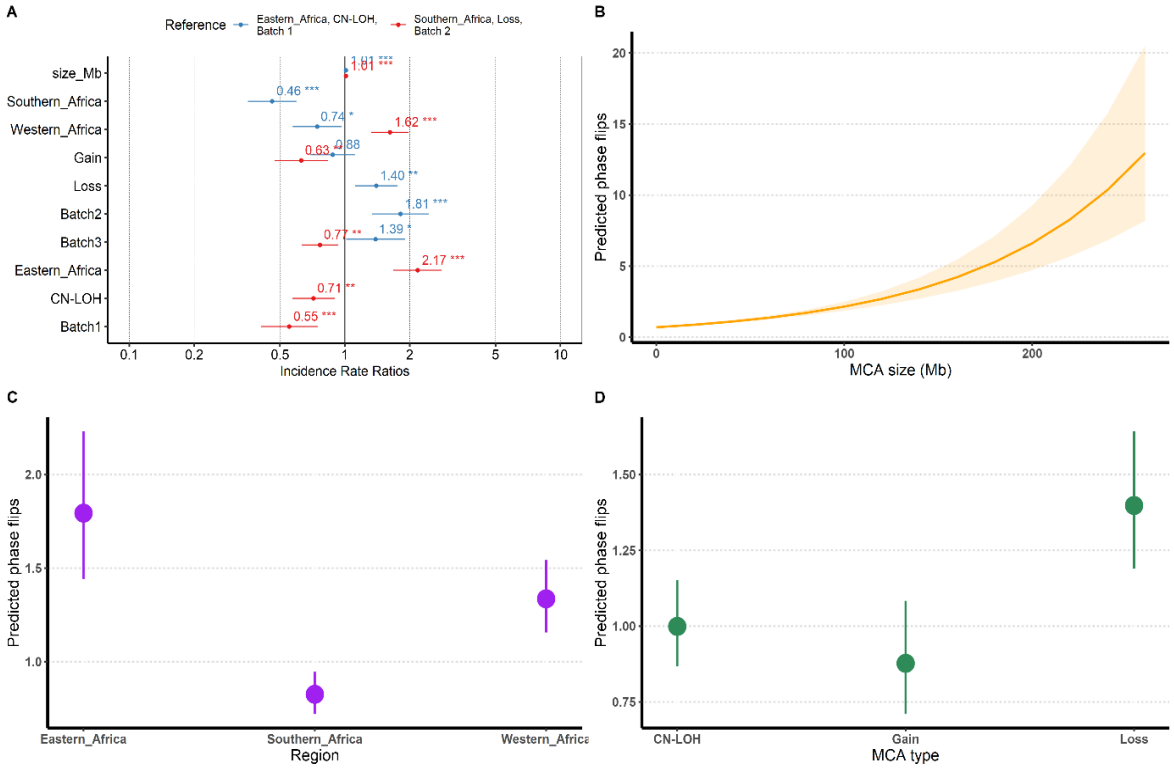


Figure 3.22: The effect of MCA size, type, and region of recruitment on phase flip counts.

A) plot of multivariate negative binomial regression incidence rate ratio estimates for MCA phase flip counts predicted by MCA size, MCA type, batch, and region of recruitment. Two model references variables are show in blue and red. B) the marginal effect of MCA size in mega bases on predicted phase flips after adjusting for type, region, and batch. Shaded areas indicate 95%CI. C) the marginal effect of region on phase flips adjusted for size, type, and batch marginal effects. Dots indicate mean predicted probability. Error bars denote 95% CI. D) the marginal effect of MCA type on predicted phase flips adjusted for size, region, and batch marginal effects. Dots indicate mean predicted probability. Error bars denote 95% CI. P-value symbols are p<0.001 = "***", p<0.01 = "**", p<0.05 = "*"

Although the regression model was not designed to explicitly predict phase flip counts, it confirmed that MCA size and region of recruitment had significant effects. The influence of MCA size on phase switch errors was expected, as larger genomic regions are more challenging to phase accurately. Similarly, the effect of region was consistent with prior knowledge of the cohort composition: Southern African participants, who were most numerous, showed the fewest phase flips, while Eastern African participants, who were least numerous, exhibited the most. This pattern suggests that phasing was most accurate in Southern Africans and least accurate in Eastern Africans. Batch effects could be explained by differences in participant composition, but the higher error counts in losses compared to other MCA types, even after adjusting for size, lacked a clear biological or technical explanation. To explore this anomaly further, phase flips were plotted against MCA size, stratified by type and region (Figure 3.23).

The scatter plots again highlighted the impact of MCA size and region (Figure 3.23). Across most categories, larger MCAs were significantly associated with increased phase flips. The exceptions were losses in Southern and Western Africans, where this relationship was absent. Instead, small losses with unexpectedly high phase switch errors were observed, likely confounding the correlation analysis and explaining why losses showed significant associations in the regression model (Figure 3.22). This clustering of small losses with high error rates pointed to an additional source of phasing inaccuracy. To investigate this further, phase flips were averaged across 1 Mb windows along the autosomes and plotted against the proportion of MCA coverage (Figure 3.24).

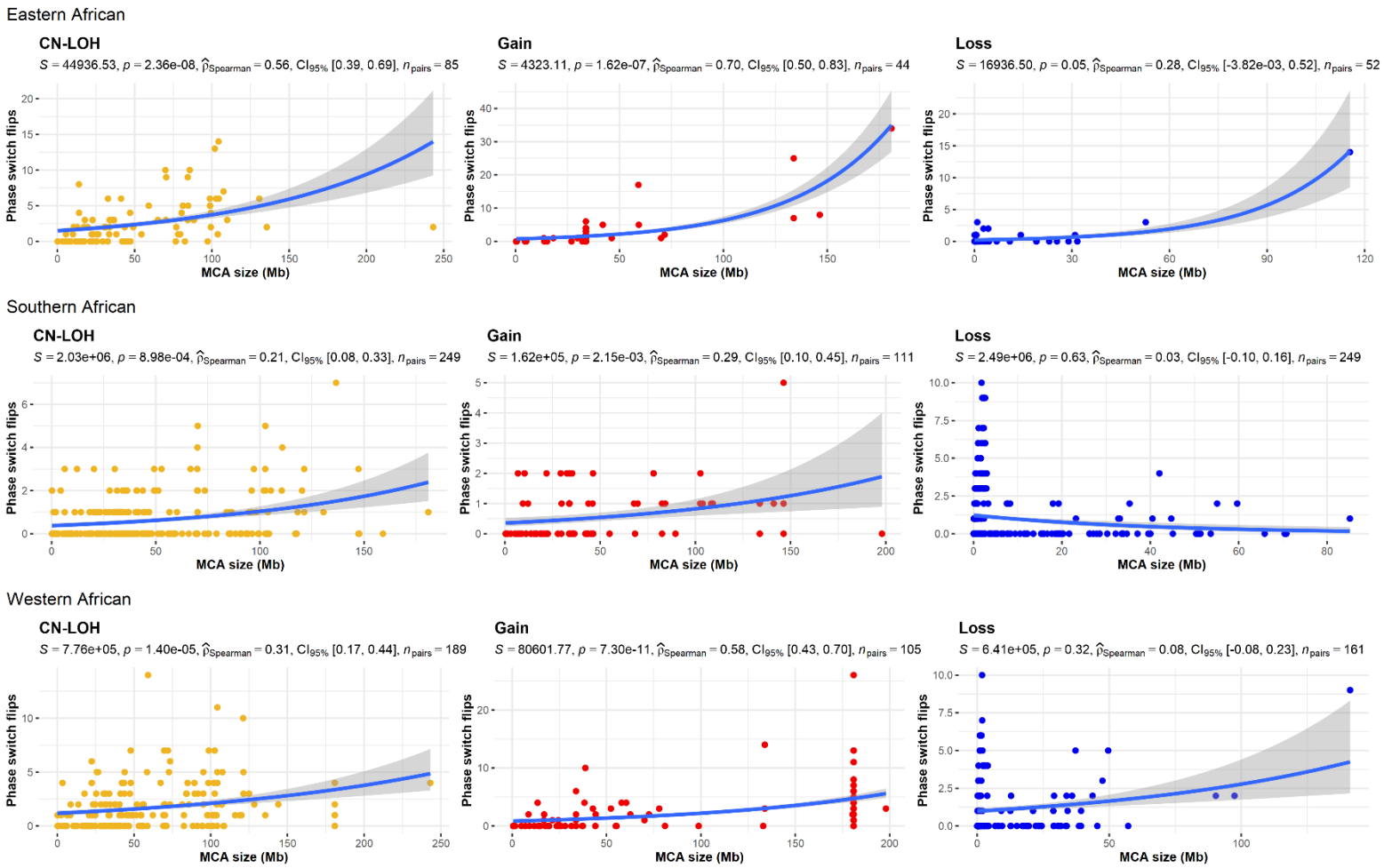


Figure 3.23: Scatter plots of MCA phase flip counts with corresponding MCA size (Mb) stratified by type and region of recruitment.

Spearman's rank correlation stats are above each plot. A Poisson line of best fit is show in blue with 95%CI in shaded area.

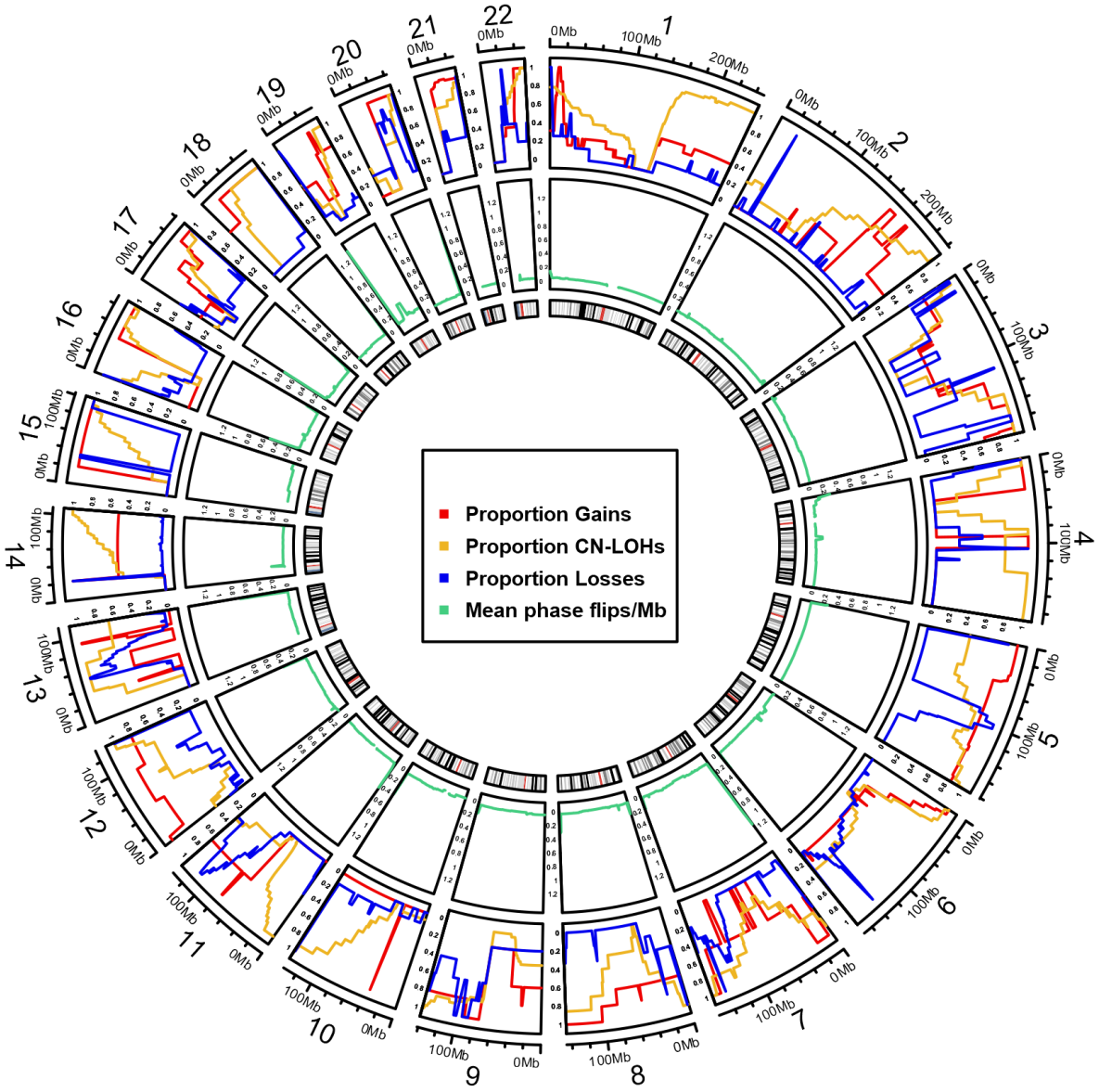


Figure 3.24: Phase switch errors are more common on the terminal ends of autosomes.

The mean phase switch flips in each 1Mb window of the autosome were calculated compared to the proportion of autosomal MCA coverage by MCA type. Proportions are scaled to the total MCA count per chromosome.

Analysis of mean phase switch errors per 1 Mb window, averaged across the cohort and across all MCA types, revealed elevated error rates at the terminal ends of most chromosome arms (Figure 3.24). Peaks in mean phase flips coincided with regions of high loss frequency, suggesting that recurrent terminal focal deletions were particularly prone to phasing inaccuracies. These spikes also overlapped with regions of reduced heterozygous probe density at chromosome ends (Figure 3.13), indicating that poor probe coverage contributed to reduced phasing accuracy. Consequently, the small

losses with high phase flip counts identified in the correlation analysis (Figure 3.23) largely reflected terminal deletions near telomeres, where phasing was least reliable. These events inflated the apparent effect of losses in the regression model (Table 3.15) (Figure 3.22).

When accounting for these terminal outliers, the multivariate regression indicated that MCAs in Eastern African participants remained most strongly associated with elevated phase flip counts. This highlights that reduced phasing accuracy at chromosome ends, where probe coverage is sparse and focal deletions are common, likely contributed to both the observed excess of errors and the regional differences in MCA detection and autosomal distribution. Given that MoChA relies on accurate long-range phasing, these technical limitations may have influenced the regional variation in MCA calls across the cohort.

3.3.7 Comparison of detected MCAs between cohorts from Africa, Europe, and Asia

Autosomal MCAs have previously been characterized in two large biobank-scale studies: the UK Biobank (UKB) and Biobank Japan (BBJ) (Loh et al., 2018, Terao et al., 2020, Loh et al., 2020). These landmark studies employed earlier versions of what would become MoChA, applied to much larger cohorts, with lower-density SNP arrays but highly accurate long-range phasing. Consequently, direct comparisons with the AWI-Gen dataset may not fully capture true similarities or differences in MCA landscapes without additional efforts to harmonize datasets and account for technical variation. Nonetheless, biological factors, environmental exposures, and pathogen burdens, all known to influence MCA formation, are expected to differ substantially between European, Asian, and African cohorts. For this reason, exploratory comparisons of basic MCA characteristics and autosomal distributions were undertaken both to evaluate the impact of technical differences and to probe for evidence of population-specific MCA patterns. UKB and BBJ MCA data were drawn directly from the extended datasets accompanying each publication and were used without modification (Loh et al., 2020, Terao et al., 2020).

3.3.7.1 Autosomal MCA characteristics differed between the AWI-Gen, UKB, and BBJ cohorts.

Autosomal MCAs were detected in 10.06% of the 9917 participants in the AWI-Gen cohort, in 3.54% of the UKB cohort and in 15.56% of the 179,417 BBJ cohort (Table 3.16). The distribution of MCAs across the autosome differed significantly between the cohorts with more losses being detected in the AWI-Gen cohort when compared to the BBJ (37.12% vs 26.55%) and UKB cohorts (37.12% vs 26.01%) ($p=1.48 \times 10^{-38}$) (Table 3.16). Fewer CN-LOH events were detected in the AWI-Gen cohort than the BBJ cohort (42.01% vs 52.93%) and the UKB cohort (42.01% vs 57.27%) (Table 3.16).

Table 3.16: Autosomal MCA characteristics summary across the AWI-Gen, U.K Biobank, and Biobank Japan cohorts.

Study	AWI-Gen	U.K Biobank	Biobank Japan
MCA type	**N (%)**	**N (%)**	**N (%)**
Gain	260 (20.88)	2389 (16.72)	4043 (20.52)
Loss	462 (37.11)	3718 (26.01)	5233 (26.55)
CN-LOH	523 (42.01)	8185 (52.27)	10431 (52.93)
Total	1245	14292	19707
Median MCA length (Mb) (Range)	27.80 (0.02 - 243.19)	35.16 (0.05-248.47)	34.90 (0.02-248.47)
MCA length group (Mb)	**N (%)**	**N (%)**	**N (%)**
0 - 0.5	148 (11.89)	211 (1.48)	556 (2.82)
0.5 - 5	237 (19.04)	1810 (12.66)	1541 (7.82)
5 - 10	55 (4.42)	654 (4.58)	934 (4.74)
>10	805 (64.66)	11617 (81.28)	16676 (84.62)
Median MCA CF (%) (Range)	2.72 (0.40- 48.02)	4.80 (0.60-103)	1.90 (0.20-93.2)
MCA CF group (%)	**N (%)**	**N (%)**	**N (%)**
0 - 1	212 (17.02)	224 (1.57)	5795 (29.41)
1 - 10	796 (63.94)	9713 (67.96)	11263 (57.15)
>10	237 (19.04)	4355 (30.47)	2649 (13.44)

CF = cell fraction. N = count.

The size distribution of MCAs in each study differed significantly ($p=3.66 \times 10^{-34}$) (Table 3.16). The median length of detected MCAs in the AWI-Gen cohort (27.80Mb) was significantly smaller than those in UKB (35.16Mb, $p=1.23 \times 10^{-28}$) or BBJ (34.90Mb, $p= 1.10 \times 10^{-34}$). A larger proportion of MCAs <0.5Mb, were detected in the AWI-Gen cohort (11.89%) than in the UKB (1.48%) or BBJ (2.82%) cohorts (Table 3.16).

Conversely, fewer MCAs in AWI-Gen (64.66%) were large chromosome spanning events compared to those om UKB (81.28%) or BBJ (84.62) (Table 3.16). The lower limit of detection was similar in AWI-Gen and BBJ (0.02Mb), whereas no MCAs smaller than 0.05Mb were detected in the UKB cohort (Table 3.16). The difference in median MCA size distribution between cohorts was mirrored across MCA types. The median length of CN-LOHs detected in AWI-Gen (42.14Mb) was significantly smaller than those detected in BBJ (53.20Mb, p = 0.01), but not UKB (43.62Mb, p = 0.51). The median length of gains in AWI-Gen (33.84Mb) was significantly smaller than those in UKB (79.36Mb, $p = 3.51 \times 10^{-46}$), but not BBJ (58.10Mb, p = 0.18). Interestingly, the median length of losses in AWI-Gen was significantly smaller than losses in both UKB (13.32Mb, $p=2.44 \times 10^{-48}$) and BBJ (13.30Mb, $p=6.43 \times 10^{-49}$).

The distribution of cell fraction estimations differed significantly between studies as well ($p < 0.0001$). The median CF of MCAs in the AWI-Gen cohort (2.72%) was significantly higher than the BBJ cohort (1.9%, $p =1.11 \times 10^{-21}$) (Table 3.16). A larger proportion of MCAs <1% were detected in the BBJ cohort (29.41%) when compared to the AWI-Gen cohort (17.0%), whereas a smaller proportion of MCAs were expanded above 10% in the BBJ cohort (13.44%) than the AWI-Gen cohort (19.1%) (Table 3.16). The highest supporting CF detected in AWI-Gen was 48% versus 93% in BBJ (Table 3.16). The median CF of MCAs was significantly lower in AWI-Gen (2.72%) compared to UKB (4.80%, $p=8.85 \times 10^{-64}$). A larger proportion of MCAs<1% were detected in AWI-Gen (17.0%) than in UKB (1.57%), while a smaller proportion were expanded above 10% in AWI-Gen (19.1%) compared to UKB (30.47%) (Table 3.16). Overall, MCAs differed significantly in the rates of MCA type, size, and CF distributions between cohorts of African, European, and Asian descent. These differences were driven by commonly occurring MCAs across the autosome which were population specific.

3.3.7.2 Autosomal MCA frequencies differed between the AWI-Gen, UKB, and BBJ cohorts.

A closer examination of the frequency of detected MCAs across the autosomes of the AWI-Gen, UKB and BBJ cohorts revealed that specific MCAs were more common in continental Africans, British Europeans, or Japanese Asians. Autosomal MCAs were stratified by chromosomal location and MCA type. The correlation of MCA frequencies across each chromosome were compared between studies (Figure 3.25). The CN-LOH frequencies across autosomes were positively and significantly correlated between the AWI-Gen cohort and the UKB cohort ($R=0.81$, $p = 4.40 \times 10^{-06}$), as well between the AWI-Gen cohort and BBJ cohorts ($R = 0.58$, $p = 0.01$). This indicated that most autosomal CN-LOH frequencies were similar. There were no significant correlations between the frequencies of gains between the AWI-Gen and UKB cohorts, suggesting that most gain frequencies were different between cohorts.

The frequencies of gains between AWI-Gen and BBJ were significantly correlated across autosomes ($R= 0.46$, $p = 0.03$). There were no significant correlations of loss frequencies across the autosome between either the UKB or BBJ cohorts. This suggested that loss frequencies were mostly different between AWI-Gen and either UKB or BBJ. Autosomes on which the frequencies differed most can be seen as deviations from the center line in Figure 3.25. The presence of outliers suggested that a number of autosomal MCA frequencies differed significantly across the cohorts. This was examined further with individual Fishers exact tests for each autosomal MCA frequency pair to identify significantly different frequencies (Table 3.17). The relative frequencies of each autosomal MCA phenotype comparison can also be seen in Figure 3.26.

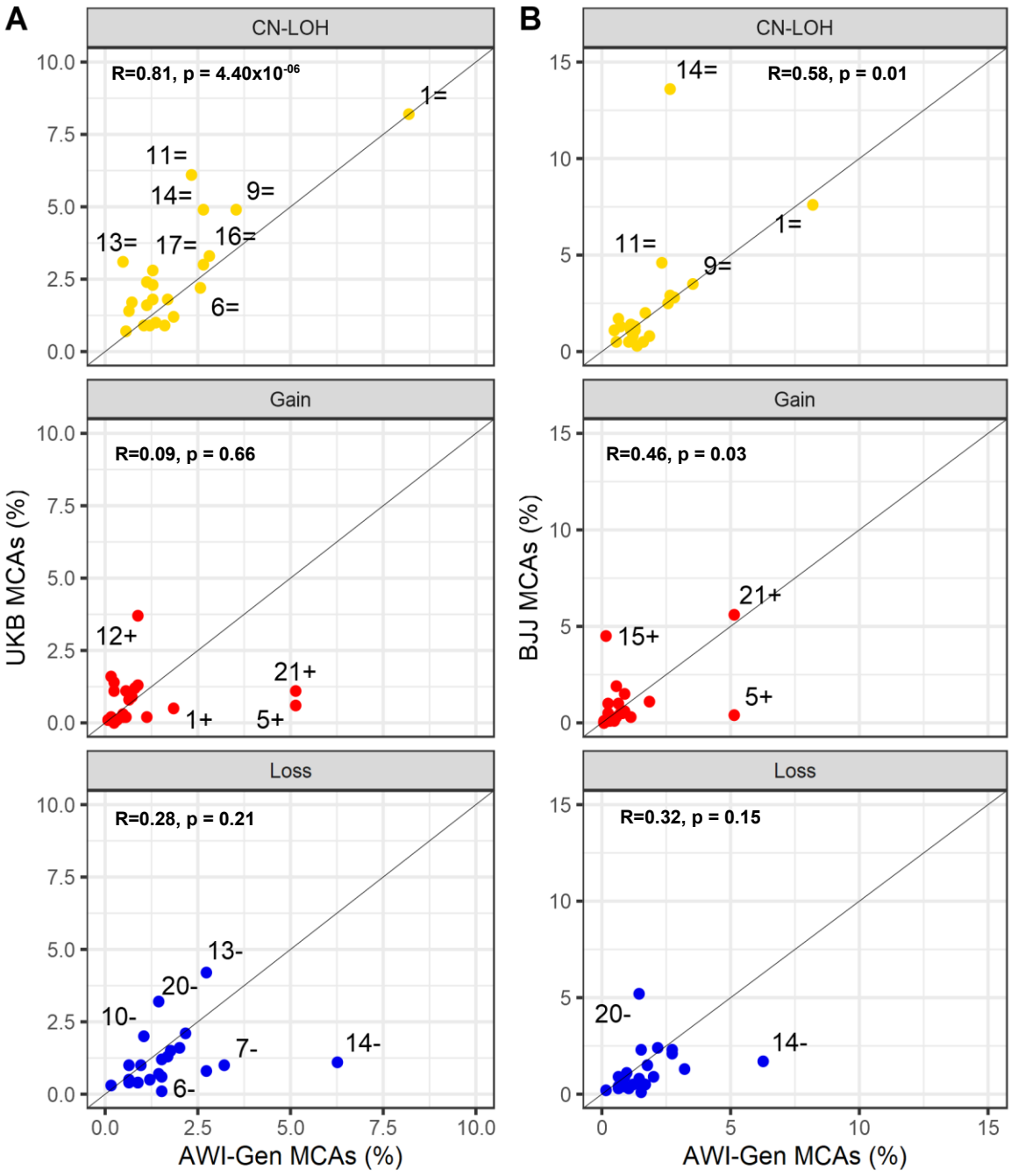


Figure 3.25: Autosomal MCA frequency correlations between AWI-Gen, UKB, and BBJ cohorts.

The frequencies of autosomal MCAs as a function of the total MCA events detected in each study were compared between the studies stratified by MCA type. **A)** AWI-Gen versus UKB. **B)** AWI-Gen versus BBJ. Individual points represent MCA frequencies for single chromosomes. Pearson's correlation statistics for each comparison by type and region are shown in each plot. AWI-Gen frequencies are on the x-axis, UKB and BBJ are on the y-axis. q = q-arm, p = p-arm; - represents a loss, = represents a CN-LOH, + represents a gain. UKB = U.K Biobank, BJJ = Biobank Japan.

Table 3.17: Autosomal MCA counts that were significantly different between the AWI-Gen and UKB or BBJ cohorts.

MCA	*AWI vs UKB*		*AWI vs BBJ*	
	OR [95% CI]	**P-value**	**OR [95% CI]**	**P-value**
Gain				
1+	3.31 [1.93-5.50]	**4.18×10^{-05}**	1.68 [1.02-2.65]	0.07
5+	8.52 [5.88-12.31]	**1.61×10^{-26}**	15.00 [10.35-21.69]	**2.34×10^{-39}**
6+	4.06 [1.98-7.92]	**2.61×10^{-04}**	3.59 [1.83-6.60]	**8.98×10^{-04}**
8+	0.40 [0.15-0.85]	**0.03**	0.27 [0.11-0.58]	**7.00×10^{-04}**
12+	0.16 [0.08-0.29]	**7.38×10^{-14}**	1.42 [0.68-2.68]	0.36
14+	0.16 [0.03-0.48]	**1.96×10^{-04}**	0.43 [0.09-1.29]	0.27
15+	0.08 [0.01-0.28]	**8.50×10^{-08}**	0.03 [0.00-0.10]	**4.25×10^{-23}**
18+	0.13 [0.03-0.38]	**7.55×10^{-06}**	0.23 [0.05-0.69]	**0.01**
19+	1.15 [0.40-2.73]	0.71	5.94 [1.89-16.15]	**6.62×10^{-03}**
21+	4.77 [3.38-6.68]	**6.30×10^{-17}**	0.88 [0.64-1.18]	0.47
Loss				
5-	0.88 [0.51-1.44]	0.80	0.45 [0.26-0.71]	**1.22×10^{-03}**
6-	2.40 [1.57-3.59]	**2.03×10^{-04}**	0.93 [0.62-1.34]	0.79
7-	1.91 [1.30-2.75]	**2.37×10^{-03}**	1.79 [1.23-2.55]	**5.99×10^{-03}**
10-	0.34 [0.18-0.60]	**1.04×10^{-04}**	2.68 [1.33-5.00]	**9.98×10^{-03}**
12-	1.93 [1.09-3.24]	**0.04**	1.78 [1.03-2.94]	0.07
13-	0.42 [0.28-0.60]	**1.02×10^{-06}**	0.84 [0.57-1.21]	0.49
14-	4.49 [3.31-6.04]	**2.73×10^{-19}**	2.97 [2.24-3.90]	**2.47×10^{-12}**
16-	0.94 [0.56-1.49]	0.91	2.66 [1.56-4.36]	**1.33×10^{-03}**
19-	9.33 [4.55-19.26]	**7.41×10^{-09}**	18.64 [8.53-42.39]	**8.19×10^{-13}**
20-	0.29 [0.17-0.47]	**2.53×10^{-08}**	0.17 [0.10-0.27]	**1.20×10^{-20}**
CN-LOH				
1=	1.44 [1.14-1.81]	**9.80×10^{-03}**	1.45 [1.15-1.82]	**5.26×10^{-03}**
5=	2.44 [1.43-3.97]	**5.54×10^{-03}**	4.03 [2.34-6.61]	**8.63×10^{-06}**
6=	1.65 [1.10-2.41]	**0.05**	1.34 [0.89-1.94]	0.25
7=	2.09 [1.28-3.28]	**9.80×10^{-03}**	3.11 [1.90-4.90]	**5.21×10^{-05}**
8=	1.62 [0.83-2.89]	0.18	2.86 [1.46-5.19]	**5.26×10^{-03}**
11=	0.49 [0.32-0.72]	**8.93×10^{-04}**	0.62 [0.41-0.90]	0.03
13=	0.20 [0.07-0.45]	**2.36×10^{-05}**	0.53 [0.19-1.18]	0.26
14=	0.72 [0.48-1.03]	0.16	0.19 [0.13-0.28]	**2.01×10^{-28}**
21=	1.96 [1.10-3.28]	**0.05**	6.11 [3.31-10.75]	**5.39×10^{-07}**

UKB = U.K Biobank, BBJ = Biobank Japan; - represents a loss, = represents a CN-LOH, + represents a gain; OR = Fischer's Exact odds ratio calculated from MCA counts as proportion of total loss, gain or CN-LOH specific to study (Tabel 3.16); CI = confidence interval; P-value from Fischer's Exact FDR adjusted. UKB and BBJ MCA counts from Loh et al. (2020) and Terao et al. (2020) extended data.

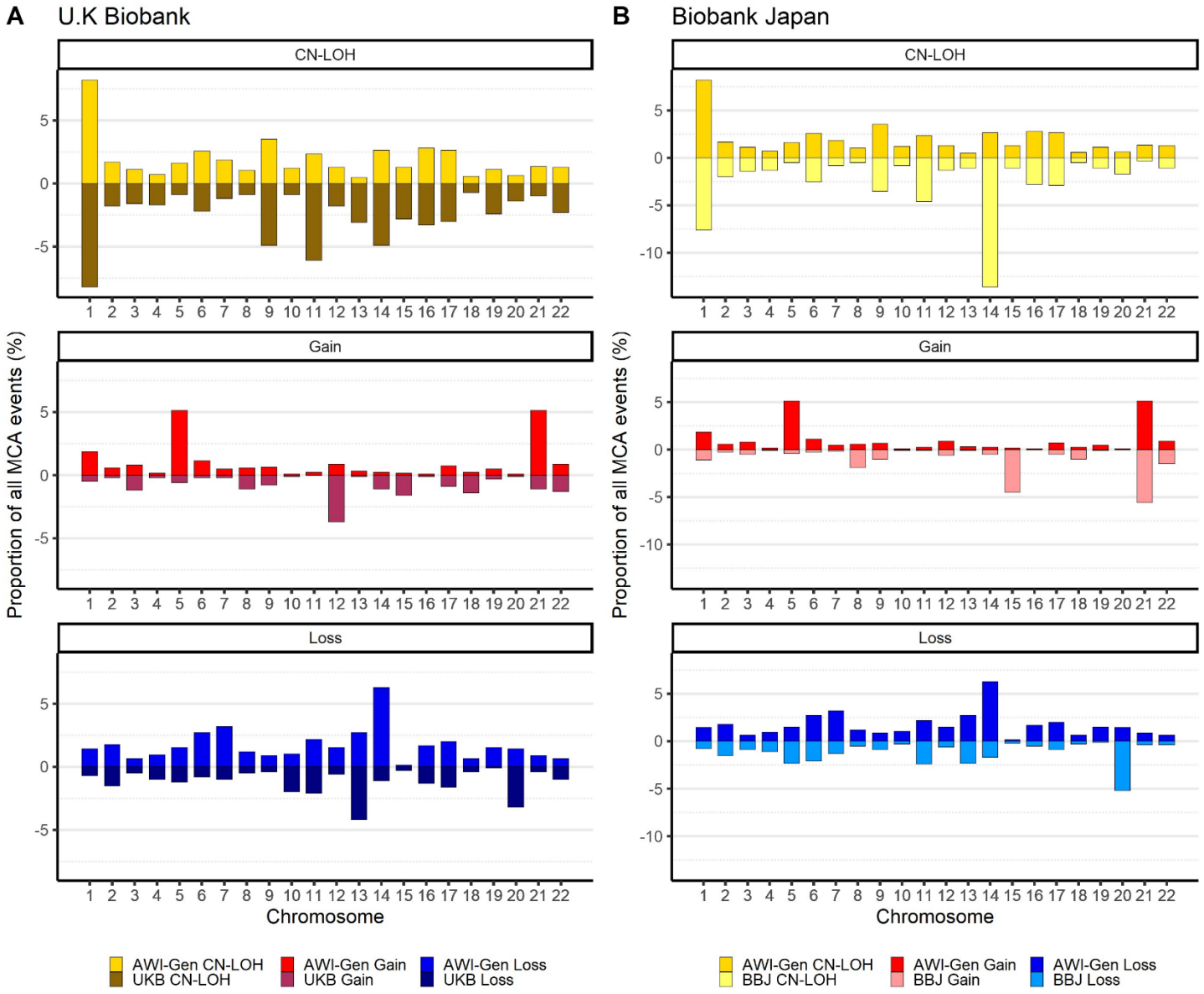


Figure 3.26: Comparison of autosomal MCA frequencies between participants from the AWI-Gen, U.K Biobank and Biobank Japan cohorts.

A) AWI-Gen versus U.K Biobank. **B)** AWI-Gen versus Biobank Japan. UKB = U.K Biobank. BBJ = Biobank Japan. Frequency data for UKB and BBJ frequencies from Loh et al. *(2020)* and Terao et al. *(2020)* extended data.

Several autosomal MCA counts, as a proportion of the total MCAs detected in each cohort, were significantly different across the AWI-Gen, UKB, and BBJ cohorts (Table 3.17) (Figure 3.26). Gain frequencies between AWI-Gen and UKB were significantly different across 9 chromosomes, while 6 gain frequencies were significantly different in AWI-Gen and BJJ (Table 3.17). Participants from AWI-Gen were significantly more likely to have gains on chromosomes 1, 5, 6, and 21 than participants from UKB, and they were significantly less likely to have gains on chromosomes 8, 12, 14,15, and 18 than participants from UKB (Table 3.17) (Figure 3.26). Participants from AWI-Gen were also significantly more likely to have gains on chromosomes 5, 6, and 19 than participants from BBJ, while they were significantly less likely to have gains on chromosomes 8, 15, and 18 than participants from BBJ (Table 3.17) (Figure 3.26).

Loss frequencies were significantly different in 8 chromosomes between AWI-Gen and UKB participants, and loss frequencies were significantly different in 7 chromosomes between AWI-Gen and BBJ participants (Table 3.17). Participants from AWI-Gen were significantly more likely to have losses on chromosomes 6, 7, 12, 14, and 19, and significantly less likely to have losses on chromosomes 10, 13, and 20 than participants from UKB (Table 3.17) (Figure 3.26). Participants from AWI-Gen were significantly more likely to have losses on chromosomes 7, 10, 14, 16, and 19, and significantly less likely to have losses on chromosomes 5 and 20 than participants from BBJ (Table 3.17) (Figure 3.26).

CN-LOH frequencies were significantly different in 7 chromosomes between AWI-Gen and UKB participants, while being significantly different in 7 chromosomes between AWI-Gen and BBJ (Table 3.17). Participants from AWI-Gen were significantly more likely to have CN-LOH on chromosomes 1, 5, 6, 7, and 21, while being significantly less likely to have CN-LOHs on chromosomes 11 and 13 than participants from UKB (Table 3.17) (Figure 3.26). Finally, participants from AWI-Gen were significantly more likely to have CN-LOHs on chromosomes 1, 5, 7, 8, and 21, while being significantly less likely to have CN-LOHs on chromosomes 11 and 14 than participants from BBJ (Table 3.17) (Figure 3.26).

Several autosomal MCAs showed compelling evidence of being African specific based off this exploratory analysis. MCAs that were the most frequent in the AWI-Gen cohort tended to show evidence of population specificity. In particular, the large gains observed in Western African participants were not commonly seen in the UKB or BBJ cohorts. Large gains of chromosome 21 more commonly seen in Eastern and Southern

African participants were also significantly more common than in UKB but not in BJJ. In contrast, gains that were commonly detected in UKB and BJJ such as on chromosomes 8, 15, and 18 were not common in any AWI-Gen participants. Additionally, large gains of chromosome 12 were common in UKB but not in AWI-Gen participants. The focal losses on chromosomes 7 and 19 which were more common in Western and Southern Africans were less common in UKB and BJJ. The small focal losses of the *TRA* region on chromosome 14 which were more likely to be seen in Southern and Eastern Africans were much less likely to be seen in UKB. They were reported to be detected in BJJ but at much lower proportions of total losses than in the AWI-Gen participants. Losses of chromosome 20 which were observed across UKB and BBJ were rare in AWI-Gen. Additionally, losses of chromosome 13 were more common in UKB than either AWI-Gen or BBJ. In AWI-Gen the cases with 13q loss were Western African. Regions of CN-LOHs that showed evidence of African specificity were found on chromosomes 1, 5, 7, and 21.

These may have novel underlying African population specific germline variation in genes that confer selective advantage that remain to be uncovered. These results add to the body of evidence for population specific MCA formation or selection that requires further investigation.

3.3.8 Participants with multiple distinct autosomal MCAs

3.3.8.1 Multiple autosomal MCAs were less common in AWI-Gen

Most of participants presented with a single MCA (83.67%). There were 116 (11.26%) participants with two MCA events, however, some of these might be a single complex structural variation (e.g., a single unbalanced translocation) that appears as two separate events. Therefore, participants were defined as having multiple MCAs if they had three or more separate MCA events. Only 47 (4.71%) participants had three or more unique non-overlapping MCA events in 47 participants (Table 3.18). The highest number of unique autosomal MCAs detected in a single individual was seven, which was seen in two different participants.

Table 3.18: Number of distinct MCAs detected in 998 participants

MCA count	1	2	3	4	5	6	7	Total
Female (%)	402 (84.28)	49 (10.27)	16 (3.35)	5 (1.05)	1 (0.21)	2 (0.42)	2 (0.42)	**477 (100)**
Male (%)	433 (83.11)	67 (12.86)	11 (2.11)	5 (0.96)	3 (0.58)	2 (0.42)	0 (0.00)	**521 (100)**
Total (%)	**835 (83.67)**	**116 (11.62)**	**27 (2.71)**	**10 (1.00)**	**4 (0.40)**	**4 (0.40)**	**2 (0.20)**	**998 (100)**

Univariate analysis (Appendix J) showed no significant associations between participant age (p = 0.63, Figure A.2A), or sex distributions (p = 0.23, FigureA.2C) and having three or more MCAs. There was weak overall association between region of recruitment and the proportion of participants with three or more MCAs (p = 0.04, Figure A.2B). However, only a single participant from Eastern Africa was found with three or more MCAs, likely biasing this analysis. Upon limiting the analysis to only Southern and Western Africans, the association was lost (p = 0.92). This was confirmed in multivariable logistic regression in which no association between age, sex, or region of recruitment was observed when further accounting for Batch (Table A.6, Appendix J). Thus, it does not appear that the cohort sociodemographic makeup had any impact on the odds of multiple distinct MCAs in this cohort. Overall, with only 4.71% of the cohort having three or more distinct MCAs, this was a rare occurrence in the AWI-Gen cohort.

3.3.8.2 Participants with multiple sub-clonal CN-LOHs

A subset of participants had multiple interstitial CN-LOH events on the same chromosome with a stepwise increase in the supporting cell fraction in the direction of the telomere. The MoChA tool identified these events as multiple interstitial CN-LOH events with boundaries directly adjacent to each other. However, each of these CN-LOHs extend from the breakpoint to the telomere and are likely distinct sub-clonal MCAs occurring successively in separate clonal expansions over time. A search for CN-LOH on the same chromosome arm in the same individual with supporting cell fractions increasing in a stepwise manner (bdev increase of more than 0.005) towards the MCA at the telomere identified six cases such cases (Table 3.19).

Table 3.19: Six participants with evidenc of multiple overlapping sub-clonal CN-LOH expasions

Study ID	Chromosome coordinates	Size (Mb)	$LOD_{BAF\text{-}phase}$	CF (%)
GQL0N	1:149942679-249250621	99.31	24.21	0.6
GQL0N	1:217911017-249250621	31.34	159.65	1.9
QLF0D	2:142359467-243199373	100.84	45.21	0.8
QLF0D	2:220334821-243199373	22.86	97.05	1.88
AB1352	6:79878217-171115067	91.24	180.36	1.92
AB1352	6:134822264-171115067	36.29	333.91	2.92
AB2670	6:0-35906389	35.91	4614.86	12.24
AB2670	6:0-57229990	57.23	1643.57	8.75
RGF0V	9:0-28221934	28.22	2152.32	39.7
RGF0V	9:0-31889350	31.89	932.96	27.95
RGF0V	9:0-35480564	35.48	718.93	22.43
RGF0V	9:0-37317241	37.32	351.92	15.94
RGF0V	9:0-44889451	44.89	37.14	4.69
AB1399	15:36651732-102531392	65.88	109.48	1.92
AB1399	15:68037578-102531392	34.49	423.83	3.26

Coordinates – chromosome (start – end); LOD – log odds of MCA versus germline using phased B allele fraction model; CF – cell fraction as a percentage

The most striking example of this can be seen in participant RGF0V, which shows five distinct sub-clonal CN-LOHs detected in a 46-year-old male from Digkale in Southern Africa (Figure 3.27). A stepwise increase in phased bdev from 0.5 at each breakpoint from the centromere to the telomere of chr9p is a tell-tale pattern of distinct CN-LOH expansions, each supported at different the fraction of cells (Figure 3.27). This points to a particularly impactful 9p allele or haplotype, within the smallest region overlapped by all five CN-LOHs, that has been made homozygous five separate times. An intriguing explanation may be somatic genetic rescue, whereby a deleterious germline heterozygous allele is nullified by the CN-LOH, leading improved selective advantage over the wild type. Thus, the CN-LOH is more likely to occur repeatedly in multiple HSC clones. The reverse can also be true. The smaller and more focused the CN-LOH event, the greater the selective advantage, which is seen as an increase in cell fraction towards the likely location of the germline mutation.

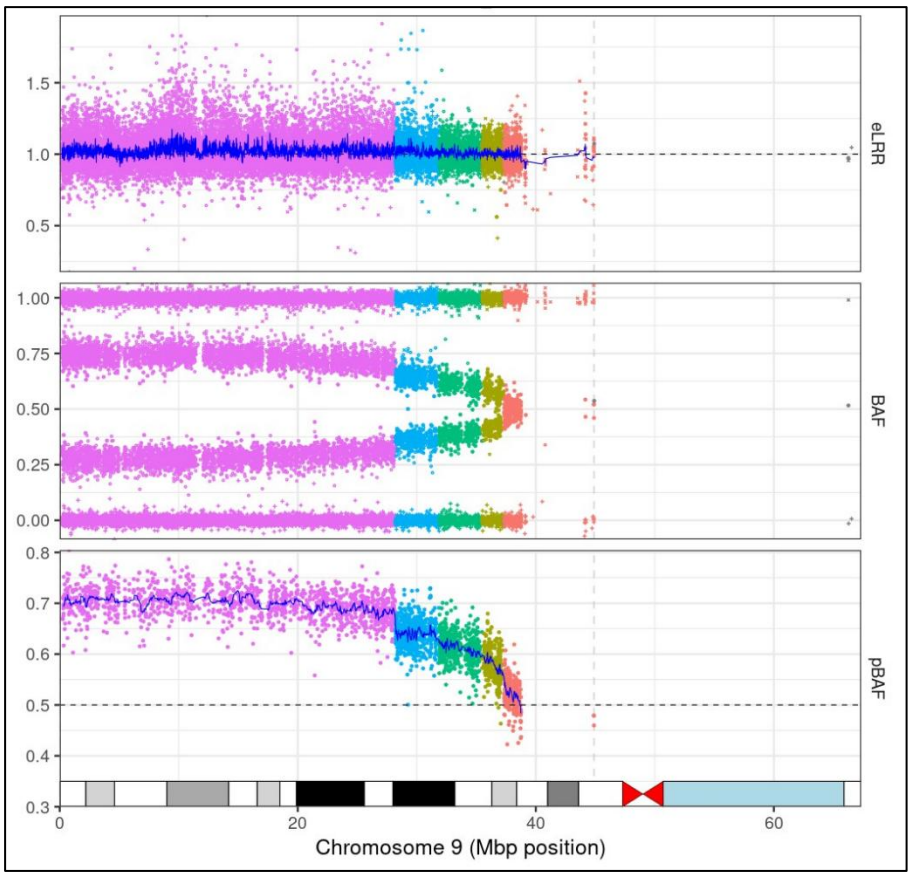


Figure 3.27: Five distinct but overlapping sub-clonal CN-LOH expasions in a 46-year-old male from Dikgale.

Plots of LRR, BAF, and phased BAF deviations for across chromosome 9. Blue lines are median LRR and phased BAF deviations for each CN-LOH. Increasing phased BAF indicate tell-tale pattern of sub-clonal events. Each sub-clonal CN-LOH (colors) extends from furthermost coordinate breakpoint to the telomere. All five overlap from 9:0-28,221,934.

3.3.8.3 Co-occurring MCAs in the AWI-Gen cohort

Autosomal MCAs were grouped into phenotypes by chromosomal arm and MCA type as done previously. Each participant with two or more MCA calls was assessed for their combination of MCA phenotype groups. This resulted in a co-occurrence MCA phenotype made up of two distinct MCAs, which was tallied and assessed for evidence of recurrence in the AWI-Gen cohort. A total of 313 combinations of MCA phenotypes were found in the 409 participants with two or more MCAs. Of these 313 combinations, only four were found in three or more participants, which was the threshold for recurrence. These combinations were investigated further to determine if they were separate sub-clonal events with potential synergistic selection advantage, singular complex MCA events, or random occurrences based on relative MCA frequency.

The four cooccurring MCAs were: 14q loss and gains in 21 (14q-21+), 5p loss and 11p loss (5p11p), 6q loss and 13 q loss (6q-13q-), and 16p loss and 19p loss (16p-19p-). Of these, 14q-21+ was seen four times in the AWI-Gen cohort, while the other three were seen three times each. There were no co-occurring CN-LOH events and only one observed gain event co-occurrence. Out of the co-occurring autosomal MCAs seen in the AWI-Gen cohort, only the 14q-21+ combination was reported as an enriched combination in the UKB and BJJ cohorts(Loh et al., 2018, Terao et al., 2020). The remaining three combinations appeared to be specific to the AWI-Gen cohort suggesting potential population specific mechanisms underlying their selection. Additional analyses of their properties, distributions, and location across autosome carried out. To determine if these were contained within the same clonal expansions, the estimated fraction of cells containing each MCA was plotted for comparison (Figure 3.28).

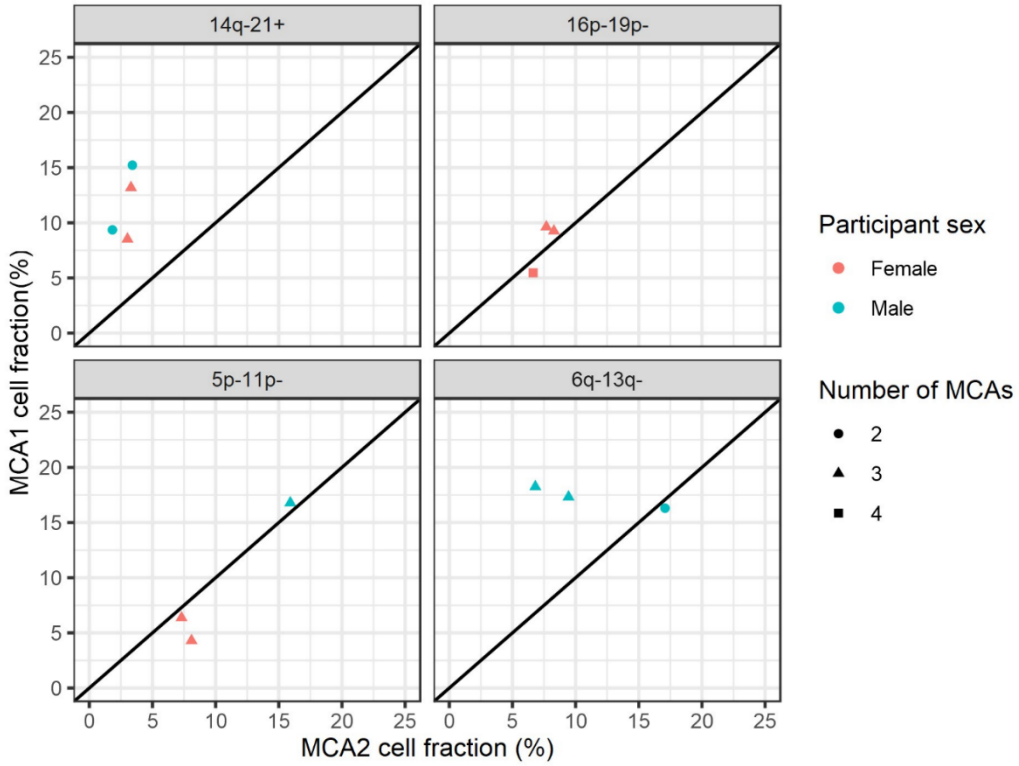


Figure 3.28: Estimated fraction of cells containing co-occurring autosomal MCAs

Each point is from a unique participant. MCA1 denotes the first MCA in the reported pair by number (e.g., 14q- in 14q-21+). MCA2 denotes the second MCA in the reported pair (e.g., 21+ in 14q-21+). Point shape represents the number of unique autosomal MCAs detected in each participant. - represents a loss, = represents a CN-LOH, + represents a gain

Interestingly, the 16p-19- combination was only observed in female participants, while the 6q-13q- combination was only observed in male participants, hinting at a possible sex bias in the mechanism of formation. However, as these combinations were found at low counts, caution is needed in their interpretation. Cell fractions of co-occurring MCAs are expected to be similar in cases where the two MCAs are present in a single expanded HSC clone. Evidence for the co-occurrence in single clonal expansion events was seen for 5p-11p- and the 16p-19p-, with cell fractions in unique participants being closely correlated (Figure 3.28). These combinations may have occurred at the same time, or they may have had similar effects on selection in the cellular niche or may be a single complex MCA event. This suggested a mechanism by which the two MCAs in the pair occur more frequently together in the same cell, possibly sequentially, leading to a synergistic effect on cell survival and proliferation. Conversely, the individual cell fractions of the combinations of 14q-21+ and 6q-13q- differed in each participant, which suggested that these were in distinct clonal expansions, and possibly random occurrences. Evidence for enrichment was determined through statistical testing of cooccurring MCA frequencies (Table 3.20).

Table 3.20: Pairwise enrichemnt of cooccuring MCAs in AWI-Gen

MCA_A	MCA_B	$N_{A\&B}$	N_A	N_B	OR	95%CI	P-value
14q-	**21+**	4	73	60	0.79	0.20-2.21	0.81
5p-	**11p-**	3	6	6	78.60	10.38-495.99	**$4.16x10^{-05}$**
6q-	**13q-**	3	29	31	3.11	0.58-10.92	0.09
16p-	**19p-**	3	11	12	21.78	3.47-98.50	**$9.07x10^{-04}$**

MAC_A – first MCA phenotype in the pair; MCA_b – second MCA phenotype in the pair; OR = odds ratio of Bs occurrence when A is present; CI = confidence interval; - is a loss; + is a gain

Pairwise enrichment analyses were performed to assess whether the presence of one MCA phenotype was significantly enriched in individuals carrying another. These results complemented the pairwise cell fraction analyses (Figure 3.28). Despite 14q− and 21+ being among the most frequently observed MCAs in the cohort, their joint occurrence in only four individuals was consistent with chance, providing no evidence of a causal or preferential association (OR = 0.79, p = 0.81). The 6q−13q− combination demonstrated a non-significant trend toward co-occurrence (OR = 3.11, p = 0.09).

However, the lack of correlation in cell fraction patterns between these two phenotypes supported the interpretation that they do not co-occur more often than expected by chance (Figure 3.28). In contrast, both 16p-19p- (OR = 21.78, p = $9.07x10^{-04}$) and 5p-11p- (OR = 78.6, p = $4.16x10^{-05}$) combinations showed strong and statistically significant co-occurrence, suggesting that these pairs preferentially arise together. This pattern may reflect biological co-selection or represent two signals of a single complex MCA involving concurrent deletions at distinct autosomal loci. To explore this further, the affected regions for each of the four recurrent co-occurring pairs were visualised across participants to assess whether the same loci were consistently impacted (Figure 3.29).

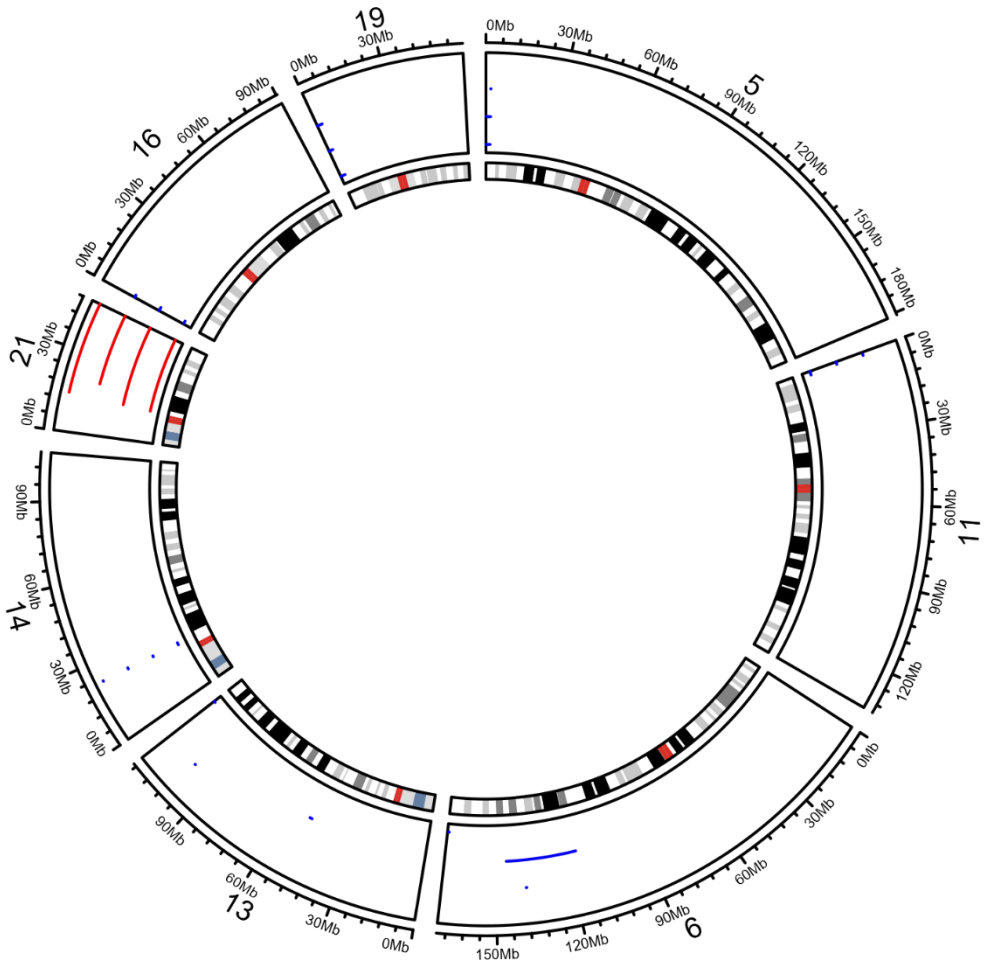


Figure 3.29: Circos plot of coocuring MCAs in the AWI-Gen cohort

Chromosomal regions impacted by putative cooccurring MCA are plotted by lines. Each line is a distinct MCA event. Cooccurring MCAs are ordered from top to bottom by the participant in which they are shared. E.g. the top deletion in chr5 and the top deletion in chr11 are shared by the same participant and for a cooccurring pair. Blue lines are deletions. Red are gains.

Plotting the genomic regions impacted by co-occurring MCA pairs provided a final layer of evidence to distinguish those with significant enrichment from those occurring by chance. Although 14q− and 21+ MCAs affected overlapping regions, both represent the most frequently observed MCAs in the cohort, and their joint occurrence is most plausibly attributable to chance. This interpretation was reinforced by the disparate cell fractions observed across all four carriers, suggesting that the two events arose independently. Similarly, the 6q−13q− MCA pair showed no evidence of non-random co-occurrence, and the locus-level visualisation revealed heterogeneity in the position, size, and cell fractions of the respective deletions, indicating that these represent distinct MCA events despite being grouped under broader phenotype categories (Figure 2.29). These findings illustrate the importance of fine-grained locus-level analyses, as phenotypically similar MCAs can differ markedly at the genomic level. In contrast, the enrichment of both 16p−19p− and 5p−11p− combinations was corroborated by the Circos plot, which demonstrated consistent terminal p-arm deletions across independent individuals (Figure 2.29). The similarity in size and cell fraction of these focal deletions supports their origin within single clonal expansions, providing strong evidence of non-random co-occurrence. Together, these results suggest that these pairs may either represent synergistic selective interactions or reflect single complex MCA events resulting in concurrent terminal p-arm loss which appear to be specific to the AWI-Gen cohort.

3.4 Discussion

In this chapter, an exploratory characterisation of autosomal MCAs in Africans from the AWI-Gen study was undertaken to address the marked absence of data from continental African cohorts in the literature. This objective was achieved through systematic characterisation of MCAs with methodology tailored to the AWI-Gen data. The exploratory nature of the investigation reflects the absence of comparable studies. To date, only a single other study has characterized MCAs in continental African cohorts using appropriate tools (Zhou et al., 2023). However, the results of this study come from a much larger and diverse continental Africa cohort. There are several reasons for this scarcity, most notably the limited availability of accessible African SNP-array or WGS datasets required for large-scale MCA characterisation. This mirrors a broader pattern across human genomics, where African cohorts remain severely underrepresented, including in population-scale GWAS (Rotimi et al., 2017).

The past decade has seen an explosion of somatic variation studies across the spectrum of fields, driven by technological breakthroughs and the establishment of biobank cohorts. Landmark studies in cohorts such as the UK Biobank, Biobank Japan, Mass General Brigham Biobank, Vanderbilt DNA Databank, the Million Veteran Programme, FinnGen, and TOPMed have collectively characterised MCAs in more than one million individuals (Loh et al., 2020, Terao et al., 2020, Uddin et al., 2022, Zekavat et al., 2021, Jakubek et al., 2023b). These efforts have provided the statistical power to uncover key relationships between MCAs, ageing, germline variation, and clonal haematopoiesis, yielding insights into disease risk and biology. However, the overwhelming majority of these samples derive from participants of European ancestry. And the African ancestry data is derived from highly admixed populations with mostly Western African genetic ancestry.

To build a global picture of MCA biology, it is essential to incorporate cohorts with diverse ancestries and geographical backgrounds. Cross-cohort analyses of this kind will clarify which patterns of MCA occurrence and behaviour are consistent across human populations, and which reflect ancestry-specific, environmental, or demographic influences. Identifying universal features of MCA biology strengthens confidence in their interpretation as intrinsic to the mechanisms of CH, while differences may provide unique insights into selective pressures, mutational processes, or health outcomes. This chapter makes a substantive contribution toward redressing this imbalance. It represents the largest and most representative

characterisation of autosomal MCA-driven CH in continental Africans to date. These findings provide a foundation for future comparative analyses across ancestries and highlight the importance of extending MCA characterisation efforts to achieve a truly global understanding of somatic variation and clonal evolution.

3.4.1 Study strengths

This study's strengths stem from three key design elements. First, the AWI-Gen cohort provided a large, diverse sample of continental Africans recruited across six sites spanning Southern, Western, and Eastern Africa. Importantly, recruitment captured not only genetic but also environmental diversity, ranging from semi-urban and urban settings in Southern Africa, to rural agrarian communities in Western Africa, and a rapidly urbanizing informal settlement in Eastern Africa (Ali et al., 2018). The MCA associated exposome, including genotoxic exposure, inflammatory modulating diet, behavioural and cultural practices, and infection burdens, very likely differs significantly between the AWI-Gen regions (Figure 1.10) (Section 1.6.2). Thus, a wider spectrum of genomic variation and environmental exposures than typically seen in North American, European, or Japanese cohorts was included here, allowing for unique insights into how both the genome and the exposome contribute to MCA mutagenesis and selection.

Second, participants were genotyped using the H3Africa SNP-array, specifically designed to capture African-specific variation (Mulder et al., 2018). Its high density and tailored representation improved sensitivity for detecting both smaller MCAs and larger lower cell fraction MCAs, leveraging the high rates of heterozygosity in African genomes (Jakubek et al., 2023b, Choudhury et al., 2020). Studies of Africans with less representative SNP-arrays do not make use of this strength. The trade-off is that lower accuracy phasing, which is inherent to smaller African LD blocks, results in lower power to detect larger and lower fraction MCAs (Jakubek et al., 2023b, Tang et al., 2025). Thus, it is essential to include as many African SNPs as possible to offset this. While high depth (30X or higher) WGS is the best dataset to call MCAs from, it is expensive and thus not a viable genotyping methodology for large African studies at this stage (Tang et al., 2025, Steventon-Jones et al., 2022). The H3Africa-SNP array provides a good balance between cost effectiveness and resolution, allowing for larger AWI-Gen cohort than would be possible WGS (Mulder et al., 2018).

Finally, the powerful algorithm employed by the MoChA tool was used to call smaller MCAs and MCAs at lower cell fractions than previously reported in African ancestry participants (Loh et al., 2018, Loh et al., 2020). Together, these components led to the most comprehensive characterization of MCAs in continental Africans to date. However, due to the exploratory nature of applying the MoChA protocol to an untested dataset, several methodological errors and limitations were encountered in this study. Certainly, the methodology should be refined in future work. Despite these challenges, we were successfully able to characterize autosomal MCAs in 10.06% of the 9,917 participants.

3.4.2 Technical considerations

Previous studies detected MCAs at rates that varied quite markedly, with autosomal MCAs detected in 3.54% of the UKB cohort, 15.56% of the BBJ cohort, 4.56% of the MGBB cohort, 1.80% of the FinnGen cohort, and 4.47% of the TOPMed cohort (Jakubek et al., 2023b, Loh et al., 2020, Terao et al., 2020, Zekavat et al., 2021). In addition to the overall detection rates differing across studies, the actual distributions of specific MCAs across the autosome differed quite markedly in some cases (Terao et al., 2020). In this study, we showed that loss, gain, and CN-LOH frequencies differed significantly between the AWI-Gen, UKB, and BJJ cohorts as well. However, direct comparisons of MCA detection rates and distributions is not recommended. The detection of MCAs from blood derived DNA is explicitly impacted by several factors that are challenging to harmonize both between studies and within cohorts. These can be broken down broadly into technical and biological factors.

These technical factors are intimately linked to the MCA detection algorithm employed by MoChA (Loh et al., 2020). At its core, MoChA calls putative MCAs from informative data points in the form of consecutively phased heterozygous SNPs (Loh et al., 2018, Loh et al., 2020). Fewer data points underlying an MCA call results in less accurate estimations of phased BAF deviations, especially across longer regions, thus resulting in more uncertainty in the call and lower resolution of and its breakpoints (Loh et al., 2018, Loh et al., 2020). This is a large part of the reason the MoChA authors state that it cannot be used to call MCAs from WES data as there are too few informative heterozygous SNP,s that are also non-uniformly distributed, to confidently call MCAs at low cell fractions (Loh et al., 2018, Loh et al., 2020)

(https://github.com/freeseek/mocha). Higher SNP density means smaller MCAs can be detected, and larger MCAs can be detected at very low cell fractions (Loh et al., 2018, Loh et al., 2020). The SNP-array used in the genotyping must be representative of the population genetic diversity to capture more informative data points and higher density arrays should perform better than lower density arrays.

The genotype data in the studies were derived from different SNP arrays with different densities or from WGS (Jakubek et al., 2023b, Loh et al., 2020, Terao et al., 2020, Zekavat et al., 2021). The UKB cohort was genotyped on the Affymetrix UK BiLEVE (~800,000 markers) and UK Biobank Axiom (~825,000 markers) arrays, the BBJ cohort was genotyped on the Illumina Infinium Omni Express (~730,000 markers) and Infinium Omni Express Exome v.1.0 (~950,000 markers) and v.1.2 arrays (~964,000 markers) but only included SNPs that intersected all three (Loh et al., 2020, Terao et al., 2020). The MGBB cohort was genotyped on the Illumina Multi-Ethnic Genotyping Array (MEGA) (~1,730,000 markers), the FinnGen cohort was genotyped on the FinnGen Thermo Fisher Axiom custom array (~650,000 markers), and the TOPMed cohort data was derived from WGS (Jakubek et al., 2023b, Zekavat et al., 2021).

In comparison the H3Africa SNP array v1 is comprised of ~2,270,000 markers which is considerably higher density. The BBJ authors found a high concordance of mean heterozygous SNP density across 1Mb windows between the BBJ and UKB studies. No similar comparison was performed in this study; however, it would be relatively safe hypothesis that the mean coverage achieved in the same 1Mb windows by the H3Africa array would be at least comparable but mostly likely higher. This is noticeable with a rudimentary comparison of the Extended Data Figure 2 from the BBJ publication and Figure 3.13 from this study (Terao et al., 2020). The higher SNP density should equate more smaller MCAs on average being detected in this study when compared to UKB and BJJ. This may partially explain the smaller median MCAs across types as well as the lower size limit in AWI-Gen when compared to UKB and BBJ Section (3.3.7) (Loh et al., 2020, Terao et al., 2020). The results from the comparison of UKB and BBJ studies and this study must be interpreted in this context. It is very likely that this study was better powered to detect smaller MCAs than UKB and BJJ. However, the phasing in this study was done on the small 1000Genomes reference panel, which few African genomes mostly from Western African and no Southern African genome representation. The phasing from UKB and BBJ was much higher accuracy over longer regions. As MoChA relies on phasing, the higher density of the H3Africa array may

have been slightly offset by lower accuracy of long-range phasing, in combination with a much smaller sample size. While the distributions of MCAs across studies are interesting and informative of future analyses, direct comparison in the absence of significant data harmonisation is not recommended.

Machiela et al. (2015) employed an older, less sensitive MCA detection algorithm to call MCAs from genotype-inferred ancestry participants on commercially available arrays which were designed to capture common variation in European and admixed populations. Thus, the resolution of informative heterozygous sites used to detect MCAs is likely reduced in the African ancestry participants, resulting in smaller MCAs being missed. Additionally, MCAs were limited to only those above 2Mb in size. If the increased proportion of smaller MCAs in this study is not due to increased probe resolution and is instead a feature of African genomes, then applying this size threshold may cause many African MCAs to be missed. Jakubek et al. (2023) called MCAs from the TOPMed WGS data. In order to make comparisons between inferred ancestry participants, the authors artificially reduced the number of heterozygous sites by random downsampling to match the population group with the lowest heterozygosity rate (Jakubek et al., 2023b). MCAs were recalled post downsampling, resulting in a decrease in African ancestry MCAs by 23% while European ancestry MCAs were only decreased by 4%. The authors note that African ancestry participants had significantly lower MCA rates than European ancestry participants both prior and post downsampling. They reported an autosomal MCA rate of 3.76% in the African American participants. This harmonisation approach was understandable, but it does appear that it impacted the African ancestry samples the most. The lower accuracy phasing of African ancestry genomes in comparison to European ancestry genomes may mean that African samples are more sensitive to loss of sites, contributing to greater rates of MCA loss.

These two studies compared African American participants to European ancestry participants. African American cohorts have variable degrees of admixture and may not fully represent the full extent of genetic diversity found in Africa (Rotimi et al., 2017). A recent study investigating MCAs in African children with Burkitt lymphoma detected autosomal MCAs in 14.2% of 674 cancer-free Ghanian men controls (Zhou et al., 2023). These MCAs were limited to >2Mb in size. For comparison we detected autosomal MCAs >2Mb in 10.4% of the 741 men from the Ghanaian town of Navrongo. The samples were genotyped on a similarly dense Infinium HumanOmni5-4v1 array

(~2,100,000 markers); however, the detection pipelines differed on many levels (Zhou et al., 2023). While a direct comparison remains challenging, the higher autosomal MCA rates in both Ghanian cohorts compared to the African American cohort suggest that MCA incidence may differ in admixed African Americans, African diaspora, and continental Africans. African ancestry populations living in Western countries may be exposed to differential lifestyle factors, environmental exposures, and infection burdens when compared to continental African cohorts, which may result in significantly different MCA distributions in across the autosome.

The next important technical factor considerations are the version of the MoChA detection, the MCA filtering strategy, and QC employed by each of the studies. The MoChA tool is regularly updated. The most important impact of this is with respect to the version deployed as MoChA and the models used in its development by the UKB and BJJ teams (Loh et al., 2018, Loh et al., 2020, Terao et al., 2020). They previous iterations used in its development used a 3 state forward-backward HMM model, while the current version of MoChA uses a Viterbi HMM model. In addition, there are several other small differences in the underlying statistics and BAF correction protocols. These differences are acknowledged and outlined on the MoChA site (https://github.com/freeseek/mocha). How this specifically would impact the results from these two studies when compared to this analysis is not clearly apparent, however, it is possible that it may be consequential and must be accounted for. Post MCA calling QC and filtering strategies also differed across the studies (Jakubek et al., 2023b, Loh et al., 2020, Terao et al., 2020, Zekavat et al., 2021). The protocol removes likely germline CNVs from putative calls in two ways. Firstly, it used a list of likely recurrent germline candidates from 1000 Genomes Phase 3 to flag and filter overlapping MCAs. How well this list applied to African samples needs to be investigated further as there may be population specific recurrent CNVs that differ in Africans. The same applied to the list of likely regions within segmental duplications. More accurate filtering may be achieved by including up to date African specific filters derived from newer long-range sequencing studies.

The second way the protocol removes germline calls is by applying MCA size specific relative intensity statistic filtering thresholds (Figure 3.1). These should be derived from the data itself and differ slightly across the studies. Additionally, putative MCA calls are assigned a $\log_{10}$-odds (LOD) score for the model based on BAF and phase by MoChA. This score is a measure of the likelihood that the call is a true MCA for the chosen

model. The MoChA authors recommend excluding all calls with an LOD < 10. However, this can be adjusted based on the needs of the study. The MGBB and FinnGen cohorts had a more stringent LOD threshold of >20 (Zekavat et al., 2021). Setting the LOD threshold to >20 in this study would have excluded an additional 411 MCAs and reduced the incidence rate to 7.22% of the AWI-Gen cohort. While this does result in an incidence closer to the MGBB and FinnGen cohorts, there is still enough variability to suggest additional factors at play. Some studies have chosen to only include MCAs above a certain size or cell fraction threshold, including only larger or high cell fraction MCAs which are more likely to impact phenotype or pathology (Jakubek et al., 2023b, Zhou et al., 2023). However, as this study did not aim to investigate the impact of autosomal MCAs on any AWI-Gen phenotype and intended to capture the full spectrum of autosomal MCAs in Africans, no additional size or cell fraction thresholds were applied to the MCA calls. All together the QC and filtering strategy employed after MCA calling can significantly impact the resulting MCA incidence rates and distributions and must be considered when comparing results. In any case, it is extremely challenging to harmonise MCA data across studies with varying technical factors to make direct comparisons. Therefore, generalisations from one cohort to the next must be cautioned against. Until such a time that a sufficiently large meta-analysis of harmonised data from diverse populations is published, the results from current studies should be considered cohort specific.

3.4.3 Future methodology improvements

Several technical and methodological factors in this study likely influenced the observed incidence and distribution of autosomal MCAs, and future analyses will benefit from targeted refinements. Some limitations reflected time constraints, the exploratory scope of the project, and an inevitable learning curve during implementation. First, the available sample size was inadvertently reduced. An incorrect file was used during QC, leading to the exclusion of ~1,000 Batch 4 samples that had otherwise passed AWI-Gen quality control. This reduction in sample size inevitably lowered power, particularly when stratifying MCAs by genomic location and event type. Future work should incorporate these excluded samples and, where possible, expand the dataset with additional cohorts. Although this error was identified during final analysis, rerunning the entire pipeline was not feasible within the timeframe of this thesis.

Second, the QC strategy applied here was not optimally tailored for MCA detection. The AWI-Gen pipeline was originally designed for association studies with different priorities, applying more stringent SNP and sample filters than recommended by MoChA. For instance, AWI-Gen excluded SNPs with >1% missingness, while MoChA can tolerate up to 3%. Relaxing thresholds would allow retention of more informative sites, thereby increasing sensitivity for MCA detection. Importantly, the exclusion of chromosome X SNPs during QC prevented analysis of MLOX and MLOY, which are the most common MCAs with well-documented phenotypic associations. While X chromosome QC requires additional steps to account for haploidy in males and X inactivation in females, its inclusion should be prioritized in future work.

Finally, the phasing strategy could be improved. In this study, SHAPEIT4 was used with the 1000 Genomes Phase 3 reference panel, as recommended in the MoChA documentation. While suitable and locally deployable, this panel is suboptimal for African populations compared with the African Genome Resources and TOPMed panels, which have been shown to outperform 1000 Genomes in imputation accuracy for African genomes (Sengupta et al., 2023b). Although data transfer and privacy constraints limited their use here, future analyses should evaluate whether improved phasing reference panels measurably enhance MCA calling accuracy, particularly in African datasets where linkage disequilibrium patterns are shorter and more complex. In summary, refinements in sample inclusion, QC thresholds, chromosome X analysis, and phasing strategies represent clear avenues to increase both the sensitivity and accuracy of MCA detection in future studies. Once accounted for, insights into the biological factors impacting MCA formation and selection can be made at the cohort level.

3.4.4 Biological factors impacting MCA distributions in AWI-Gen

The biological factors are the inherent similarities or differences in MCA formation mechanisms impacting mutation rates, and the biological or environmental exposures impacting MCA selection. It is these biological differences or similarities that we are interested in which would improve our understanding of MCA formation, selection, and their impact on health and aging. The continued publication of large-scale MCA in CH studies have confirmed previously suspected MCA risk factors and have begun to

identify important factors impacting MCA mutagenesis and selection across diverse cohorts.

This analysis confirms prior associations of increasing MCA risk with age and increased MCA risk in males. The age-related increase in MCA risk observed in this study reflects the convergence of multiple degenerative processes that fundamentally compromise genomic stability within the haematopoietic system. The primary driver involves progressive DNA damage accumulation in HSCs. This genomic deterioration results from intrinsic cellular processes including oxidative stress and replication errors, as well as extrinsic exposures to inflammation, genotoxins, and environmental mutagens (Section 1.5.2). Critically, age-related decline in DNA repair capacity exacerbates genomic instability, with aging HSCs demonstrating declining efficiency DSB leading to the increasing risk of non-random DSB accumulation (Rübe et al., 2011, White and Vijg, 2016). It has been shown that NHEJ efficiency decreases ~3 fold in aged tissues with a compensatory shift toward the more error-prone MMEJ pathway (Rall-Scharpf et al., 2021). Simultaneously, the aging bone marrow niche undergoes profound microenvironmental changes characterized by fatty bone marrow accumulation, increased inflammatory cytokine production, and chronic "inflammaging" that creates selective pressures favouring HSC clones with resistance mutations (Cain et al., 2025). The theory of the ‘selective shadow’ provides a plausible evolutionary framework for understanding why MCA-driven clonal expansion accelerates dramatically after reproductive age, as selection pressure diminishes after reproductive age, allowing previously neutral or deleterious mutations to expand later in life (Kirkwood and Melov, 2011).

The observed male predominance in overall MCA frequency and the distinct sex-specific MCA patterns may be a result of biological differences in HSC regulation, DDR capacity, and hormonal influences on genomic stability. Males demonstrate consistently higher autosomal MCA rates with up to 12-fold differences for specific MCAs, attributable to testosterone's promotion of oxidative stress and chromosomal instability contrasted with oestrogen’s protective effects through antioxidant gene induction, enhanced mitochondrial function, and activation of DNA repair pathways (Watson and Blundell, 2023, Ray et al., 2008, Cui et al., 2022). The mechanistic basis involves sex hormone-mediated differential regulation of DDR efficiency, with females showing greater age-related decline in NHEJ, MMEJ, and critical repair proteins ATM, BLM, and KU70, particularly following menopause when oestrogen’s protective effects

are abruptly withdrawn (Rall-Scharpf et al., 2021). HSC biology itself exhibits profound sex-specific regulation, with ERα directly expressed on HSCs mediating enhanced proliferation and self-renewal in females, while the abrupt hormonal transition at menopause creates a vulnerable period for genomic instability compared to males' gradual testosterone decline (Cui et al., 2022, Ray et al., 2008). The sex-specific MCA enrichment patterns likely reflect differential selective pressures within aging bone marrow niches, where oestrogen loss promotes fatty bone marrow accumulation and T-cell activation leading to altered cytokine environments, while sex chromosome complement influences expression of DDR and inflammatory responses that ultimately shape the fitness landscape for MCAs (Bacharach et al., 2024).

The apparent paradox between females demonstrating greater age-related decline in DDR pathways yet males consistently showing higher MCA risk is interesting, and its explanation lies in the understanding of MCA mutagenesis being a product of DNA damage and repair efficiency. While females show greater NHEJ decline, testosterone directly increases ROS production and promotes oxidative stress in HSCs, creating a chronic pro-mutagenic environment that operates continuously from early adulthood, whereas female DDR decline occurs primarily post-menopause (Cui et al., 2022, Ray et al., 2008). Crucially, oestrogen maintains protective effects through damage prevention mechanisms including antioxidant gene induction and ROS reduction, meaning that even with declining repair efficiency, the overall genomic damage burden requiring repair remains lower in women due to reduced damage generation (Cui et al., 2022, Ray et al., 2008).

The specific sex-enriched MCA patterns observed in this study (Section 3.3.5.3) likely reflect locus-specific mechanisms involving both sex hormone-mediated mutagenesis at DSB fragile sites and sex-specific selective pressures driven from altered gene dosage within these regions. Sex hormones may influence DSB formation at particular genomic hotspots through their effects on DNA repair checkpoint controls (Cocchi et al., 2005). Further research is required to elucidate whether these sex-specific MCA enrichment patterns result from differential chromosomal fragility at hormone-sensitive DSB hotspots, sex-specific bone marrow niche selective pressures favouring clones with altered metabolism or stress resistance pathways, or population-specific genetic architecture that predisposes certain chromosomal regions to sex-biased instability in African ancestry individuals.

The differential autosomal MCA distributions observed across Western, Eastern, and Southern African participants likely reflect region-specific environmental factors impacting mutagenesis and selection interacting with genetically diverse haematopoietic systems over decades of exposure (Section 3.3.5.4) (Figure 1.10). These factors that are discussed in detail in section 1.6.2, include infection burdens, diets and behavioural intake, and occupational and environmental genotoxic exposure, often modulated through inflammation. In Southern Africa, the unique combination of the world's highest HIV burden and TB incidence creates a chronic inflammatory environment characterized by persistent immune activation, CD4+ T-cell depletion, and altered bone marrow niche dynamics that favour specific clonal expansions Conversely, Western and Eastern African populations face high malaria transmission rates that induce fundamentally different haematopoietic stresses including HSC hyperproliferation, lineage skewing toward myelopoiesis/erythropoiesis, and inflammatory profiles that differ from HIV-associated cytokine patterns. Western African participants, predominantly from rural agrarian communities with lower industrial pollution, higher exposure to biomass smoke, and diets richer in traditional anti-inflammatory staples, contrast sharply with South and East African participants residing in urban and semi-urban settings characterized by elevated industrial emissions, processed foods, and lifestyle-associated genotoxins. These differences likely further modulate MCA selection pressures through distinct inflammatory and oxidative environments.

The observed regional MCA enrichment patterns likely also reflect the interaction these factors with population-specific DNA damage repair variants and CH driver variants. The complex population structure within Africa, ranging from ancestral West African Bantu origins to East African Bantu-Nilotic admixture to South African Khoe-San, Bantu, European, and Asian admixture, likely results in region-specific haplotype architectures, recombination hotspots, and chromosomal fragile sites that may predispose to different MCA formation patterns (Choudhury et al., 2020). Gene-environment interactions likely amplify these genetic differences, as population-specific DDR variants may confer differential responses to region-specific stressors. Hypothetical examples include West African DDR/CH variants interacting with malaria-induced replication stress, East African variants responding to mixed infectious and environmental pressures, and South African admixed populations harbouring variants that respond uniquely to HIV-associated inflammation or industrial genotoxic

exposures. The observation that 5p gains predominate in West Africans while 21q gains show the reverse pattern, alongside sex-specific regional differences in losses (7q-, 13q-), suggests that the intersection of ancestry-specific genetic architecture with regionally distinct environmental selective pressures creates the complex mosaic of MCA patterns observed across African populations, highlighting the importance of population-specific analysis approaches rather than broad continental ancestry classifications. While comprehensive identification of region-specific germline risk alleles and precise exposome factors remains an imperative for future research, these findings represent a novel demonstration of how diverse environmental and genetic contexts within a single continental cohort can reveal underlying selection pressures shaping MCA landscapes, offering new insights into the multifactorial determinants of clonal haematopoiesis.

Chapter 4: Conclusion

It is now well-established that PZMs are a normal part of the spectrum of human variation. Despite this consensus, the contribution of PZMs to variability in the human phenotype is often overlooked. There are several reasons for this. Most notably, PZMs are technically challenging to detect from the conventional assays commonly utilized in clinical studies. Implicitly, you are trying to detect a mutation that is present in only a fraction of the total cells under investigation that may appear to share properties with sequencing artifacts (Dou et al., 2018). Thus, additional specialised assays, specialised analysis pipelines, and specially trained interdisciplinary personnel may be needed to confidently call and interpret PZMs. This can add significant costs and increase turnaround time in large clinical studies. Groups might look at the literature and see that historically, PZMs have accounted for few positive diagnoses in comparison to germline variants. Additionally, even if you can confidently detect PZMs, their interpretation is challenging (Zemet et al., 2022).

Interpreting the phenotypic impact of PZMs presents a significant challenge. Investigating the contribution of PZMs adds complexity that some research groups may deem unjustifiable given budgetary and time constraints. This thesis argues that, even under resource limitations, awareness of PZM detection challenges and cohort-specific protocol limitations enables effective incorporation of GM analyses in African studies. This is particularly urgent for African cohorts, where data remains scarce. A comprehensive literature review prompted several research questions aimed at bridging this gap in the knowledge.

The central question was whether advanced bioinformatics tools could identify GM in existing datasets of continental African ancestry, with an affirmative hypothesis. Recent advances in bioinformatics pipelines, designed for both clinical and epidemiological applications, now permit reliable GM characterization with appropriate datasets. These datasets have been historically absent for Africans. However, this is changing. Collaborative initiatives such as the H3Africa consortium have catalysed the establishment of diverse, high-quality African cohorts such as DDD-Africa and AWI-Gen which are derived from NGS sequencing and microarray platforms required for GM characterisation. This confluence of technological progress and enhanced cohort diversity facilitates a more comprehensive understanding of GM on the African continent.

Inclusion of African GM data is essential for refining pathogenic mutation classification and understanding somatic mutagenesis amid diverse environmental and germline backgrounds. Not only in African populations, but across the global context. Consistently, expanding cohort diversity has proven beneficial for interpreting genomic data across all populations. The findings presented herein provide robust evidence that these benefits extend to GM. Accordingly, this thesis was designed to determine whether GM could be effectively screened for in continental African datasets, and the results confirmed that such characterisation is achievable.

4.1 Key findings

The aim of this thesis was to identify appropriate detection strategies to explore the extent and impact of GM in large continental African ancestry trio-WES and SNP-array datasets. Two objectives were designed to achieve this aim. Firstly, to implement a bioinformatics pipeline to screen for pathogenic mosaic small variants that may contribute to DDs in South African trio-WES data from the DDD-Africa study. Secondly, implement a bioinformatics pipeline for the exploratory characterisation of autosomal MCAs in diverse continental African populations using H3Africa SNP-array data from the AWI-Gen study. Both objectives were successfully achieved.

The data from Chapter 2 demonstrates that mosaic small variants can be identified in African trio-WES datasets by incorporating statistical tests, such as those used in the EM-Mosaic tool, into variant prioritization pipelines to detect significant deviations in VAF distributions while accounting for confounding factor impact. Integrating a pipeline for detecting PZMs in affected probands extends standard DNM pipelines. While it relaxes VAF thresholds, it incorporates more stringent filtering to mitigate known false positives. This approach ensures the detection of conventional DNMs while also identifying potential PZMs that would otherwise be missed by standard VAF filtering. This approach has the potential to significantly enhance the yield of a trio-WES strategy with minimal drawbacks. In Chapter 2, 182 *de novo* SNPs and 29 *de novo* INDELs were identified in 62 DD probands from African trio-WES data. Of these, 36 SNPs and 11 INDELs showed strong evidence of mosaicism, variants that would have been missed using a standard DNM approach with stricter VAF thresholds. Both PZMs and DNMs identified through the variant prioritization pipeline were subsequently curated for potential contributions to the DD phenotype.

The pipeline in Chapter 2 leverages the strengths of large trio-WES datasets for DNM detection. Expanding VAF thresholds increases sensitivity but also enriches for false positives due to methodological artifacts, necessitating additional filtering steps. The *de novo* origin of variants in affected probands was assigned by ensuring no read support in either parent, substantially reducing false positives. Although confirmation would require validation through secondary analysis. Additionally, sequencing data from 186 individuals processed in the same facility, batch, and protocol allowed for the removal of sequencing artifacts by requiring candidate variants to be absent in all other participants, whether affected or healthy. Expanding this approach to the full DDD-Africa cohort, including non-trios, would further improve artifact filtering and variant prioritization. The EM-Mosaic tool enhances confidence in mosaic SNP detection by accounting for confounding factors affecting VAF distributions. It employs a beta-binomial model instead of a standard binomial model, allowing for more accurate statistical testing of VAF deviations. While this method is limited to SNPs and cannot be applied to INDELs, it significantly strengthens the reliability of mosaic SNP candidate identification. As a result, INDELs required more careful evaluation due to the overlap between their formation mechanisms and sources of false positives in WES. This limitation is inherent to WES, where variant callers like GATK HaplotypeCaller are known to detect fewer INDELs, further necessitating stringent filtering criteria such for the binomial p-value, resulting in a higher DP requirement. These limitations must be carefully considered, as they depend on the source and characteristics of the dataset. Nonetheless, expanding the DNM pipeline to include PZMs significantly increased the number of candidate variants available for evaluation in relation to DDs.

Another advantage of trio-WES was its ability to screen for gonadosomatic variation in the parents of affected probands. These variants are particularly prone to false positives and therefore require careful evaluation and validation. A larger cohort would further strengthen filtering by ensuring that such variants are present only in a single parent and the proband and absent in all other participants in the sequencing run. Standard heterozygous variant pipelines rely on variable expressivity and penetrance to prevent segregation analysis from excluding certain variants. However, since gonadosomatic variants in unaffected parents are not necessarily pathogenic, they can be more readily included in this analysis. Detecting gonadosomatic variation also has important implications for recurrence risk counselling in autosomal dominant disorders.

Unlike fully inherited variants, the recurrence risk for siblings is reduced from the expected 50% for monoallelic dominant disorders, aligning instead with the parental VAF, which reflects the level of mosaicism in the germline.

These variants are usually excluded from standard trio pipelines to reduce the risk of false positives, which are enriched in these VAF distributions. However, screening for candidate gonadosomatic variants significantly increases the potential for increased yield, by making use of trio-WES. In Chapter 2 it resulted in the inclusion of 372 additional SNP candidates and 20 INDEL candidates. Candidate SNPs and INDELs from both the proband DNM and gonadosomatic mosaic parent pipelines were further prioritized using virtual gene panels and downstream curation analyses. In Chapter 2, this process included filtering through DDG2P and LOF-intolerant gene panels, followed by clinical curation using ACMG/AMP guidelines and a prioritization strategy tailored to the challenges of variant interpretation in African DD presentations. This approach led to the identification of 12 candidate variants with strong evidence for contribution to DD in the DDD-Africa cohort. Among these, three novel variants exhibited strong evidence of somatic mosaicism in probands, three were previously undetected germline DNMs, and one was a known variant with strong mosaicism evidence. Additionally, five inherited variants were identified in affected probands, originating from gonadosomatic parents. If confirmed using orthogonal approaches, these variants would significantly enhance the yield of trio-WES in this DD cohort while also providing a definitive diagnosis for the 10 affected families and ending their diagnostic odyssey. Notably, the pipelines developed in Chapter 2 were seamlessly integrated into standard variant prioritization workflows without incurring additional costs.

The data from Chapter 3 demonstrated that autosomal MCAs could be detected across diverse continental African populations using H3Africa SNP-array data from the AWI-Gen study. The strength of this analysis lies in the combination of MoChA, the most accurate MCA detection tool, with the large and diverse AWI-Gen cohort, which includes participants from Southern, Eastern, and Western Africa, genotyped on the H3Africa array to provide comprehensive coverage of African genomes. This resulted in the largest characterization of African autosomal MCAs to date, identifying autosomal MCAs in 10.1% of the AWI-Gen cohort.

The odds of autosomal MCAs increased with age and were independently associated with males, confirming previously reported associations. However, certain MCAs, such

as small 5p terminal losses and 10q losses, were more frequent in females. Notably, the odds of at least one autosomal MCA were significantly higher in Western Africans, although this was not true for all autosomal loci. Specific MCAs were more common in distinct populations: gains on chromosome 5 in Western Africans, small 17q losses in Southern Africans, and gains on chromosome 21 in Eastern Africans. This suggests that unique factors may predispose individuals from each region to MCA mutagenesis, such as germline haplotypes in DNA damage response DDR pathways or region-specific selection pressures. These could include factors like infection burdens (e.g., HIV, malaria, TB), environmental exposures (e.g., pollutants, occupational hazards), or dietary/lifestyle differences, which may drive the selection of certain autosomal MCA loci in participants from each region. Importantly, the high incidence of autosomal MCAs in the AWI-Gen cohort, compared to African ancestry and Arican diaspora populations in large studies from other cohorts, suggests that unique factors within the African continent are influencing MCA formation and selection. These factors are likely distinct from those in the global North. Thus, MCA distributions from those populations cannot be generalised to continental African cohorts. The success of this analysis lies in the size and diversity of the AWI-Gen cohort, which enabled the identification of novel autosomal MCA hotspots and drivers in participants from continental Africa. The analysis was conducted retrospectively on the AWI-Gen study cohort, which was not originally recruited for MCA detection. Nevertheless, the findings demonstrate that with a sufficiently large, high-density SNP-array dataset representative of African genetic variation, MCAs can be effectively detected in African cohorts. Moreover, the novel hotspots and drivers identified may unveil significant biological factors influencing CH in both African and non-African populations in future studies.

4.2 Study limitations

While both objectives were achieved, successfully enriching the knowledge base on African GM, they were not without limitations.

The primary limitation of Chapter 2 was the WES data itself. While trio-WES is highly effective for GM screening due to its ability to filter false positives, its read depth and uniformity of coverage impose significant constraints on mosaic variant detection. The relatively low coverage of 45X meant that, according to EM-Mosaic's power calculations, only about a third of mosaic SNPs were detectable. In many cases, even

this coverage was not achieved, restricting detection to high-level mosaicism and reducing confidence in calls within low-coverage regions, such as exon-intron boundaries.

Although a VAF threshold of 0.1 is standard for WES, most somatic mutations fall below this threshold, meaning many true mosaic variants potentially contributing to DD pathology were likely missed. Additionally, because this was a retrospective screening strategy using existing data, a dedicated mosaic variant caller was not applied at the outset. Instead, the GATK ploidy setting was configured to exclude variants at extreme VAF distributions. While this is standard for trio-WES, it likely led to the omission of lower VAF candidates and the inclusion of false positives—variants that would typically be filtered out by dedicated mosaic variant callers such as Mutect2 or MosaicForecast.

Furthermore, INDELs and parental gonadosomatic variants could not be analysed using EM-Mosaic, meaning that overdispersion of VAF distributions for these variants was not accounted for. Most critically, none of the identified candidate variants have been experimentally validated, as this fell outside the scope of the current objective. As a result, all conclusions must be interpreted with caution, acknowledging the potential for false positives. The goal of this objective was to identify the most promising candidate variants influencing DDs within the cohort, which could then be validated using higher-depth targeted sequencing or digital droplet PCR methods that are not currently available on site. As a result, these variants may well be false-positive artifacts, particularly given that variants in these VAF regions are known to be enriched for such errors. Additionally, the non-trio WES data from DDD-Africa could not be analysed for GM contributions using these pipelines. Future analyses will need to incorporate a panel-of-normals approach and tools like MosaicForecast to detect GM in WES duos, though this approach comes with its own limitations and considerations.

The curation of variants in African DDs presents unique challenges. Phenotypic heterogeneity introduces variability in the clinical data, making it more difficult to confidently associate candidate variants with specific phenotypes. The more detailed and precise the phenotypic information, the easier it becomes to assign pathogenicity in accordance with ACMG/AMP guidelines. There are currently no clear guidelines for interpreting GM in variant curation. For instance, there are no established recommendations on how to assess atypical phenotypes or the unclear segregation of partial phenotypes in cases of gonadosomatic variation. As a result, GM interpretation remains highly subjective, adding another layer of complexity to variant classification.

Finally, the contribution of other classes of GM to DDs, such as MCAs was not assessed in these analyses.

In Chapter 3, limitations mirrored the strengths, as they were inherently linked to the MoChA tool, the H3Africa SNP array, and the AWI-Gen cohort. The MoChA tool depends on accurate phasing of heterozygous SNPs to detect allelic imbalances. In this analysis, phasing was performed using the 1000 Genomes Phase 3 reference panel, which lacks comprehensive representation of African populations, particularly Southern Africans. This limitation was evident in the association between phase switch errors and the region of recruitment. Improved phasing using the African Genome Resources Panel or the H3Africa panel could enhance the sensitivity of MCA detection in this cohort.

Additional limitations include the unintentional exclusion of approximately 1,000 Batch 4 samples from the AWI-Gen cohort due to sample filtering errors, which should be incorporated in future analyses. Moreover, the estimation of MCA size and breakpoints was constrained by the coverage of the H3Africa SNP array. Finally, the unequal distribution of sub-cohorts, with significantly fewer Eastern Africans compared to Western and Southern Africans, limits direct comparisons between these groups. Additionally, comparing MCA distributions across AWI-Gen, BBJ, and UKB is complicated by differences in array design and methodological approaches. A direct comparison would require harmonizing these datasets or controlling for technological factors such as array density and quality control strategies, which is extremely challenging.

While many novel drivers of MCA expansion were identified, the selection of gene sources was limited to those previously implicated in blood malignancies, potentially overlooking drivers outside these lists. Furthermore, although the AWI-Gen cohort includes extensive phenotypic data, key variables, such as smoking status and infection history, were not incorporated into the association analyses in Chapter 3. For instance, smoking prevalence is significantly higher among South African men than women, and smoking is known to increase the risk of autosomal MCA occurrence. Similarly, HIV infection, which is associated with an increased risk of MCA, may have confounded the observed associations. Although the potential links between tuberculosis, malaria, and CH remain theoretical yet biologically plausible, they could also contribute to the differential MCA distributions observed between regions and sexes. Investigating these associations further may offer valuable insights into the

pathophysiology of CH and inform potential preventive strategies in regions where these infections are prevalent. The impact of autosomal MCAs on AWI-Gen health outcomes was not assessed in this analysis. Finally, the quality control resulted in the exclusion of sec chromosome SNPs. Thus, MLOY and MLOX were not assessed in this analysis.

4.3 Future directions

The most pressing future direction for Chapter 2 is the orthogonal confirmation and validation of candidate variants using alternative methodologies, such as ultra-high depth targeted sequencing or digital droplet PCR. This step is crucial for subsequent statistical analyses aimed at assessing gene set burdens, selection pressures, and the enrichment of DNMs or PZMs in DD genes. Additionally, phasing variants in the trio-WES data could be incorporated into mosaicism calls, thereby increasing confidence and reducing false positives. Expanding the trio-WES dataset with additional trio samples from DDD-Africa is necessary to further improve EM-Mosaic's estimates and false-positive filtering and improve diagnostic yields. Finally, the contribution of MCAs to DDs in the DDD-Africa cohort should be evaluated, either through array-based methods with MoChA or using MrMosaic, with the understanding that MoChA requires data from at least 100 probands for effective analysis. Incorporating duos into the analysis using parallel methodologies that do not require trios, such as MosaicForecast coupled with a panel of normal, could significantly enhance yield and provide independent confirmation of variants identified in this study. Finally, sequencing the saliva swab samples from probands will enable confirmation of systemic mosaicism in candidate variants and facilitate comparisons between different DNA sources.

The AWI-Gen dataset has recently been expanded by several thousand samples, providing an important opportunity to refine our understanding of MCA distribution across Africa. To maximize the power of this analysis, the cohort should be further expanded by incorporating additional samples genotyped with the H3Africa array from other consortium studies. Increasing the sample size, particularly with participants from diverse African sites, will enhance the representativeness and resolution of MCA analyses. Phasing accuracy remains a critical factor in MCA detection, and improvements can be made by utilizing a larger cohort and phasing with more representative reference panels, such as the African Genome Resources panel or the

newly released H3Africa reference panel. This would improve the precision of MCA calling and allow for more reliable detection of mosaic events.

Beyond improving technical accuracy, integrating AWI-Gen phenotype data will be essential in expanding our understanding of African MCAs and their potential impact on health outcomes. Identifying additional risk factors associated with MCAs could provide new insights into their biological and clinical significance. Furthermore, longitudinal analyses using AWI-Gen Phase 2 five-year follow-up samples will allow for the characterization of clonal dynamics over time in African populations, shedding light on how MCAs evolve under different environmental and genetic contexts. Finally, future analyses must also incorporate sex chromosome MCAs, which were not included in this study. Given the biological relevance of sex chromosome anomalies in hematopoietic clonal expansion and disease risk, their inclusion will provide a more comprehensive picture of mosaicism across African populations.

There remains substantial scope to build upon the analyses presented in Chapters 2 and 3. The work described here is exploratory in nature and represents only an initial step toward uncovering the full landscape of genetic mosaicism in African populations. Future studies with larger sample sizes, refined methodological pipelines, and integrative analyses that combine genomic, environmental, and clinical data will be essential to deepen and broaden these findings. Despite these limitations in methodology and scope, this thesis has successfully achieved its primary aim. By establishing both the feasibility and the value of such analyses, this work lays a foundation for future research that will advance our understanding of GM, and its broader implications for human health and disease in Africa and globally.

References

AALTONEN, L. A., ABASCAL, F., ABESHOUSE, A., ABURATANI, H., ADAMS, D. J., AGRAWAL, N., AHN, K. S., AHN, S.-M., AIKATA, H., AKBANI, R., AKDEMIR, K. C., AL-AHMADIE, H., AL-SEDAIRY, S. T., AL-SHAHROUR, F., ALAWI, M., ALBERT, M., ALDAPE, K., ALEXANDROV, L. B., ALLY, A., ... & THE, I. T. P.-C. A. O. W. G. C. 2020. Pan-cancer analysis of whole genomes. *Nature,* 578**,** 82-93.

ABASCAL, F., HARVEY, L. M. R., MITCHELL, E., LAWSON, A. R. J., LENSING, S. V., ELLIS, P., RUSSELL, A. J. C., ALCANTARA, R. E., BAEZ-ORTEGA, A. & WANG, Y. 2021. Somatic mutation landscapes at single-molecule resolution. *Nature,* 593**,** 405-410.

ABOU TAYOUN, A. N., PESARAN, T., DISTEFANO, M. T., OZA, A., REHM, H. L., BIESECKER, L. G., HARRISON, S. M. & CLINGEN SEQUENCE VARIANT INTERPRETATION WORKING, G. 2018. Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. *Human Mutation,* 39**,** 1517-1524.

ACHA-SAGREDO, A., GANGULI, P. & CICCARELLI, F. D. 2022. Somatic variation in normal tissues: friend or foe of cancer early detection? *Ann Oncol,* 33**,** 1239-1249.

ACUNA-HIDALGO, R., VELTMAN, J. A. & HOISCHEN, A. 2016. New insights into the generation and role of de novo mutations in health and disease. *Genome biology,* 17**,** 1-19.

ADAR, S., HU, J., LIEB, J. D. & SANCAR, A. 2016. Genome-wide kinetics of DNA excision repair in relation to chromatin state and mutagenesis. *Proc Natl Acad Sci U S A,* 113**,** E2124-33.

AGYEMANG, C., MEEKS, K., BEUNE, E., OWUSU-DABO, E., MOCKENHAUPT, F. P., ADDO, J., DE GRAFT AIKINS, A., BAHENDEKA, S., DANQUAH, I., SCHULZE, M. B., SPRANGER, J., BURR, T., AGYEI-BAFFOUR, P., AMOAH, S. K., GALBETE, C., HENNEMAN, P., KLIPSTEIN-GROBUSCH, K., NICOLAOU, M., ADEYEMO, A., ... & STRONKS, K. 2016. Obesity and type 2 diabetes in sub-Saharan Africans – Is the burden in today's Africa similar to African migrants in Europe? The RODAM study. *BMC Medicine,* 14**,** 166.

ALEXANDROV, L. B., KIM, J., HARADHVALA, N. J., HUANG, M. N., TIAN NG, A. W., WU, Y., BOOT, A., COVINGTON, K. R., GORDENIN, D. A., BERGSTROM, E. N., ISLAM, S. M. A., LOPEZ-BIGAS, N., KLIMCZAK, L. J., MCPHERSON, J. R., MORGANELLA, S., SABARINATHAN, R., WHEELER, D. A., MUSTONEN, V., ALEXANDROV, L. B., ... & CONSORTIUM, P. 2020a. The repertoire of mutational signatures in human cancer. *Nature,* 578**,** 94-101.

ALEXANDROV, L. B., KIM, J., HARADHVALA, N. J., HUANG, M. N., TIAN NG, A. W., WU, Y., BOOT, A., COVINGTON, K. R., GORDENIN, D. A., BERGSTROM, E. N., ISLAM, S. M. A., LOPEZ-BIGAS, N., KLIMCZAK, L. J., MCPHERSON, J. R., MORGANELLA, S., SABARINATHAN, R., WHEELER, D. A., MUSTONEN, V., BOUTROS, P., ... & CONSORTIUM, P. 2020b. The repertoire of mutational signatures in human cancer. *Nature,* 578**,** 94-101.

ALEXANDROV, L. B., NIK-ZAINAL, S., WEDGE, D. C., APARICIO, S. A. J. R., BEHJATI, S., BIANKIN, A. V., BIGNELL, G. R., BOLLI, N., BORG, A. &

BØRRESEN-DALE, A.-L. 2013. Signatures of mutational processes in human cancer. *Nature,* 500**,** 415-421.
ALEXANDROV, L. B. & STRATTON, M. R. 2014. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Current opinion in genetics & development,* 24**,** 52-60.
ALEXANDROV, L. B. & ZHIVAGUI, M. 2019. Mutational Signatures and the Etiology of Human Cancers. *In:* BOFFETTA, P. & HAINAUT, P. (eds.) *Encyclopedia of Cancer (Third Edition).* Oxford: Academic Press.
ALI, S. A., SOO, C., AGONGO, G., ALBERTS, M., AMENGA-ETEGO, L., BOUA, R. P., CHOUDHURY, A., CROWTHER, N. J., DEPUUR, C., GÓMEZ-OLIVÉ, F. X., GUIRAUD, I., HAREGU, T. N., HAZELHURST, S., KAHN, K., KHAYEKA-WANDABWA, C., KYOBUTUNGI, C., LOMBARD, Z., MASHINYA, F., MICKLESFIELD, L., ... & RAMSAY, M. 2018. Genomic and environmental risk factors for cardiometabolic diseases in Africa: methods used for Phase 1 of the AWI-Gen population cross-sectional study. *Global Health Action,* 11**,** 1507133.
ALISSA, M. D. G. & WALSH, C. A. 2018. Somatic mosaicism and neurodevelopmental disease. *Nature Neuroscience,* 21**,** 1504-1514.
AMERICAN PSYCHIATRIC ASSOCIATION, D. & AMERICAN PSYCHIATRIC, A. 2013. *Diagnostic and statistical manual of mental disorders: DSM-5*, American psychiatric association Washington, DC.
AMOS, W. 2013. Variation in heterozygosity predicts variation in human substitution rates between populations, individuals and genomic regions. *PLoS One,* 8**,** e63048.
ANDERSON, D. & LASSMANN, T. 2022. An expanded phenotype centric benchmark of variant prioritisation tools. *Human Mutation,* 43**,** 539-546.
ANDRIANI, G. A., VIJG, J. & MONTAGNA, C. 2017. Mechanisms and consequences of aneuploidy and chromosome instability in the aging brain. *Mechanisms of Ageing and Development,* 161**,** 19-36.
ARAI, F., HIRAO, A., OHMURA, M., SATO, H., MATSUOKA, S., TAKUBO, K., ITO, K., KOH, G. Y. & SUDA, T. 2004. Tie2/Angiopoietin-1 Signaling Regulates Hematopoietic Stem Cell Quiescence in the Bone Marrow Niche. *Cell,* 118**,** 149-161.
ASHENAFI, S., LORETI, M. G., BEKELE, A., ASEFFA, G., AMOGNE, W., KASSA, E., ADERAYE, G. & BRIGHENTI, S. 2023. Inflammatory immune profiles associated with disease severity in pulmonary tuberculosis patients with moderate to severe clinical TB or anemia. *Front Immunol,* 14**,** 1296501.
AUTON, A., ABECASIS, G. R., ALTSHULER, D. M., DURBIN, R. M., ABECASIS, G. R., BENTLEY, D. R., CHAKRAVARTI, A., CLARK, A. G., DONNELLY, P., EICHLER, E. E., FLICEK, P., GABRIEL, S. B., GIBBS, R. A., GREEN, E. D., HURLES, M. E., KNOPPERS, B. M., KORBEL, J. O., LANDER, E. S., LEE, C., ... & NATIONAL EYE INSTITUTE, N. I. H. 2015. A global reference for human genetic variation. *Nature,* 526**,** 68-74.
AVAGYAN, S. & ZON, L. I. 2023. Clonal hematopoiesis and inflammation – the perpetual cycle. *Trends in Cell Biology,* 33**,** 695-707.
BACHARACH, T., KAUSHANSKY, N. & SHLUSH, L. I. 2024. Age-related micro-environmental changes as drivers of clonal hematopoiesis. *Curr Opin Hematol,* 31**,** 53-57.
BAILEY, J. A., YAVOR, A. M., MASSA, H. F., TRASK, B. J. & EICHLER, E. E. 2001. Segmental Duplications: Organization and Impact Within the Current Human Genome Project Assembly. *Genome Research,* 11**,** 1005-1017.
BAINE-SAVANHU, F., MACAULAY, S., LOUW, N., BOLLWEG, A., FLYNN, K., MOLATOLI, M., NEVONDWE, P., SEYMOUR, H., CARSTENS, N., KRAUSE,

A. & LOMBARD, Z. 2023. Identifying the genetic causes of developmental disorders and intellectual disability in Africa: a systematic literature review. *Frontiers in Genetics,* 14.

BAKER, T. M., WAISE, S., TARABICHI, M. & VAN LOO, P. 2024. Aneuploidy and complex genomic rearrangements in cancer evolution. *Nature Cancer,* 5**,** 228-239.

BALAGUÉ-DOBÓN, L., CÁCERES, A. & GONZÁLEZ, J. R. 2022. Fully exploiting SNP arrays: a systematic review on the tools to extract underlying genomic structure. *Briefings in Bioinformatics,* 23**,** bbac043.

BALDRIDGE, M. T., KING, K. Y., BOLES, N. C., WEKSBERG, D. C. & GOODELL, M. A. 2010. Quiescent haematopoietic stem cells are activated by IFN-γ in response to chronic infection. *Nature,* 465**,** 793-797.

BANERJEE, S., ZHAO, Q., WANG, B., QIN, J., YUAN, X., LOU, Z., ZHENG, W., LI, H., WANG, X., CHENG, X., ZHU, Y., LIN, F., YANG, F., XU, J., MUNSHI, A., DAS, P., ZHOU, Y., MANDAL, K., WANG, Y., ... & LI, C. 2024. A novel in-frame deletion in KIF5C gene causes infantile onset epilepsy and psychomotor retardation. *MedComm,* 5**,** e469.

BAO, E. L., NANDAKUMAR, S. K., LIAO, X., BICK, A. G., KARJALAINEN, J., TABAKA, M., GAN, O. I., HAVULINNA, A. S., KIISKINEN, T. T. J., LAREAU, C. A., DE LAPUENTE PORTILLA, A. L., LI, B., EMDIN, C., CODD, V., NELSON, C. P., WALKER, C. J., CHURCHHOUSE, C., DE LA CHAPELLE, A., KLEIN, D. E., ... & FINNGEN ANALYSIS WORKING, G. 2020. Inherited myeloproliferative neoplasm risk affects haematopoietic stem cells. *Nature,* 586**,** 769-775.

BARYAWNO, N., SEVERE, N. & SCADDEN, D. T. 2017. Hematopoiesis: Reconciling Historic Controversies about the Niche. *Cell Stem Cell,* 20**,** 590-592.

BENJAMIN, D., SATO, T., CIBULSKIS, K., GETZ, G., STEWART, C. & LICHTENSTEIN, L. 2019. Calling Somatic SNVs and Indels with Mutect2. *bioRxiv***,** 861054.

BERNKOPF, M., ABDULLAH, U. B., BUSH, S. J., WOOD, K. A., GHAFFARI, S., GIANNOULATOU, E., KOELLING, N., MAHER, G. J., THIBAUT, L. M., WILLIAMS, J., BLAIR, E. M., KELLY, F. B., BLOSS, A., BURKITT-WRIGHT, E., CANHAM, N., DENG, A. T., DIXIT, A., EASON, J., ELMSLIE, F., ... & GORIELY, A. 2023. Personalized recurrence risk assessment following the birth of a child with a pathogenic de novo mutation. *Nature Communications,* 14**,** 853.

BHATTACHARYA, R., UDDIN, M. M., PATEL, A. P., NIROULA, A., FINNERAN, P., BERNARDO, R., FITCH, K. V., LU, M. T., BLOOMFIELD, G. S., MALVESTUTTO, C., ABERG, J. A., FICHTENBAUM, C. J., HORNSBY, W., RIBAUDO, H. J., LIBBY, P., EBERT, B. L., ZANNI, M. V., DOUGLAS, P. S., GRINSPOON, S. K. & NATARAJAN, P. 2024. Risk factors for clonal hematopoiesis of indeterminate potential in people with HIV: a report from the REPRIEVE trial. *Blood Advances,* 8**,** 959-967.

BHATTACHARYA, R., ZEKAVAT, S. M., HAESSLER, J., FORNAGE, M., RAFFIELD, L., UDDIN, M. M., BICK, A. G., NIROULA, A., YU, B. & GIBSON, C. 2022. Clonal hematopoiesis is associated with higher risk of stroke. *Stroke,* 53**,** 788-797.

BICK, A. G., PIRRUCCELLO, J. P., GRIFFIN, G. K., GUPTA, N., GABRIEL, S., SALEHEEN, D., LIBBY, P., KATHIRESAN, S. & NATARAJAN, P. 2020a. Genetic Interleukin 6 Signaling Deficiency Attenuates Cardiovascular Risk in Clonal Hematopoiesis. *Circulation,* 141**,** 124-131.

BICK, A. G., WEINSTOCK, J. S., NANDAKUMAR, S. K., FULCO, C. P., BAO, E. L., ZEKAVAT, S. M., SZETO, M. D., LIAO, X., LEVENTHAL, M. J., NASSER, J., CHANG, K., LAURIE, C., BURUGULA, B. B., GIBSON, C. J., LIN, A. E., TAUB, M. A., AGUET, F., ARDLIE, K., MITCHELL, B. D., ... & NATARAJAN, P. 2020b.

Inherited causes of clonal haematopoiesis in 97,691 whole genomes. *Nature,* 586**,** 763-768.
BIESECKER, L. G. & SPINNER, N. B. 2013. A genomic view of mosaicism and human disease. *Nature Reviews Genetics,* 14**,** 307-320.
BLESSON, A. & COHEN, J. S. 2020. Genetic counseling in neurodevelopmental disorders. *Cold Spring Harbor Perspectives in Medicine,* 10.
BLEVE, A., MOTTA, F., DURANTE, B., PANDOLFO, C., SELMI, C. & SICA, A. 2023. Immunosenescence, Inflammaging, and Frailty: Role of Myeloid Cells in Age-Related Diseases. *Clinical Reviews in Allergy & Immunology,* 64**,** 123-144.
BOGESKA, R., MIKECIN, A.-M., KASCHUTNIG, P., FAWAZ, M., BÜCHLER-SCHÄFF, M., LE, D., GANUZA, M., VOLLMER, A., PAFFENHOLZ, S. V., ASADA, N., RODRIGUEZ-CORREA, E., FRAUHAMMER, F., BUETTNER, F., BALL, M., KNOCH, J., STÄBLE, S., WALTER, D., PETRI, A., CARREÑO-GONZALEZ, M. J., ... & MILSOM, M. D. 2022. Inflammatory exposure drives long-lived impairment of hematopoietic stem cell self-renewal activity and accelerated aging. *Cell Stem Cell,* 29**,** 1273-1284.e8.
BOLTON, K. L., PTASHKIN, R. N., GAO, T., BRAUNSTEIN, L., DEVLIN, S. M., KELLY, D., PATEL, M., BERTHON, A., SYED, A., YABE, M., COOMBS, C. C., CALTABELLOTTA, N. M., WALSH, M., OFFIT, K., STADLER, Z., MANDELKER, D., SCHULMAN, J., PATEL, A., PHILIP, J., ... & PAPAEMMANUIL, E. 2020. Cancer therapy shapes the fitness landscape of clonal hematopoiesis. *Nature Genetics,* 52**,** 1219-1226.
BOPE, C. D., CHIMUSA, E. R., NEMBAWARE, V., MAZANDU, G. K., DE VRIES, J. & WONKAM, A. 2019. Dissecting in silico Mutation Prediction of Variants in African Genomes: Challenges and Perspectives. *Frontiers in Genetics,* 10.
BORGES, P., PASQUALIM, G. & MATTE, U. 2021. Which Is the Best In Silico Program for the Missense Variations in IDUA Gene? A Comparison of 33 Programs Plus a Conservation Score and Evaluation of 586 Missense Variants. *Frontiers in Molecular Biosciences,* 8.
BOVERI, T. 1914. *Zur frage der entstehung maligner tumoren,* Jena,, G. Fischer.
BOWLING, S., SRITHARAN, D., OSORIO, F. G., NGUYEN, M., CHEUNG, P., RODRIGUEZ-FRATICELLI, A., PATEL, S., YUAN, W.-C., FUJIWARA, Y., LI, B. E., ORKIN, S. H., HORMOZ, S. & CAMARGO, F. D. 2020. An Engineered CRISPR-Cas9 Mouse Line for Simultaneous Readout of Lineage Histories and Gene Expression Profiles in Single Cells. *Cell,* 181**,** 1693-1694.
BRENNER, S., JACOB, F. & MESELSON, M. 1961. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature,* 190**,** 576-581.
BREUSS, M. W., ANTAKI, D., GEORGE, R. D., KLEIBER, M., JAMES, K. N., BALL, L. L., HONG, O., MITRA, I., YANG, X., WIRTH, S. A., GU, J., GARCIA, C. A. B., GUJRAL, M., BRANDLER, W. M., MUSAEV, D., NGUYEN, A., MCEVOY-VENNERI, J., KNOX, R., STICCA, E., ... & GLEESON, J. G. 2020. Autism risk in offspring can be assessed through quantification of male sperm mosaicism. *Nat Med,* 26**,** 143-150.
BROMHAM, L. 2009. Darwin would have loved DNA: celebrating Darwin 200. *Biology Letters,* 5**,** 503-505.
BROWNING, S. R. & BROWNING, B. L. 2011. Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics,* 12**,** 703-714.
BULATI, M., BUFFA, S., CANDORE, G., CARUSO, C., DUNN-WALTERS, D. K., PELLICANÒ, M., WU, Y.-C. & ROMANO, G. C. 2011. B cells and immunosenescence: a focus on IgG+ IgD− CD27−(DN) B cells in aged humans. *Ageing research reviews,* 10**,** 274-284.

BUONINFANTE, O. A., PILZECKER, B., SPANJAARD, A., DE GROOT, D., PREKOVIC, S., SONG, J.-Y., LIEFTINK, C., AYIDAH, M., PRITCHARD, C. E. J., VIVIÉ, J., MCGRATH, K. E., HUIJBERS, I. J., PHILIPSEN, S., VON LINDERN, M., ZWART, W., BEIJERSBERGEN, RODERICK L., PALIS, J., VAN DEN BERK, P. C. M. & JACOBS, H. 2023. Mammalian life depends on two distinct pathways of DNA damage tolerance. *Proceedings of the National Academy of Sciences,* 120**,** e2216055120.

BURSSED, B., ZAMARIOLLI, M., BELLUCCO, F. T. & MELARAGNO, M. I. 2022. Mechanisms of structural chromosomal rearrangement formation. *Molecular Cytogenetics,* 15**,** 23.

BUSCH, K., KLAPPROTH, K., BARILE, M., FLOSSDORF, M., HOLLAND-LETZ, T., SCHLENNER, S. M., RETH, M., HÖFER, T. & RODEWALD, H.-R. 2015. Fundamental properties of unperturbed haematopoiesis from stem cells in vivo. *Nature,* 518**,** 542-546.

CAGAN, A., BAEZ-ORTEGA, A., BRZOZOWSKA, N., ABASCAL, F., COORENS, T. H. H., SANDERS, M. A., LAWSON, A. R. J., HARVEY, L. M. R., BHOSLE, S., JONES, D., ALCANTARA, R. E., BUTLER, T. M., HOOKS, Y., ROBERTS, K., ANDERSON, E., LUNN, S., FLACH, E., SPIRO, S., JANUSZCZAK, I., ... & MARTINCORENA, I. 2022. Somatic mutation rates scale with lifespan across mammals. *Nature,* 604**,** 517-524.

CAIADO, F., PIETRAS, E. M. & MANZ, M. G. 2021. Inflammation as a regulator of hematopoietic stem cell function in disease, aging, and clonal selection. *Journal of Experimental Medicine,* 218**,** e20201541.

CAIN, T. L., DERECKA, M. & MCKINNEY-FREEMAN, S. 2025. The role of the haematopoietic stem cell niche in development and ageing. *Nat Rev Mol Cell Biol,* 26**,** 32-50.

CALDECOTT, K. W. 2024. Causes and consequences of DNA single-strand breaks. *Trends in Biochemical Sciences,* 49**,** 68-78.

CAMPBELL, I. M., SHAW, C. A., STANKIEWICZ, P. & LUPSKI, J. R. 2015. Somatic mosaicism: implications for disease and transmission genetics. *Trends in Genetics,* 31**,** 382-392.

CAMPBELL, M. C., HIRBO, J. B., TOWNSEND, J. P. & TISHKOFF, S. A. 2014. The peopling of the African continent and the diaspora into the new world. *Current opinion in genetics & development,* 29**,** 120-132.

CANNON, S., WILLIAMS, M., GUNNING, A. C. & WRIGHT, C. F. 2023. Evaluation of in silico pathogenicity prediction tools for the classification of small in-frame indels. *BMC Medical Genomics,* 16**,** 36.

CARMICHAEL, N., TSIPIS, J., WINDMUELLER, G., MANDEL, L. & ESTRELLA, E. 2015. “Is it Going to Hurt?”: The Impact of the Diagnostic Odyssey on Children and Their Families. *Journal of Genetic Counseling,* 24**,** 325-335.

CASTIGLIONI, S., DI FEDE, E., BERNARDELLI, C., LETTIERI, A., PARODI, C., GRAZIOLI, P., COLOMBO, E. A., ANCONA, S., MILANI, D., OTTAVIANO, E., BORGHI, E., MASSA, V., GHELMA, F., VIGNOLI, A., LESMA, E. & GERVASINI, C. 2022. KMT2A: Umbrella Gene for Multiple Diseases. *Genes (Basel),* 13.

CHALLEN, G. A. & GOODELL, M. A. 2020. Clonal hematopoiesis: mechanisms driving dominance of stem cell clones. *Blood,* 136**,** 1590-1598.

CHANG, C. C., CHOW, C. C., TELLIER, L. C., VATTIKUTI, S., PURCELL, S. M. & LEE, J. J. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 4, s13742–015–0047–8.

CHANG VICKY, C., ZHOU, W., BERNDT SONJA, I., ANDREOTTI, G., YEAGER, M., PARKS CHRISTINE, G., SANDLER DALE, P., ROTHMAN, N., BEANE FREEMAN LAURA, E., MACHIELA MITCHELL, J. & HOFMANN JONATHAN,

N. 2023. Glyphosate Use and Mosaic Loss of Chromosome Y among Male Farmers in the Agricultural Health Study. *Environmental Health Perspectives,* 131**,** 127006.

CHATTERJEE, N. & WALKER, G. C. 2017. Mechanisms of DNA damage, repair, and mutagenesis. *Environmental and Molecular Mutagenesis,* 58**,** 235-263.

CHEN, S., FRANCIOLI, L. C., GOODRICH, J. K., COLLINS, R. L., KANAI, M., WANG, Q., ALFÖLDI, J., WATTS, N. A., VITTAL, C., GAUTHIER, L. D., POTERBA, T., WILSON, M. W., TARASOVA, Y., PHU, W., GRANT, R., YOHANNES, M. T., KOENIG, Z., FARJOUN, Y., BANKS, E., ... & GENOME AGGREGATION DATABASE, C. 2024. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature,* 625**,** 92-100.

CHEN, X., LIGUMSKY, H., AMBROSE, C., SIBRIAN, D., TRAN, B., ARIF, D., CASTELLANOS, O., KESSNER, D., LUO, H., UBALE, M., COLEMAN, A., MAHAGANAPATHY, V., JÖNSSON, T. J., BASHO, R. K., LEE, J. S. H., MATASCI, N. & AGUS, D. B. 2025. Monitoring the rate and variability of somatic genomic alterations using long-read sequencing. *Sci Rep,* 15**,** 18397.

CHEN, Z., YUAN, Y., CHEN, X., CHEN, J., LIN, S., LI, X. & DU, H. 2020. Systematic comparison of somatic variant calling performance among different sequencing depth and mutation frequency. *Sci Rep,* 10**,** 3501.

CHENG, C., HONG, W., LI, Y., XIAO, X., MCKAY, J., HAN, Y., BYUN, J., PENG, B., ALBANES, D., LAM, S., TARDON, A., CHEN, C., BOJESEN, S. E., LANDI, M. T., JOHANSSON, M., RISCH, A., BICKEBÖLLER, H., WICHMANN, H. E., CHRISTIANI, D. C., ... & AMOS, C. I. 2023. Mosaic Chromosomal Alterations Are Associated With Increased Lung Cancer Risk: Insight From the INTEGRAL-ILCCO Cohort Analysis. *Journal of Thoracic Oncology,* 18**,** 1003-1016.

CHENG, H., ZHENG, Z. & CHENG, T. 2020. New paradigms on hematopoietic stem cell differentiation. *Protein & Cell,* 11**,** 34-44.

CHENG, K. C., CAHILL, D. S., KASAI, H., NISHIMURA, S. & LOEB, L. A. 1992. 8-Hydroxyguanine, an abundant form of oxidative DNA damage, causes GT and AC substitutions. *Journal of biological chemistry,* 267**,** 166-172.

CHENG, T., RODRIGUES, N., SHEN, H., YANG, Y.-G., DOMBKOWSKI, D., SYKES, M. & SCADDEN, D. T. 2000. Hematopoietic Stem Cell Quiescence Maintained by p21cip1/waf1. *Science,* 287**,** 1804-1808.

CHESHIER, S. H., PROHASKA, S. S. & WEISSMAN, I. L. 2007. The Effect of Bleeding on Hematopoietic Stem Cell Cycling and Self-Renewal. *Stem Cells and Development,* 16**,** 707-718.

CHIURAZZI, P., DESTRO-BISOL, G., GENUARDI, M., OOSTRA, B., SPEDINI, G. & NERI, G. 1996. Extended gene diversity at the FMR1 locus and neighbouring CA repeats in a sub-Saharan population. *American journal of medical genetics,* 64**,** 216-219.

CHO, NAM W., DILLEY, ROBERT L., LAMPSON, MICHAEL A. & GREENBERG, ROGER A. 2014. Interchromosomal Homology Searches Drive Directional ALT Telomere Movement and Synapsis. *Cell,* 159**,** 108-121.

CHOUDHURY, A., ARON, S., BOTIGUÉ, L. R., SENGUPTA, D., BOTHA, G., BENSELLAK, T., WELLS, G., KUMUTHINI, J., SHRINER, D. & FAKIM, Y. J. 2020. High-depth African genomes inform human migration and health. *Nature,* 586**,** 741-748.

CHOUDHURY, A., BRANDENBURG, J.-T., CHIKOWORE, T., SENGUPTA, D., BOUA, P. R., CROWTHER, N. J., AGONGO, G., ASIKI, G., GÓMEZ-OLIVÉ, F. X., KISIANGANI, I., MAIMELA, E., MASEMOLA-MAPHUTHA, M., MICKLESFIELD, L. K., NONTERAH, E. A., NORRIS, S. A., SORGHO, H., TINTO, H., TOLLMAN, S., GRAHAM, S. E., ... & CONSORTIUM, H. A. 2022.

Meta-analysis of sub-Saharan African studies provides insights into genetic architecture of lipid traits. *Nature Communications,* 13**,** 2578.

CHU, J. S., BAILLIE, D. L. & CHEN, N. 2010. Convergent evolution of RFX transcription factors and ciliary genes predated the origin of metazoans. *BMC evolutionary biology,* 10**,** 1-10.

CIBULSKIS, K., LAWRENCE, M. S., CARTER, S. L., SIVACHENKO, A., JAFFE, D., SOUGNEZ, C., GABRIEL, S., MEYERSON, M., LANDER, E. S. & GETZ, G. 2013. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology,* 31**,** 213-219.

COCCHI, L., SCARCELLI, V., PULITI, A., BARALE, R. & SBRANA, I. 2005. Endogenous sex hormones affect the mutagen-induced chromosome damage by altering a caffeine-sensitive checkpoint. *Mutat Res,* 570**,** 281-8.

COOK, C. B., ARMSTRONG, L., BOERKOEL, C. F., CLARKE, L. A., DU SOUICH, C., DEMOS, M. K., GIBSON, W. T., GILL, H., LOPEZ, E., PATEL, M. S., SELBY, K., ABU-SHARAR, Z., STUDY, C., ELLIOTT, A. M. & FRIEDMAN, J. M. 2021. Somatic mosaicism detected by genome-wide sequencing in 500 parent–child trios with suspected genetic disease: clinical and genetic counseling implications. *Molecular Case Studies,* 7.

COOMBS, C. C., ZEHIR, A., DEVLIN, S. M., KISHTAGARI, A., SYED, A., JONSSON, P., HYMAN, D. M., SOLIT, D. B., ROBSON, M. E. & BASELGA, J. 2017. Therapy-related clonal hematopoiesis in patients with non-hematologic cancers is common and associated with adverse clinical outcomes. *Cell stem cell,* 21**,** 374-382.

COORENS, T. H. H., COLLORD, G., JUNG, H., WANG, Y., MOORE, L., HOOKS, Y., MAHBUBANI, K., SAEB-PARSY, K., CAMPBELL, P. J., MARTINCORENA, I., LEUNG, S. Y. & STRATTON, M. R. 2024. The somatic mutation landscape of normal gastric epithelium. *bioRxiv***,** 2024.03.17.585238.

COORENS, T. H. H., MOORE, L., ROBINSON, P. S., SANGHVI, R., CHRISTOPHER, J., HEWINSON, J., PRZYBILLA, M. J., LAWSON, A. R. J., SPENCER CHAPMAN, M. & CAGAN, A. 2021. Extensive phylogenies of human development inferred from somatic mutations. *Nature,* 597**,** 387-392.

COPELAND, H., KIVUVA, E., FIRTH, H. V. & WRIGHT, C. F. 2021. Systematic assessment of outcomes following a genetic diagnosis identified through a large-scale research study into developmental disorders. *Genetics in Medicine,* 23**,** 1058-1064.

COSENZA, M. R., RODRIGUEZ-MARTIN, B. & KORBEL, J. O. 2022. Structural variation in cancer: role, prevalence, and mechanisms. *Annual Review of Genomics and Human Genetics,* 23**,** 123-152.

CUI, X., ZHAO, X. & LIANG, Y. 2022. Sex differences in normal and malignant hematopoiesis. *Blood Sci,* 4**,** 185-191.

CURRIE, C. E., FORD, E., BENHAM WHYTE, L., TAYLOR, D. M., MIHALAS, B. P., ERENT, M., MARSTON, A. L., HARTSHORNE, G. M. & MCAINSH, A. D. 2022. The first mitotic division of human embryos is highly error prone. *Nature Communications,* 13**,** 6755.

DAI, X., YAN, X., WINTERGERST, K. A., CAI, L., KELLER, B. B. & TAN, Y. 2020. Nrf2: Redox and Metabolic Regulator of Stem Cell State and Function. *Trends in Molecular Medicine,* 26**,** 185-200.

DANECEK, P., BONFIELD, J. K., LIDDLE, J., MARSHALL, J., OHAN, V., POLLARD, M. O., WHITWHAM, A., KEANE, T., MCCARTHY, S. A., DAVIES, R. M. & LI, H. 2021. Twelve years of SAMtools and BCFtools. *GigaScience,* 10**,** giab008.

DARWIN, C. 1859. *On the origin of species by means of natural selection, or, The preservation of favoured races in the struggle for life.,* London, John Murray.

DAWES, R., BOURNAZOS, A. M., BRYEN, S. J., BOMMIREDDIPALLI, S., MARCHANT, R. G., JOSHI, H. & COOPER, S. T. 2023. SpliceVault predicts the precise nature of variant-associated mis-splicing. *Nature Genetics,* 55**,** 324-332.

DAWOUD, A. A. Z., GILBERT, R. D., TAPPER, W. J. & CROSS, N. C. P. 2022. Clonal myelopoiesis promotes adverse outcomes in chronic kidney disease. *Leukemia,* 36**,** 507-515.

DE ALMEIDA, A. J. P. O., DE OLIVEIRA, J. C. P. L., DA SILVA PONTES, L. V., DE SOUZA JÚNIOR, J. F., GONÇALVES, T. A. F., DANTAS, S. H., DE ALMEIDA FEITOSA, M. S., SILVA, A. O. & DE MEDEIROS, I. A. 2022. ROS: Basic Concepts, Sources, Cellular Signaling, and its Implications in Aging Pathways. *Oxidative Medicine and Cellular Longevity,* 2022**,** 1225578.

DE BONT, R. & VAN LAREBEKE, N. 2004. Endogenous DNA damage in humans: a review of quantitative data. *Mutagenesis,* 19**,** 169-185.

DELANEAU, O., ZAGURY, J.-F., ROBINSON, M. R., MARCHINI, J. L. & DERMITZAKIS, E. T. 2019. Accurate, scalable and integrative haplotype estimation. *Nature Communications,* 10**,** 5436.

DEMANELIS, K., DELGADO, D. A., TONG, L., JASMINE, F., AHMED, A., ISLAM, T., PARVEZ, F., KIBRIYA, M. G., GRAZIANO, J. H., AHSAN, H. & PIERCE, B. L. 2023. Somatic loss of the Y chromosome is associated with arsenic exposure among Bangladeshi men. *International Journal of Epidemiology,* 52**,** 1035-1046.

DESAI, P., MENCIA-TRINCHANT, N., SAVENKOV, O., SIMON, M. S., CHEANG, G., LEE, S., SAMUEL, M., RITCHIE, E. K., GUZMAN, M. L., BALLMAN, K. V., ROBOZ, G. J. & HASSANE, D. C. 2018. Somatic mutations precede acute myeloid leukemia years before diagnosis. *Nature Medicine,* 24**,** 1015-1023.

DESDÍN-MICÓ, G., SOTO-HEREDERO, G., ARANDA, J. F., OLLER, J., CARRASCO, E., GABANDÉ-RODRÍGUEZ, E., BLANCO, E. M., ALFRANCA, A., CUSSÓ, L. & DESCO, M. 2020. T cells with dysfunctional mitochondria induce multimorbidity and premature senescence. *Science,* 368**,** 1371-1376.

DEVILLERS, R., DOS SANTOS, A., DESTOMBES, Q., LAPLANTE, M. & ELOWE, S. 2024. Recent insights into the causes and consequences of chromosome mis-segregation. *Oncogene,* 43**,** 3139-3150.

DHARAN, N. J., YEH, P., BLOCH, M., YEUNG, M. M., BAKER, D., GUINTO, J., ROTH, N., FTOUNI, S., OGNENOVSKA, K. & SMITH, D. 2021. HIV is associated with an increased risk of age-related clonal hematopoiesis among older adults. *Nature Medicine,* 27**,** 1006-1011.

DIANOV, G. L. & HÜBSCHER, U. 2013. Mammalian Base Excision Repair: the Forgotten Archangel. *Nucleic Acids Research,* 41**,** 3483-3490.

DOMOGALA, D. D., GAMBIN, T., ZEMET, R., WU, C. W., SCHULZE, K. V., YANG, Y., WILSON, T. A., MACHOL, I., LIU, P. & STANKIEWICZ, P. 2021. Detection of low-level parental somatic mosaicism for clinically relevant SNVs and indels identified in a large exome sequencing dataset. *Human Genomics,* 15**,** 1-11.

DOU, Y., GOLD, H. D., LUQUETTE, L. J. & PARK, P. J. 2018. Detecting Somatic Mutations in Normal Cells. *Trends in Genetics,* 34**,** 545-557.

DOU, Y., KWON, M., RODIN, R. E., CORTÉS-CIRIANO, I., DOAN, R., LUQUETTE, L. J., GALOR, A., BOHRSON, C., WALSH, C. A. & PARK, P. J. 2020. Accurate detection of mosaic variants in sequencing data without matched controls. *Nature Biotechnology,* 38**,** 314-319.

DOU, Y., YANG, X., LI, Z., WANG, S., ZHANG, Z., YE, A. Y., YAN, L., YANG, C., WU, Q., LI, J., ZHAO, B., HUANG, A. Y. & WEI, L. 2017. Postzygotic single-nucleotide mosaicisms contribute to the etiology of autism spectrum disorder and autistic traits and the origin of mutations. *Human Mutation,* 38**,** 1002-1013.

DUBOIS, F., SIDIROPOULOS, N., WEISCHENFELDT, J. & BEROUKHIM, R. 2022. Structural variations in cancer and the 3D genome. *Nature Reviews Cancer,* 22**,** 533-546.

DUFFY, K. A., HATHAWAY, E. R., KLEIN, S. D., GANGULY, A. & KALISH, J. M. 2021. Epigenetic mosaicism and cell burden in Beckwith–Wiedemann syndrome due to loss of methylation at imprinting control region 2. *Molecular Case Studies,* 7.

DURKIE, M., CASSIDY, E.-J., BERRY, I., OWENS, M., TURNBULL, C., SCOTT, R. H., TAYLOR, R. W., DEANS, Z. C., ELLARD, S., BAPLE, E. L. & MCMULLAN, D. J. 2024. ACGS Best Practice Guidelines for Variant Classification in Rare Disease 2024. *ACGS Best Practice Guidelines for Variant Classification in Rare Disease.* United Kingdom: Association for Clinical Genomic Science.

DUTTA, A. & SCHACHERER, J. 2025. The dynamics of loss of heterozygosity events in genomes. *EMBO Rep,* 26**,** 602-612.

EBERHARDT, R. Y., WRIGHT, C. F., FITZPATRICK, D. R., HURLES, M. E. & FIRTH, H. V. 2023. Detection of mosaic chromosomal alterations in children with severe developmental disorders recruited to the DDD study. *Genetics in Medicine Open,* 1**,** 100836.

ELLIS, P., MOORE, L., SANDERS, M. A., BUTLER, T. M., BRUNNER, S. F., LEE-SIX, H., OSBORNE, R., FARR, B., COORENS, T. H. H. & LAWSON, A. R. J. 2021. Reliable detection of somatic mutations in solid tissues by laser-capture microdissection and low-input DNA sequencing. *Nature Protocols,* 16**,** 841-871.

ELTAWANSY, S. A., BAKOS, A. & CHECTON, J. 2015. Noncompaction Cardiomyopathy with Charcot-Marie-Tooth Disease. *Case Reports in Cardiology,* 2015**,** 646890.

ESSADDAM, L., ZITOUNI, O., KRAOUA, L., TRABELSI, M., SASSI, H., KMIHA, S., CHARFI, F., EL GUICHE, D., KEBAÏLI, R., JABALLAH, N., RJEB, M., ZOUARI, N., EL ARIBI, Y., HIZEM, S., WANNES, S., FKIH ROMDHANE, I., SFAR, M. T., BEN HAMOUDA, H., HADJ SALEM, R., ... & BEN BECHER, S. 2023. Turner syndrome: results of the first Tunisian study group on Turner syndrome (TuSGOT). *J Pediatr Endocrinol Metab,* 36**,** 577-583.

ESSOP, F. B. & KRAUSE, A. 2013. Diagnostic, carrier and prenatal genetic testing for fragile X syndrome and other FMR-1-related disorders in Johannesburg, South Africa: a 20-year review: Festschrift: Professor Trefor Jenkins-molecular genetics. *South African medical journal,* 103**,** 994-998.

EVANS, M. A. & WALSH, K. 2022. Clonal hematopoiesis, somatic mosaicism, and age-associated disease. *Physiological Reviews,* 103**,** 649-716.

FABRE, M. A., DE ALMEIDA, J. G., FIORILLO, E., MITCHELL, E., DAMASKOU, A., RAK, J., ORRÙ, V., MARONGIU, M., CHAPMAN, M. S., VIJAYABASKAR, M. S., BAXTER, J., HARDY, C., ABASCAL, F., WILLIAMS, N., NANGALIA, J., MARTINCORENA, I., CAMPBELL, P. J., MCKINNEY, E. F., CUCCA, F., ... & VASSILIOU, G. S. 2022. The longitudinal dynamics and natural history of clonal haematopoiesis. *Nature,* 606**,** 335-342.

FAILLA, G. 1958. The aging process and cancerogenesis.

FASINU, P. & ORISAKWE, O. E. 2013. Heavy metal pollution in sub-Saharan Africa and possible implications in cancer epidemiology. *Asian Pac J Cancer Prev,* 14**,** 3393-402.

FATUMO, S., CHIKOWORE, T., CHOUDHURY, A., AYUB, M., MARTIN, A. R. & KUCHENBAECKER, K. 2022. A roadmap to increase diversity in genomic studies. *Nature Medicine,* 28**,** 243-250.

FAYOMI, A. P. & ORWIG, K. E. 2018. Spermatogonial stem cells and spermatogenesis in mice, monkeys and men. *Stem cell research,* 29**,** 207-214.

FENG, B.-J. 2017. PERCH: A Unified Framework for Disease Gene Prioritization. *Human Mutation,* 38**,** 243-251.
FIRTH, H. V., RICHARDS, S. M., BEVAN, A. P., CLAYTON, S., CORPAS, M., RAJAN, D., VOOREN, S. V., MOREAU, Y., PETTETT, R. M. & CARTER, N. P. 2009. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources. *The American Journal of Human Genetics,* 84**,** 524-533.
FLACH, J., BAKKER, S. T., MOHRIN, M., CONROY, P. C., PIETRAS, E. M., REYNAUD, D., ALVAREZ, S., DIOLAITI, M. E., UGARTE, F., FORSBERG, E. C., LE BEAU, M. M., STOHR, B. A., MÉNDEZ, J., MORRISON, C. G. & PASSEGUÉ, E. 2014. Replication stress is a potent driver of functional decline in ageing haematopoietic stem cells. *Nature,* 512**,** 198-202.
FLEMMING, W. 1882. *Zellsubstanz, kern und zelltheilung,* Leipzig, Vogel.
FLIEDNER, T. M., GRAESSLE, D., PAULSEN, C. & REIMERS, K. 2002. Structure and Function of Bone Marrow Hemopoiesis: Mechanisms of Response to Ionizing Radiation Exposure. *Cancer Biotherapy and Radiopharmaceuticals,* 17**,** 405-426.
FORD, C. E., POLANI, P. E., BRIGGS, J. H. & BISHOP, P. M. F. 1959. A Presumptive Human XXY/XX Mosaic. *Nature,* 183**,** 1030-1032.
FORTES-LIMA, C. A., BURGARELLA, C., HAMMARÉN, R., ERIKSSON, A., VICENTE, M., JOLLY, C., SEMO, A., GUNNINK, H., PACCHIAROTTI, S., MUNDEKE, L., MATONDA, I., MULUWA, J. K., COUTROS, P., NYAMBE, T. S., CIKOMOLA, J. C., COETZEE, V., DE CASTRO, M., EBBESEN, P., DELANGHE, J., ... & SCHLEBUSCH, C. M. 2024. The genetic legacy of the expansion of Bantu-speaking peoples in Africa. *Nature,* 625**,** 540-547.
FRANCESCHI, C., BONAFÈ, M., VALENSIN, S., OLIVIERI, F., DE LUCA, M., OTTAVIANI, E. & DE BENEDICTIS, G. 2000. Inflamm-aging: An Evolutionary Perspective on Immunosenescence. *Annals of the New York Academy of Sciences,* 908**,** 244-254.
FRANCIS, M., GORMAN, B. R., BIGDELI, T. B., GENOVESE, G., VOLOUDAKIS, G., BENDL, J., ZENG, B., VENKATESH, S., CHATZINAKOS, C., MCAULEY, E., JI, S.-G., MARKIANOS, K., SCHREINER, P. A., PARTAN, E., SHI, Y., DEVINENI, P., PROGRAM, V. A. M. V., MOSER, J., MURALIDHAR, S., ... & PYARAJAN, S. 2024. Multi-ancestry genome-wide association meta-analysis of mosaic loss of chromosome Y in the Million Veteran Program identifies 167 novel loci. *medRxiv***,** 2024.04.24.24306301.
FRANKLIN, R. E. & GOSLING, R. G. 1953. Molecular configuration in sodium thymonucleate. *Nature,* 171**,** 740-741.
FREED, D. & PEVSNER, J. 2016. The Contribution of Mosaic Variants to Autism Spectrum Disorder. *PLoS Genet,* 12**,** e1006245.
FUJINO, T., ASADA, S., GOYAMA, S. & KITAMURA, T. 2022. Mechanisms involved in hematopoietic stem cell aging. *Cellular and Molecular Life Sciences,* 79**,** 473.
GAMBIN, T., LIU, Q., KAROLAK, J. A., GROCHOWSKI, C. M., XIE, N. G., WU, L. R., YAN, Y. H., CAO, Y., COBAN AKDEMIR, Z. H. & WILSON, T. A. 2020. Low-level parental somatic mosaic SNVs in exomes from a large cohort of trios with diverse suspected Mendelian conditions. *Genetics in Medicine,* 22**,** 1768-1776.
GAO, D., ZHU, B., CAO, X., ZHANG, M. & WANG, X. 2019. Roles of NIPBL in maintenance of genome stability. *Seminars in Cell & Developmental Biology,* 90**,** 181-186.
GAO, T., KASTRITI, M. E., LJUNGSTRÖM, V., HEINZEL, A., TISCHLER, A. S., OBERBAUER, R., LOH, P. R., ADAMEYKO, I., PARK, P. J. & KHARCHENKO,

P. V. 2023. A pan-tissue survey of mosaic chromosomal alterations in 948 individuals. *Nat Genet,* 55**,** 1901-1911.

GAO, T., PTASHKIN, R., BOLTON, K. L., SIRENKO, M., FONG, C., SPITZER, B., MENGHRAJANI, K., OSSA, J. E. A., ZHOU, Y., BERNARD, E., LEVINE, M., MARTINEZ, J. S. M., ZHANG, Y., FRANCH-EXPÓSITO, S., PATEL, M., BRAUNSTEIN, L. Z., KELLY, D., YABE, M., BENAYED, R., ... & PAPAEMMANUIL, E. 2021. Interplay between chromosomal alterations and gene mutations shapes the evolutionary trajectory of clonal hematopoiesis. *Nature Communications,* 12**,** 338.

GARCIA, P., FERNANDEZ-HERNANDEZ, R., CUADRADO, A., COCA, I., GOMEZ, A., MAQUEDA, M., LATORRE-PELLICER, A., PUISAC, B., RAMOS, F. J., SANDOVAL, J., ESTELLER, M., MOSQUERA, J. L., RODRIGUEZ, J., PIÉ, J., LOSADA, A. & QUERALT, E. 2021. Disruption of NIPBL/Scc2 in Cornelia de Lange Syndrome provokes cohesin genome-wide redistribution with an impact in the transcriptome. *Nature Communications,* 12**,** 4551.

GAZIANO, J. M., CONCATO, J., BROPHY, M., FIORE, L., PYARAJAN, S., BREELING, J., WHITBOURNE, S., DEEN, J., SHANNON, C., HUMPHRIES, D., GUARINO, P., ASLAN, M., ANDERSON, D., LAFLEUR, R., HAMMOND, T., SCHAA, K., MOSER, J., HUANG, G., MURALIDHAR, S., ... & O'LEARY, T. J. 2016. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol,* 70**,** 214-23.

GEIGER, H., FURUTA, Y., VAN WYK, S., PHILLIPS, J. A. & TINKER, R. J. 2024. The Clinical Spectrum of Mosaic Genetic Disease. *Genes (Basel),* 15.

GELMAN, A. & AUERBACH, J. 2016. Age-aggregation bias in mortality trends. *Proc Natl Acad Sci U S A,* 113**,** E816-7.

GENOVESE, G., KÄHLER, A. K., HANDSAKER, R. E., LINDBERG, J., ROSE, S. A., BAKHOUM, S. F., CHAMBERT, K., MICK, E., NEALE, B. M. & FROMER, M. 2014. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *New England Journal of Medicine,* 371**,** 2477-2487.

GIL-SANTANA, L., CRUZ, L. A. B., ARRIAGA, M. B., MIRANDA, P. F. C., FUKUTANI, K. F., SILVEIRA-MATTOS, P. S., SILVA, E. C., OLIVEIRA, M. G., MESQUITA, E. D. D., RAUWERDINK, A., COBELENS, F., OLIVEIRA, M. M., KRITSKI, A. & ANDRADE, B. B. 2019. Tuberculosis-associated anemia is linked to a distinct inflammatory profile that persists after initiation of antitubercular therapy. *Scientific Reports,* 9**,** 1381.

GLESSNER, J. T., CHANG, X., LIU, Y., LI, J., KHAN, M., WEI, Z., SLEIMAN, P. M. A. & HAKONARSON, H. 2021. MONTAGE: a new tool for high-throughput detection of mosaic copy number variation. *BMC Genomics,* 22**,** 133.

GOLDMAN, E. A., SPELLMAN, P. T. & AGARWAL, A. 2023. Defining clonal hematopoiesis of indeterminate potential: evolutionary dynamics and detection under aging and inflammation. *Molecular Case Studies,* 9**,** a006251.

GOLDMAN, S. L., MACKAY, M., AFSHINNEKOO, E., MELNICK, A. M., WU, S. & MASON, C. E. 2019. The Impact of Heterogeneity on Single-Cell Sequencing. *Front Genet,* 10**,** 8.

GOMBAULT, C., GRENET, G., SEGUREL, L., DURET, L., GUEYFFIER, F., CATHÉBRAS, P., PONTIER, D., MAINBOURG, S., SANCHEZ-MAZAS, A. & LEGA, J.-C. 2023. Population designations in biomedical research: Limitations and perspectives. *HLA,* 101**,** 3-15.

GOMES, M. C. M., GOMES, B. J., CHRISTIANSON, A. L., BAILLY, C., MCKERROW, N. & MALHERBE, H. L. 2024. A Capacity Audit of Medical Geneticists and Genetic Counsellors in South Africa, 2024: A National Crisis. *Genes* [Online], 15.

GONZÁLEZ, J. R., RODRÍGUEZ-SANTIAGO, B., CÁCERES, A., PIQUE-REGI, R., ROTHMAN, N., CHANOCK, S. J., ARMENGOL, L. & PÉREZ-JURADO, L. A. 2011. A fast and accurate method to detect allelic genomic imbalances underlying mosaic rearrangements using SNP array data. *BMC bioinformatics,* 12**,** 166-166.

GOSHO, M., OHIGASHI, T., NAGASHIMA, K., ITO, Y. & MARUO, K. 2023. Bias in Odds Ratios From Logistic Regression Methods With Sparse Data Sets. *J Epidemiol,* 33**,** 265-275.

GREGORY, T. R. 2009. Understanding Natural Selection: Essential Concepts and Common Misconceptions. *Evolution: Education and Outreach,* 2**,** 156-175.

GROVER, A., SANJUAN-PLA, A., THONGJUEA, S., CARRELHA, J., GIUSTACCHINI, A., GAMBARDELLA, A., MACAULAY, I., MANCINI, E., LUIS, T. C., MEAD, A., JACOBSEN, S. E. W. & NERLOV, C. 2016. Single-cell RNA sequencing reveals molecular and functional platelet bias of aged haematopoietic stem cells. *Nature Communications,* 7**,** 11075.

GRUCA-STRYJAK, K., DODA-NOWAK, E., DZIERLA, J., WRÓBEL, K., SZYMANKIEWICZ-BRĘBOROWICZ, M. & MAZELA, J. 2024. Advancing the Clinical and Molecular Understanding of Cornelia de Lange Syndrome: A Multidisciplinary Pediatric Case Series and Review of the Literature. *J Clin Med,* 13.

GRUHN, J. R., ZIELINSKA, A. P., SHUKLA, V., BLANSHARD, R., CAPALBO, A., CIMADOMO, D., NIKIFOROV, D., CHAN, A. C.-H., NEWNHAM, L. J., VOGEL, I., SCARICA, C., KRAPCHEV, M., TAYLOR, D., KRISTENSEN, S. G., CHENG, J., ERNST, E., BJØRN, A.-M. B., COLMORN, L. B., BLAYNEY, M., ... & HOFFMANN, E. R. 2019. Chromosome errors in human eggs shape natural fertility over reproductive life span. *Science,* 365**,** 1466-1469.

GRUND, B., BAKER, J. V., DEEKS, S. G., WOLFSON, J., WENTWORTH, D., COZZI-LEPRI, A., COHEN, C. J., PHILLIPS, A., LUNDGREN, J. D. & NEATON, J. D. 2016. Relevance of interleukin-6 and D-dimer for serious non-AIDS morbidity and death among HIV-positive adults on suppressive antiretroviral therapy. *PloS one,* 11**,** e0155100.

GU, Z., GU, L., EILS, R., SCHLESNER, M. & BRORS, B. 2014. circlize implements and enhances circular visualization in R. *Bioinformatics,* 30**,** 2811-2812.

GUAN, X., MENG, X., ZHONG, G., ZHANG, Z., WANG, C., XIAO, Y., FU, M., ZHAO, H., ZHOU, Y., HONG, S., XU, X., BAI, Y., KAN, H., CHEN, R., WU, T. & GUO, H. 2024. Particulate matter pollution, polygenic risk score and mosaic loss of chromosome Y in middle-aged and older men from the Dongfeng-Tongji cohort study. *Journal of Hazardous Materials,* 471**,** 134315.

GUDMUNDSSON, S., SINGER-BERK, M., WATTS, N. A., PHU, W., GOODRICH, J. K., SOLOMONSON, M., GENOME AGGREGATION DATABASE, C., REHM, H. L., MACARTHUR, D. G. & O'DONNELL-LURIA, A. 2022. Variant interpretation using population databases: Lessons from gnomAD. *Human Mutation,* 43**,** 1012-1030.

GURDASANI, D., CARSTENSEN, T., FATUMO, S., CHEN, G., FRANKLIN, C. S., PRADO-MARTINEZ, J., BOUMAN, H., ABASCAL, F., HABER, M. & TACHMAZIDOU, I. 2019. Uganda genome resource enables insights into population history and genomic discovery in Africa. *Cell,* 179**,** 984-1002.

HA, Y.-J., KANG, S., KIM, J., KIM, J., JO, S.-Y. & KIM, S. 2023. Comprehensive benchmarking and guidelines of mosaic variant calling strategies. *Nature Methods,* 20**,** 2058-2067.

HAAS, S., TRUMPP, A. & MILSOM, M. D. 2018. Causes and Consequences of Hematopoietic Stem Cell Heterogeneity. *Cell Stem Cell,* 22**,** 627-638.

HAMILTON, M. L., GUO, Z., FULLER, C. D., VAN REMMEN, H., WARD, W. F., AUSTAD, S. N., TROYER, D. A., THOMPSON, I. & RICHARDSON, A. 2001. A reliable assessment of 8-oxo-2-deoxyguanosine levels in nuclear and mitochondrial DNA using the sodium iodide method to isolate DNA. *Nucleic Acids Research,* 29**,** 2117-2126.

HAMOSH, A., SCOTT, A. F., AMBERGER, J. S., BOCCHINI, C. A. & MCKUSICK, V. A. 2005. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research,* 33**,** D514-D517.

HAN, W. M., SAZZAD, H. M. S., BLOCH, M., BAKER, D. A., ROTH, N., BOWDEN-REID, E., SMITH, D. E., HOY, J. F., WOOLLEY, I., FINLAYSON, R., TEMPLETON, D. J., MATTHEWS, G. V., COSTELLO, J., DAWSON, M. A., DAWSON, S.-J., POLIZZOTTO, M. N., PETOUMENOS, K., YEH, P., DHARAN, N. J., ... & DHARAN, N. J. 2024. Age-related clonal hematopoiesis and HIV infection are associated with geriatric outcomes: The ARCHIVE study. *Cell Reports Medicine,* 5.

HAPPLE, R. 1993. Mosaicism in Human Skin: Understanding the Patterns and Mechanisms. *Archives of Dermatology,* 129**,** 1460-1470.

HARA, T., HASHIMOTO, Y., AKUZAWA, T., HIRAI, R., KOBAYASHI, H. & SATO, K. 2014. Rer1 and calnexin regulate endoplasmic reticulum retention of a peripheral myelin protein 22 mutant that causes type 1A Charcot-Marie-Tooth disease. *Scientific Reports,* 4**,** 6992.

HARRIS, P. A., TAYLOR, R., MINOR, B. L., ELLIOTT, V., FERNANDEZ, M., O'NEAL, L., MCLEOD, L., DELACQUA, G., DELACQUA, F., KIRBY, J. & DUDA, S. N. 2019. The REDCap consortium: Building an international community of software platform partners. *Journal of Biomedical Informatics,* 95**,** 103208.

HATJE, K., MÜHLHAUSEN, S., SIMM, D. & KOLLMAR, M. 2019. The Protein-Coding Human Genome: Annotating High-Hanging Fruits. *BioEssays,* 41**,** 1900066.

HATTON, I. A., GALBRAITH, E. D., MERLEAU, N. S. C., MIETTINEN, T. P., SMITH, B. M. & SHANDER, J. A. 2023. The human cell count and size distribution. *Proc Natl Acad Sci U S A,* 120**,** e2303077120.

HE, J., KOU, S. H., LI, J., DING, X. & WANG, S. M. 2024. Pathogenic variants in human DNA damage repair genes mostly arose after the latest human out-of-Africa migration. *Front Genet,* 15**,** 1408952.

HEALTH EFFECTS INSTITUTE 2022. The state of air quality and health impacts in Africa. A report from the state of global air initiative. Boston: Health Effects Institute.

HERMS, A. & JONES, P. H. 2023. Somatic Mutations in Normal Tissues: New Perspectives on Early Carcinogenesis. *Annual Review of Cancer Biology,* 7**,** 189-205.

HERRICK, J. 2024. DNA Damage, Genome Stability, and Adaptation: A Question of Chance or Necessity? *Genes* [Online], 15.

HIROKAWA, N., NIWA, S. & TANAKA, Y. 2010. Molecular Motors in Neurons: Transport Mechanisms and Roles in Brain Function, Development, and Disease. *Neuron,* 68**,** 610-638.

HIRSCHHORN, K., DECKER WAYNE, H. & COOPER HERBERT, L. 1960. Human Intersex with Chromosome Mosaicism of Type XY/XO. *New England Journal of Medicine,* 263**,** 1044-1048.

HOLLAND, A. J. & CLEVELAND, D. W. 2012. Losing balance: the origin and impact of aneuploidy in cancer: 'Exploring aneuploidy: the significance of chromosomal imbalance'review series. *EMBO reports,* 13**,** 501-514.

HOLM, V. A. 1989. Developmental Disabilities: Delivery of Medical Care for Children and Adults. *JAMA,* 262**,** 2935-2936.

HORMOZDIARI, F., PENN, O., BORENSTEIN, E. & EICHLER, E. E. 2015. The discovery of integrated gene networks for autism and related disorders. *Genome research,* 25**,** 142-154.

HSIEH, A., MORTON, S. U., WILLCOX, J. A. L., GORHAM, J. M., TAI, A. C., QI, H., DEPALMA, S., MCKEAN, D., GRIFFIN, E., MANHEIMER, K. B., BERNSTEIN, D., KIM, R. W., NEWBURGER, J. W., PORTER, G. A., SRIVASTAVA, D., TRISTANI-FIROUZI, M., BRUECKNER, M., LIFTON, R. P., GOLDMUNTZ, E., ... & SHEN, Y. 2020. EM-mosaic detects mosaic point mutations that contribute to congenital heart disease. *Genome Medicine,* 12**,** 42.

HSIEH, P. & ZHANG, Y. 2017. The Devil is in the details for DNA mismatch repair. *Proc Natl Acad Sci U S A,* 114**,** 3552-3554.

HSU, D. C., BREGLIO, K. F., PEI, L., WONG, C.-S., ANDRADE, B. B., SHEIKH, V., SMELKINSON, M., PETROVAS, C., RUPERT, A. & GIL-SANTANA, L. 2018. Emergence of polyfunctional cytotoxic CD4+ T cells in Mycobacterium avium immune reconstitution inflammatory syndrome in human immunodeficiency virus-infected patients. *Clinical Infectious Diseases,* 67**,** 437-446.

HUANG, A. Y. & LEE, E. A. 2022. Identification of Somatic Mutations From Bulk and Single-Cell Sequencing Data. *Frontiers in Aging,* 2.

HUANG, A. Y., ZHANG, Z., YE, A. Y., DOU, Y., YAN, L., YANG, X., ZHANG, Y. & WEI, L. 2017. MosaicHunter: accurate detection of postzygotic single-nucleotide mosaicism through next-generation sequencing of unpaired, trio, and paired samples. *Nucleic Acids Res,* 45**,** e76.

HUBBARD, A. K., BROWN, D. W. & MACHIELA, M. J. 2023. Clonal hematopoiesis due to mosaic chromosomal alterations: Impact on disease risk and mortality. *Leukemia Research,* 126**,** 107022.

HUISMAN, S. A., REDEKER, E. J. W., MAAS, S. M., MANNENS, M. M. & HENNEKAM, R. C. M. 2013. High rate of mosaicism in individuals with Cornelia de Lange syndrome. *Journal of Medical Genetics,* 50**,** 339.

HUPPKE, P., WEISSBACH, S., CHURCH, J. A., SCHNUR, R., KRUSEN, M., DREHA-KULACZEWSKI, S., KÜHN-VELTEN, W. N., WOLF, A., HUPPKE, B., MILLAN, F., BEGTRUP, A., ALMUSAFRI, F., THIELE, H., ALTMÜLLER, J., NÜRNBERG, P., MÜLLER, M. & GÄRTNER, J. 2017. Activating de novo mutations in NFE2L2 encoding NRF2 cause a multisystem disorder. *Nature Communications,* 8**,** 818.

IOANNIDIS, N. M., ROTHSTEIN, J. H., PEJAVER, V., MIDDHA, S., MCDONNELL, S. K., BAHETI, S., MUSOLF, A., LI, Q., HOLZINGER, E. & KARYADI, D. 2016. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *The American Journal of Human Genetics,* 99**,** 877-885.

ITO, K., TAKAHASHI, A., MORITA, M., SUZUKI, T. & YAMAMOTO, T. 2011. The role of the CNOT1 subunit of the CCR4-NOT complex in mRNA deadenylation and cell viability. *Protein & Cell,* 2**,** 755-763.

JACKSON, S. P. & BARTEK, J. 2009. The DNA-damage response in human biology and disease. *Nature,* 461**,** 1071-1078.

JAGANATHAN, K., KYRIAZOPOULOU PANAGIOTOPOULOU, S., MCRAE, J. F., DARBANDI, S. F., KNOWLES, D., LI, Y. I., KOSMICKI, J. A., ARBELAEZ, J., CUI, W., SCHWARTZ, G. B., CHOW, E. D., KANTERAKIS, E., GAO, H., KIA, A., BATZOGLOU, S., SANDERS, S. J. & FARH, K. K.-H. 2019. Predicting Splicing from Primary Sequence with Deep Learning. *Cell,* 176**,** 535-548.e24.

JAGANNATHAN-BOGDAN, M. & ZON, L. I. 2013. Hematopoiesis. *Development,* 140**,** 2463-2467.

JAISWAL, S. & EBERT, B. L. 2019. Clonal hematopoiesis in human aging and disease. *Science,* 366**,** eaan4673.

JAISWAL, S., NATARAJAN, P., SILVER, A. J., GIBSON, C. J., BICK, A. G., SHVARTZ, E., MCCONKEY, M., GUPTA, N., GABRIEL, S., ARDISSINO, D., BABER, U., MEHRAN, R., FUSTER, V., DANESH, J., FROSSARD, P., SALEHEEN, D., MELANDER, O., SUKHOVA, G. K., NEUBERG, D., ... & EBERT, B. L. 2017. Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. *New England Journal of Medicine,* 377**,** 111-121.

JAKUBEK, Y. A., MA, X., STILP, A. M., YU, F., BACON, J., WONG, J. W., AGUET, F., ARDLIE, K., ARNETT, D. K., BARNES, K., BIS, J. C., BLACKWELL, T., BECKER, L. C., BOERWINKLE, E., BOWLER, R. P., BUDOFF, M. J., CARSON, A. P., CHEN, J., CHO, M. H., ... & AUER, P. L. 2025. Genomic and phenotypic correlates of mosaic loss of chromosome Y in blood. *The American Journal of Human Genetics,* 112**,** 276-290.

JAKUBEK, Y. A., REINER, A. P. & HONIGBERG, M. C. 2023a. Risk factors for clonal hematopoiesis of indeterminate potential and mosaic chromosomal alterations. *Translational Research,* 255**,** 171-180.

JAKUBEK, Y. A., ZHOU, Y., STILP, A., BACON, J., WONG, J. W., OZCAN, Z., ARNETT, D., BARNES, K., BIS, J. C., BOERWINKLE, E., BRODY, J. A., CARSON, A. P., CHASMAN, D. I., CHEN, J., CHO, M., CONOMOS, M. P., COX, N., DOYLE, M. F., FORNAGE, M., ... & AUER, P. L. 2023b. Mosaic chromosomal alterations in blood across ancestries using whole-genome sequencing. *Nature Genetics,* 55**,** 1912-1919.

JANG, W., PARK, J., CHAE, H. & KIM, M. 2022. Comparison of *In Silico* Tools for Splice-Altering Variant Prediction Using Established Spliceogenic Variants: An End-User's Point of View. *International Journal of Genomics,* 2022**,** 5265686.

JU, Y. S., MARTINCORENA, I., GERSTUNG, M., PETLJAK, M., ALEXANDROV, L. B., RAHBARI, R., WEDGE, D. C., DAVIES, H. R., RAMAKRISHNA, M. & FULLAM, A. 2017. Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature,* 543**,** 714-718.

JÓNSSON, H., SULEM, P., KEHR, B., KRISTMUNDSDOTTIR, S., ZINK, F., HJARTARSON, E., HARDARSON, M. T., HJORLEIFSSON, K. E., EGGERTSSON, H. P. & GUDJONSSON, S. A. 2017. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature,* 549**,** 519-522.

KAMP, M., KRAUSE, A. & RAMSAY, M. 2021. Has translational genomics come of age in Africa? *Human Molecular Genetics,* 30**,** R164-R173.

KANDARAKOV, O., BELYAVSKY, A. & SEMENOVA, E. 2022. Bone Marrow Niches of Hematopoietic Stem and Progenitor Cells. *International Journal of Molecular Sciences* [Online], 23.

KAPADIA, C. D. & GOODELL, M. A. 2024. Tissue mosaicism following stem cell aging: blood as an exemplar. *Nat Aging,* 4**,** 295-308.

KARACA, E., POSEY, J. E., COBAN AKDEMIR, Z., PEHLIVAN, D., HAREL, T., JHANGIANI, S. N., BAYRAM, Y., SONG, X., BAHRAMBEIGI, V. & YUREGIR, O. O. 2018. Phenotypic expansion illuminates multilocus pathogenic variation. *Genetics in Medicine,* 20**,** 1528-1537.

KARCZEWSKI, K. J., FRANCIOLI, L. C., TIAO, G., CUMMINGS, B. B., ALFÖLDI, J., WANG, Q., COLLINS, R. L., LARICCHIA, K. M., GANNA, A., BIRNBAUM, D. P., GAUTHIER, L. D., BRAND, H., SOLOMONSON, M., WATTS, N. A., RHODES, D., SINGER-BERK, M., ENGLAND, E. M., SEABY, E. G., KOSMICKI, J. A., ... & GENOME AGGREGATION DATABASE, C. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature,* 581**,** 434-443.

KAROLCHIK, D., HINRICHS, A. S., FUREY, T. S., ROSKIN, K. M., SUGNET, C. W., HAUSSLER, D. & KENT, W. J. 2004. The UCSC Table Browser data retrieval tool. *Nucleic acids research,* 32**,** D493-D496.

KAUSHANSKY, K. 2006. Lineage-Specific Hematopoietic Growth Factors. *New England Journal of Medicine,* 354**,** 2034-2045.

KENT, W. J., SUGNET, C. W., FUREY, T. S., ROSKIN, K. M., PRINGLE, T. H., ZAHLER, A. M. & HAUSSLER, D. 2002. The human genome browser at UCSC. *Genome research,* 12**,** 996-1006.

KERN, A. D. & HAHN, M. W. 2018. The Neutral Theory in Light of Natural Selection. *Molecular Biology and Evolution,* 35**,** 1366-1371.

KESSLER, M. D., DAMASK, A., O'KEEFFE, S., BANERJEE, N., LI, D., WATANABE, K., MARKETTA, A., VAN METER, M., SEMRAU, S., HOROWITZ, J., TANG, J., KOSMICKI, J. A., RAJAGOPAL, V. M., ZOU, Y., HOUVRAS, Y., GHOSH, A., GILLIES, C., MBATCHOU, J., WHITE, R. R., ... & COLLABORATION, G.-R. D. 2022. Common and rare variant associations with clonal haematopoiesis phenotypes. *Nature,* 612**,** 301-309.

KESSLER, M. D., LOESCH, D. P., PERRY, J. A., HEARD-COSTA, N. L., TALIUN, D., CADE, B. E., WANG, H., DAYA, M., ZINITI, J., DATTA, S., CELEDÓN, J. C., SOTO-QUIROS, M. E., AVILA, L., WEISS, S. T., BARNES, K., REDLINE, S. S., VASAN, R. S., JOHNSON, A. D., MATHIAS, R. A., ... & ZOELLNER, S. 2020. De novo mutations across 1,465 diverse genomes reveal mutational insights and reductions in the Amish founder population. *Proceedings of the National Academy of Sciences,* 117**,** 2560-2569.

KHAN, N., DOWNEY, J., SANZ, J., KAUFMANN, E., BLANKENHAUS, B., PACIS, A., PERNET, E., AHMED, E., CARDOSO, S., NIJNIK, A., MAZER, B., SASSETTI, C., BEHR, M. A., SOARES, M. P., BARREIRO, L. B. & DIVANGAHI, M. 2020. M. tuberculosis Reprograms Hematopoietic Stem Cells to Limit Myelopoiesis and Impair Trained Immunity. *Cell,* 183**,** 752-770.e22.

KIM, J., KIM, D., LIM, J. S., MAENG, J. H., SON, H., KANG, H. C., NAM, H., LEE, J. H. & KIM, S. 2019. The use of technical replication for detection of low-level somatic mutations in next-generation sequencing. *Nat Commun,* 10**,** 1047.

KIM, P. G., NIROULA, A., SHKOLNIK, V., MCCONKEY, M., LIN, A. E., SŁABICKI, M., KEMP, J. P., BICK, A., GIBSON, C. J. & GRIFFIN, G. 2021. Dnmt3a-mutated clonal hematopoiesis promotes osteoporosis. *Journal of Experimental Medicine,* 218**,** e20211872.

KIM, S. H., KWON, S. S., PARK, M. R., LEE, H. A., KIM, J. H., CHA, J., KIM, S., BAEK, S. T., LEE, J. S., KIM, H. D., CHOI, J. R., LEE, S. T. & KANG, H. C. 2023. Detecting Low-Variant Allele Frequency Mosaic Pathogenic Variants of NF1, TSC2, and AKT3 Genes from Blood in Patients with Neurodevelopmental Disorders. *J Mol Diagn,* 25**,** 583-591.

KING, D. A., JONES, W. D., CROW, Y. J., DOMINICZAK, A. F., FOSTER, N. A., GAUNT, T. R., HARRIS, J., HELLENS, S. W., HOMFRAY, T., INNES, J., JONES, E. A., JOSS, S., KULKARNI, A., MANSOUR, S., MORRIS, A. D., PARKER, M. J., PORTEOUS, D. J., SHIHAB, H. A., SMITH, B. H., ... & STUDY, T. D. D. D. 2015. Mosaic structural variation in children with developmental disorders. *Human Molecular Genetics,* 24**,** 2733-2745.

KING, D. A., SIFRIM, A., FITZGERALD, T. W., RAHBARI, R., HOBSON, E., HOMFRAY, T., MANSOUR, S., MEHTA, S. G., SHEHLA, M., TOMKINS, S. E., VASUDEVAN, P. C., HURLES, M. E. & STUDY, T. D. D. D. 2017. Detection of structural mosaicism from targeted and whole-genome sequencing data. *Genome Research,* 27**,** 1704-1714.

KIRCHER, M., WITTEN, D. M., JAIN, P., O'ROAK, B. J., COOPER, G. M. & SHENDURE, J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics,* 46**,** 310-315.
KIRKWOOD, THOMAS B. L. & MELOV, S. 2011. On the Programmed/Non-Programmed Nature of Ageing within the Life History. *Current Biology,* 21**,** R701-R707.
KLOOSTERMAN, W. P., FRANCIOLI, L. C., HORMOZDIARI, F., MARSCHALL, T., HEHIR-KWA, J. Y., ABDELLAOUI, A., LAMEIJER, E.-W., MOED, M. H., KOVAL, V. & RENKENS, I. 2015. Characteristics of de novo structural changes in the human genome. *Genome research,* 25**,** 792-801.
KNUDSON, A. G. 1971. Mutation and Cancer: Statistical Study of Retinoblastoma. *Proceedings of the National Academy of Sciences,* 68**,** 820-823.
KOBOLDT, D. C. 2020. Best practices for variant calling in clinical sequencing. *Genome Medicine,* 12**,** 91.
KOGUT, I., SCHOLZ, J. L., CANCRO, M. P. & CAMBIER, J. C. 2012. B cell maintenance and function in aging. *Seminars in Immunology,* 24**,** 342-349.
KOPANOS, C., TSIOLKAS, V., KOURIS, A., CHAPPLE, C. E., ALBARCA AGUILERA, M., MEYER, R. & MASSOURAS, A. 2019. VarSome: the human genomic variant search engine. *Bioinformatics,* 35**,** 1978-1980.
KOVURI, P., YADAV, A. & SINHA, H. 2023. Role of genetic architecture in phenotypic plasticity. *Trends in Genetics,* 39**,** 703-714.
KRAUSE, A., SEYMOUR, H. & RAMSAY, M. 2018. Common and Founder Mutations for Monogenic Traits in Sub-Saharan African Populations. *Annual Review of Genomics and Human Genetics,* 19**,** 149-175.
KRUPP, D. R., BARNARD, R. A., DUFFOURD, Y., EVANS, S. A., MULQUEEN, R. M., BERNIER, R., RIVIÈRE, J.-B., FOMBONNE, E. & O'ROAK, B. J. 2017. Exonic Mosaic Mutations Contribute Risk for Autism Spectrum Disorder. *The American Journal of Human Genetics,* 101**,** 369-390.
KRUSZKA, P., BERGER, S. I., WEISS, K., EVERSON, J. L., MARTINEZ, A. F., HONG, S., ANYANE-YEBOA, K., LIPINSKI, R. J. & MUENKE, M. 2019. A CCR4-NOT Transcription Complex, Subunit 1, CNOT1, Variant Associated with Holoprosencephaly. *The American Journal of Human Genetics,* 104**,** 990-993.
KRZYWINSKI, M. I., SCHEIN, J. E., BIROL, I., CONNORS, J., GASCOYNE, R., HORSMAN, D., JONES, S. J. & MARRA, M. A. 2009. Circos: An information aesthetic for comparative genomics. *Genome Research*.
KULKARNI, R. & KALE, V. 2020. Physiological Cues Involved in the Regulation of Adhesion Mechanisms in Hematopoietic Stem Cell Fate Decision. *Frontiers in Cell and Developmental Biology,* 8.
KUNKEL, T. A. 2009. Evolving views of DNA replication (in)fidelity. *Cold Spring Harb Symp Quant Biol,* 74**,** 91-101.
LANDRUM, M. J., LEE, J. M., RILEY, G. R., JANG, W., RUBINSTEIN, W. S., CHURCH, D. M. & MAGLOTT, D. R. 2014. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids research,* 42**,** D980-D985.
LAPLANTE, M. & SABATINI, DAVID M. 2012. mTOR Signaling in Growth Control and Disease. *Cell,* 149**,** 274-293.
LAURIE, C. C., LAURIE, C. A., RICE, K., DOHENY, K. F., ZELNICK, L. R., MCHUGH, C. P., LING, H., HETRICK, K. N., PUGH, E. W., AMOS, C., WEI, Q., WANG, L. E., LEE, J. E., BARNES, K. C., HANSEL, N. N., MATHIAS, R., DALEY, D., BEATY, T. H., SCOTT, A. F., ... & WEIR, B. S. 2012. Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nat Genet,* 44**,** 642-50.

LEE, B. C. H., ROBINSON, P. S., COORENS, T. H. H., YAN, H. H. N., OLAFSSON, S., LEE-SIX, H., SANDERS, M. A., SIU, H. C., HEWINSON, J., YUE, S. S. K., TSUI, W. Y., CHAN, A. S. Y., CHAN, A. K. W., HO, S. L., CAMPBELL, P. J., MARTINCORENA, I., BUCZACKI, S. J. A., YUEN, S. T., LEUNG, S. Y. & STRATTON, M. R. 2022. Mutational landscape of normal epithelial cells in Lynch Syndrome patients. *Nature Communications,* 13**,** 2710.

LEE, M., LUI, A. C. Y., CHAN, J. C. K., DOONG, P. H. L., KWONG, A. K. Y., MAK, C. C. Y., LI, R. H. W., KAN, A. S. Y. & CHUNG, B. H. Y. 2023. Revealing parental mosaicism: the hidden answer to the recurrence of apparent de novo variants. *Human Genomics,* 17**,** 91.

LEE, M. S. J., MATSUO-DAPAAH, J., DEL ROSARIO ZORRILLA, C., OMATSU, Y., NAGASAWA, T., UEMURA, S., IWAMA, A., ISHII, K. J. & COBAN, C. 2024. Acute malaria suppresses the B lymphocytic niche in the bone marrow through the alteration of CXCL12-abundant reticular cells. *International Immunology,* 36**,** 339-352.

LEE-SIX, H., OLAFSSON, S., ELLIS, P., OSBORNE, R. J., SANDERS, M. A., MOORE, L., GEORGAKOPOULOS, N., TORRENTE, F., NOORANI, A., GODDARD, M., ROBINSON, P., COORENS, T. H. H., O'NEILL, L., ALDER, C., WANG, J., FITZGERALD, R. C., ZILBAUER, M., COLEMAN, N., SAEB-PARSY, K., ... & STRATTON, M. R. 2019. The landscape of somatic mutation in normal colorectal epithelial cells. *Nature,* 574**,** 532-537.

LEVINE, M. S. & HOLLAND, A. J. 2018. The impact of mitotic errors on cell proliferation and tumorigenesis. *Genes Dev,* 32**,** 620-638.

LEZAJA, A. & ALTMEYER, M. 2021. Dealing with DNA lesions: When one cell cycle is not enough. *Curr Opin Cell Biol,* 70**,** 27-36.

LI, H. 2014. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics,* 30**,** 2843-2851.

LI, H. & DURBIN, R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics,* 25**,** 1754-1760.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G., DURBIN, R. & GENOME PROJECT DATA PROCESSING, S. 2009. The sequence alignment/map format and SAMtools. *bioinformatics,* 25**,** 2078-2079.

LI, R., DI, L., LI, J., FAN, W., LIU, Y., GUO, W., LIU, W., LIU, L., LI, Q., CHEN, L., CHEN, Y., MIAO, C., LIU, H., WANG, Y., MA, Y., XU, D., LIN, D., HUANG, Y., WANG, J., ... & WU, C. 2021. A body map of somatic mutagenesis in morphologically normal human tissues. *Nature,* 597**,** 398-403.

LI, S., GARRETT-BAKELMAN, F. E., CHUNG, S. S., SANDERS, M. A., HRICIK, T., RAPAPORT, F., PATEL, J., DILLON, R., VIJAY, P., BROWN, A. L., PERL, A. E., CANNON, J., BULLINGER, L., LUGER, S., BECKER, M., LEWIS, I. D., TO, L. B., DELWEL, R., LÖWENBERG, B., ... & MASON, C. E. 2016. Distinct evolution and dynamics of epigenetic and genetic heterogeneity in acute myeloid leukemia. *Nat Med,* 22**,** 792-9.

LI, X., LI, C., ZHANG, W., WANG, Y., QIAN, P. & HUANG, H. 2023. Inflammation and aging: signaling pathways and intervention therapies. *Signal Transduction and Targeted Therapy,* 8**,** 239.

LI, Y., ROBERTS, N. D., WALA, J. A., SHAPIRA, O., SCHUMACHER, S. E., KUMAR, K., KHURANA, E., WASZAK, S., KORBEL, J. O., HABER, J. E., IMIELINSKI, M., AKDEMIR, K. C., ALVAREZ, E. G., BAEZ-ORTEGA, A., BEROUKHIM, R., BOUTROS, P. C., BOWTELL, D. D. L., BRORS, B., BURNS, K. H., ... & CONSORTIUM, P. 2020. Patterns of somatic structural variation in human cancer genomes. *Nature,* 578**,** 112-121.

LIM, E. T., UDDIN, M., DE RUBEIS, S., CHAN, Y., KAMUMBU, A. S., ZHANG, X., D'GAMA, A. M., KIM, S. N., HILL, R. S., GOLDBERG, A. P., POULTNEY, C., MINSHEW, N. J., KUSHIMA, I., ALEKSIC, B., OZAKI, N., PARELLADA, M., ARANGO, C., PENZOL, M. J., CARRACEDO, A., ... & AUTISM SEQUENCING, C. 2017. Rates, distribution and implications of postzygotic mosaic mutations in autism spectrum disorder. *Nature Neuroscience,* 20**,** 1217-1224.

LIN, M., WHITMIRE, S., CHEN, J., FARREL, A., SHI, X. & GUO, J.-T. 2017. Effects of short indels on protein structure and function in human genomes. *Scientific Reports,* 7**,** 9313.

LIN, S.-H., BROWN, D. W., ROSE, B., DAY, F., LEE, O. W., KHAN, S. M., HISLOP, J., CHANOCK, S. J., PERRY, J. R. B. & MACHIELA, M. J. 2021. Incident disease associations with mosaic chromosomal alterations on autosomes, X and Y chromosomes: insights from a phenome-wide association study in the UK Biobank. *Cell & Bioscience,* 11**,** 143.

LIN, S.-H., KHAN, S. M., ZHOU, W., BROWN, D. W., VERGARA, C., WOLINSKY, S. M., MARTÍNEZ-MAZA, O., MARGOLICK, J. B., MARTINSON, J. J., HUSSAIN, S. K., ENGELS, E. A. & MACHIELA, M. J. 2023. Mosaic chromosomal alterations detected in men living with HIV and the relationship to non-Hodgkin lymphoma. *AIDS,* 37.

LINDAHL, T. 1993. Instability and decay of the primary structure of DNA. *Nature,* 362**,** 709-715.

LINTON, P. J. & DORSHKIND, K. 2004. Age-related changes in lymphocyte development and function. *Nat Immunol,* 5**,** 133-9.

LIPTON, J. O. & SAHIN, M. 2014. The neurology of mTOR. *Neuron,* 84**,** 275-291.

LIU, A., GENOVESE, G., ZHAO, Y., PIRINEN, M., ZEKAVAT, S. M., KENTISTOU, K. A., YANG, Z., YU, K., VLASSCHAERT, C., LIU, X., BROWN, D. W., HUDJASHOV, G., GORMAN, B. R., DENNIS, J., ZHOU, W., MOMOZAWA, Y., PYARAJAN, S., TUZOV, V., PAJUSTE, F.-D., ... & MILLION VETERAN, P. 2024. Genetic drivers and cellular selection of female mosaic X chromosome loss. *Nature,* 631**,** 134-141.

LIU, X., KAMATANI, Y. & TERAO, C. 2021. Genetics of autosomal mosaic chromosomal alteration (mCA). *Journal of Human Genetics,* 66**,** 879-885.

LIU, X., LI, C., MOU, C., DONG, Y. & TU, Y. 2020a. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Medicine,* 12**,** 103.

LIU, Y., BAI, Y., WU, X., LI, G., WEI, W., FU, W., WANG, G., FENG, Y., MENG, H. & LI, H. 2020b. Polycyclic aromatic hydrocarbons exposure and their joint effects with age, smoking, and TCL1A variants on mosaic loss of chromosome Y among coke-oven workers. *Environmental Pollution,* 258**,** 113655.

LIU, Y., LUO, M., JIN, Z., ZHAO, M. & QU, H. 2018. dbLGL: an online leukemia gene and literature database for the retrospective comparison of adult and childhood leukemia genetics with literature evidence. *Database,* 2018**,** bay062.

LOEWE, L. & HILL, W. G. 2010. The population genetics of mutations: good, bad and indifferent. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 365**,** 1153-1167.

LOFTFIELD, E., ZHOU, W., GRAUBARD, B. I., YEAGER, M., CHANOCK, S. J., FREEDMAN, N. D. & MACHIELA, M. J. 2018. Predictors of mosaic chromosome Y loss and associations with mortality in the UK Biobank. *Scientific Reports,* 8**,** 12316.

LOGAN, C. W., MIGUEL DE LA, H., GEORGE, A. R. W., AMANDA, L., LISA, M. V., MICHAEL, T. P., DAFFODIL, M. C., DANA, B.-B., ASHLEY, C., ALEXANDER, T., HEATHER, Z., ALICIA, B. B., TINA, P., RACHID, K., STEVEN, H. &

AMANDA, B. S. 2023. APPLICATION OF THE ACMG/AMP FRAMEWORK TO CAPTURE EVIDENCE RELEVANT TO PREDICTED AND OBSERVED IMPACT ON SPLICING: RECOMMENDATIONS FROM THE CLINGEN SVI SPLICING SUBGROUP. *medRxiv*, 2023.02.24.23286431.

LOH, P.-R., GENOVESE, G., HANDSAKER, R. E., FINUCANE, H. K., RESHEF, Y. A., PALAMARA, P. F., BIRMANN, B. M., TALKOWSKI, M. E., BAKHOUM, S. F., MCCARROLL, S. A. & PRICE, A. L. 2018. Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature,* 559**,** 350-355.

LOH, P.-R., GENOVESE, G. & MCCARROLL, S. A. 2020. Monogenic and polygenic inheritance become instruments for clonal selection. *Nature,* 584**,** 136-141.

LUBALA, T. K., LUMAKA, A., MBUYI-MUSANZAYI, S., KAYEMBE, T., SHONGO, M. Y. P., MUKUKU, O., LUBALA, N., MALAMBA-LEZ, D., LUBOYA, O. N. & LUKUSA-TSHILOBO, P. 2018. Fragile X syndrome with mosaic size mutation in a Bantu patient from Central Africa. *Clin Dysmorphol,* 27**,** 66-69.

LUMAKA, A., CARSTENS, N., DEVRIENDT, K., KRAUSE, A., KULOHOMA, B., KUMUTHINI, J., MUBUNGU, G., MUKISA, J., NEL, M., OLANREWAJU, T. O., LOMBARD, Z., LANDOURÉ, G. & AS MEMBERS OF THE RARE DISEASE WORKING GROUP OF THE, H. A. C. 2022. Increasing African genomic data generation and sharing to resolve rare and undiagnosed diseases in Africa: a call-to-action by the H3Africa rare diseases working group. *Orphanet Journal of Rare Diseases,* 17**,** 230.

LUMAKA, A., COSEMANS, N., LULEBO MAMPASI, A., MUBUNGU, G., MVUAMA, N., LUBALA, T., MBUYI-MUSANZAYI, S., BRECKPOT, J., HOLVOET, M. & DE RAVEL, T. 2017. Facial dysmorphism is influenced by ethnic background of the patient and of the evaluator. *Clinical genetics,* 92**,** 166-171.

LYNCH, M., ACKERMAN, M. S., GOUT, J.-F., LONG, H., SUNG, W., THOMAS, W. K. & FOSTER, P. L. 2016. Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics,* 17**,** 704-714.

LÖBRICH, M. & JEGGO, P. 2017. A Process of Resection-Dependent Nonhomologous End Joining Involving the Goddess Artemis. *Trends in Biochemical Sciences,* 42**,** 690-701.

MA, W., TANG, W., KWOK, J. S. L., TONG, A. H. Y., LO, C. W. S., CHU, A. T. W. & CHUNG, B. H. Y. 2024. A review on trends in development and translation of omics signatures in cancer. *Computational and Structural Biotechnology Journal,* 23**,** 954-971.

MACHIELA, M. J., ZHOU, W., SAMPSON, J. N., DEAN, M. C., JACOBS, K. B., BLACK, A., BRINTON, L. A., CHANG, I. S., CHEN, C., CHEN, C., CHEN, K., COOK, L. S., CROUS BOU, M., DE VIVO, I., DOHERTY, J., FRIEDENREICH, C. M., GAUDET, M. M., HAIMAN, C. A., HANKINSON, S. E., ... & CHANOCK, S. J. 2015. Characterization of large structural genetic mosaicism in human autosomes. *American journal of human genetics,* 96**,** 487-497.

MACIA, A., MUÑOZ-LOPEZ, M., CORTES, J. L., HASTINGS, R. K., MORELL, S., LUCENA-AGUILAR, G., MARCHAL, J. A., BADGE, R. M. & GARCIA-PEREZ, J. L. 2011. Epigenetic Control of Retrotransposon Expression in Human Embryonic Stem Cells. *Molecular and Cellular Biology,* 31**,** 300-316.

MADDEN-HENNESSEY, K., GUPTA, D., RADECKI, A. A., GUILD, C., RATH, A. & HEINEN, C. D. 2022. Loss of mismatch repair promotes a direct selective advantage in human stem cells. *Stem Cell Reports,* 17**,** 2661-2673.

MAHER, G. J., MCGOWAN, S. J., GIANNOULATOU, E., VERRILL, C., GORIELY, A. & WILKIE, A. O. M. 2016. Visualizing the origins of selfish de novo mutations in individual seminiferous tubules of human testes. *Proceedings of the National Academy of Sciences,* 113**,** 2454-2459.

MALHERBE, H. L., ALDOUS, C., CHRISTIANSON, A. L., DARLISON, M. W. & MODELL, B. 2021. Modelled epidemiological data for selected congenital disorders in South Africa. *J Community Genet,* 12**,** 357-376.
MALONGANE, F., PHOSWA, W. N. & BEREJENA, T. 2024. The effect of indigenous African Diet on inflammatory markers linked to Type 2 Diabetic Mellitus. *Human Nutrition & Metabolism,* 35**,** 200236.
MAREES, A. T., DE KLUIVER, H., STRINGER, S., VORSPAN, F., CURIS, E., MARIE-CLAIRE, C. & DERKS, E. M. 2018. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research,* 27**,** e1608.
MARTIN, A. R., KANAI, M., KAMATANI, Y., OKADA, Y., NEALE, B. M. & DALY, M. J. 2019. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics,* 51**,** 584-591.
MARTINCORENA, I. & CAMPBELL, P. J. 2015. Somatic mutation in cancer and normal cells. *Science,* 349**,** 1483-1489.
MARTINCORENA, I., FOWLER, J. C., WABIK, A., LAWSON, A. R. J., ABASCAL, F., HALL, M. W. J., CAGAN, A., MURAI, K., MAHBUBANI, K., STRATTON, M. R., FITZGERALD, R. C., HANDFORD, P. A., CAMPBELL, P. J., SAEB-PARSY, K. & JONES, P. H. 2018. Somatic mutant clones colonize the human esophagus with age. *Science,* 362**,** 911-917.
MARTINCORENA, I., RAINE, K. M., GERSTUNG, M., DAWSON, K. J., HAASE, K., VAN LOO, P., DAVIES, H., STRATTON, M. R. & CAMPBELL, P. J. 2017. Universal patterns of selection in cancer and somatic tissues. *Cell,* 171**,** 1029-1041.
MARTÍNEZ-GLEZ, V., TENORIO, J., NEVADO, J., GORDO, G., RODRÍGUEZ-LAGUNA, L., FEITO, M., DE LUCAS, R., PÉREZ-JURADO, L. A., RUIZ PÉREZ, V. L., TORRELO, A., SPINNER, N. B., HAPPLE, R., BIESECKER, L. G. & LAPUNZINA, P. 2020. A six-attribute classification of genetic mosaicism. *Genetics in Medicine,* 22**,** 1743-1757.
MARZÀ-FLORENSA, A., BOATENG, D., AGYEMANG, C., BEUNE, E., MEEKS, K. A. C., BAHENDEKA, S., LEVITT, N. & KLIPSTEIN-GROBUSCH, K. 2021. Multimorbidity Among Migrant and Non-Migrant Ghanaians: The RODAM Study. *Int J Public Health,* 66**,** 1604056.
MATHIESON, I. & SCALLY, A. 2020. What is ancestry? *PLoS Genet,* 16**,** e1008624.
MATTHEWS, H. K., BERTOLI, C. & DE BRUIN, R. A. M. 2022. Cell cycle control in cancer. *Nature Reviews Molecular Cell Biology,* 23**,** 74-88.
MCKENNA, A., HANNA, M., BANKS, E., SIVACHENKO, A., CIBULSKIS, K., KERNYTSKY, A., GARIMELLA, K., ALTSHULER, D., GABRIEL, S. & DALY, M. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research,* 20**,** 1297-1303.
MCLAREN, W., GIL, L., HUNT, S. E., RIAT, H. S., RITCHIE, G. R. S., THORMANN, A., FLICEK, P. & CUNNINGHAM, F. 2016. The Ensembl Variant Effect Predictor. *Genome Biology,* 17**,** 122.
MCRAE, J. F., CLAYTON, S., FITZGERALD, T. W., KAPLANIS, J., PRIGMORE, E., RAJAN, D., SIFRIM, A., AITKEN, S., AKAWI, N., ALVI, M., AMBRIDGE, K., BARRETT, D. M., BAYZETINOVA, T., JONES, P., JONES, W. D., KING, D., KRISHNAPPA, N., MASON, L. E., SINGH, T., ... & DECIPHERING DEVELOPMENTAL DISORDERS, S. 2017. Prevalence and architecture of de novo mutations in developmental disorders. *Nature,* 542**,** 433-438.
MESCHEDE, I. P., BURGOYNE, T., TOLMACHOVA, T., SEABRA, M. C. & FUTTER, C. E. 2020. Chronically shortened rod outer segments accompany photoreceptor cell death in Choroideremia. *PLOS ONE,* 15**,** e0242284.

MIESCHER, J. F. 1871. *Ueber die chemische Zusammensetzung der Eiterzellen,* Berlin, August Hirschwald.
MILLER, P. G., QIAO, D., ROJAS-QUINTERO, J., HONIGBERG, M. C., SPERLING, A. S., GIBSON, C. J., BICK, A. G., NIROULA, A., MCCONKEY, M. E. & SANDOVAL, B. 2022. Association of clonal hematopoiesis with chronic obstructive pulmonary disease. *Blood, The Journal of the American Society of Hematology,* 139**,** 357-368.
MILLER, P. G. & STEENSMA, D. P. 2020. Implications of clonal hematopoiesis for precision oncology. *JCO Precision Oncology,* 4**,** 639-646.
MILLER, S. A., DYKES, D. D. & POLESKY, H. F. 1988. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Research,* 16**,** 1215-1215.
MISTELI, T. 2020. The Self-Organizing Genome: Principles of Genome Architecture and Function. *Cell,* 183**,** 28-45.
MISTRY, J. J., MARLEIN, C. R., MOORE, J. A., HELLMICH, C., WOJTOWICZ, E. E., SMITH, J. G. W., MACAULAY, I., SUN, Y., MORFAKIS, A., PATTERSON, A., HORTON, R. H., DIVEKAR, D., MORRIS, C. J., HAESTIER, A., DI PALMA, F., BERAZA, N., BOWLES, K. M. & RUSHWORTH, S. A. 2019. ROS-mediated PI3K activation drives mitochondrial transfer from stromal cells to hematopoietic stem cells in response to infection. *Proceedings of the National Academy of Sciences,* 116**,** 24610-24619.
MITCHELL, E., SPENCER CHAPMAN, M., WILLIAMS, N., DAWSON, K. J., MENDE, N., CALDERBANK, E. F., JUNG, H., MITCHELL, T., COORENS, T. H. H., SPENCER, D. H., MACHADO, H., LEE-SIX, H., DAVIES, M., HAYLER, D., FABRE, M. A., MAHBUBANI, K., ABASCAL, F., CAGAN, A., VASSILIOU, G. S., ... & CAMPBELL, P. J. 2022. Clonal dynamics of haematopoiesis across the human lifespan. *Nature,* 606**,** 343-350.
MOHIUDDIN, M., KOOY, R. F. & PEARSON, C. E. 2022. De novo mutations, genetic mosaicism and human disease. *Frontiers in Genetics,* 13.
MONTGOMERY, S. B., GOODE, D. L., KVIKSTAD, E., ALBERS, C. A., ZHANG, Z. D., MU, X. J., ANANDA, G., HOWIE, B., KARCZEWSKI, K. J. & SMITH, K. S. 2013. The origin, evolution, and functional impact of short insertion–deletion variants identified in 179 human genomes. *Genome research,* 23**,** 749-761.
MOOG, U., FELBOR, U., HAS, C. & ZIRN, B. 2020. Disorders Caused by Genetic Mosaicism. *Dtsch Arztebl International,* 117**,** 119-125.
MOORE, L., CAGAN, A., COORENS, T. H. H., NEVILLE, M. D. C., SANGHVI, R., SANDERS, M. A., OLIVER, T. R. W., LEONGAMORNLERT, D., ELLIS, P. & NOORANI, A. 2021. The mutational landscape of human somatic and germline cells. *Nature,* 597**,** 381-386.
MORENO-RUIZ, N., LAO, O., ARÓSTEGUI, J. I., LAAYOUNI, H., CASALS, F. & CONSORTIUM, G. E. R. 2022. Assessing the digenic model in rare disorders using population sequencing data. *Eur J Hum Genet,* 30**,** 1439-1443.
MORGAN, T. H. 1910. Sex limited inheritance in Drosophila. *Science,* 32**,** 120-122.
MORRISON, S. J. & SCADDEN, D. T. 2014. The bone marrow niche for haematopoietic stem cells. *Nature,* 505**,** 327-334.
MOYER, R., BRILEY, D., JOHNSEN, A., STEWART, U. & SHAW, B. R. 1993. Echinomycin, a bis-intercalating agent, induces C→ T mutations via cytosine deamination. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis,* 288**,** 291-300.
MULDER, N., ABIMIKU, A., ADEBAMOWO, S. N., DE VRIES, J., MATIMBA, A., OLOWOYO, P., RAMSAY, M., SKELTON, M. & STEIN, D. J. 2018. H3Africa: current perspectives. *Pharmgenomics Pers Med,* 11**,** 59-66.

MUNISHA, M. & SCHIMENTI, J. C. 2021. Genome maintenance during embryogenesis. *DNA Repair (Amst),* 106**,** 103195.

MURAKI, K., NYHAN, K., HAN, L. & MURNANE, J. P. 2012. Mechanisms of telomere loss and their consequences for chromosome instability. *Frontiers in Oncology,* 2.

MUSTJOKI, S. & YOUNG NEAL, S. 2021. Somatic Mutations in “Benign” Disease. *New England Journal of Medicine,* 384**,** 2039-2052.

NAM, C. H., YOUK, J., KIM, J. Y., LIM, J., PARK, J. W., OH, S. A., LEE, H. J., PARK, J. W., WON, H., LEE, Y., JEONG, S.-Y., LEE, D.-S., OH, J. W., HAN, J., LEE, J., KWON, H. W., KIM, M. J. & JU, Y. S. 2023. Widespread somatic L1 retrotransposition in normal colorectal epithelium. *Nature,* 617**,** 540-547.

NEPH, S., KUEHN, M. S., REYNOLDS, A. P., HAUGEN, E., THURMAN, R. E., JOHNSON, A. K., RYNES, E., MAURANO, M. T., VIERSTRA, J., THOMAS, S., SANDSTROM, R., HUMBERT, R. & STAMATOYANNOPOULOS, J. A. 2012. BEDOPS: high-performance genomic feature operations. *Bioinformatics,* 28**,** 1919-1920.

NESTA, A. V., TAFUR, D. & BECK, C. R. 2021. Hotspots of Human Mutation. *Trends in Genetics,* 37**,** 717-729.

NG, S. W. K., ROUHANI, F. J., BRUNNER, S. F., BRZOZOWSKA, N., AITKEN, S. J., YANG, M., ABASCAL, F., MOORE, L., NIKITOPOULOU, E., CHAPPELL, L., LEONGAMORNLERT, D., IVOVIC, A., ROBINSON, P., BUTLER, T., SANDERS, M. A., WILLIAMS, N., COORENS, T. H. H., TEAGUE, J., RAINE, K., ... & CAMPBELL, P. J. 2021. Convergent somatic mutations in metabolism genes in chronic liver disease. *Nature,* 598**,** 473-478.

NIOHURU, I. 2023. Disease Burden and Mortality. *In:* NIOHURU, I. (ed.) *Healthcare and Disease Burden in Africa: The Impact of Socioeconomic Factors on Public Health.* Cham: Springer International Publishing.

NIRENBERG, M. W. & MATTHAEI, J. H. 1961. The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. *Proceedings of the National Academy of Sciences,* 47**,** 1588-1602.

NIROULA, A., SEKAR, A., MURAKAMI, M. A., TRINDER, M., AGRAWAL, M., WONG, W. J., BICK, A. G., UDDIN, M. M., GIBSON, C. J., GRIFFIN, G. K., HONIGBERG, M. C., ZEKAVAT, S. M., PARUCHURI, K., NATARAJAN, P. & EBERT, B. L. 2021. Distinction of lymphoid and myeloid clonal hematopoiesis. *Nature Medicine,* 27**,** 1921-1927.

NOHMI, T. 2018. Thresholds of Genotoxic and Non-Genotoxic Carcinogens. *Toxicological Research,* 34**,** 281-290.

NOWELL, P. C. 1976. The Clonal Evolution of Tumor Cell Populations. *Science,* 194**,** 23-28.

NUMANAGIĆ, I., GÖKKAYA, A. S., ZHANG, L., BERGER, B., ALKAN, C. & HACH, F. 2018. Fast characterization of segmental duplications in genome assemblies. *Bioinformatics,* 34**,** i706-i714.

NÍ LEATHLOBHAIR, M. & LENSKI, R. E. 2022. Population genetics of clonally transmissible cancers. *Nature Ecology & Evolution,* 6**,** 1077-1089.

O'KEEFE, C., MCDEVITT, M. A. & MACIEJEWSKI, J. P. 2010. Copy neutral loss of heterozygosity: a novel chromosomal lesion in myeloid malignancies. *Blood,* 115**,** 2731-2739.

O'LEARY, N. A., WRIGHT, M. W., BRISTER, J. R., CIUFO, S., HADDAD, D., MCVEIGH, R., RAJPUT, B., ROBBERTSE, B., SMITH-WHITE, B., AKO-ADJEI, D., ASTASHYN, A., BADRETDIN, A., BAO, Y., BLINKOVA, O., BROVER, V., CHETVERNIN, V., CHOI, J., COX, E., ERMOLAEVA, O., ... & PRUITT, K. D. 2016. Reference sequence (RefSeq) database at NCBI: current status,

taxonomic expansion, and functional annotation. *Nucleic Acids Research,* 44**,** D733-D745.
OBENG-ABOAGYE, E., FRIMPONG, A., AMPONSAH, J. A., DANSO, S. E., OWUSU, E. D. A. & OFORI, M. F. 2023. Inflammatory cytokines as potential biomarkers for early diagnosis of severe malaria in children in Ghana. *Malaria Journal,* 22**,** 220.
ODGIS, J. A., GALLAGHER, K. M., REHMAN, A. U., MARATHE, P. N., BONINI, K. E., SEBASTIN, M., DI BIASE, M., BROWN, K., KELLY, N. R., RAMOS, M. A., THOMAS-WILSON, A., GUHA, S., OKUR, V., GANAPATHI, M., ELKHOURY, L., EDELMANN, L., ZINBERG, R. E., ABUL-HUSN, N. S., DIAZ, G. A., ... & GELB, B. D. 2023. Detection of mosaic variants using genome sequencing in a large pediatric cohort. *American Journal of Medical Genetics, Part A,* 191**,** 699-710.
OKAYASU, R. & ILIAKIS, G. 1994. Evidence That the Product of the xrs Gene Is Predominantly Involved in the Repair of a Subset of Radiation-Induced Interphase Chromosome Breaks Rejoining with Fast Kinetics. *Radiation Research,* 138**,** 34-43.
OMOTOSO, O. E., TEIBO, J. O., ATIBA, F. A., OLADIMEJI, T., ADEBESIN, A. O. & BABALGHITH, A. O. 2022. Bridging the genomic data gap in Africa: implications for global disease burdens. *Globalization and Health,* 18**,** 103.
PANG, W. W., PRICE, E. A., SAHOO, D., BEERMAN, I., MALONEY, W. J., ROSSI, D. J., SCHRIER, S. L. & WEISSMAN, I. L. 2011. Human bone marrow hematopoietic stem cells are increased in frequency and myeloid-biased with age. *Proc Natl Acad Sci U S A,* 108**,** 20012-7.
PARENTI, I., RABANEDA, L. G., SCHOEN, H. & NOVARINO, G. 2020. Neurodevelopmental Disorders: From Genetics to Functional Pathways. *Trends in Neurosciences,* 43**,** 608-621.
PARK, S., MALI, N. M., KIM, R., CHOI, J. W., LEE, J., LIM, J., PARK, J. M., PARK, J. W., KIM, D., KIM, T., YI, K., CHOI, J. H., KWON, S. G., HONG, J. H., YOUK, J., AN, Y., KIM, S. Y., OH, S. A., KWON, Y., ... & JU, Y. S. 2021. Clonal dynamics in early human embryogenesis inferred from somatic mutation. *Nature,* 597**,** 393-397.
PATIL, I. 2021. Visualizations with statistical details: The'ggstatsplot'approach. *Journal of Open Source Software,* 6**,** 3167.
PEI, W., SHANG, F., WANG, X., FANTI, A.-K., GRECO, A., BUSCH, K., KLAPPROTH, K., ZHANG, Q., QUEDENAU, C., SAUER, S., FEYERABEND, T. B., HÖFER, T. & RODEWALD, H.-R. 2020. Resolving Fates and Single-Cell Transcriptomes of Hematopoietic Stem Cell Clones by PolyloxExpress Barcoding. *Cell Stem Cell,* 27**,** 383-395.e8.
PEJAVER, V., BYRNE, A. B., FENG, B.-J., PAGEL, K. A., MOONEY, S. D., KARCHIN, R., O'DONNELL-LURIA, A., HARRISON, S. M., TAVTIGIAN, S. V., GREENBLATT, M. S., BIESECKER, L. G., RADIVOJAC, P., BRENNER, S. E., BIESECKER, L. G., HARRISON, S. M., TAYOUN, A. A., BERG, J. S., BRENNER, S. E., CUTTING, G. R., ... & TOPPER, S. 2022. Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. *The American Journal of Human Genetics,* 109**,** 2163-2177.
PEPRAH, E. K., ALLEN, E. G., WILLIAMS, S. M., WOODARD, L. M. & SHERMAN, S. L. 2010. Genetic diversity of the fragile X syndrome gene (FMR1) in a large Sub-Saharan West African population. *Annals of human genetics,* 74**,** 316-325.
PFEIFER, G. P., YOU, Y.-H. & BESARATINIA, A. 2005. Mutations induced by ultraviolet light. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis,* 571**,** 19-31.

PINCEZ, T., LEE, S. S. K., ILBOUDO, Y., PREUSS, M., PHAM HUNG D'ALEXANDRY D'ORENGIANI, A.-L., BARTOLUCCI, P., GALACTÉROS, F., JOLY, P., BAUER, D. E., LOOS, R. J. F., LINDSLEY, R. C. & LETTRE, G. 2021. Clonal hematopoiesis in sickle cell disease. *Blood,* 138**,** 2148-2152.

PINHO, S. & FRENETTE, P. S. 2019. Haematopoietic stem cell activity and interactions with the niche. *Nature Reviews Molecular Cell Biology,* 20**,** 303-320.

POSEY, J. E., ROSENFELD, J. A., JAMES, R. A., BAINBRIDGE, M., NIU, Z., WANG, X., DHAR, S., WISZNIEWSKI, W., AKDEMIR, Z. H. C., GAMBIN, T., XIA, F., PERSON, R. E., WALKIEWICZ, M., SHAW, C. A., SUTTON, V. R., BEAUDET, A. L., MUZNY, D., ENG, C. M., YANG, Y., ... & PLON, S. E. 2016. Molecular diagnostic experience of whole-exome sequencing in adult patients. *Genetics in Medicine,* 18**,** 678-685.

POURMONTASERI, H. & KHANMOHAMMADI, S. 2024. Demographic risk factors of pro-inflammatory diet: a narrative review. *Front Nutr,* 11**,** 1448806.

PÖS, O., RADVANSZKY, J., BUGLYÓ, G., PÖS, Z., RUSNAKOVA, D., NAGY, B. & SZEMES, T. 2021. DNA copy number variation: Main characteristics, evolutionary significance, and pathological aspects. *Biomedical Journal,* 44**,** 548-559.

QUINLAN, A. R. & HALL, I. M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics,* 26**,** 841-842.

RAHBARI, R., WUSTER, A., LINDSAY, S. J., HARDWICK, R. J., ALEXANDROV, L. B., TURKI, S. A., DOMINICZAK, A., MORRIS, A., PORTEOUS, D., SMITH, B., STRATTON, M. R., HURLES, M. E. & CONSORTIUM, U. K. 2016. Timing, rates and spectra of human germline mutation. *Nat Genet,* 48**,** 126-133.

RALL-SCHARPF, M., FRIEDL, T. W. P., BIECHONSKI, S., DENKINGER, M., MILYAVSKY, M. & WIESMÜLLER, L. 2021. Sex-specific differences in DNA double-strand break repair of cycling human lymphocytes during aging. *Aging (Albany NY),* 13**,** 21066-21089.

RAMABOLI, M. C., OCVIRK, S., KHAN MIRZAEI, M., EBERHART, B. L., VALDIVIA-GARCIA, M., METWALY, A., NEUHAUS, K., BARKER, G., RU, J., NESENGANI, L. T., MAHDI-JOEST, D., WILSON, A. S., JONI, S. K., LAYMAN, D. C., ZHENG, J., MANDAL, R., CHEN, Q., PEREZ, M. R., FORTUIN, S., ... & O'KEEFE, S. J. D. 2024. Diet changes due to urbanization in South Africa are linked to microbiome and metabolome signatures of Westernization and colorectal cancer. *Nature Communications,* 15**,** 3379.

RAMSAY, M., CROWTHER, N., TAMBO, E., AGONGO, G., BALOYI, V., DIKOTOPE, S., GÓMEZ-OLIVÉ, X., JAFF, N., SORGHO, H., WAGNER, R., KHAYEKA-WANDABWA, C., CHOUDHURY, A., HAZELHURST, S., KAHN, K., LOMBARD, Z., MUKOMANA, F., SOO, C., SOODYALL, H., WADE, A., ... & SANKOH, O. 2016. H3Africa AWI-Gen Collaborative Centre: a resource to study the interplay between genomic and environmental risk factors for cardiometabolic diseases in four sub-Saharan African countries. *Global Health, Epidemiology and Genomics,* 1**,** e20.

RAMÍREZ-BELLO, J. & JIMÉNEZ-MORALES, M. 2017. [Functional implications of single nucleotide polymorphisms (SNPs) in protein-coding and non-coding RNA genes in multifactorial diseases]. *Gaceta Médica de México,* 153**,** 238-250.

RAY, R., NOVOTNY, N. M., CRISOSTOMO, P. R., LAHM, T., ABARBANELL, A. & MELDRUM, D. R. 2008. Sex steroids and stem cell function. *Mol Med,* 14**,** 493-501.

REHM, H. L., BERG, J. S., BROOKS, L. D., BUSTAMANTE, C. D., EVANS, J. P., LANDRUM, M. J., LEDBETTER, D. H., MAGLOTT, D. R., MARTIN, C. L.,

NUSSBAUM, R. L., PLON, S. E., RAMOS, E. M., SHERRY, S. T. & WATSON, M. S. 2015. ClinGen — The Clinical Genome Resource. *New England Journal of Medicine,* 372**,** 2235-2242.
REN, P., DONG, X. & VIJG, J. 2022. Age-related somatic mutation burden in human tissues. *Frontiers in Aging,* 3.
RENTZSCH, P., SCHUBACH, M., SHENDURE, J. & KIRCHER, M. 2021. CADD-Splice—improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Medicine,* 13**,** 31.
REVY, P., KANNENGIESSER, C. & FISCHER, A. 2019. Somatic genetic rescue in Mendelian haematopoietic diseases. *Nature Reviews Genetics,* 20**,** 582-598.
RIBIERRE, T., DELEUZE, C., BACQ, A., BALDASSARI, S., MARSAN, E., CHIPAUX, M., MURACA, G., ROUSSEL, D., NAVARRO, V. & LEGUERN, E. 2018. Second-hit mosaic mutation in mTORC1 repressor DEPDC5 causes focal cortical dysplasia–associated epilepsy. *The Journal of clinical investigation,* 128**,** 2452-2458.
RICHARDS, S., AZIZ, N., BALE, S., BICK, D., DAS, S., GASTIER-FOSTER, J., GRODY, W. W., HEGDE, M., LYON, E., SPECTOR, E., VOELKERDING, K. & REHM, H. L. 2015. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine,* 17**,** 405-424.
RIEPE, T. V., KHAN, M., ROOSING, S., CREMERS, F. P. M. & T HOEN, P. A. C. 2021. Benchmarking deep learning splice prediction tools using functional splice assays. *Human Mutation,* 42**,** 799-810.
ROBINSON, J. T., THORVALDSDÓTTIR, H., WINCKLER, W., GUTTMAN, M., LANDER, E. S., GETZ, G. & MESIROV, J. P. 2011. Integrative genomics viewer. *Nature biotechnology,* 29**,** 24-26.
ROBINSON, P. N., KÖHLER, S., BAUER, S., SEELOW, D., HORN, D. & MUNDLOS, S. 2008. The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. *The American Journal of Human Genetics,* 83**,** 610-615.
ROBINSON, P. S., COORENS, T. H. H., PALLES, C., MITCHELL, E., ABASCAL, F., OLAFSSON, S., LEE, B. C. H., LAWSON, A. R. J., LEE-SIX, H., MOORE, L., SANDERS, M. A., HEWINSON, J., MARTIN, L., PINNA, C. M. A., GALAVOTTI, S., RAHBARI, R., CAMPBELL, P. J., MARTINCORENA, I., TOMLINSON, I. & STRATTON, M. R. 2021. Increased somatic mutation burdens in normal human cells due to defective DNA polymerases. *Nature Genetics,* 53**,** 1434-1442.
ROCCO, J. M., ZHOU, Y., LIU, N. S., LAIDLAW, E., GALINDO, F., ANDERSON, M. V., RUPERT, A., LAGE, S. L., ORTEGA-VILLA, A. M., YU, S., LISCO, A., MANION, M., VASSILIOU, G. S., DUNBAR, C. E. & SERETI, I. 2024. Clonal hematopoiesis in people with advanced HIV and associated inflammatory syndromes. *JCI Insight,* 9.
RODIN, R. E., DOU, Y., KWON, M., SHERMAN, M. A., D'GAMA, A. M., DOAN, R. N., RENTO, L. M., GIRSKIS, K. M., BOHRSON, C. L., KIM, S. N., NADIG, A., LUQUETTE, L. J., GULHAN, D. C., WALSH, C. A., GANZ, J., WOODWORTH, M. B., LI, P., RODIN, R. E., HILL, R. S., ... & BRAIN SOMATIC MOSAICISM, N. 2021. The landscape of somatic mutation in cerebral cortex of autistic and neurotypical individuals revealed by ultra-deep whole-genome sequencing. *Nature Neuroscience,* 24**,** 176-185.
RODRIGUEZ-FRATICELLI, A. E. & CAMARGO, F. 2021. Systems analysis of hematopoiesis using single-cell lineage tracing. *Current Opinion in Hematology,* 28.

ROTIMI, C. N., BENTLEY, A. R., DOUMATEY, A. P., CHEN, G., SHRINER, D. & ADEYEMO, A. 2017. The genomic landscape of African populations in health and disease. *Human Molecular Genetics,* 26**,** R225-R236.

ROUX, P., SALORT, D. & XU, Z. 2021. Adaptation to DNA damage as a bet-hedging mechanism in a fluctuating environment. *R Soc Open Sci,* 8**,** 210460.

ROWLANDS, C., THOMAS, H. B., LORD, J., WAI, H. A., ARNO, G., BEAMAN, G., SERGOUNIOTIS, P., GOMES-SILVA, B., CAMPBELL, C. & GOSSAN, N. 2021. Comparison of in silico strategies to prioritize rare genomic variants impacting RNA splicing for the diagnosis of genomic disorders. *Scientific Reports,* 11**,** 20607.

RUNDBERG NILSSON, A., SONEJI, S., ADOLFSSON, S., BRYDER, D. & PRONK, C. J. 2016. Human and Murine Hematopoietic Stem Cell Aging Is Associated with Functional Impairments and Intrinsic Megakaryocytic/Erythroid Bias. *PLOS ONE,* 11**,** e0158369.

RÜBE, C. E., FRICKE, A., WIDMANN, T. A., FÜRST, T., MADRY, H., PFREUNDSCHUH, M. & RÜBE, C. 2011. Accumulation of DNA damage in hematopoietic stem and progenitor cells during human aging. *PLoS One,* 6**,** e17487.

SAINI, N., GIACOBONE, C. K., KLIMCZAK, L. J., PAPAS, B. N., BURKHOLDER, A. B., LI, J.-L., FARGO, D. C., BAI, R., GERRISH, K., INNES, C. L., SCHURMAN, S. H. & GORDENIN, D. A. 2021. UV-exposure, endogenous DNA damage, and DNA replication errors shape the spectra of genome changes in human skin. *PLOS Genetics,* 17**,** e1009302.

SAKUMI, K. 2019. Germline mutation: de novo mutation in reproductive lineage cells. *Genes Genet Syst,* 94**,** 3-12.

SANDRAN, N. G., FORNARINO, D. L., CORBETT, M. A., KROES, T., GARDNER, A. E., MACLENNAN, A. H., GÉCZ, J. & VAN EYK, C. L. 2024. Application of multiple mosaic callers improves post-zygotic mutation detection from exome sequencing data. *Genet Med,* 26**,** 101220.

SANMIGUEL, J. M., YOUNG, K. & TROWBRIDGE, J. J. 2020. Hand in hand: intrinsic and extrinsic drivers of aging and clonal hematopoiesis. *Experimental Hematology,* 91**,** 1-9.

SANO, S., HORITANI, K., OGAWA, H., HALVARDSON, J., CHAVKIN, N. W., WANG, Y., SANO, M., MATTISSON, J., HATA, A., DANIELSSON, M., MIURA-YURA, E., ZAGHLOOL, A., EVANS, M. A., FALL, T., DE HOYOS, H. N., SUNDSTRÖM, J., YURA, Y., KOUR, A., ARAI, Y., ... & WALSH, K. 2022. Hematopoietic loss of Y chromosome leads to cardiac fibrosis and heart failure mortality. *Science,* 377**,** 292-297.

SARKAR, H. & MOOSAJEE, M. 2022. Choroideremia: molecular mechanisms and therapies. *Trends in Molecular Medicine,* 28**,** 378-387.

SASANI, T. A., PEDERSEN, B. S., GAO, Z., BAIRD, L., PRZEWORSKI, M., JORDE, L. B. & QUINLAN, A. R. 2019. Large, three-generation human families reveal post-zygotic mosaicism and variability in germline mutation accumulation. *Elife,* 8**,** e46922.

SAVATT, J. M. & MYERS, S. M. 2021. Genetic Testing in Neurodevelopmental Disorders. *Frontiers in Pediatrics,* 9.

SCHMAUCK-MEDINA, T., MOLIERE, A., LAUTRUP, S., ZHANG, J., CHLOPICKI, S., MADSEN, H. B., CAO, S., SOENDENBROE, C., MANSELL, E., VESTERGAARD, M. B., LI, Z., SHILOH, Y., OPRESKO, P. L., EGLY, J. M., KIRKWOOD, T., VERDIN, E., BOHR, V. A., COX, L. S., STEVNSNER, T., ... & FANG, E. F. 2022. New hallmarks of ageing: a 2022 Copenhagen ageing meeting summary. *Aging (Albany NY),* 14**,** 6829-6839.

SCULLY, R., GANESAN, S., VLASAKOVA, K., CHEN, J., SOCOLOVSKY, M. & LIVINGSTON, D. M. 1999. Genetic Analysis of BRCA1 Function in a Defined Tumor Cell Line. *Molecular Cell,* 4**,** 1093-1099.

SCULLY, R., PANDAY, A., ELANGO, R. & WILLIS, N. A. 2019. DNA double-strand break repair-pathway choice in somatic mammalian cells. *Nature Reviews Molecular Cell Biology,* 20**,** 698-714.

SEGAL, F., SPIRO, F. & WILTON, E. 1967. Mosaic XY-XXY in a Bantu--a case report. *S Afr Med J,* 41**,** 701-3.

SENGUPTA, D., BOTHA, G., MEINTJES, A., MBIYAVANGA, M., HAZELHURST, S., MULDER, N., RAMSAY, M. & CHOUDHURY, A. 2023a. Performance and accuracy evaluation of reference panels for genotype imputation in sub-Saharan African populations. *Cell Genomics,* 3.

SENGUPTA, D., BOTHA, G., MEINTJES, A., MBIYAVANGA, M., HAZELHURST, S., MULDER, N., RAMSAY, M. & CHOUDHURY, A. 2023b. Performance and accuracy evaluation of reference panels for genotype imputation in sub-Saharan African populations. *Cell Genomics,* 3**,** 100332.

SEPLYARSKIY, V. B. & SUNYAEV, S. 2021. The origin of human mutation in light of genomic data. *Nat Rev Genet,* 22**,** 672-686.

SHAO, L., LI, H., PAZHANISAMY, S. K., MENG, A., WANG, Y. & ZHOU, D. 2011. Reactive oxygen species and hematopoietic stem cell senescence. *International Journal of Hematology,* 94**,** 24-32.

SHERMAN, M. A., RODIN, R. E., GENOVESE, G., DIAS, C., BARTON, A. R., MUKAMEL, R. E., BERGER, B., PARK, P. J., WALSH, C. A. & LOH, P.-R. 2021. Large mosaic copy number variations confer autism risk. *Nature Neuroscience,* 24**,** 197-203.

SHIRAI, Y.-T., SUZUKI, T., MORITA, M., TAKAHASHI, A. & YAMAMOTO, T. 2014. Multifunctional roles of the mammalian CCR4–NOT complex in physiological phenomena. *Frontiers in Genetics,* 5.

SIES, H. & JONES, D. P. 2020. Reactive oxygen species (ROS) as pleiotropic physiological signalling agents. *Nature Reviews Molecular Cell Biology,* 21**,** 363-383.

SIMON, A. & COOP, G. 2024. The contribution of gene flow, selection, and genetic drift to five thousand years of human allele frequency change. *Proceedings of the National Academy of Sciences,* 121**,** e2312377121.

SNIEGOWSKI, P. D., GERRISH, P. J., JOHNSON, T. & SHAVER, A. 2000. The evolution of mutation rates: separating causes from consequences. *Bioessays,* 22**,** 1057-1066.

SNIPES, G. J., SUTER, U., WELCHER, A. A. & SHOOTER, E. M. 1992. Characterization of a novel peripheral nervous system myelin protein (PMP-22/SR13). *Journal of Cell Biology,* 117**,** 225-238.

SOLARSH, G. & HOFMAN, K. J. 2006. Developmental Disabilities. 2nd ed.: The International Bank for Reconstruction and Development / The World Bank, Washington (DC).

SOLLIS, E., MOSAKU, A., ABID, A., BUNIELLO, A., CEREZO, M., GIL, L., GROZA, T., GÜNEŞ, O., HALL, P., HAYHURST, J., IBRAHIM, A., JI, Y., JOHN, S., LEWIS, E., MACARTHUR, JACQUELINE A. L., MCMAHON, A., OSUMI-SUTHERLAND, D., PANOUTSOPOULOU, K., PENDLINGTON, Z., ... & HARRIS, LAURA W. 2023. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Research,* 51**,** D977-D985.

SONG, M., HAN, Y., ZHAO, Y., LV, J., YU, C., PEI, P., YANG, L., MILLWOOD, I. Y., WALTERS, R. G., CHEN, Y., DU, H., YANG, X., YAO, W., CHEN, J., CHEN, Z., GENOVESE, G., TERAO, C., LI, L., SUN, D. & GROUP, C. K. B. C. 2024.

Association of autosomal mosaic chromosomal alterations with risk of bladder cancer in Chinese adults: a prospective cohort study. *Cell Death Dis,* 15**,** 706.
SPENCER CHAPMAN, M., MITCHELL, E., YOSHIDA, K., WILLIAMS, N., FABRE, M. A., RANZONI, A. M., ROBINSON, P. S., KREGAR, L. D., WILK, M., BOETTCHER, S., MAHBUBANI, K., SAEB PARSY, K., GOWERS, K. H. C., JANES, S. M., NG, S. W. K., HOARE, M., GREEN, A. R., VASSILIOU, G. S., CVEJIC, A., ... & CAMPBELL, P. J. 2025. Prolonged persistence of mutagenic DNA lesions in somatic cells. *Nature,* 638**,** 729-738.
SREENIVASAN, V. K. A., YUMICEBA, V. & SPIELMANN, M. 2025. Structural variants in the 3D genome as drivers of disease. *Nat Rev Genet*.
SRIVASTAVA, S., LOVE-NICHOLS, J. A., DIES, K. A., LEDBETTER, D. H., MARTIN, C. L., CHUNG, W. K., FIRTH, H. V., FRAZIER, T., HANSEN, R. L. & PROCK, L. 2019. Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders. *Genetics in Medicine,* 21**,** 2413-2421.
SROUR, M. & SHEVELL, M. 2014. Genetics and the investigation of developmental delay/intellectual disability. *Archives of Disease in Childhood,* 99**,** 386.
STEENSMA, D. P., BEJAR, R., JAISWAL, S., LINDSLEY, R. C., SEKERES, M. A., HASSERJIAN, R. P. & EBERT, B. L. 2015. Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. *Blood,* 126**,** 9-16.
STEIN, D., KARS, M. E., WU, Y., BAYRAK Ç, S., STENSON, P. D., COOPER, D. N., SCHLESSINGER, A. & ITAN, Y. 2023. Genome-wide prediction of pathogenic gain- and loss-of-function variants from ensemble learning of a diverse feature set. *Genome Med,* 15**,** 103.
STEINHAUS, R., PROFT, S., SCHUELKE, M., COOPER, D. N., SCHWARZ, J. M. & SEELOW, D. 2021. MutationTaster2021. *Nucleic Acids Research,* 49**,** W446-W451.
STEVENTON-JONES, V., STAVISH, D., HALLIWELL, J. A., BAKER, D. & BARBARIC, I. 2022. Single Nucleotide Polymorphism (SNP) Arrays and Their Sensitivity for Detection of Genetic Changes in Human Pluripotent Stem Cell Cultures. *Curr Protoc,* 2**,** e606.
STOWELL, JAMES A. W., WEBSTER, MICHAEL W., KÖGEL, A., WOLF, J., SHELLEY, KATHRYN L. & PASSMORE, LORI A. 2016. Reconstitution of Targeted Deadenylation by the Ccr4-Not Complex and the YTH Domain Protein Mmi1. *Cell Reports,* 17**,** 1978-1989.
STRAUCH, Y., LORD, J., NIRANJAN, M. & BARALLE, D. 2022. CI-SpliceAI—improving machine learning predictions of disease causing splicing variants using curated alternative splice sites. *Plos one,* 17**,** e0269159.
STUMPF, F., KELLER, B., GRESSIES, C. & SCHUETZ, P. 2023. Inflammation and Nutrition: Friend or Foe? *Nutrients,* 15.
SUGIAMAN-TRAPMAN, D., VITEZIC, M., JOUHILAHTI, E.-M., MATHELIER, A., LAUTER, G., MISRA, S., DAUB, C. O., KERE, J. & SWOBODA, P. 2018. Characterization of the human RFX transcription factor family by regulatory and target gene analysis. *BMC genomics,* 19**,** 1-15.
SUN, L., WANG, Z., LU, T., MANOLIO, T. A. & PATERSON, A. D. 2023. eXclusionarY: 10 years later, where are the sex chromosomes in GWASs? *The American Journal of Human Genetics,* 110**,** 903-912.
SUN, Y., RUIVENKAMP, C. A., HOFFER, M. J., VRIJENHOEK, T., KRIEK, M., VAN ASPEREN, C. J., DEN DUNNEN, J. T. & SANTEN, G. W. 2015. Next-generation diagnostics: gene panel, exome, or whole genome? *Human mutation,* 36**,** 648-655.

SUPEK, F. & LEHNER, B. 2019. Scales and mechanisms of somatic mutation rate variation across the human genome. *DNA Repair (Amst),* 81**,** 102647.
SWENBERG, J. A., LU, K., MOELLER, B. C., GAO, L., UPTON, P. B., NAKAMURA, J. & STARR, T. B. 2011. Endogenous versus Exogenous DNA Adducts: Their Role in Carcinogenesis, Epidemiology, and Risk Assessment. *Toxicological Sciences,* 120**,** S130-S145.
SZILARD, L. 1959. ON THE NATURE OF THE AGING PROCESS. *Proceedings of the National Academy of Sciences,* 45**,** 30-45.
TABREZ, S., PRIYADARSHINI, M., PRIYAMVADA, S., KHAN, M. S., NA, A. & ZAIDI, S. K. 2014. Gene–environment interactions in heavy metal and pesticide carcinogenesis. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis,* 760**,** 1-9.
TAKAGI, M., KANEKO-SCHMITT, S., SUZUMORI, N., NISHIMURA, G. & HASEGAWA, T. 2012. Atypical achondroplasia due to somatic mosaicism for the common thanatophoric dysplasia mutation R248C. *American Journal of Medical Genetics Part A,* 158A**,** 247-250.
TALIUN, D., HARRIS, D. N., KESSLER, M. D., CARLSON, J., SZPIECH, Z. A., TORRES, R., TALIUN, S. A. G., CORVELO, A., GOGARTEN, S. M. & KANG, H. M. 2021. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature,* 590**,** 290-299.
TANG, D., KAMITAKI, N., MUKAMEL, R. E., RUBINACCI, S. & LOH, P.-R. 2025. Patterns and drivers of 43,617 mosaic chromosomal alterations in blood. *medRxiv***,** 2025.07.30.25332451.
TAVTIGIAN, S. V., GREENBLATT, M. S., HARRISON, S. M., NUSSBAUM, R. L., PRABHU, S. A., BOUCHER, K. M. & BIESECKER, L. G. 2018. Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genetics in Medicine,* 20**,** 1054-1060.
TAVTIGIAN, S. V., HARRISON, S. M., BOUCHER, K. M. & BIESECKER, L. G. 2020. Fitting a naturally scaled point system to the ACMG/AMP variant classification guidelines. *Human Mutation,* 41**,** 1734-1737.
TEAM, R. 2022a. RStudio: integrated development environment for R. RStudio, Inc.
TEAM, R. C. 2022b. R: A language and environment for statistical computing. R Foundation for Statistical Computing.
TERAO, C., SUZUKI, A., MOMOZAWA, Y., AKIYAMA, M., ISHIGAKI, K., YAMAMOTO, K., MATSUDA, K., MURAKAMI, Y., MCCARROLL, S. A., KUBO, M., LOH, P.-R. & KAMATANI, Y. 2020. Chromosomal alterations among age-related haematopoietic clones in Japan. *Nature,* 584**,** 130-135.
THE UNIPROT, C. 2023. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research,* 51**,** D523-D531.
THORMANN, A., HALACHEV, M., MCLAREN, W., MOORE, D. J., SVINTI, V., CAMPBELL, A., KERR, S. M., TISCHKOWITZ, M., HUNT, S. E., DUNLOP, M. G., HURLES, M. E., WRIGHT, C. F., FIRTH, H. V., CUNNINGHAM, F. & FITZPATRICK, D. R. 2019. Flexible and scalable diagnostic filtering of genomic variants using G2P with Ensembl VEP. *Nature Communications,* 10**,** 2373.
THORPE, J., OSEI-OWUSU, I. A., AVIGDOR, B. E., TUPLER, R. & PEVSNER, J. 2020. Mosaicism in Human Health and Disease. *Annual Review of Genetics,* 54**,** 487-510.
TIAN, Y., PESARAN, T., CHAMBERLIN, A., FENWICK, R. B., LI, S., GAU, C.-L., CHAO, E. C., LU, H.-M., BLACK, M. H. & QIAN, D. 2019. REVEL and BayesDel outperform other in silico meta-predictors for clinical variant classification. *Scientific Reports,* 9**,** 12752.

TOMSO, D. J. & BELL, D. A. 2003. Sequence context at human single nucleotide polymorphisms: overrepresentation of CpG dinucleotide at polymorphic sites and suppression of variation in CpG islands. *Journal of molecular biology,* 327**,** 303-308.

TORAL-LOPEZ, J., GONZALEZ-HUERTA, L. M. & CUEVAS-COVARRUBIAS, S. A. 2012. Complete monosomy mosaic of chromosome 21: case report and review of literature. *Gene,* 510**,** 175-9.

TOUTAIN, J., GOUTTE-GATTAT, D., HOROVITZ, J. & SAURA, R. 2018. Confined placental mosaicism revisited: Impact on pregnancy characteristics and outcome. *PloS one,* 13**,** e0195905.

TREANGEN, T. J. & SALZBERG, S. L. 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics,* 13**,** 36-46.

TUBBS, A. & NUSSENZWEIG, A. 2017. Endogenous DNA Damage as a Source of Genomic Instability in Cancer. *Cell,* 168**,** 644-656.

TUDEK, B., WINCZURA, A., JANIK, J., SIOMEK, A., FOKSINSKI, M. & OLINSKI, R. 2010. Involvement of oxidatively damaged DNA and repair in cancer development and aging. *Am J Transl Res,* 2**,** 254-84.

UDDIN, M. M., SAADATAGAH, S., NIROULA, A., YU, B., HORNSBY, W. E., GANESH, S., LANNERY, K., SCHUERMANS, A., HONIGBERG, M. C., BICK, A. G., LIBBY, P., EBERT, B. L., BALLANTYNE, C. M. & NATARAJAN, P. 2024. Long-term longitudinal analysis of 4,187 participants reveals insights into determinants of clonal hematopoiesis. *Nature Communications,* 15**,** 7858.

UDDIN, M. M., YU, Z., WEINSTOCK, J. S., NAKAO, T., NIROULA, A., URBUT, S. M., KOYAMA, S., ZEKAVAT, S. M., PARUCHURI, K., SILVER, A. J., MACK, T. M., WONG, M. Y., HAIDERMOTA, S. M., BHATTACHARYA, R., JORSHERY, S. D., RADDATZ, M. A., HONIGBERG, M. C., HORNSBY, W. E., ZHANG, M. J., ... & NATARAJAN, P. 2022. Germline genomic and phenomic landscape of clonal hematopoiesis in 323,112 individuals. *medRxiv***,** 2022.07.29.22278015.

VAINIERI, M. L., BLAGBOROUGH, A. M., MACLEAN, A. L., HALTALLI, M. L. R., RUIVO, N., FLETCHER, H. A., STUMPF, M. P. H., SINDEN, R. E. & LO CELSO, C. 2016. Systematic tracking of altered haematopoiesis during sporozoite-mediated malaria development reveals multiple response points. *Open Biology,* 6**,** 160038.

VALECHA, M. & POSADA, D. 2022. Somatic variant calling from single-cell DNA sequencing data. *Comput Struct Biotechnol J,* 20**,** 2978-2985.

VALLETTA, S., THOMAS, A., MENG, Y., REN, X., DRISSEN, R., SENGÜL, H., DI GENUA, C. & NERLOV, C. 2020. Micro-environmental sensing by bone marrow stroma identifies IL-6 and TGFβ1 as regulators of hematopoietic ageing. *Nature Communications,* 11**,** 4075.

VAN BERGEN, N. J., MASSEY, S., QUIGLEY, A., ROLLO, B., HARRIS, A. R., KAPSA, R. M. I. & CHRISTODOULOU, J. 2022. CDKL5 deficiency disorder: molecular insights and mechanisms of pathogenicity to fast-track therapeutic development. *Biochem Soc Trans,* 50**,** 1207-1224.

VAN DER AUWERA, G. A., CARNEIRO, M. O., HARTL, C., POPLIN, R., DEL ANGEL, G., LEVY-MOONSHINE, A., JORDAN, T., SHAKIR, K., ROAZEN, D., THIBAULT, J., BANKS, E., GARIMELLA, K. V., ALTSHULER, D., GABRIEL, S. & DEPRISTO, M. A. 2013. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics,* 43**,** 11.10.1-11.10.33.

VAN VEGHEL-PLANDSOEN, M. M., WOUTERS, C. H., KROMOSOETO, J. N., DEN RIDDER-KLÜNNEN, M. C., HALLEY, D. J. & VAN DEN OUWELAND, A. M.

2011. Multiplex ligation-depending probe amplification is not suitable for detection of low-grade mosaicism. *Eur J Hum Genet,* 19**,** 1009-12.
VERHOEF, S., BAKKER, L., TEMPELAARS, A. M. P., HESSELING-JANSSEN, A. L. W., MAZURCZAK, T., JOZWIAK, S., FOIS, A., BARTALINI, G., ZONNENBERG, B. A., VAN ESSEN, A. J., LINDHOUT, D., HALLEY, D. J. J. & VAN DEN OUWELAND, A. M. W. 1999. High Rate of Mosaicism in Tuberous Sclerosis Complex. *The American Journal of Human Genetics,* 64**,** 1632-1637.
VIGUERA, E., CANCEILL, D. & EHRLICH, S. D. 2001. Replication slippage involves DNA polymerase pausing and dissociation. *EMBO J,* 20**,** 2587-95.
VIJG, J. 2014. Somatic mutations, genome mosaicism, cancer and aging. *Current Opinion in Genetics & Development,* 26**,** 141-149.
VIJG, J. 2021. From DNA damage to mutations: All roads lead to aging. *Ageing Research Reviews,* 68**,** 101316.
VIJG, J., SCHUMACHER, B., ABAKIR, A., ANTONOV, M., BRADLEY, C., CAGAN, A., CHURCH, G., GLADYSHEV, V. N., GORBUNOVA, V., MASLOV, A. Y., REIK, W., SHARIFI, S., SUH, Y. & WALSH, K. 2023. Mitigating age-related somatic mutation burden. *Trends Mol Med,* 29**,** 530-540.
VISSERS, L. E. L. M., KALVAKURI, S., DE BOER, E., GEUER, S., OUD, M., VAN OUTERSTERP, I., KWINT, M., WITMOND, M., KERSTEN, S., POLLA, D. L., WEIJERS, D., BEGTRUP, A., MCWALTER, K., RUIZ, A., GABAU, E., MORTON, J. E. V., GRIFFITH, C., WEISS, K., GAMBLE, C., ... & DE BROUWER, A. P. M. 2020. De Novo Variants in CNOT1, a Central Component of the CCR4-NOT Complex Involved in Gene Expression and RNA and Protein Stability, Cause Neurodevelopmental Delay. *The American Journal of Human Genetics,* 107**,** 164-172.
VLASSCHAERT, C., MACK, T., HEIMLICH, J. B., NIROULA, A., UDDIN, M. M., WEINSTOCK, J., SHARBER, B., SILVER, A. J., XU, Y., SAVONA, M., GIBSON, C., LANKTREE, M. B., RAUH, M. J., EBERT, B. L., NATARAJAN, P., JAISWAL, S. & BICK, A. G. 2023. A practical approach to curate clonal hematopoiesis of indeterminate potential in human genetic data sets. *Blood,* 141**,** 2214-2223.
VON ELM, E., ALTMAN, D. G., EGGER, M., POCOCK, S. J., GØTZSCHE, P. C. & VANDENBROUCKE, J. P. 2007. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *The lancet,* 370**,** 1453-1457.
WAGGONER, D., WAIN, K. E., DUBUC, A. M., CONLIN, L., HICKEY, S. E., LAMB, A. N., MARTIN, C. L., MORTON, C. C., RASMUSSEN, K. & SCHUETTE, J. L. 2018. Yield of additional genetic testing after chromosomal microarray for diagnosis of neurodevelopmental disability and congenital anomalies: a clinical practice resource of the American College of Medical Genetics and Genomics (ACMG). *Genetics in Medicine,* 20**,** 1105-1113.
WALDVOGEL, S. M., POSEY, J. E. & GOODELL, M. A. 2024. Human embryonic genetic mosaicism and its effects on development and disease. *Nat Rev Genet,* 25**,** 698-714.
WANG, T., RUAN, S., ZHAO, X., SHI, X., TENG, H., ZHONG, J., YOU, M., XIA, K., SUN, Z. & MAO, F. 2021a. OncoVar: an integrated database and analysis platform for oncogenic driver variants in cancers. *Nucleic Acids Research,* 49**,** D1289-D1301.
WANG, Y., BAE, T., THORPE, J., SHERMAN, M. A., JONES, A. G., CHO, S., DAILY, K., DOU, Y., GANZ, J., GALOR, A., LOBON, I., PATTNI, R., ROSENBLUH, C., TOMASI, S., TOMASINI, L., YANG, X., ZHOU, B., AKBARIAN, S., BALL, L. L., ... & BRAIN SOMATIC MOSAICISM, N. 2021b. Comprehensive identification of somatic nucleotide variants in human brain tissue. *Genome Biology,* 22**,** 92.

WATSON, C. J. & BLUNDELL, J. R. 2023. Mutation rates and fitness consequences of mosaic chromosomal alterations in blood. *Nature Genetics,* 55**,** 1677-1685.
WATSON, C. J., PAPULA, A. L., POON, G. Y. P., WONG, W. H., YOUNG, A. L., DRULEY, T. E., FISHER, D. S. & BLUNDELL, J. R. 2020. The evolutionary dynamics and fitness landscape of clonal hematopoiesis. *Science,* 367**,** 1449-1454.
WATSON, J. D. & CRICK, F. H. C. 1953. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature,* 171**,** 737-738.
WEISKOPF, D., WEINBERGER, B. & GRUBECK-LOEBENSTEIN, B. 2009. The aging of the immune system. *Transplant International,* 22**,** 1041-1050.
WELCH, J. S., LEY, T. J., LINK, D. C., MILLER, C. A., LARSON, D. E., KOBOLDT, D. C., WARTMAN, L. D., LAMPRECHT, T. L., LIU, F. & XIA, J. 2012. The origin and evolution of mutations in acute myeloid leukemia. *Cell,* 150**,** 264-278.
WHIFFIN, N., MINIKEL, E., WALSH, R., O'DONNELL-LURIA, A. H., KARCZEWSKI, K., ING, A. Y., BARTON, P. J. R., FUNKE, B., COOK, S. A., MACARTHUR, D. & WARE, J. S. 2017. Using high-resolution variant frequencies to empower clinical genome interpretation. *Genetics in Medicine,* 19**,** 1151-1158.
WHITE, R. R. & VIJG, J. 2016. Do DNA Double-Strand Breaks Drive Aging? *Mol Cell,* 63**,** 729-38.
WILFERT, A. B., SULOVARI, A., TURNER, T. N., COE, B. P. & EICHLER, E. E. 2017. Recurrent de novo mutations in neurodevelopmental disorders: properties and clinical implications. *Genome Medicine,* 9**,** 101.
WONG, J. Y. Y., MARGOLIS, H. G., MACHIELA, M., ZHOU, W., ODDEN, M. C., PSATY, B. M., ROBBINS, J., JONES, R. R., ROTTER, J. I., CHANOCK, S. J., ROTHMAN, N., LAN, Q. & LEE, J. S. 2018. Outdoor air pollution and mosaic loss of chromosome Y in older men from the Cardiovascular Health Study. *Environment International,* 116**,** 239-247.
WONKAM, A., MUNUNG, N. S., DANDARA, C., ESOH, K. K., HANCHARD, N. A. & LANDOURE, G. 2022. Five Priorities of African Genomics Research: The Next Frontier. *Annu Rev Genomics Hum Genet,* 23**,** 499-521.
WORLD BANK. 2020a. *Prevalence of current tobacco use (% of adults)* [Online]. Available: https://genderdata.worldbank.org/en/indicator/sh-prv-smok?gender=total [Accessed 27 February 2025].
WORLD BANK. 2020b. *Total alcohol consumption per capita (liters of pure alcohol, projected estimates, 15+ years of age)* [Online]. Available: https://data.worldbank.org/indicator/SH.ALC.PCAP.LI?view=map [Accessed 27 February 2025].
WORLD BANK. 2022. *Prevalence of HIV, total (% of population ages 15-49)* [Online]. Available: https://data.worldbank.org/indicator/SH.DYN.AIDS.ZS?view=map [Accessed 27 February 2025].
WORLD BANK. 2024a. *Research and development expenditure (% of GDP)* [Online]. Available: https://data.worldbank.org/indicator/GB.XPD.RSDV.GD.ZS [Accessed 26 February 2025].
WORLD BANK. 2024b. *Rural population (% of total population) - Sub-Saharan Africa* [Online]. Available: https://data.worldbank.org/indicator/SP.RUR.TOTL.ZS?locations=ZG [Accessed 26 February 2025].
WORLD BANK. 2024c. *The World Bank in South Africa* [Online]. Available: https://www.worldbank.org/en/country/southafrica/overview [Accessed 26 February 2025].
WORLD HEALTH ORGANIZATION 2024a. *Global tuberculosis report 2024,* Geneva.

WORLD HEALTH ORGANIZATION 2024b. *World malaria report 2024: addressing inequity in the global malaria response. ,* Geneva.

WRIGHT, C. F., CAMPBELL, P., EBERHARDT, R. Y., AITKEN, S., PERRETT, D., BRENT, S., DANECEK, P., GARDNER, E. J., CHUNDRU, V. K., LINDSAY, S. J., ANDREWS, K., HAMPSTEAD, J., KAPLANIS, J., SAMOCHA, K. E., MIDDLETON, A., FOREMAN, J., HOBSON, R. J., PARKER, M. J., MARTIN, H. C., ... & FIRTH, H. V. 2023. Genomic Diagnosis of Rare Pediatric Disease in the United Kingdom and Ireland. *New England Journal of Medicine,* 388**,** 1559-1571.

WRIGHT, C. F., PRIGMORE, E., RAJAN, D., HANDSAKER, J., MCRAE, J., KAPLANIS, J., FITZGERALD, T. W., FITZPATRICK, D. R., FIRTH, H. V. & HURLES, M. E. 2019. Clinically-relevant postzygotic mosaicism in parents and children with developmental disorders in trio exome sequencing data. *Nat Commun,* 10**,** 2985.

WRIGHT, D. E., WAGERS, A. J., GULATI, A. P., JOHNSON, F. L. & WEISSMAN, I. L. 2001. Physiological migration of hematopoietic stem and progenitor cells. *Science,* 294**,** 1933-1936.

XIANG, J., PENG, J., BAXTER, S. & PENG, Z. 2020. AutoPVS1: An automatic classification tool for PVS1 interpretation of null variants. *Human Mutation,* 41**,** 1488-1498.

XIU, N. N., YANG, X. D., XU, J., JU, B., SUN, X. Y. & ZHAO, X. C. 2023. Leukemic transformation during anti-tuberculosis treatment in aplastic anemia-paroxysmal nocturnal hemoglobinuria syndrome: A case report and review of literature. *World J Clin Cases,* 11**,** 6908-6919.

YADAV, V. K., DEGREGORI, J. & DE, S. 2016. The landscape of somatic mutations in protein coding genes in apparently benign human tissues carries signatures of relaxed purifying selection. *Nucleic Acids Research,* 44**,** 2075-2084.

YANG, X., BREUSS, M. W., XU, X., ANTAKI, D., JAMES, K. N., STANLEY, V., BALL, L. L., GEORGE, R. D., WIRTH, S. A., CAO, B., NGUYEN, A., MCEVOY-VENNERI, J., CHAI, G., NAHAS, S., VAN DER KRAAN, L., DING, Y., SEBAT, J. & GLEESON, J. G. 2021. Developmental and temporal characteristics of clonal sperm mosaicism. *Cell,* 184**,** 4772-4783.e15.

YANG, X., XU, X., BREUSS, M. W., ANTAKI, D., BALL, L. L., CHUNG, C., SHEN, J., LI, C., GEORGE, R. D., WANG, Y., BAE, T., CHENG, Y., ABYZOV, A., WEI, L., ALEXANDROV, L. B., SEBAT, J. L., AVERBUJ, D., ROY, S., COURCHESNE, E., ... & NETWORK, N. B. S. M. 2023. Control-independent mosaic single nucleotide variant detection with DeepMosaic. *Nature Biotechnology,* 41**,** 870-877.

YOUNG, C. D., HUBBARD, A. K., SAINT-MAURICE, P. F., CHAN, I. C. C., CAO, Y., TRAN, D., BOLTON, K. L., CHANOCK, S. J., MATTHEWS, C. E., MOORE, S. C., LOFTFIELD, E. & MACHIELA, M. J. 2024. Social, Behavioral, and Clinical Risk Factors Are Associated with Clonal Hematopoiesis. *Cancer Epidemiology, Biomarkers & Prevention,* 33**,** 1423-1432.

YOUNG, K., BORIKAR, S., BELL, R., KUFFLER, L., PHILIP, V. & TROWBRIDGE, J. J. 2016. Progressive alterations in multipotent hematopoietic progenitors underlie lymphoid cell loss in aging. *Journal of Experimental Medicine,* 213**,** 2259-2267.

YU, Z., COORENS, T. H. H., UDDIN, M. M., ARDLIE, K. G., LENNON, N. & NATARAJAN, P. 2024. Genetic variation across and within individuals. *Nature Reviews Genetics,* 25**,** 548-562.

ZEIDAN, A. M., SHALLIS, R. M., WANG, R., DAVIDOFF, A. & MA, X. 2019. Epidemiology of myelodysplastic syndromes: Why characterizing the beast is a prerequisite to taming it. *Blood Reviews,* 34**,** 1-15.

ZEKAVAT, S. M., LIN, S.-H., BICK, A. G., LIU, A., PARUCHURI, K., WANG, C., UDDIN, M. M., YE, Y., YU, Z., LIU, X., KAMATANI, Y., BHATTACHARYA, R., PIRRUCCELLO, J. P., PAMPANA, A., LOH, P.-R., KOHLI, P., MCCARROLL, S. A., KIRYLUK, K., NEALE, B., ... & FINNGEN, C. 2021. Hematopoietic mosaic chromosomal alterations increase the risk for diverse types of infection. *Nature Medicine,* 27**,** 1012-1024.

ZEMET, R., VAN DEN VEYVER, I. B. & STANKIEWICZ, P. 2022. Parental mosaicism for apparent de novo genetic variants: Scope, detection, and counseling challenges. *Prenatal Diagnosis,* 42**,** 811-821.

ZHANG, C. R. C., NIX, D., GREGORY, M., CIORBA, M. A., OSTRANDER, E. L., NEWBERRY, R. D., SPENCER, D. H. & CHALLEN, G. A. 2019. Inflammatory cytokines promote clonal hematopoiesis with specific mutations in ulcerative colitis patients. *Exp Hematol,* 80**,** 36-41.e3.

ZHAO, K., PERSHAD, Y., POISNER, H. M., MA, X., QUADE, K., VLASSCHAERT, C., MACK, T., KHANKARI, N. K., VON BECK, K., BROGAN, J., KISHTAGARI, A., CORTY, R. W., LI, Y., XU, Y., REINER, A. P., SCHEET, P., AUER, P. L. & BICK, A. G. 2025. Genetic drivers and clinical consequences of mosaic chromosomal alterations in 1 million individuals. *medRxiv***,** 2025.03.05.25323443.

ZHOU, W., FISCHER, A., OGWANG, M. D., LUO, W., KERCHAN, P., REYNOLDS, S. J., TENGE, C. N., WERE, P. A., KUREMU, R. T., WEKESA, W. N., MASALU, N., KAWIRA, E., KINYERA, T., OTIM, I., LEGASON, I. D., NABALENDE, H., AYERS, L. W., BHATIA, K., GOEDERT, J. J., ... & MBULAITEYE, S. M. 2023. Mosaic chromosomal alterations in peripheral blood leukocytes of children in sub-Saharan Africa. *Nature Communications,* 14**,** 8081.

Appendices

Appendix A:PhD ethics certificate

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

R49 Mr J Evans

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
CLEARANCE CERTIFICATE NO. M2011144

NAME: (Principal Investigator)	Mr J Evans
DEPARTMENT:	School of Pathology Department of Human Genetics Sydney Brenner Institute for Molecular Biosciences Medical School University
PROJECT TITLE:	Exploring the extent and impact of genetic mosaicism in African populations
DATE CONSIDERED:	Ad hoc
DECISION:	Approved unconditionally
CONDITIONS:	Sub-study under M160830
SUPERVISOR:	Professor Z Lombard and Dr N Carstens
APPROVED BY:	Dr CB Penny, Chairperson, HREC (Medical)
DATE OF APPROVAL:	2020/12/18

This Clearance Certificate is valid for 5 years from the date of approval. An extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office secretariat on the 3rd floor, Phillip Tobias Building, Parktown, University of the Witwatersrand, Johannesburg.

I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated from the research protocol as approved, I/we undertake to submit details to the Committee. **I agree to submit a yearly progress report**. When a funder requires annual re-certification, the application date will be one year after the date when the study was initially reviewed. In this case, the study was initially reviewed in **November** and therefore reports and re-certification will be due in the month of **November** each year. Unreported changes to the study may invalidate the clearance given by the HREC (Medical).

Signature of Principal Investigator

18/12/2020
Date

Appendix B: DDD-Africa ethics certificate

UNIVERSITY OF THE WITWATERSRAND JOHANNESBURG

R49 Professors A Krause & Z Lombard; Dr N Carstens; Ms N Louw

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
CLEARANCE CERTIFICATE NO. M230567

NAME: (Principal Investigator)	Professors A Krause & Z Lombard; Dr N Carstens; Ms N Louw
DEPARTMENT:	School of Pathology Division of Human Genetics National Health Laboratory Service
PROJECT TITLE:	*Decyphering Development Disorders in Africa (DDD-Africa) - evaluating clinical exome sequencing in an African setting*
DATE CONSIDERED:	Ad hoc
DECISION:	Approved unconditionally
CONDITIONS:	Renewal of M18/06/78 - expired on 2023/07/05 NIH Study No. UO1MH115483
NOTE:	If contact information regarding student study participants is required, please contact the Registrar's office - <Nicolsen.Potgieter@wits.ac.za>
SUPERVISOR:	Not applicable
APPROVED BY:	Dr M Vorster, Co-Chairperson, HREC (Medical)
DATE OF APPROVAL:	2023/06/20 **EXPIRY DATE:** 2028/06/19

This Clearance Certificate is valid for 5 years from the date of approval. An extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office secretariat on the 3rd floor, Phillip Tobias Building, Parktown, University of the Witwatersrand, Johannesburg.

I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated from the research protocol as approved, I/we undertake to submit details to the Committee. I agree to submit a yearly progress report. When a funder requires annual re-certification, the application date will be one year after the date when the study was initially reviewed. In this case, the study was initially reviewed in **May** and therefore reports and re-certification will be due in the month of **May** each year. Unreported changes to the study may invalidate the clearance given by the HREC (Medical).

Signature of Principal Investigator — Date: 2023/08/31

Appendix C: Plagiarism declaration

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

FACULTY OF HEALTH SCIENCES

PLAGIARISM DECLARATION TO BE SIGNED BY ALL HIGHER DEGREE STUDENTS

SENATE PLAGIARISM POLICY: APPENDIX ONE

I Jonathan Evans (Student number: 2415960) am a student registered for the degree of Doctor of Philosophy in the academic year 2025.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work except where I have explicitly indicated otherwise.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.
- I have included as an appendix a report from "Turnitin" (or other approved plagiarism detection) software indicating the level of plagiarism in my research document.

Signature:  Date: 29-09-2025

Appendix D: Genetic mosaicism in African rare disease review papers

ABOUSSAIR, N., JAOUAD, I. C., DEQUAQUI, S. C., et al. 2012. Cytogenetic analysis of 5572 patients referred for suspected chromosomal abnormalities in Morocco. Genetic Testing and Molecular Biomarkers, 16, 569-573.

ABU-AMERO, K. K., HELLANI, A., BENDER, P., et al. 2009. High-resolution analysis of DNA copy number alterations in patients with primary open-angle glaucoma. Mol Vis, 15, 1594-8.

ALLY, F. E. & GRACE, H. J. 1979. Chromosome abnormalities in South African mental retardates. S Afr Med J, 55, 710-2.

ASMAL, A. & EVANS WILLIAM, G. 2020. McCune-Albright syndrome: a case report and review. South African Radiographer, 58, 39-42.

BECK, K. D. & HARPER, C. A. 2016. Bilateral Optic Nerve Colobomas in Infant with Tetraploidy. Retinal Cases and Brief Reports, 10, 252-253.

BEIGHTON, P., KOZLOWSKI, K. S., GARDNER, J., et al. 1999. Broad clavicles in trisomy 8 mosaicism: a new sign. Skeletal Radiol, 28, 359-61.

BELKADY, B., ELKHATTABI, L., ELKARHAT, Z., et al. 2019. Chromosomal abnormalities in patients with intellectual disability: A 21-year retrospective study. Human Heredity, 83, 274-282.

BELMOKHTAR, F., BELMOKHTAR, R. & KERFOUF, A. 2016. Cytogenetic study of down syndrome in Algeria: Report and review. Journal of Medical Sciences (Taiwan), 36, 46-52.

BEN KHELIFA, M., COUTTON, C., ZOUARI, R., et al. 2014. Mutations in DNAH1, which encodes an inner arm heavy chain dynein, lead to male infertility from multiple morphological abnormalities of the sperm flagella. Am J Hum Genet, 94, 95-104.

BOUALI, N., HMIDA, D., MOUGOU, S., et al. 2015. Analysis of FMR1 gene premutation and X chromosome cytogenetic abnormalities in 100 Tunisian patients presenting premature ovarian failure. Annales d'Endocrinologie, 76, 671-678.

BOUGACHA-ELLEUCH, N., ELLEUCH, M., CHARFI, N., et al. 2016. Unusual association of turner syndrome and hypopituitarism in a Tunisian family. Current Research in Translational Medicine, 64, 9-13.

BRESLOW, N., OLSHAN, A., BECKWITH, J. B., et al. 1993. Epidemiology of Wilms tumor. Medical and Pediatric Oncology, 21, 172-181.

CHARLES, A. L. 2008. Fragile X allelemorphism among the mentally retarded and affected families. Research Journal of Medical Sciences, 2, 224-230.

CHATEAU, A., KUTSCHE, K., FUCHS, S., et al. 2020. Microphthalmia with linear skin defects syndrome associated with hypopigmented mosaic lesions and ptosis: two siblings from Africa. Int J Dermatol, 59, 864-866.

CHEN, A. & CHMELL, S. 2011. Heterotopic ossification after patellar tendon repair in a man with trisomy 8 mosaicism: A case report and literature review. Journal of Medical Case Reports, 5.

COLLERAN, J. A., DAYKIN, E. C., HERNANDEZ, C., et al. 2023. Novel mosaic TRAF7 likely pathogenic variant in an African American family. American Journal of Medical Genetics, Part A, 191, 1990-1993.

DAVID, G., DÜRR, A., STEVANIN, G., et al. 1998. Molecular and clinical correlations in autosomal dominant cerebellar ataxia with progressive macular dystrophy (SCA7). Hum Mol Genet, 7, 165-70.

DE BOER, A., VERMEULEN, K., EGGER, J. I. M., et al. 2018. EHMT1 mosaicism in apparently unaffected parents is associated with autism spectrum disorder and neurocognitive dysfunction. Molecular Autism, 9.

DE VILLIERS, J. N. P., KOTZE, M. J., VAN HEERDEN, C. J., et al. 2005. Overrepresentation of the founder PPOX gene mutation R59W in a South African patient with severe clinical manifestation of porphyria. Experimental Dermatology, 14, 50-55.

DUMONT, M., HICKMAN, G., HOVNANIAN, A., et al. 2023. "African" acral keratoderma: Further evidence for a single entity. A retrospective study of 42 patients. Ann Dermatol Venereol, 150, 16-23.

EL-BASSYOUNI, H. T., AFIFI, H. H., AGLAN, M. S., et al. 2012. Growth curves of Egyptian patients with Turner syndrome. American Journal of Medical Genetics, Part A, 158 A, 2687-2691.

EL-DAHTORY, F. A. M. 2013. Triple X Egyptian woman and a Down's syndrome offspring. Indian Journal of Human Genetics, 19, 111-112.

EL-GILANY, A. H., YAHIA, S., SHOKER, M., et al. 2011. Cytogenetic and comorbidity profile of down syndrome in Mansoura university children's hospital, Egypt. Indian Journal of Human Genetics, 17, 157-163.

ELLEUCH, M., MNIF FEKI, M., KAMMOUN, M., et al. 2010. Descriptive analyses of Turner syndrome: 49 cases in Tunisia. Annales d'Endocrinologie, 71, 111-116.

ESSADDAM, L., ZITOUNI, O., KRAOUA, L., et al. 2023. Turner syndrome: results of the first Tunisian study group on Turner syndrome (TuSGOT). J Pediatr Endocrinol Metab, 36, 577-583.

FINDLAY, G. 1993. The genetic mosaic. J R Soc Med, 86, 212-6.

GILTAY, J. C., KLIJN, A. J., DE JONG, T. P. V. M., et al. 2016. Tetraploid/Diploid Mosaicism in Cultured Genital Skin Fibroblasts: Is It Causally Related to Penoscrotal Hypospadias. Molecular Syndromology, 7, 153-159.

GNAMEY, D. K., KOFFI, K. S., NAGALO, K., et al. 2010. Aplasia cutis congenita associated with Goltz syndrome in a male neonate. Genetic Counseling, 21, 41-47.

GRACE, H. J. & LIGHT, J. K. 1974. A South African Indian hermaphrodite. S Afr Med J, 48, 1349-50.

GUEDENON, K. M., DOUBAJ, Y., AKOLLY, D. A. E., et al. 2020. Hutchinson-Gilford Progeria syndrome: Report of the first Togolese case. Am J Med Genet A, 182, 1316-1320.

HAFEZ, M., EL-TAHAN, M., ZEDAN, M., et al. 1984. Demographic trends of Down's syndrome in Egypt. Human Biology, 56, 703-712.

HAMMOND, L. S., MACIAS, M. M., TARLETON, J. C., et al. 1997. Fragile X syndrome and deletions in FMR1: new case and review of the literature. Am J Med Genet, 72, 430-4.

HAPPLE, R. 2009. Monoallelic expression on autosomes may explain an unusual heritable form of pigmentary mosaicism: a historical case revisited. Clin Exp Dermatol, 34, 834-7.

HEGAZY, S., MARQUES, M. C., CANNA, S. W., et al. 2022. NEMO-NDAS: A Panniculitis in the Young Representing an Autoinflammatory Disorder in Disguise. Am J Dermatopathol, 44, e64-e66.

HIRSHFELD, A. B., THOMPSON, W. R., PATEL, A., et al. 2001. Proximal trisomy of 1q mosaicism in a girl with hypertrophic cardiomyopathy associated with Wolff-Parkinson-White syndrome and multiple congenital anomalies. Am J Med Genet, 100, 264-8.

HU, X., AYALA, S. S., DYER, L., et al. 2021. A rare case of postnatal mosaic trisomy 12 with severe congenital heart disease and literature review. American Journal of Medical Genetics, Part A, 185, 1864-1869.

JACYK, K. & SMITH, A. 1990. Mosaic acral keratosis. Clin Exp Dermatol, 15, 361-2.

JANSEN, S., KRUGER, A. J. & LIEBENBERG, G. 1991. Turner/Down mosaicism. A case report. S Afr Med J, 79, 731-2.

JONARD, L., FELDMANN, D., PARSY, C., et al. 2008. A familial case of Keratitis-Ichthyosis-Deafness (KID) syndrome with the GJB2 mutation G45E. Eur J Med Genet, 51, 35-43.

JONES, K. L., MCNAMARA, E. A., LONGONI, M., et al. 2018. Dual diagnoses in 152 patients with Turner syndrome: Knowledge of the second condition may lead to modification of treatment and/or surveillance. American Journal of Medical Genetics, Part A, 176, 2435-2445.

JORDAN, M. A., MARQUES, I., ROSENDORFF, J., et al. 1998. Trisomy 8 mosaicism: a further five cases illustrating marked clinical and cytogenetic variability. Genet Couns, 9, 139-46.

KAMMOUN, I., CHAABOUNI, M., TRABELSI, M., et al. 2008. Genetic analysis of Turner syndrome: 89 cases in Tunisia. Annales d'Endocrinologie, 69, 440-445.

KHALIFA, M., STEIN, J., GRAU, L., et al. 2013. Partial deletion of ANKRD11 results in the KBG phenotype distinct from the 16q24.3 microdeletion syndrome. Am J Med Genet A, 161a, 835-40.

KLEKOWSKI, N. & SHWAYDER, T. 2011. Unilateral acrokeratoelastoidosis-second reported case. Pediatric Dermatology, 28, 20-22.

KROB, G., BRAUN, A. & KUHNLE, U. 1994. True hermaphroditism: geographical distribution, clinical findings, chromosomes and gonadal histology. Eur J Pediatr, 153, 2-10.

LAZEBNIK, N. & COHEN, L. 2015. Prenatal diagnosis and findings of tetrasomy 9p. Journal of Obstetrics and Gynaecology Research, 997-1002.

LUBALA, T. K., LUMAKA, A., MBUYI-MUSANZAYI, S., et al. 2018. Fragile X syndrome with mosaic size mutation in a Bantu patient from Central Africa. Clin Dysmorphol, 27, 66-69.

LYONS, N. F. 1973. Results of chromosome studies of 200 African patients referred for cytogenetic analysis at Harare Hospital. Central African Journal of Medicine, 19, 199-200.

MANASSE, B. F., LEKGATE, N., PFAFFENZELLER, W. M., et al. 2000. The Pallister-Killian syndrome is reliably diagnosed by FISH on buccal mucosa. Clin Dysmorphol, 9, 163-5.

MENSAH-ADO, I., LAUHBOUET, M. D., LOKROU, A., et al. 1994. Turner's syndrome: Mosaic and isochromosome q in a young African. Medecine Tropicale, 54, 349-351.

MITRE, V., HEYM, K., CLARK, G. D., et al. 2021. Neurocutaneous melanocytosis and leptomeningeal melanoma. Journal of Pediatric Hematology/Oncology, 43, e195-e197.

MNIF-FEKI, M., SAFI, W., BOUGACHA-ELLEUCH, N., et al. 2021. Occurrence of Hypopituitarism in Tunisian Turner Syndrome patients: familial versus sporadic cases. Gynecological Endocrinology, 37, 848-852.

MOHAMMED, A. M., KAMEL, A. K., HAMMAD, S. A., et al. 2009. Constitutional retinoblastoma gene deletion in Egyptian patients. World Journal of Pediatrics, 5, 222-225.

MOORE, S. W. & ZAAHL, M. G. 2012. Intronic RET gene variants in Down syndrome-associated Hirschsprung disease in an African population. J Pediatr Surg, 47, 299-302.

MOORE, S. W. & ZAAHL, M. G. 2014. Tissue specific somatic mutations and aganglionosis in Hirschsprung's disease. J Pediatr Surg, 49, 258-61; discussion 261.

MOORES, P. 1973. Blood group and tissue mosaicism in a Natal Indian woman. Acta Haematol, 50, 299-304.

MOUAFO TAMBO, F. F., DAHOUN, S., KAMADJOU, C., et al. 2016. Mixed gonadal dysgenesis in Yaoundé: A preliminary experience about three cases. Afr J Paediatr Surg, 13, 145-9.

NAICKER, T. & ALDOUS, C. 2014. Two trisomy 22 live births in one hospital in 15 months: Is it as rare as we thought? Fetal and Pediatric Pathology, 33, 35-41.

ODGIS, J. A., GALLAGHER, K. M., REHMAN, A. U., et al. 2023. Detection of mosaic variants using genome sequencing in a large pediatric cohort. American Journal of Medical Genetics, Part A, 191, 699-710.

OSENI, G. O., JAIN, D., MOSSEY, P. A., et al. 2018. Identification of paternal uniparental disomy on chromosome 22 and a de novo deletion on chromosome 18 in individuals with orofacial clefts. Mol Genet Genomic Med, 6, 924-932.

PELLACANI, A., ROZELL-SHANNON, L., NEGRELLO, S., et al. 2022. The vanishing port-wine stain birthmark—consideration for a rare type of congenital vascular anomaly. European Journal of Plastic Surgery, 45, 997-1013.

PILLAY, T., WINSHIP, W. S. & RAMDIAL, P. K. 1998. Pigmentary abnormalities in trisomy of chromosome 13. Clin Dysmorphol, 7, 191-4.

POSEY, J. E., ROSENFELD, J. A., JAMES, R. A., et al. 2016. Molecular diagnostic experience of whole-exome sequencing in adult patients. Genetics in Medicine, 18, 678-685.

RETIEF, A. E., KOTZE, G. M. & FOX, M. F. 1979. An analysis of cytogenetic studies. South African Medical Journal, 55, 1076-1081.

ROPERS, H. H., HITZEROTH, H. W., HAMEISTER, H., et al. 1978. Distribution of G6PD phenotypes in red blood cells of Southern African Negroids: evidence for somatic selection. Hum Genet, 42, 215-21.

SAAD, K., ABDELRAHMAN, A. A., ABDEL-RAHEEM, Y. F., et al. 2014. Turner syndrome: Review of clinical, neuropsychiatric, and EEG status: An experience of tertiary center. Acta Neurologica Belgica, 114, 1-9.

SAP, S. N. U., MBONO BETOKO, R., ETOA ETOGA, M., et al. 2020. Observational study of disorders of sex development in Yaounde, Cameroon. Journal of Pediatric Endocrinology and Metabolism, 33, 417-423.

SASSI, H., ELARIBI, Y., JILANI, H., et al. 2021. Beckwith–Wiedemann syndrome: Clinical, histopathological and molecular study of two Tunisian patients and review of literature. Molecular Genetics and Genomic Medicine, 9.

SBIDIAN, E., FELDMANN, D., BENGOA, J., et al. 2010. Germline mosaicism in keratitis-ichthyosis-deafness syndrome: pre-natal diagnosis in a familial lethal form. Clin Genet, 77, 587-92.

SEGAL, F., SPIRO, F. & WILTON, E. 1967. Mosaic XY-XXY in a Bantu--a case report. S Afr Med J, 41, 701-3.

SHALABY, H. M. A. 2011. A study of new potential risk factors for Down syndrome in Upper Egypt. Egyptian Journal of Medical Human Genetics, 12, 15-19.

SHAWKY, R. M., ELKHALEK, H. S. A., GAD, S., et al. 2014. Unilateral proximal focal femoral deficiency, fibular aplasia, tibial campomelia and oligosyndactyly in an Egyptian child - Probable FFU syndrome. Egyptian Journal of Medical Human Genetics, 15, 299-303.

SILE MEFO, H., ABENA OBAMA, M. T., DAHOUN, S., et al. 1996. Down syndrome in Yaoundé, Cameroon. Incidence, karyotype and congenital defects. Annales de Pediatrie, 43, 785-788.

SLIMANI, W., BEN KHELIFA, H., DIMASSI, S., et al. 2019. Clinical and molecular findings in nine new cases of tetrasomy 18p syndrome: FISH and array CGH characterization. Molecular Cytogenetics, 12.

SOARES, C. A., TKACHENKO, N., VALE-FERNANDES, E., et al. 2023. Genetic counseling and carrier screening in candidates for gamete donation at a Portuguese center. Jornal Brasileiro de Reproducao Assistida, 27, 180-184.

SPURDLE, A. B., SHANKMAN, S. & RAMSAY, M. 1995. XX true hermaphroditism in southern African blacks: exclusion of SRY sequences and uniparental disomy of the X chromosome. Am J Med Genet, 55, 53-6.

STARKMAN, M. N. & SHAW, M. W. 1967. Atypical acrocentric chromosomes in Negro and Caucasian Mongols. Am J Hum Genet, 19, 162-73.

SYKES, L. & BRADFIELD, C. 2023. Proteus Syndrome – A one in a million occurrence. South African Dental Journal, 78, 86-91.

TEMTAMY, S. A., ABOUL-EZZ, E. H., EL-HADIDI, S. M., et al. 1994. Oral manifestations and histochemical studies of the gingiva in Down syndrome. The Journal of the Egyptian Public Health Association, 69, 305-326.

TORRES, L., LÓPEZ, M., MÉNDEZ, J. P., et al. 1996. Molecular analysis in true hermaphrodites with different karyotypes and similar phenotypes. Am J Med Genet, 63, 348-55.

TURLEAU, C., TAILLARD, F., DOUSSAU DE BAZIGNAN, M., et al. 1986. Hypomelanosis of Ito (incontinentia pigmenti achromians) and mosaicism for a microdeletion of 15q1. Hum Genet, 74, 185-7.

VAN DYK, E. & PRETORIUS, P. J. 2012. Point mutation instability (PIN) mutator phenotype as model for true back mutations seen in hereditary tyrosinemia type 1 - a hypothesis. J Inherit Metab Dis, 35, 407-11.

VERMA, I. C., MATHEWS, A. R., FAQUIH, A., et al. 1990. Cytogenetic analysis of down syndrome in Libya. The Indian Journal of Pediatrics, 57, 245-248.

VORSTER, A. A., REBELLO, M. T., COUTTS, N., et al. 2004. Arg120stop nonsense mutation in the RP2 gene: mutational hotspot and germ line mosaicism? Clin Genet, 65, 7-10.

WELLS, G. L., BARKER, S. E., FINLEY, S. C., et al. 1994. Congenital heart disease in infants with Down's syndrome. South Med J, 87, 724-7.

WONKAM, A., VEIGNE, S. W., ABASS, A., et al. 2015. Features of Turner syndrome among a group of Cameroonian patients. Int J Gynaecol Obstet, 129, 264-6.

WOODMAN, B. F., JORDAN, M. A., MOLLER, L. I., et al. 1995. The Pallister-Killian syndrome in an African individual. Genetic Counseling, 6, 33-36.

XIAO, A. Y., KANAAN, H. D., LAI, Z., et al. 2022. Role of progenitor cell marker CD133 in supporting diagnosis of collapsing glomerulopathy. Int Urol Nephrol, 54, 1957-1968.

ZAKI, M. S., KAMEL, A. A. & EL-RUBY, M. 2005. Double aneuploidy in three Egyptian patients: Down-Turner and Down-Llinefelter syndromes. Genetic Counseling, 16, 393-402.

ZASLAV, A. L., BLUMENTHAL, D., WILLNER, J. P., et al. 2000. Prenatal diagnosis of trisomy 4 mosaicism. Am J Med Genet, 95, 381-4.

Appendix E: DDD-Africa recruitment criteria and demographic distribution

Table A. 2: DDD-Africa inclusion and exclusion criteria categories.

Inclusion criteria:
A) Moderate to profound developmental delay; with or without dysmorphic features.
B) Multiple major malformations in two or more different organ systems; with or without developmental delay.
C) One major malformation with additional clinically relevant minor anomalies and/or dysmorphic features; with or without developmental delay.
D) Mild developmental delay and additional clinically relevant minor anomalies and/or dysmorphic features; medical geneticist consensus on relevance required.
Exclusion criteria
A) Known conditions with no strong evidence for monogenic aetiology.
B) Suspicion of an acquired brain lesion.
C) Suspected cases with a multifactorial aetiology.
D) Suspected environmental contribution including extreme prematurity and teratogen exposure.
E) Mild intellectual disability or developmental delay with no other clinically relevant abnormalities or dysmorphic features.

Table A. 1: Reasons for recruitment in 62 probands.

Total inclusion criteria category	**N (%)**
A - Moderate to profound developmental delay; with or without dysmorphic features.	37 (59.6)
B - Multiple major malformations in two or more different organ systems; with or without developmental delay.	12 (19.35)
C - One major malformation with additional clinically relevant minor anomalies and/or dysmorphic features; with or without developmental delay.	21 (33.87)
D - Mild developmental delay and additional clinically relevant minor anomalies and/or dysmorphic features; medical geneticist consensus on relevance required.	21 (33.87)
Inclusion criteria combinations	**N (%)**
A only	16 (25.81)
A and B	7 (11.29)
A and C	14 (22.58)
A and D	0 (0)
B only	2 (3.23)
B and C	0 (0)
B and D	3 (4.84)
C only	2 (3.23)
C and D	5 (8.06)
D only	13 (20.97)

Appendix F: Variants identified in a parallel DDD-Africa study

Table A. 3: Pathogenic small variants previously detected in parallel DDD-Africa analysis

Proband ID	Gene	Variant	Consequence	VAF
De novo variants				
13	*MTOR*	ENST00000361445.9:c.5395G>A	missense variant	0.58
14	*SETBP1*	ENST00000649279.2:c.2612T>C	missense variant	0.41
15	*NFIX*	ENST00000592199.6:c.293del	frameshift variant	0.44
21	*NSD2*	ENST00000508803.6:c.3160_3183del	In-frame deletion	0.24
28	*PPP2R1A*	ENST00000322088.11:c.773G>A	missense variant	0.51
34	*CHD7*	ENST00000423902.7:c.8528del	frameshift variant	0.54
38	*CHAMP1*	ENST00000361283.4:c.775_782del	frameshift variant	0.51
43	*CDH2*	ENST00000269141.8:c.1885G>C	missense variant	0.50
47	*ARID1B*	ENST00000636930.2:c.2780del	frameshift variant	0.47
50	*CHD7*	ENST00000423902.7:c.232C>T	stop gained	0.42
51	*CHD7*	ENST00000423902.7:c.5833C>T	stop gained	0.42
53	*AUTS2*	ENST00000342771.10:c.376C>T	stop gained	0.40
66	*EHMT1*	ENST00000460843.6:c.2073dup	frameshift variant	0.56
81	*SCN2A*	ENST00000375437.7:c.727C>T	stop gained	0.43
86	*STXBP1*	ENST00000373299.5:c.416C>T	missense variant	0.45
92	*FOXC1*	ENST00000645831.2:c.392C>G	missense variant	0.32
117	*NSD1*	ENST00000439151.7:c.4913A>G	missense variant	0.39
Inherited variants				
52	*MECP2*	ENST00000453960.7:c.961C>T	missense variant	1.00
118	*FLNA*	ENST00000369850.10:c.620C>T	missense variant	1.00

Del – deletion; dup – duplication; ID – trio identifier; VAF – variant allele fraction in proband; ENST – transcript identifier

Appendix G: Table of variants classified as VUS

Table A. 4: Candidate proband PZMs and parental PZMs classified as variants of uncertain significance.

ID	Variant	Consequence	Panel	BayesDel[$]	SpliceAIΔ	ACMG/AMP code	Class[^]	VAF_P	FDR_P	EM_{Post}	VAF_F	VAF_M
Proband PZM candidates												
17	*RORB*:c.601G>T	missense	DDG2P	0.04	-	PM2, PS2, PP2	VUS	0.18	$9.67x10^{-04}$	22.15	0.00	0.00
46	*MATN3*:c.1379C>A	stop gained	DDG2P	0.59	-	PVS1, PM2	VUS	0.16	$9.67x10^{-04}$	39.13	0.00	0.00
High-level parental gonadosomatic candidates												
58	*POLR3B*:c.2714-16A>G	splice polypyrimidine tract	DDG2P	-	AG-0.65	PM2, PP3, PM6	VUS	0.48	$2.04x10^{-02}$	-	0.00	0.32
108	*TSC2*:c.3037G>C	missense	DDG2P	0.34	-	PM2, PP3, PM6	VUS	0.51	$9.65x10^{-02}$	-	0.00	0.33
75	*OTX2*::c.421C>T	missense	DDG2P	0.13	-	PM1, BS2, PP3, PM6	VUS	0.54	$1.74x10^{-01}$	-	0.00	0.33

ID –Family identifier; $ BayesDel no allele fraction model; ^Class – ACMG/AMP classification; FDRP – FDR adjusted binomial p-value; # EMPost – EM-mosaic posterior odds; VUS – variant of unknown significance; VAFP – proband variant allele fraction; VAFF – father variant allele fraction; VAFM – mother variant allele fraction; AG – acceptor gain

Appendix H: IGV plots for candidate VUS variants

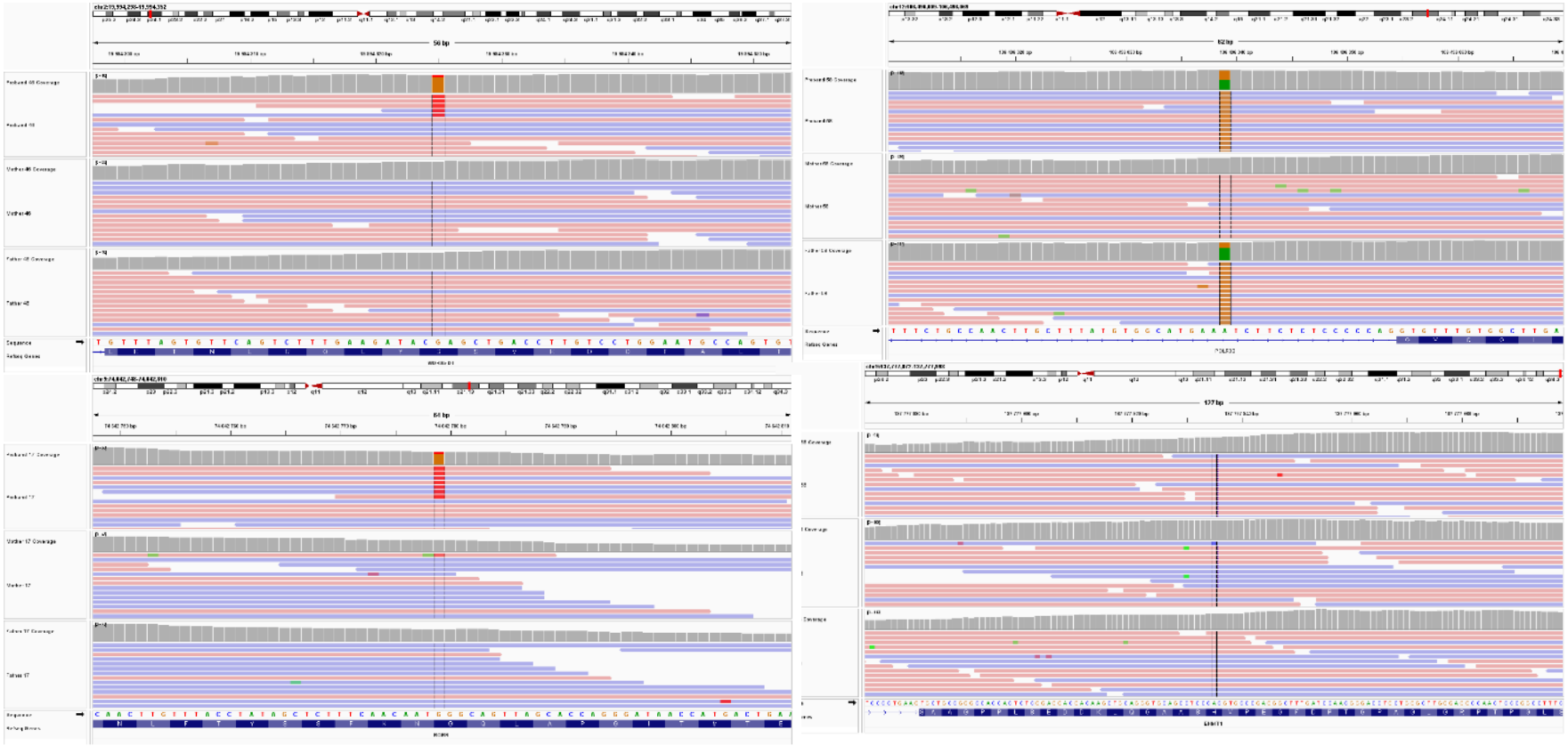


Figure A. 1: IGV plots for VUS candidate varaints

Appendix I: Stratified MCA phenotypes for bubble graph

Table A. 5: The characteristics of 82 MCA phenotypes stratified by chromosomal arm and type

MCA	Count	Mean age (SE)	Mean CF (SE)	Fraction M	EA (%)	SA (%)	WA (%)
1+	23	55.74 (2.16)	5.48 (1.25)	0.43	13.04	43.48	43.48
1p-	14	52.00 (1.60)	9.09 (1.37)	0.50	0.00	64.29	35.71
1p=	45	54.53 (1.46)	3.03 (0.74)	0.51	13.33	46.67	40.00
1q-	4	56.00 (1.35)	11.12 (5.70)	1.00	0.00	25.00	75.00
1q=	57	53.89 (1.02)	1.74 (0.32)	0.46	17.54	31.58	50.88
2+	7	55.29 (3.83)	5.83 (1.77)	0.57	14.29	28.57	57.14
2p-	10	58.40 (3.13)	10.40 (1.40)	0.60	10.00	90.00	0.00
2p=	6	54.17 (4.00)	1.29 (0.31)	0.50	50.00	16.67	33.33
2q-	12	51.50 (2.84)	13.84 (1.95)	0.42	33.33	16.67	50.00
2q=	17	48.94 (1.42)	2.13 (0.67)	0.29	17.65	41.18	41.18
3+	10	51.80 (2.89)	3.94 (1.12)	0.60	10.00	40.00	50.00
3p-	4	57.50 (1.32)	7.39 (2.25)	0.00	50.00	0.00	50.00
3p=	7	58.14 (5.55)	1.85 (0.41)	0.57	14.29	71.43	14.29
3q-	4	60.00 (6.40)	19.45 (5.90)	0.00	0.00	25.00	75.00
3q=	7	57.43 (3.65)	2.23 (0.69)	0.71	42.86	42.86	14.29
4p-	5	53.00 (2.28)	12.60 (3.36)	0.60	20.00	80.00	0.00
4p=	4	51.00 (5.08)	1.61 (0.29)	0.25	25.00	50.00	25.00
4q-	7	58.00 (4.30)	10.73 (2.92)	0.57	14.29	57.14	28.57
4q=	5	55.80 (6.54)	1.21 (0.30)	0.80	0.00	20.00	80.00
5+	64	52.33 (0.91)	2.47 (0.27)	0.58	4.69	15.63	79.69
5p-	9	52.44 (1.68)	9.41 (1.60)	0.11	0.00	55.56	44.44
5p=	8	51.38 (1.88)	0.63 (0.05)	0.63	0.00	25.00	75.00
5q-	10	53.30 (2.49)	4.28 (1.09)	0.30	0.00	70.00	30.00
5q=	19	51.68 (1.57)	2.01 (0.65)	0.53	5.26	26.32	68.42
6+	14	53.29 (1.22)	13.45 (4.31)	0.57	14.29	35.71	50.00
6p=	23	54.48 (1.96)	2.19 (0.62)	0.48	21.74	56.52	21.74
6q-	32	55.56 (1.39)	10.09 (1.32)	0.53	6.25	56.25	37.5
6q=	9	56.44 (3.61)	1.65 (0.38)	0.67	22.22	55.56	22.22
7+	6	55.00 (2.37)	7.20 (3.30)	0.50	0.00	83.33	16.67
7p-	14	49.86 (1.89)	10.53 (1.48)	0.71	14.29	50.00	35.71

7p=	12	56.58 (2.14)	1.61 (0.65)	0.33	25.00	41.67	33.33
7q-	26	54.88 (1.67)	11.09 (1.31)	0.54	3.85	50.00	46.15
7q=	13	54.08 (2.84)	2.40 (0.77)	0.62	15.38	61.54	23.08
8+	7	62.86 (4.08)	5.38 (2.37)	0.57	14.29	57.14	28.57
8p-	10	54.40 (1.19)	9.60 (1.86)	0.40	0.00	80.00	20.00
8p=	7	54.00 (3.09)	4.28 (2.02)	0.71	14.29	57.14	28.57
8q-	5	48.40 (0.93)	8.92 (1.69)	0.60	0.00	40.00	60.00
8q=	6	53.33 (1.69)	0.64 (0.05)	0.33	16.67	66.67	16.67
9+	8	50.25 (4.12)	10.05 (5.76)	0.25	0.00	62.5	37.5
9p=	12	54.83 (3.53)	10.04 (3.84)	0.83	0.00	83.33	16.67
9q-	11	53.09 (1.59)	8.75 (2.29)	0.73	0.00	45.45	54.55
9q=	33	55.58 (1.65)	1.90 (0.34)	0.64	12.12	54.55	33.33
10q-	10	50.50 (2.24)	8.39 (1.94)	0.20	0.00	60.00	40.00
10q=	14	53.71 (2.27)	1.79 (0.42)	0.86	14.29	50.00	35.71
11p-	9	47.56 (1.97)	11.06 (1.30)	0.44	11.11	33.33	55.56
11p=	22	52.91 (1.20)	4.32 (1.46)	0.50	0.00	59.09	40.91
11q-	18	55.94 (3.01)	6.46 (1.12)	0.56	11.11	72.22	16.67
11q=	7	51.00 (2.12)	2.81 (1.12)	0.29	0.00	57.14	42.86
12+	11	55.64 (1.19)	6.10 (2.19)	0.36	18.18	45.45	36.36
12p-	4	52.75 (3.64)	9.84 (2.58)	0.50	0.00	50.00	50.00
12p=	5	53.40 (1.86)	3.14 (0.71)	0.40	40.00	20.00	40.00
12q-	15	52.87 (2.83)	9.80 (1.49)	0.47	0.00	60.00	40.00
12q=	11	52.09 (1.42)	1.62 (0.27)	0.36	36.36	36.36	27.27
13+	4	58.00 (4.81)	6.36 (1.45)	0.00	0.00	75.00	25.00
13q-	34	53.82 (1.57)	7.18 (1.10)	0.76	11.76	35.29	52.94
13q=	6	57.50 (4.22)	1.19 (0.23)	0.83	0.00	66.67	33.33
14q-	78	54.03 (0.93)	11.48 (0.49)	0.46	29.49	52.56	17.95
14q=	33	52.91 (1.49)	2.26 (0.33)	0.42	24.24	51.52	24.24
15q=	16	53.56 (2.64)	2.23 (0.90)	0.56	25.00	37.5	37.5
16p-	15	54.87 (2.00)	8.25 (1.40)	0.33	0.00	46.67	53.33
16p=	15	53.80 (2.02)	5.77 (2.19)	0.40	13.33	53.33	33.33
16q-	7	53.86 (3.50)	7.35 (2.30)	0.86	0.00	42.86	57.14
16q=	20	52.95 (1.99)	1.65 (0.29)	0.60	15.00	50.00	35.00
17+	9	52.56 (3.48)	4.89 (1.29)	0.44	0.00	33.33	66.67
17p-	10	59.00 (4.21)	10.36 (1.67)	0.40	0.00	80.00	20.00
17p=	4	58.75 (5.94)	2.66 (1.22)	0.50	0.00	100.00	0.00
17q-	15	54.00 (2.90)	9.94 (1.32)	0.53	13.33	80.00	6.67
17q=	29	49.34 (1.33)	2.72 (0.57)	0.66	13.79	37.93	48.28

18q-	8	51.38 (1.82)	9.96 (2.19)	0.63	0.00	25.00	75.00
18q=	7	48.43 (2.84)	1.67 (0.47)	0.57	28.57	57.14	14.29
19+	6	53.50 (4.89)	7.56 (3.00)	0.33	50.00	33.33	16.67
19p-	16	54.69 (2.25)	8.46 (1.21)	0.44	12.5	43.75	43.75
19p=	5	53.80 (2.99)	2.89 (0.87)	0.60	0.00	40.00	60.00
19q=	9	57.67 (4.06)	2.81 (0.53)	0.44	11.11	44.44	44.44
20q-	16	58.88 (2.43)	7.26 (1.23)	0.69	12.5	50.00	37.5
20q=	6	53.33 (4.79)	3.86 (1.82)	0.67	16.67	66.67	16.67
21+	64	54.44 (0.95)	2.37 (0.15)	0.61	34.38	57.81	7.81
21q-	11	54.45 (3.16)	8.69 (1.83)	0.45	0.00	63.64	36.36
21q=	17	53.94 (2.29)	1.96 (0.81)	0.53	17.65	47.06	35.29
22+	11	56.91 (2.96)	3.26 (0.67)	0.27	9.09	72.73	18.18
22q-	8	53.50 (3.81)	12.06 (2.00)	0.63	0.00	50.00	50.00
22q=	16	52.06 (1.81)	2.09 (0.46)	0.50	18.75	50.00	31.25

1-20 = chromosome; q = q -arm, p = p-arm; - represents a loss, = represents a CN-LOH, + represents a gain; M = male; CF = cell fraction (%); SE = standard error of mean; SA = Southern Africa; EA = Eastern Africa; WA = Western Africa

Appendix J:Multiple MCA association analysis results

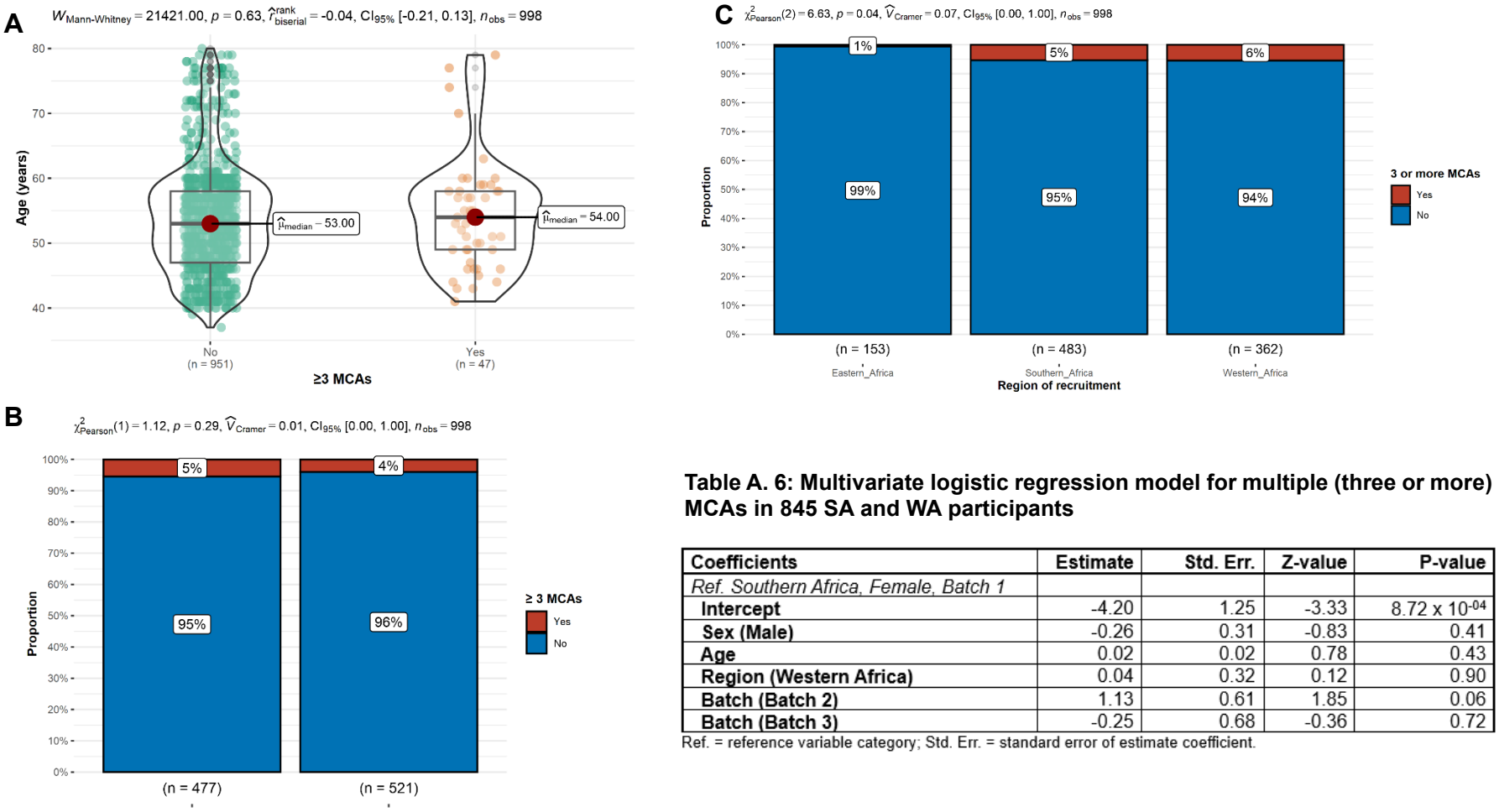


Table A. 6: Multivariate logistic regression model for multiple (three or more) MCAs in 845 SA and WA participants

Coefficients	Estimate	Std. Err.	Z-value	P-value
Ref. Southern Africa, Female, Batch 1				
Intercept	-4.20	1.25	-3.33	8.72×10^{-04}
Sex (Male)	-0.26	0.31	-0.83	0.41
Age	0.02	0.02	0.78	0.43
Region (Western Africa)	0.04	0.32	0.12	0.90
Batch (Batch 2)	1.13	0.61	1.85	0.06
Batch (Batch 3)	-0.25	0.68	-0.36	0.72

Ref. = reference variable category; Std. Err. = standard error of estimate coefficient.

Figure A. 2: Univariate analysis of multiple MCAs by age, sex, and region of recruitement for 998 participants with at least one autosomal MCA

Appendix K:Expanded autosomal MCAs

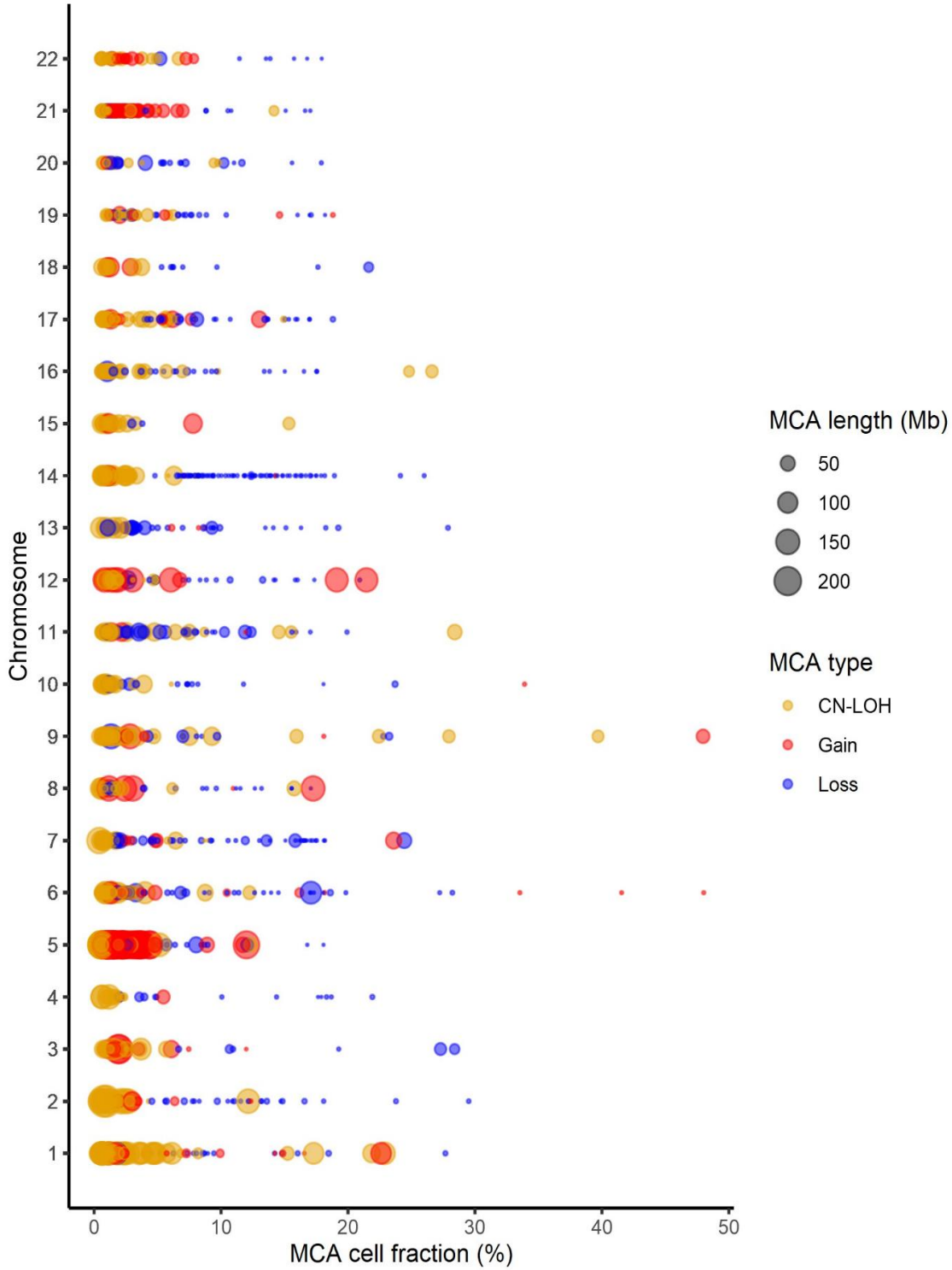


Figure A. 3: The cell fraction distribution of detected autosomal MCAs by type, size, and chromosomal location

Bubbles represent discrete MCA events plotted by genome location over the estimated fraction of cells with the MCA event. Size of the MCA event is reflected by the size of the bubble. Colour illustrates the predicted MCA type.

Appendix L: Turnitin Report

Turnitin Originality Report

Processed on: 30-Sep-2025 2:14 PM SAST
ID: 2766709274
Word Count: 118965
Submitted: 1

Jonathan Evans PhD Final Submission Corrected.pdf By JONATHAN EVANS

Similarity Index	Similarity by Source	
7%	Internet Sources:	4%
	Publications:	6%
	Student Papers:	1%

< 1% match (Internet from 18-Jan-2025)
https://www.frontiersin.org/research-topics/8265/epub

< 1% match (Internet from 06-Jan-2025)
https://www.frontiersin.org/research-topics/19699/epub

< 1% match (Internet from 14-Mar-2025)
https://www.frontiersin.org/research-topics/18181/epub

< 1% match (Internet from 24-May-2025)
https://www.frontiersin.org/research-topics/8472/epub

< 1% match (Internet from 10-Jul-2024)
https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2022.815759/epub?isPublishedV2=false

< 1% match (Internet from 21-Jul-2024)
https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2023.1085210/full

< 1% match (Internet from 08-Jul-2024)
https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2024.1408952/full

< 1% match (Internet from 19-Mar-2025)
https://www.frontiersin.org/research-topics/9393/epub

< 1% match (Internet from 14-Mar-2025)
https://www.frontiersin.org/research-topics/1960/epub