

https://dspace.org/

Establishing a baseline for developmental disorder diagnosis by evaluating current processes and mapping common benign copy number variation in Africa

Wiener, Emma Karin

2022

https://hdl.handle.net/10539/34335

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

Establishing a Baseline for Developmental Disorder Diagnosis by Evaluating Current Processes and Mapping Common Benign Copy Number Variation in Africa

Emma Karin Wiener

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

February 2022

Supervisors: Prof. Amanda Krause, Prof. Zané Lombard, Prof. Scott Hazelhurst

Declaration

I Emma Wiener declare that this thesis is my own, unaided work, unless otherwise specified in the text. 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 this or any other university.



…….. **04/06/2022** …….

Emma Karin Wiener Date

Publications and Presentations

Poster Presentations:

- **European Society of Human Genetics Meeting (Virtual conference) – 6–9th June 2020**
  - Wiener E.K., Lombard Z., Krause A. Exome sequencing as a valuable first-line investigation for developmental disorders in low- & middle-income countries: Insights from a South African Cohort
- **Faculty of Health Sciences Virtual Research Day – 15th October 2020**
  - Wiener E.K., Lombard Z., Krause A. Exome sequencing as a valuable first-line investigation for developmental disorders in low- & middle-income countries: Insights from a South African Cohort

Oral Presentations:

- **16th H3Africa Consortium Meeting (Virtual conference) – 7–10th September 2020**
  - Wiener E.K., Establishing a Baseline for Developmental Disorder Diagnosis by Evaluating Current Processes and Mapping Common Benign Copy Number Variation in Africa
  - Awarded First Prize for Best Fellows presentation
- **American Society of Human Genetics Virtual Meeting – 18–22nd October 2021**
  - Wiener E.K., Cottino L.A., Jakubosky D., Macleod A., Noyes H., Nyangiri., Krause A., Hazelhurst S., Lombard Z., as members of the H3Africa Consortium. An analysis of population copy number variation in sub-Saharan African genomes

Publication submitted for review:

- Wiener E.K., Lombard Z., Buchanan J., Krause A. for the DDD-Africa research group as members of the H3Africa Consortium. Genetic services in a resource-constrained African setting – diagnostic odysseys, non-individualised testing strategies and the need for exome sequencing

Abstract

Developmental disorders (DDs) are life altering and debilitating conditions believed to be present at high rates in sub-Saharan Africa. The non-specific presentation of DDs and their heterogenous genetic aetiologies make them challenging to diagnose. Guidelines to assist in diagnosing DDs currently recommend whole exome sequencing (WES) as the first-line investigation, as WES can detect various types of genetic variants including copy number variants (CNVs). The high cost of WES means strong motivation for its necessity is required, but little research has been conducted and published in Africa about patients with DD and their diagnostic process. Baseline population variants are required for interpretation of WES results, but baseline population CNVs in African individuals have not been well characterised. We therefore aimed to address these two knowledge gaps. To address the first, a file audit was performed on 934 patient records, presenting over a year period, to a medical genetics clinic in South Africa. We found that 83% of patients presented with DDs. The largest group of patients (57%) presented with rare, less recognizable conditions and 92% of these patients remained undiagnosed after available testing was performed. Patients with DDs are the largest group of patients seen, and diagnostic testing approaches are limited, so where clinical features are not distinct, the diagnostic yield is low. Therefore, a significant number of these patients would benefit from a higher yield test like WES. To address the second gap, 1027 high coverage whole genome sequences of individuals from west, central, southern and east Africa, were analysed using Manta and Graphtyper2, and 919 of the samples with Genome STRiP. A high confidence set of 9001 variants called by both tools was produced, conforming to expected parameters for population CNVs. CNVs were found to differ between African regions. 17% of variants were novel, and a number of these were high copy number multi-allelic CNVs. CNVs involving DD genes uncovered African specific allele frequencies and some novel CNVs. This dataset, and continued work, will serve to enrich reference databases with African CNVs that are important for genetic research in African ancestry individuals.

Acknowledgments

I would like to express my gratitude to the following people and entities:

Firstly, to my supervisors Prof Amanda Krause, Prof Zané Lombard and Prof Scott Hazelhurst for all their guidance, input, encouragement and help at each and every stage of this PhD journey. Thank you for always pushing me to improve and helping me to persevere at the hardest times.

The DDD-Africa Study and by extension the National Institute of Health and National Institute of Mental Health for the PhD Fellowship that enabled me to pursue this PhD project. Research reported in this publication was supported by the National Institute Of Mental Health of the National Institutes of Health under Award Number U01MH115483. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

The PV Tobias Educational Bursary of which I was a grateful recipient for two years.

To Lydia Wu and Gege Ran for their assistance in capturing many patient files and the pilot analyses they performed on this data.

To Dr James Buchanan who gave me valuable insight and assistance on the paper that emanated from Chapter two.

To Dr David Jakubosky for his merging and stitching pipeline for Genome STRiP and all his assistance on performing these quality control steps.

Table of Contents

Declaration ..... **ii**

Publications and Presentations ..... **iii**

Abstract ..... **iv**

Acknowledgments ..... **v**

Table of Contents ..... **vi**

List of Figures ..... **xi**

List of Tables ..... **xiv**

Nomenclature ..... **xv**

Preface ..... **xvi**

Introduction **1**

1.1 Developmental Disorders 2

1.1.1 Aetiology of developmental disorders 2

1.2 Developmental Disorders in Africa 3

1.3 Genetic Mechanisms of Developmental Disorders 4

1.3.1 Single nucleotide variants 4

1.3.2 Structural variants and copy number variants 4

1.4 Diagnostic Process for Developmental Disorders 7

1.5 Diagnostic Tests for Copy Number Variation 9

1.5.1 Karyotyping 9

1.5.2 Fluorescent *in situ* hybridization 10

1.5.3 Multiplex ligation-dependent probe amplification 10

1.5.4 Chromosomal microarray 11

A. Array comparative genomic hybridization 11

B. Single nucleotide polymorphism array 11

1.5.5 Next generation sequencing 12

A. Whole exome sequencing 12
B. Whole genome sequencing 14
1.6 Barriers to Implementation of Whole Exome Sequencing 15
1.6.1 Technological and cost barrier 15
1.6.2 Need for African population variant data 16
1.7 Structural Variation Discovery Globally 17
1.7.1 Structural variation in southern African populations 20
1.8 Detection of Structural Variation from Whole Genome Sequences 21
1.8.1 Read pair method 22
1.8.2 Split read method 23
1.8.3 Read depth method 23
1.8.4 Assembly-based method 24
1.8.5 Combined tools 24
1.8.6 Structural variant calling on multiple genomes 25
1.8.7 Application of structural variation algorithms 27
1.9 Aims and Outline of the Study 28

Characterization of a Genetics Clinic Cohort and Evaluation of the Diagnostic Process in South Africa 30
2.1 Introduction 31
2.2 Materials & Methods 32
2.2.1 Study Design and Case Review Selection 32
2.2.2 The Medical Genetics Clinic 32
2.2.3 Ethics 33
2.2.4 Data capture & management 33
2.2.5 Diagnostic costing 34
2.2.6 Data analysis 34
2.3 Results 35
2.3.1 Overview of the cohort 35
2.3.2 Presentation and diagnostic outcomes 36
2.3.3 Patient diagnoses 37

2.3.4 Characteristics of patient groups A, B and C 39
A. Group A – Aneuploidies 39
B. Group B – Easily recognisable conditions 39
C. Group C – Rare, less recognizable conditions 40
2.3.5 Cost of Genetic Investigations 43
2.4 Discussion 44

An Analysis of Population Copy Number Variation in Sub-Saharan African Genomes 48

3.1 Introduction 49
3.2 Methods 50
3.2.1 Datasets 50
A. Cell Biology Research Laboratory HIV Study 50
B. H3Africa-Baylor 51
C. African Wits-INDEPTH Partnership for the Genomic Study (AWI-Gen) project 51
D. Southern African Human Genome Programme 52
E. Simons Genome Diversity Project 52
F. 1000 Genomes Project Consortium 52
3.2.2 Whole genome sequence alignment to Human Build 38 52
3.2.3 Ethical clearance 53
3.2.4 Copy number variant calling algorithms 54
3.2.5 Genome STRiP 55
A. Implementation 55
B. Quality control of detected sites 55
3.2.6 Manta and Graphytyper2 57
3.2.7 Number of copy number variants per sample 58
3.2.8 Intersection of Manta and Genome STRiP datasets 58
3.2.9 Comparison to established databases 59
3.2.10 Functional analyses 59
3.2.11 Principal component analysis 59
3.2.12 Variants overlapping developmental disorder genes 60

3.2.13 Regional comparison of bi-allelic developmental disorder gene variants 60
3.2.14 Pathogenicity prediction 61
3.3 Results 61
3.3.1 Final number of samples included 61
3.3.2 Overview of the results section 62
3.3.3 Entire Manta and Genome STRiP call sets 63
A. Variant type breakdown of each entire call set 63
B. Size profile of each entire call set 63
C. Variant allele frequency profile of each entire call set 66
D. Principal component analysis of the Manta call set reveals batch effects and some regional differences 70
3.3.4 Intersection of Manta and Genome STRiP call sets 72
A. Rationale 72
B. Variant type breakdown of the intersection call set 72
C. Size profile of the intersection call set 73
D. Variant allele frequency profile of the intersection call set 73
E. Distribution of copy number variants across the genome 74
F. Functional analysis of intersection call set 76
G. Novel African variants 77
3.3.5 Copy number variants overlapping genes known to cause developmental disorders 80
A. Rationale 80
B. Variant type breakdown of the developmental disorder gene call set 81
C. Novel variants that overlap developmental disorder genes 82
D. Differences of copy number variants in developmental disorder genes between African regions 83
E. Functional analysis of the developmental disorder gene variants 84
F. Predicted pathogenicity of variants 88
3.4 Discussion 89
3.4.1 Overview 89

3.4.2 Profiles of separate datasets 89
3.4.3 The intersection variant set 90
3.4.4 Novel African variants 92
3.4.5 African regional differences 93
3.4.6 Baseline population variants involving developmental disorder genes 93
3.4.7 Conclusion 95

Conclusion 96
4.1 Improvement Required in the Diagnosis of Developmental Disorders 97
4.2 Diagnostic Test with Higher Yield Needed for Patients with Developmental Disorders 97
4.3 A Catalogue of African Population Copy Number Variants 99
4.4 Challenges and Study Limitations 101
4.5 Future Research 103
4.6 Concluding Remarks 105

References 105

Appendix I Human Genetics Retrospective Data Ethics Certificate 118

Appendix II Emma Wiener PhD Study Ethics Certificate 120

Appendix III Phenotypic Category List 122

Appendix IV CBRL HIV Study Ethics Certificate 124

Appendix V AWIGen Ethics Certificate 126

Appendix VI SAHGP Data Permission 128

Appendix VII Plagiarism Declaration 130

Appendix VIII Turnitin Report 132

List of Figures

Chapter 1 – Introduction

Figure 1. 1: Types of SVs compared to a reference region.. 6

Figure 1. 2: Flow chart of decision making for investigations for GDD.. 8

Figure 1. 3: Four basic SV detection methods from short read WGS.. 22

Chapter 2 – Characterization of a Genetics Clinic Cohort and Evaluation of the Diagnostic Process in South Africa

Figure 2. 1: Overview of the functional structure of the genetics clinic and the two main types of patient referrals 33

Figure 2. 2: Overall percentages of patient diagnostic outcomes.. 37

Figure 2. 3: Top 10 genetic diagnoses made in the cohort.. 38

Figure 2. 4: Three main groups of patients from the file review identified with differing diagnostic profiles. 38

Figure 2. 5: Top ten phenotypes observed in patients in Group C.. 41

Figure 2. 6: Genetic Tests of undiagnosed patients in Group C in an attempt to reach a diagnosis 42

Figure 2. 7: Number of genetic investigations undergone by undiagnosed patients in Group C.. 42

Figure 2. 8: Total cost of genetics investigations for each Group C patient against the length of time the patient has been in the clinic.. 43

Chapter 3 – An Analysis of Population Copy Number Variation in Sub-Saharan African Genomes

Figure 3. 1: Distribution of samples across Africa 53

Figure 3. 2: Final number of samples included for the Manta (a) and Genome STRiP (b) pipelines 62

Figure 3. 3: Number of variants called by Manta and Genome STRiP in each of the size classes.. 64

Figure 3. 4: Violin plot of CNV size distribution. 65

Figure 3. 5: Number of CNVs called by Manta in four VAF classes.. 67

Figure 3. 6: Number of CNVs called by Genome STRiP in three VAF classes.. 68

Figure 3. 7: Number of CNVs per VAF class called by Manta and Genome STRiP.. 69

Figure 3. 8: Manta variants plotted by VAF and size (bp).. 69

Figure 3. 9: PCA of Manta variants showing batch effects.. 70

Figure 3. 10: PCA of all Manta variants showing some regional differences.. 71

Figure 3. 11: PCA of H3Africa data showing regional differences between east, west and southern Africa.. 71

Figure 3. 12: Intersection of the comparable Manta and Genome STRiP variants... 73

Figure 3. 13: Violin plot of CNV size distribution of intersection variant set.. 73

Figure 3. 14: Number of CNVs in intersection variant set in each of the four VAF classes.. 74

Figure 3. 15: Genomic representation of CNV density of intersection Manta and Genome STRiP variant set.. 75

Figure 3. 16: Functional analysis of intersection variant dataset. 10% of all variants in the combined dataset overlapped exons or entire genes. 76

Figure 3. 17: Percentage of variants involving coding transcripts for each VAF class. 77

Figure 3. 18: Genomic representation of CNV density of novel variants..... 78

Figure 3. 19: Distribution of diploid copy number for chr1:125175980–125177652..79

Figure 3. 20: Functional analysis of novel variants.. 80

Figure 3. 21: Genomic locations of variants overlapping DD genes.....82

Figure 3. 22: Comparison of intersection DD gene CNVs found in individuals from each African region.....84

Figure 3. 23: Functional analysis of variants in intersection DD gene variant set.....85

Figure 3. 24: Gene location of intersection DD gene variants that overlap coding transcripts.. 88

List of Tables

Chapter 2 – Characterization of a Genetics Clinic Cohort and Evaluation of the Diagnostic Process in South Africa

Table 2. 1: Summary of information captured on each patient 34

Table 2. 2: Demographic profile of the genetics clinic patient cohort 36

Table 2. 3 Comparison of test use and costs between Groups A, B and C 44

Chapter 3 – An Analysis of Population Copy Number Variation in Sub-Saharan African Genomes

Table 3. 1: Breakdown of variant types detected by each tool 63

Table 3. 2: Low frequency and common population CNVs involving whole genes or exons of DD genes. 87

Nomenclature

ACMG – American College of Medical Genetics

ADME – Absorption, Distribution, Metabolism and Excretion

ATP – Adenosine Triphosphate

CMA - Chromosomal Microarray

CNV – Copy Number Variant

DD – Developmental Disorder

DDG2P – DD Gene2Phenotype

DGV – Database of Genomic Variants

DDD-Africa – Deciphering Developmental Disorders in Africa

DDD-UK - Deciphering Developmental Disorders in the United Kingdom

DNA – Deoxyribonucleotide Acid

FISH – Fluorescent *In Situ* Hybridization

GDD – Global Developmental Delay

GSK – GlaxoSmithKline

HIV – Human Immunodeficiency Virus

NHLS – National Health Laboratory Services

MLPA – Multiplex Ligation-dependent Probe Amplification

NAD – Nicotinamide Adenine Dinucleotide

NDD – Neurodevelopmental Disorder

NGS – Next Generation Sequencing

PCA – Principal Component Analysis

PCR – Polymerase Chain Reaction

RNA – Ribonucleic Acid

SNV – Single Nucleotide Variant

SV – Structural Variant

VAF – Variant Allele Frequency

VCF – Variant Call Format

WES – Whole Exome Sequencing

WGS – Whole Genome Sequences

Preface

Developmental disorders (DDs) are a group of conditions characterized by abnormal growth and development. They usually present from infancy to early childhood with physical abnormalities, delayed developmental milestones and sometimes both. DDs have many aetiologies, but it is believed that about half have a genetic aetiology (Srour and Shevell, 2014). The prevalence of these disorders in South Africa is unknown, but the Global Burden of Disease Study 2016 reports that developmental disabilities in sub-Saharan Africa have increased by 70% since 1990 (Olusanya *et al.*, 2018). This suggests that there are a large number of children in sub-Saharan Africa with genetic-based DDs. This group of patients, however, have not been prioritized in term of research or appropriate healthcare services (Kromberg *et al.*, 2013). The genetic aetiologies of DDs are highly heterogenous, making these conditions difficult to diagnose. Even in developed countries with ample testing capabilities the genetic aetiology may not be found in up to 50% of cases. In a resource-constrained setting such as Africa this number is likely even higher, though little has been published about patients with DDs in Africa in terms of aetiologies, prevalence or diagnostic outcomes.

This PhD research was conducted as part of the Deciphering Developmental Disorders in Africa (DDD-Africa) study, that has been initiated to investigate DDs in sub-Saharan Africa with a suspected genetic aetiology (https://h3africa.org/index.php/ddd-africa/). This study is being done in collaboration with scientists from the Welcome Sanger Institute who conducted a large-scale investigation into DD in the United Kingdom. The Deciphering Developmental Disorders in the United Kingdom (DDD-UK) Study (Firth *et al.*, 2011) aimed to improve understanding of genetic causes of DD and advance clinical genetic practice for diagnosing children with DD. This study is an excellent example of the research into DDs globally that has started to uncover some of the heterogeneous and complex genetic aetiologies and phenotypes of this diverse group of conditions.

The DDD-UK study used chromosomal microarrays (genome wide screening for copy number variants (CNVs)) and whole exome sequencing (WES) (sequencing of the exome (all protein coding regions ~1% the genome)) to establish a diagnosis in patients enrolled in the study. WES can detect both single nucleotide variants (SNVs) and CNVs and is now the recommended first-line investigation for patients with DDs,

having a higher yield of 30-53% compared to ~12% chromosomal microarray (Srivastava *et al.*, 2019; Manickam *et al.*, 2021).

The DDD-Africa study is using a similar approach. The study aims to assess and characterise 500 patients with undiagnosed DDs. These children have a thorough clinical assessment and deep phenotyping by a medical geneticist, and non-genetic causes are excluded as far as possible. Secondly, blood samples are being collected from trios (the patient and their parents) and WES performed on the DNA from these samples. This WES data will be analysed with bioinformatic pipelines (for SNVs and CNVs) to try to discover the genetic cause of the DD in these patients. These pipelines filter and prioritize variants from the full set of variants, detected in a patient's WES, to a smaller set of candidate variants with the highest likelihood of causing the patients disease (Smedley *et al.*, 2015). One important set of variants that need to be filtered out are likely benign variants that are found in healthy individuals in the population from which the patient is from.

The goals of the DDD-Africa study are to uncover phenotypes and genetic aetiologies of DD in Africa, to improve research capabilities into such conditions in Africa and to produce a practical plan for the implementation of WES in an African setting. Achievement of these goals will result in increased knowledge of genetic DDs and increased capacity to improve the diagnostic yield in patients with DD in Africa.

At the outset of the DDD-Africa Study some key gaps in baseline knowledge were identified. Addressing these gaps will significantly aid in the successful completion of the goals of DDD-Africa study. This PhD study aims to address two of these areas:

The first is a lack in the knowledge of the current diagnostic process for patients with DD. The yield and cost of current testing methods is unknown, and additionally a review and summary of the population of patients presenting to the genetics clinic has not been systematically studied. *Chapter Two* describes the characterisation of the genetics clinic patient cohort and the yield of current diagnostic processes in use.

The second area is the lack of knowledge of baseline African CNVs. Baseline population SNVs have been more comprehensively assessed (Choudhury *et al.*, 2020) but baseline African CNVs have not been well described. Knowledge of these variants is important for successful filtering and interpretation of potentially pathogenic novel CNVs detected from WES. However the value of the dataset of population CNVs has

value beyond its use for diagnostic WES interpretation in DDD-Africa. Population CNVs from a diverse African cohort will be a valuable addition as a source of African CNVs to a database of genomic variation for broad use. This aim was well aligned to an aim of the H3Africa GlaxoSmithKline (GSK) absorption, distribution, metabolism and excretion (ADME) collaboration study (da Rocha *et al.*, 2021), to characterise baseline CNVs present in the ADME genes. Permission had also been given to analyse baseline CNVs from the H3Africa Baylor WGS dataset (Choudhury *et al.*, 2020), by the H3Africa Consortium, to the TrypanoGen study, who had previously investigated baseline CNVs in their study cohort (Nyangiri *et al.*, 2020). The CNV analysis for this study was therefore done in collaboration with both the H3Africa GSK ADME project and the TrypanoGen group under the banner of the H3Africa Consortium. *Chapter Three* describes the production and analysis of a dataset of population CNVs from diverse African populations.

Chapter 1

Introduction

1.1 Developmental Disorders

Developmental Disorders (DDs) is a broad term used to describe conditions that affect normal growth and development. It encompasses physical malformations and abnormalities as well as abnormal neurological development. Physical abnormalities and malformations such as organ malformations or limb anomalies are evident even from embryonic development or infancy. Abnormal neurological development usually manifests in early childhood as delayed developmental milestones, or later in life as intellectual disability. Global developmental delay (GDD) is the term used if intellectual disability is observed before the age of five years, when standardised tests to diagnose intellectual disability are not yet possible. GDD is one of the most common presentations of DDs and is defined as significant functional delay in milestones of at least two domains (motor, speech/language, cognition, personal–social and activities of daily living). Intellectual disability involves significant limitations of intellectual function and adaptive behaviour in conceptual, social and practical areas (American Psychiatric Association, 2013). Associated with intellectual disability in these children are other conditions such as autism spectrum disorders, seizures and other behavioural disorders, which make DDs very pervasive, chronic and life altering conditions (Srour and Shevell, 2014).

1.1.1 Aetiology of developmental disorders

DDs can either be congenital or acquired and are usually classified by the time of insult as prenatal, perinatal or postnatal. Prenatal insults include all genetic abnormalities as well as teratogenic exposures. Perinatal onset causes are largely due to neonatal and birth complications, such as birth asphyxia, and postnatal causes include infant and childhood infections, trauma, severe nutritional deficiencies and protein energy malnutrition. Prenatal causes are considered to account for 50–70% cases of DDs with perinatal and postnatal accounting for similar remaining proportions (Strømme, 2000). Studies assessing intellectual disability and developmental delay have shown that around 25–50% of DDs have a genetic aetiology (Srour and Shevell, 2014).

Another mechanism that is emerging as a cause for DDs may not fall into either congenital or acquired causes – multifactorial disease mechanism. Multifactorial diseases are caused by a combination of genetic predisposition or vulnerability and environmental exposures (Parenti *et al.*, 2020). In the context of DD this has not been

completely understood, but it is well known that the brain is especially susceptible to environmental exposures not only in utero but in childhood as well. These exposures are likely not as overt as teratogenic exposures but may result in epigenetic effects or create an environment that causes disease in a genetically vulnerable person (Tran and Miyake, 2017).

1.2 Developmental Disorders in Africa

A study into the prevalence of GDD by Grantham-McGregor *et al.* (2007) estimated that there were 200 million children worldwide at risk for DDs and that the majority of them lived in sub-Saharan Africa and South Asia. Olusanya *et al.* (2018) performed an analysis of the global burden of disease for children under five in 195 countries globally and found that developmental disabilities are found at disproportionally higher rates in low- and middle-income countries (LMICs). They discuss that this is partially the result of decrease in under five mortality, which has halved in the past 20 years globally, including a sharp decline in LMIC countries such as South Africa. Disabilities globally have decreased, but in sub-Saharan Africa they have increased by 70% since 1990. This high burden of disabilities means there is great need for increasing systematic research of and care for children with disabilities. Another issue they raise is that many countries with high burdens of disabilities, often have a dearth of population-based data to inform health policy and interventions (Olusanya *et al.*, 2018). This lack of population-based data on the prevalence of childhood disabilities such as DDs has resulted in services to diagnose and treat such conditions being poorly funded and prioritised in South Africa (Kromberg *et al.*, 2013; Malherbe *et al.*, 2021).

There have been very few studies to establish the overall prevalence and dominant aetiologies of DD in South Africa. Two smaller studies investigating the prevalence of DDs in South Africa were conducted in Kwa-Zulu Natal (Couper, 2002) and Mpumalanga (Christianson *et al.*, 2002) and they found the prevalence of DDs as 83/1000 and 35.6/1000 respectively. This number is far higher than those from similar types of studies done in other countries where most countries reported <10/1000 the highest being Pakistan who reported 24.3/1000 (Durkin, 2002). Studies investigating cerebral palsy, a common DD in South Africa, have shown a larger proportion of perinatal insult events than expected (van Toorn *et al.*, 2007). Additionally, studies into

foetal alcohol syndrome have shown very high rates in South Africa (Olivier *et al.*, 2016).

Given the fact that ~50% of DDs are believed to have a genetic aetiology (Srour and Shevell, 2014), it is likely that genetic DDs form a significant portion of the greatly increased burden of DDs in sub-Saharan Africa (Olusanya *et al.*, 2018). However, there have been no large-scale investigations into genetically-based DD and its prevalence, and the true distribution of aetiologies is unknown. As mentioned in the preface, the DDD-Africa study is investigating DDs of genetic aetiology in sub-Saharan Africa, to begin to understand the aetiologies and phenotypes of these conditions in Africa.

1.3 Genetic Mechanisms of Developmental Disorders

The genetic mechanisms underlying DDs are highly heterogeneous and include variants of almost all sizes and types. As the focus of this project is on CNVs, SNVs will not be dealt with below in detail, but they are indisputably a very important genetic cause of DDs.

1.3.1 Single nucleotide variants

Single nucleotide variants (SNVs) are responsible for a large proportion of DDs with genetic aetiology (Deciphering Developmental Disorders, 2015). They cause disease by altering the protein coding sequence of a gene, or altering a binding splice or regulatory site, ultimately affecting the final protein product. The consequences of these different types of variants are known as non-synonymous, missense, nonsense or loss of function variants, depending on the way in which the protein product is altered. SNVs are easier to detect than CNVs (De Coster and Van Broeckhoven, 2019) and so have been well characterised in the context of DDs, which is why they are not the focus of this study.

1.3.2 Structural variants and copy number variants

Structural Variants (SVs) underlie a large proportion of variation between individuals. Alkan *et al.* (2011) and Sudmant *et al.* (2015b) report that SV are responsible for 3-10 times more difference between genomes than SNV. Although disease causing SNVs

are relatively more common, SVs have also been shown to play a very large role in the causation of diseases. Studies into the mechanisms by which SVs cause disease have found that SVs can cause disease by numerous mechanisms, partially due to their comparatively larger size than SNVs (Collins *et al.*, 2020).

SVs are variants where a portion larger than 50bp of DNA is deleted, duplicated, inverted, translocated or inserted as depicted in Figure 1.1. Inversions and translocations are balanced and do not result in an overall change to genomic content but may still disrupt the transcription of a protein and thus be disease causing. Deletions, duplications and insertions are unbalanced resulting in a net gain or loss of genetic material, and thus often change the number of copies of a gene or genes and are therefore referred to as copy number variants (CNVs) (Pos *et al.*, 2021). CNVs thus form a subset of SVs and were originally defined with a minimum size of 1000bp (Feuk *et al.*, 2006). This definition, however, has changed over time, and they now are defined along with SVs with a minimum size of 50bp (Zhang *et al.*, 2009).

CNVs are generally believed to form during recombination, replication and repair events within the genome. The study of breakpoint regions of CNVs, have shown that these areas are enriched in repeat sequences such as low copy repeats, short or long interspersed nuclear elements. These sequence motifs play a role in triggering non-allelic homologous recombination, a common mechanism by which CNVs form. This type of CNV formation mechanism usually results in recurrent CNVs that have common breakpoints between individuals (Hastings *et al.*, 2009). Other non-recurrent CNVs that have different breakpoints between individuals usually form by other mechanisms such as non-homologous end joining, micro-homologous end joining or replication associated mechanisms such as fork stalling, replication slippage or template switching (Pos *et al.*, 2021).

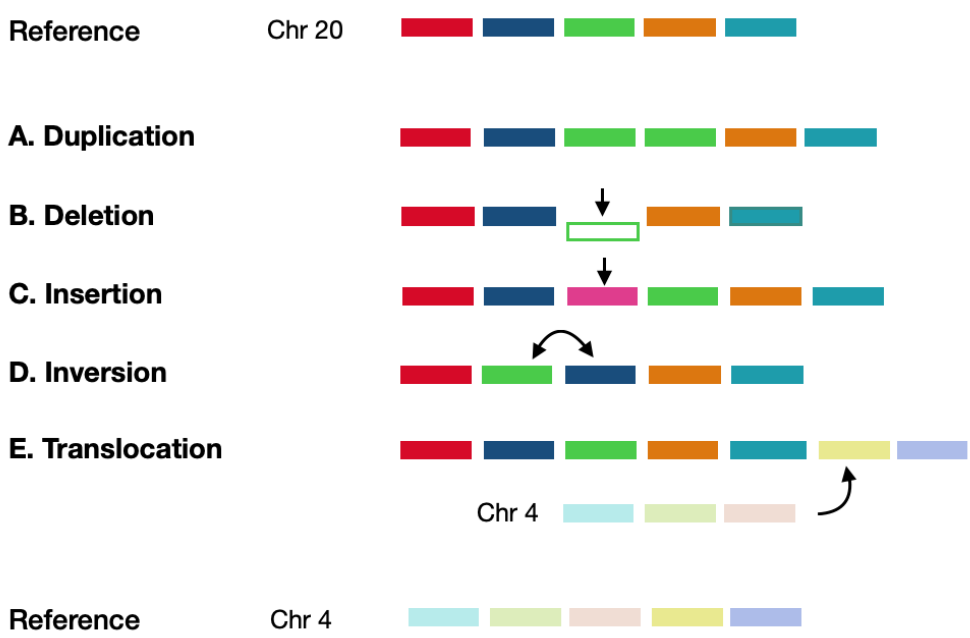


Figure 1. 1: Types of SVs compared to a reference region. A reference of segments of chromosome 20 with a single green segment. A. duplication of the green segment increases copy number to two, B. deletion of the green segment decreases copy number to zero, C. insertion of the pink segment. A, B and C are unbalanced types of SVs with copy number variation seen in duplications and deletions. D. inversions and E. translocations are balanced types of SV with no net loss or gain of genetic material from the genome.

Rare, recurrent and *de novo* SVs have been implicated in many diseases including autism spectrum disorders, schizophrenia, NDDs and cancer (International Schizophrenia, 2008; Prasad *et al.*, 2012; Kaminsky *et al.*, 2011; Li *et al.*, 2020). Although all types of SVs can cause disease the most common mechanism is through copy number change, where a gene is duplicated or deleted changing the dosage or amount of that protein being produced. CNVs become disease causing if a gene involved is sensitive to dosage or copy number change. The change in dosage may involve the gene directly or its regulatory elements up or downstream. A gene sensitive to dosage loss is referred to as haplo-insufficient and a gene sensitive to dosage gain is referred to as triplo-sensitive (Riggs *et al.*, 2020). Gene dosage alteration by CNVs is reported to be the cause in ~15% of neurodevelopmental disorders (NDDs) (Kaminsky *et al.*, 2011).

Balanced rearrangements, although they usually don't alter gene dosage, may alter 3D structures and the position of regulatory features, and thereby cause disease (D'Haene and Vergult, 2021; Short *et al.*, 2018). SVs in the non-coding genome have been found to be responsible for many DDs, and the mechanisms by which this occurs are gradually being understood (D'Haene and Vergult, 2021). These non-coding

regions are mostly regulatory features such as micro-RNA, long non-coding RNA, promoters and enhancers. Given the larger size of SVs, compared to SNVs, regulatory regions appear to be more vulnerable to SVs than they are to SNVs and so are being studied with increasing interest (D'Haene and Vergult, 2021).

The impact of SVs on complex traits and multifactorial diseases also remains largely unknown. The full spectrum of SVs, functional implications and mutation rates for SVs present in populations, has been hindered by the complexity and technical challenges of SV discovery (Collins *et al.*, 2020).

The introduction of microarray technology and more recently whole exome sequencing (WES) into genetic diagnostics has gradually uncovered the varied and important role of CNVs in the causation of a large number of DDs (Coe *et al.*, 2014; Uddin *et al.*, 2016). As a result, CNVs are being considered an increasingly important part of the genetic diagnostic process when investigating DDs (Srivastava *et al.*, 2019). Databases such as ClinVar (Landrum *et al.*, 2018) and DECIPHER (Firth *et al.*, 2009) also now have a great number of CNVs known to be the cause of a DD. Together these show the importance of CNVs as a genetic mechanism underlying DDs.

1.4 Diagnostic Process for Developmental Disorders

The process of diagnosing DDs is particularly challenging for two main reasons. Firstly, the presentation of these conditions is varied and often non-specific and secondly the aetiologies, as mentioned earlier, are highly heterogeneous. Clinical assessment of a child with a DD is vitally important in assessing the possible aetiology. A medical geneticist can determine the likelihood of teratogenic or acquired causes as opposed to a genetic cause, as well as assessing the different phenotypes observed in the child (Srour and Shevell, 2014). After clinical assessment directed investigations are performed, which include biochemical, metabolic and radiological tests, but genetic tests are the most important tests for determining the definitive aetiology of DDs (Mithyantha *et al.*, 2017). This clinical diagnostic process is depicted in Figure 1.2 showing a decision-making flow chart for investigating GDD (a common non-specific presenting feature of many DDs).

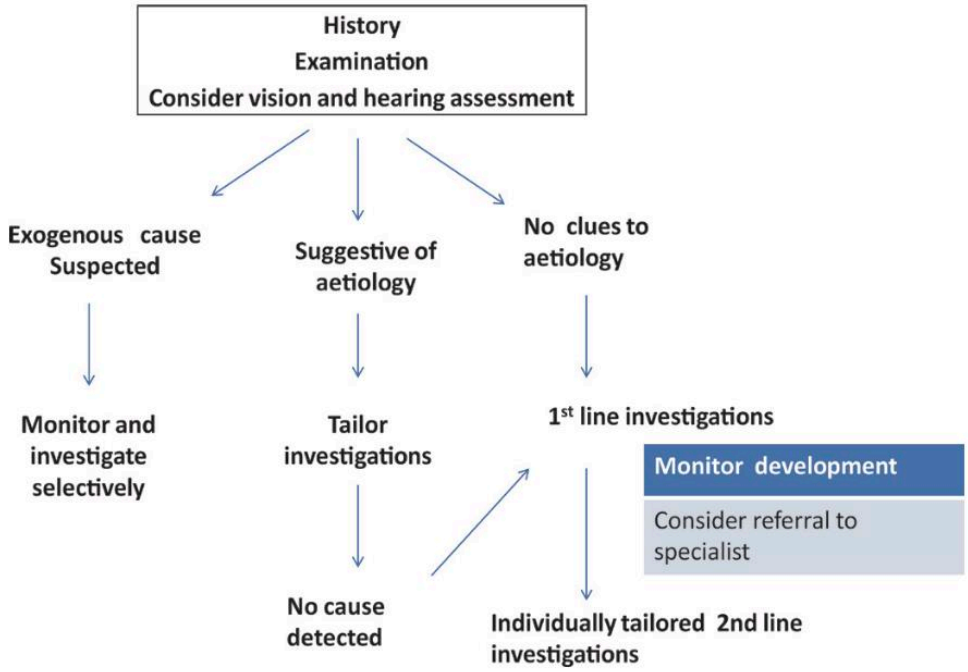


Figure 1. 2: Flow chart of decision making for investigations for GDD. Taken from Mithyantha *et al.* (2017).

Establishing a definitive genetic diagnosis for DDs, even if the condition cannot be cured, is very important (Carmichael *et al.*, 2015). It helps the healthcare provider with treatment decisions for associated co-morbidities and pre-emptive management and informs the need for prenatal testing and relative risks to subsequent pregnancies. It is also valuable for parents, enabling them to better understand prognosis and be involved in relevant support groups (Flore and Milunsky, 2012). Additionally, confirmation of a diagnosis ends the diagnostic odyssey of successive investigations. This diagnostic odyssey is exceptionally hard for patient's families. It causes psychosocial stress due to the uncertainly of not having a definitive diagnosis, as well as financial stress from indirect costs over time (Makela *et al.*, 2009; Dragojlovic *et al.*, 2020). In studies performed before the introduction of NGS, it was reported that no diagnosis is established in up to 60% of cases of GDD, even in developed countries (Kaufman *et al.*, 2010).

Due to the difficulty of diagnosing DDs, diagnostic guidelines have been developed over time to assist clinicians in choosing the best tests to diagnose patients with DD in the most timeous and effective way possible. These recommendations have changed over time as new technologies became available (Mithyantha *et al.*, 2017). Guidelines

usually include multiple types of tests, as mentioned above, but the choice of first-line genetic investigation has the greatest effect on the diagnostic yield.

The recommendation for genetic investigations has undergone many changes in the past 20 years as molecular testing technologies have evolved. As discussed above genetic causes account for a large proportion of cases, and so the changes in genetic testing capabilities have radically changed the outlook of the diagnostic process for DDs (Manickam *et al.*, 2021). Due to the diverse aetiologies of DDs where either SNVs or CNVs may be the cause of a patient's condition, both types of variation need to be considered in terms of testing methodologies. Before the introduction of NGS there was no unbiased genome-wide test able to detect both SNVs and CNVs. WES can detect both these types of variation without any bias caused by design based on existing knowledge of variation. Given that this thesis is focused on CNVs and their role in DDs, the coming sections will focus on diagnostic testing for this type of variation.

1.5 Diagnostic Tests for Copy Number Variation

1.5.1 Karyotyping

Conventional karyotype analysis with Giemsa-stained banding was one of the first genetic tests used as a first-line investigation for suspected developmental or intellectual disability (Shevell *et al.*, 2003). Obtaining a karyotype involves culturing cells from a blood sample, stimulating mitosis and then arresting cells mid-cycle to obtain cells in metaphase. Cells are smeared on a slide and fixed and then stained with Giemsa stain that reveals bands on the chromosomes. It is then analysed microscopically and the cells in metaphase (chromosomes are best visualised in this phase) are isolated and photographed. This image must then be analysed and the chromosomes identified and arranged in order. Chromosomes are then scrutinized for any abnormalities. One can detect large chromosomal translocations, aneuploidies, as well as large duplications, deletions or insertions. Its resolution however only enables detection of alterations in the 3-10Mb range (Shevell *et al.*, 2003). This technique also relies on subjective assessment of losses or gains, meaning results are not standard between different centres (Geiersbach *et al.*, 2014).

1.5.2 Fluorescent *in situ* hybridization

Fluorescent *in situ* hybridization (FISH) technology has been used for molecular diagnostics since the 1990s. It works on the basis of complementary hybridization of fluorescently labelled probes to chromosomes in metaphase (Levsky and Singer, 2003). The resolution obtained by FISH depends on probe size but is in the 100-200kb range which is significantly smaller than that obtained by karyotyping (Cui *et al.*, 2016). Its use in DD diagnosis has mostly been for detection of aneuploidy, micro-deletion or -duplication and sub-telomeric rearrangements. The major limitation of FISH is that a probe must be designed for a specific target, so it is not a technique to use for discovery of novel pathogenic loci. If however a cause with an available probe is strongly suspected, it is an efficient way of diagnosing the condition, as only a single locus is tested instead of looking at the whole genome (Manning *et al.*, 2010). FISH tests are also time consuming, expensive and labour intensive, which needs to be considered in the value of the test to the process of DD diagnosis.

1.5.3 Multiplex ligation-dependent probe amplification

Multiplex ligation-dependent probe amplification (MLPA) is an assay designed to assess CNV. It uses a multiplex Polymerase Chain Reaction (PCR) system with sets of 5' and 3' probes, specific to exons or target sequences of interest. Primers anneal to the target sequences and a universal primer then amplifies all probes. The amplicons are then separated by capillary electrophoresis, and the amount of DNA amplified by each set is derived by the intensity of the fluorescence signal. From this the CNV for the targeted regions can be assessed (Stuppia *et al.*, 2012). MLPA is able to accurately detect CNVs in the regions included in the test, but it is a limited test, as each test can only look at ~40 different loci. This means it is helpful to screen for common disorders caused by CNVs, in targeted regions, but not genome-wide copy number detection. One of the main advantages of MLPA is that it is a very cost effective test that does not require expensive equipment (Stuppia *et al.*, 2012). Jehee *et al.* (2011) showed that MLPA is more cost-effective test for congenital anomalies and intellectual disability than chromosomal microarray (CMA) (Jehee *et al.*, 2011). For this reason it has been used widely in resource limited environments such as in South Africa (Fieggen *et al.*, 2019).

1.5.4 Chromosomal microarray

CMA, including both array CGH and single nucleotide polymorphism arrays, has been the gold standard for DD diagnosis for many years and is still widely used as a first-line test for the diagnosis of DDs.

A. Array comparative genomic hybridization

Array comparative genomic hybridization (CGH) technology works on the principle of hybridization of complementary strands of DNA. A chip contains immobilised control probes or sequences to which patient DNA is applied. Patient and control probes are labelled with different coloured fluorophores, enabling the detection of the differences of colour intensity either more or less of the patient DNA compared to the control DNA. These fluorescent signals are detected and analysed by the array reader. These changes are quantified using the log R ratio that either increases for duplications or decreases for deletions enabling copy number detection. Initial comparative genomic hybridization arrays had low resolution, using bacterial artificial chromosome probes, that were 75–100kb long. The results had resolution similar to that of karyotypes of 3–10Mb. However, the technology rapidly improved using much smaller oligonucleotide probes 50–60bp long, which enabled improved resolution of arrays. The resolution depends on the density of probes in covering the targeted region, or whole genome, depending on the type of array (Manning *et al.*, 2010).

B. Single nucleotide polymorphism array

Another type of array is designed with common SNVs densely distributed across the genome. This technology uses small oligonucleotide probes that tag common SNVs, enabling genotyping at each SNV location as either Allele A or B (LaFramboise, 2009). In a diploid genome each location will yield a genotype of AA, AB or BB. CNVs are detected by observing either loss of heterozygosity with no AB sites for a region (for a deletion) or a split heterozygous band with AAB or BBA (for a duplication). This is quantified by calculating a B allele frequency defined as a normalised ratio of the quantity of B allele to the total of both alleles (Zhao *et al.*, 2004). Combination arrays utilising both SNVs and oligonucleotides have also been developed and these are very versatile and have high resolution. High density CMAs now have resolution down to the ~1kb level, and based on multiple adjacent probes are able to detect CNVs very accurately (Haraksingh *et al.*, 2017).

In a consensus statement involving multiple genomic centres, Miller *et al.* (2010) published a review of the literature showing that CMA testing offered a diagnostic yield of ~12% compared to the yield of ~5% for karyotyping in cases of unexplained DDs. This improvement is primarily due to its ability to detect sub-microscopic deletions and insertions genome wide. They concluded at that time that CMA should replace karyotyping as the first-line investigation for cases of unexplained DDs or congenital anomalies, and it remains the most widely implemented first-line investigation for DDs.

1.5.5 Next generation sequencing

Next generation sequencing (NGS) as a technique has been in use since the early 2000s. The basic technology is massively parallel sequencing of multiple fragmented DNA strands (Ronaghi *et al.*, 1996). The technique involves fragmenting of genomes into sequencing libraries of short DNA fragments, followed by ligation of adaptors or paired ends. The strands are then clonally amplified before each strand is sequenced. There are many different methods used to detect the sequence of bases being incorporated into the strand, and in paired end methods this is done in both directions on each strand (Rothberg *et al.*, 2011). The raw signals from each short DNA read are digitized to single base calls that are then interpreted by computational analysis and assembled. An essential aspect of this technology is that each region is sequenced multiple times resulting in a number of short reads covering each region. The depth coverage is important for accurately detecting SNVs, and variation in depth coverage is used to estimate CNVs (Ledergerber and Dessimoz, 2011). This technique has revolutionized genetics in a research context, but also diagnostic testing as it is substantially faster, cheaper and easier than traditional Sanger sequencing. Being a sequencing technique, it has been used widely for detecting SNVs diagnostically, but also for the detection of CNVs in finer scale than by previous techniques (Zhao *et al.*, 2013).

A. Whole exome sequencing

Sequencing the exome with NGS involves sequencing of all coding transcripts or exons, with short flanking regions. During library preparation, when probes or adaptors are ligated, these probes are designed to target and hybridize to exonic regions. Thus when clonal amplification occurs exonic regions are amplified and sequenced. It is an efficient approach and has many advantages. It focuses on protein coding regions, as

opposed to the rest of the less characterized non-coding genome. It is genome wide rather than choosing specific gene targets, enabling potential discovery of new regions of interest. The exome amounts to ~1,5% of the entire genome and is thus considerably cheaper to sequence than a whole genome, as well as being a much smaller data file to analyse and store (Teer and Mullikin, 2010). Given the assumption that mutations and variation that cause many Mendelian disorders occur in or near protein coding regions, WES is ideal to identify this variation and has proved a very powerful tool for discovering new disease associated genes (Smedley *et al.*, 2015).

Another significant advantage of this technology is that both SNVs and CNVs can be detected from the data, making it a very efficient, powerful diagnostic tool (Pfundt *et al.*, 2017). Although SNVs, being easier to detect from the sequences, were the early focus of diagnostic exome analysis, pipelines to extract CNVs from the data have been developed. There are some features of WES that make detecting CNV very challenging; mainly the discontinuity of exon sequences with the accompanying altered read depth distribution changes. This means that a CNV is highly likely to span exonic and non-exonic regions. Low complexity and high GC content regions also add to this difficulty. These challenges have been overcome using many different strategies by different tools resulting in tools with very varied capabilities, strengths and weaknesses (Zhao *et al.*, 2020).

The main approach for detecting CNVs from WES uses the depth of short reads over a region or exon to detect copy number changes. Steps in the process of WES; such as capturing of exonic regions with targeted probes and problems encountered during the amplification of these regions (low complexity and high GC content regions) result in over or under representation of exonic regions, that may then be interpreted falsely as a CNV. Different tools however vary greatly in methods of analysing and interpreting read depth variability to overcome these challenges (Gordeeva *et al.*, 2021). Some tools like CANOES (Backenroth *et al.*, 2014) and exomeCopy (Love *et al.*, 2011) apply binomial models to the assumption of read depth distribution to overcome false read depth variability. Other tools like EXCAVATOR2 (D'Aurizio *et al.*, 2016) and CNVKit (Talevich *et al.*, 2016) use both normalized on-target (exonic) reads and off-target reads to call CNVs where CLAMMS (Packer *et al.*, 2016) and ExomeDepth (Plagnol *et al.*, 2012) use reference set optimisation to improve CNV calling accuracy. Despite these varied approaches no tool has yet achieved very high sensitivity and specificity given the significant challenges that have to be overcome (Gordeeva *et al.*, 2021).

A recent benchmarking publication compared 16 of these tools and found that tools have highly varying capabilities and state that an appropriate tool must be chosen according to the specific size range of CNVs desired and the specific use of the tool (Gordeeva *et al.*, 2021). Despite these challenges CNVs can be successfully detected from WES and studies have shown that including CNV analysis in a WES pipeline can increase in the diagnostic yield by 6–17% (Charng *et al.*, 2016; Vissers *et al.*, 2017).

The analysis of WES for both SNVs and CNVs is a complex and time consuming process. One important step involves a process of filtering out likely benign variants from the entire set of variants detected in the individual before a set of possible causative variants is left (Smedley *et al.*, 2014). There are a number of different strategies that can be employed, but most include pathogenicity prediction on all variants, pedigree information as well as the filtering out variants that have been found in populations at a frequency >0,1%. These variants are considered highly unlikely to be the cause of a very rare condition. This means that databases with population variant frequencies are important resources for successful WES analysis to identify disease causing variants (Smedley *et al.*, 2015).

In 2019 a consensus statement was published by Srivastava *et al.* (2019) indicating that WES should replace microarray as the first-line investigation for NDD. A multidisciplinary expert working group was gathered to assess the diagnostic yield of WES for NDDs. After a literature scoping review they concluded that WES consistently outperforms CMA, with a diagnostic yield of 30-53% compared to 12% for CMA. They therefore concluded WES should be the recommended first-line investigation for neurodevelopmental disorders (Srivastava *et al.*, 2019). In 2021 the American College of Medical Genetics (ACMG) similarly published an evidence-based guideline for congenital anomalies, developmental delay and intellectual disability recommending WES as a first-line test for these conditions (Manickam *et al.*, 2021).

B. Whole genome sequencing

Whole genome sequencing (WGS) by NGS is an established, valuable research method which has become widely used as it has become more readily available at lower costs. Due to its potential to accurately identify all forms of variation, including CNVs, WGS has been proposed as a possible diagnostic test for many years. This would make it a potentially ideal tool for diagnosing DDs, which often have varied, very rare genetic causes. There have, however, been several obstacles to its

implementation as a routine diagnostic tool (Stavropoulos *et al.*, 2016). Despite the lowering costs of WGS, it still remains considerably more expensive than CMA or WES. Additionally, the large amount of data produced requires large amounts of secure storage and analysing that amount of data is more time consuming than the analysis for CMA or WES. Due to the greater scope of detection possible with WGS, it has the potential to be a very powerful tool and so projects have been implemented to increase evidence of its effectiveness and clinical utility as diagnostic tool (Turnbull *et al.*, 2018; Lee *et al.*, 2021). In the 2021 ACMG guideline mentioned above, they recommended WES as the first-line investigation for DDs. They further suggest WGS should be considered as a first- or second-line investigation for DD, as it has a superior ability to detect CNVs compared to WES. They describe that WGS has shown greater impacts in clinical management, even higher yields than WES, and that it may be more cost effective when implemented at an early stage of the diagnostic process (Manickam *et al.*, 2021). So far WES has been more widely implemented than WGS for the diagnosis of DDs.

1.6 Barriers to Implementation of Whole Exome Sequencing

1.6.1 Technological and cost barrier

The implementation of WES in a low resource setting such as Africa poses significant multifaceted challenges. Implementation of new technologies such as WES requires new infrastructure, laboratory validation and policy changes as well as new skills in diagnostic laboratory staff for running and interpreting results (Krause, 2019). Expensive infrastructure and instituting new practices in laboratories can be barriers to implementation of new technology tests, especially in resource constrained countries. It is therefore often necessary to motivate strongly for these changes, through research, to show the need for change and the effectiveness of the new technology, to improve diagnostic practice over time.

The guidelines published by Srivastava *et al.* (2019) and the ACMG guideline (Manickam *et al.*, 2021) adds validity to the approach being taken in DDD-Africa to utilise WES to research DDs in Africa and through this to establish a practical plan for WES implementation in Africa. In most African countries, including South Africa, WES is not available as a routine test (Krause, 2019; Kamp *et al.*, 2021). Tests currently in

use routinely are mainly karyotyping, MLPA and FISH and a limited number of microarrays. Increasing evidence and guidelines indicate that tests with higher yields, most notably WES, although initially more expensive would ultimately reduce costs. This is because a single test with a higher yield would enable the diagnosis of more patients earlier, which would reduce the number of investigations, both genetic and clinical (Singleton, 2011; Strauss *et al.*, 2018). Additionally, research by Dragojlovic and colleagues shows that the indirect costs for patients over time, who remain undiagnosed, must be taken into account when assessing the cost effectiveness of new testing methodologies, such as WES. Their research shows that even if initial costs are higher, having a genetic diagnosis, results in lower indirect costs compared to costs of patients who remain undiagnosed (Dragojlovic *et al.*, 2020).

1.6.2 Need for African population variant data

Another barrier to implementation of such guidelines in Africa is a lack of baseline population variant information from African individuals for variant filtering. Reference datasets of common population SNVs and CNVs are well characterised for populations of European ancestry in databases such as gnomAD (Karczewski *et al.*, 2020), gnomAD-SV (Collins *et al.*, 2020), dbSNP (Sherry *et al.*, 2001) and dbVar (Lappalainen *et al.*, 2013), but they contain very little data from continental African individuals. Databases with good representation of diverse African populations are of vital importance for the interpretation of variants found in clinical and disease association studies (Bope *et al.*, 2019). This is especially important as African genomes are known to harbour more genetic diversity than other population groups (Gurdasani *et al.*, 2015; Choudhury *et al.*, 2020). This greater diversity makes the dearth of common population variation an even greater problem, as it is more likely that African variants and frequency information is missing from existing databases. The lack of this information poses a significant problem when common population variation must be filtered out as part of the variant filtering process for WES or WGS testing.

A recent landmark paper by Choudhury *et al.* (2020) provided insight into SNVs from the most diverse set of African genomes to date. This study showed the value of studying diverse African populations, finding 3 million novel single nucleotide polymorphisms from African populations not previously represented in global population studies. They describe 54 variants that had been previously classified as ‘likely pathogenic’ or pathogenic in ClinVar, but were seen at frequency >5% in one of

populations included in the study, 13 of which were seen at a frequency >5% across all African populations. This indicates possible misclassification of such variants, without the inclusion of data from African populations. Findings such as these highlight the importance of including African datasets into global variant databases. This research has made a significant contribution of African SNV to global databases such as gnomAD and dbSNP which will be valuable in African disease studies. They did not perform CNV analysis on the varied African cohort, meaning there is still a significant lack of similar African CNV knowledge.

The technical complexities and highly varied nature of SVs has resulted in research into CNVs and SVs lagging significantly behind SNV research. Consequently, there is a lack of high quality, publicly available variant maps from large population studies. Available datasets have been variable in terms of the technologies and analysis methods employed, as well as limits on public access (Abel *et al.*, 2020; Collins *et al.*, 2020). These challenges have affected CNV research globally, including in Africa, and have resulted in baseline population CNVs being understudied in southern African populations. Without a significant set of African population CNVs, with southern African representation, interpretation of potentially pathogenic CNV found in African genetic disease studies will be very challenging (Bope *et al.*, 2019; Krause, 2019).

Baseline common CNV databases are produced from the study of CNVs in populations so we will now focus on how all types of SV are described at a population level.

1.7 Structural Variation Discovery Globally

Knowledge of human SVs has grown substantially, as research using first CMA and later NGS have uncovered the depth and scope of the human SV landscape (Auton *et al.*, 2015; Sudmant *et al.*, 2015a; Collins *et al.*, 2020). Both normal variation and SVs associated with disease have been studied in order to attempt to understand the role of SVs in health and disease (McCarroll and Altshuler, 2007; International Schizophrenia, 2008; Zhang *et al.*, 2009; Coe *et al.*, 2014; Li *et al.*, 2020). Studying SVs has been considerably more difficult than the study of SNVs and has subsequently lagged behind the study of SNVs (Abel *et al.*, 2020; Collins *et al.*, 2020). This difference is due to the fact that SNV are uniform in nature where SV vary in both size and type meaning that their detection is significantly more challenging (Kosugi *et al.*, 2019). Early ground-breaking studies to map human CNVs used fosmid DNA genomic

libraries mapped against the reference human genome. These studies started to elucidate loci of variation, but were limited in resolution, to define breakpoints, due to the size of the fosmid clones. They were also only measuring variation in a few individuals (Tuzun *et al.*, 2005).

Three landmark studies after this by Conrad *et al.* (2006), McCarroll *et al.* (2006) and Redon *et al.* (2006) used cell lines from the HapMap Consortium which contains ~270 individuals from different global populations. They performed high density oligonucleotide and SNP arrays and each found multiple novel CNVs. Validation of variants was performed by various methods, such as custom microarrays with high coverage over detected loci, quantitative PCR and multiple analysis methodologies, to minimise false positive identification of CNV loci. The CNVs found in these studies were released to an early version of the Database of Genomic Variants (DGV) (MacDonald *et al.*, 2014). Although microarray methods were successful at SV discovery, they were not able to provide sequence level information of SV breakpoints. With WGS by NGS becoming cheaper and more readily available, large cohorts could be sequenced to uncover both SVs and SNVs from the same technology.

The first large-scale study of human genetic variation including SVs from WGS was performed as part of the objectives of the 1000 Genomes Project (Auton *et al.*, 2015). In 2015 at completion of the project, they had sequenced 2504 individuals from 26 populations including 500 African individuals from east and west Africa but no southern African individuals. They reported 3.6 million short insertions or deletions (<50bp) and 60 000 structural variants (>50bp). They also report that the typical genome differs from the first human reference genome at between 4.1 and 5 million sites, indicating the large amount of variation present globally (Auton *et al.*, 2015). The SV analysis group of the 1000 Genomes Consortium produced a more detailed SV analysis where they report 42 279 biallelic deletions, 6 025 biallelic duplications, 2 929 multi-allelic CNVs, 786 inversions, 168 nuclear mitochondrial insertions, and 16 631 mobile element insertions (Sudmant *et al.*, 2015b).

In 2019 Chaisson *et al.* (2019) did a multi-platform SV analysis of three trios from Nigeria, China and Puerto Rico to detect a complete set of human SVs found in these individuals, with accurate breakpoints that were phased to resolve individual's haplotypes. They used a combination of short and long read technologies. The advantage of long read technology in this context is that low complexity and repetitive

regions, and those with segmental duplications, can be assembled from the longer contigs in a way not possible with short read technology. Additionally, it enabled phasing of haplotypes from the long reads not possible from short reads. They also tested and used large number of different algorithms, assessing the most effective and sensitive combinations of algorithms to detect all classes of SVs. Using these multiple technologies and methods they found on average 27 622 SVs per genome, 3-7 fold more variants per genome than most short read high-throughput studies (Chaisson *et al.*, 2019).

In 2020 two large scale studies were published investigating SVs in diverse population groups. Abel *et al.* (2020) designed a scalable pipeline for SV detection in populations and then mapped and characterised SV for 17 795 individuals from diverse world ethnicities (Abel *et al.*, 2020). ~34% of this cohort were African samples, but almost all these samples were African American individuals not Africans from the African continent. They looked particularly at rare SVs and found that the average person has 2.9 rare SVs that alter coding regions. In a similar study Collins *et al.* (2020) designed a pipeline for population SV detection and produced a reference SV dataset from 14 891 individuals to be included in gnomAD as gnomAD-SV. Similarly, this paper included many African American samples but very few continental Africans. They described that the SV landscape is even richer than previously thought and that 25–29% of rare protein-truncating events are caused by SVs. These two papers were landmark papers with regards to population SV reference datasets both in terms of the pipelines that can now be implemented by others for similar studies and the datasets they produced.

Both these studies highlighted the frequency of rare SVs with potentially deleterious functional impacts. But in highlighting this they uncovered how little knowledge we have of the functional impact of SVs. They discuss that significant functional studies are required, to better understand the heterogeneous impacts of SVs, before a comprehensive and accurate functional prediction tool will be possible. Only recently was an interpretation guideline comparable to that for SNVs released by the ACMG to standardize the interpretation of CNVs between laboratories. These standards enable scoring of CNVs in a clinical diagnostic setting classifying them as Benign, Likely Benign, Variant of Unknown significance, Likely Pathogenic and Pathogenic. There are also specific recommendations regarding the inheritance pattern of the gene

contained in the CNV in terms of whether two pathogenic variants are required to be disease causing in autosomal recessive cases.

The pipeline designed by Collins *et al.* (2020) is available as a ready to run cloud based service on the Broad Institute's Firecloud service (https://portal.firecloud.org), which is an invaluable tool for enabling high quality SV detection globally, even for those who do not have the computational resources to run SV detection. The production of the gnomAD-SV database is an important step forward, as there is an increasing reliance on such databases for disease association studies and clinical genetic testing, as a reference of normal population variation.

1.7.1 Structural variation in southern African populations

A small number of studies have analysed CNVs or SVs in Southern African populations. Schuster *et al.* (2010) investigated genetic variation in five Southern African genomes of Khoisan and Bantu individuals and described SNPs and a limited set of SVs in these individuals. They performed copy number estimation for autosomal Refseq genes, comparing the number of copies of these genes between a Southern Kalahari individual and a Yoruba Nigerian, Han Chinese and European individuals. They found a number of genes with higher copy numbers in the Southern Kalahari individual compared to the other two individuals. Sudmant *et al.* (2015a) did a global comparison of CNV including 22 African genomes in their cohort (including the Southern African individuals sequenced by Schuster *et al.* (2010). In their article they report that African individuals didn't have a higher overall load of CNVs but have the greatest CNV diversity. Principal component analysis (PCA) of deletions and duplications, both showed distinct separation of African genomes from other continents, especially in the PCA of deletions (Sudmant *et al.*, 2015a).

Another study into African CNVs by Nyangiri *et al.* (2020) analysed CNVs from 232 medium coverage WGS from individuals from three different countries representing 3 African ethnolinguistic groups (Two Niger-Congo and one Nilo-Saharan). They found 7608 CNV regions using Genome STRiP and cn.MOPS. 224 of the CNV regions found were novel, but the majority of these novel CNV regions were found at low frequencies and were not shared between populations. Using SNP signatures they showed that CNVs were actual targets of selection at some loci. They also performed a PCA on the data but did not find any clustering within Africa (Nyangiri *et al.*, 2020).

The lack of comprehensive knowledge of SV and CNV in Southern African populations is a significant problem for disease association and clinical diagnostics in Africa as a knowledge of population CNV is critical for variant interpretation (Bope *et al.*, 2019). The challenges that have slowed SV discovery and mapping, and the production of high-quality reference datasets for SV globally are even more acutely felt when considering Southern Africa. Africa has a lack of resources in terms of finance and hardware, but also a lack of experience in SV discovery pipelines, required to perform such studies. Reference databases for SV, like the new gnomAD-SV database, are a significant step for SV discovery globally, but are less useful for disease studies in Southern African individuals, as they lack allele frequencies for Southern African populations and specific Southern African variants. This lack of Southern African SV reference datasets is a significant obstacle to interpretation of SVs in clinical diagnostics and disease discovery studies (Bope *et al.*, 2019; Krause, 2019).

1.8 Detection of Structural Variation from Whole Genome Sequences

The technology that has been most widely used for SV discovery in a research setting is short read WGS. It enables SV discovery of genome wide SV with good breakpoint resolution. Because of this many SV calling algorithms have been designed to detect SV from WGS.

Many different methodologies have been used in the design of SV calling algorithms to extract the most accurate SVs from WGS and algorithms have greatly improved over time. The detection of SV from short-read WGS is a challenging process. A WGS file has a finalised base sequence compiled from the multiple reads mapped to that region. In order to call SVs from this file one has to look at the reads themselves, to infer the variation in structure or copy number in that specific region. SV calling algorithms have been designed to make use of the multiple attributes of reads and read pairs of WGS files to call SVs accurately (Zhao *et al.*, 2013). There are still many shortcomings of these methodologies and tools are often error prone, having high false positive and negative rates. These algorithms have been greatly improved using combinations of methods and through the incorporation of statistical methods to assist in determining accurate SVs (Zhao *et al.*, 2013; Kosugi *et al.*, 2019).

There are four basic strategies that algorithms use to detect SV from WGS, shown in Figure 1.3. There are also tools that use combinations of these four methods SV (Zhao *et al.*, 2013; Kosugi *et al.*, 2019).

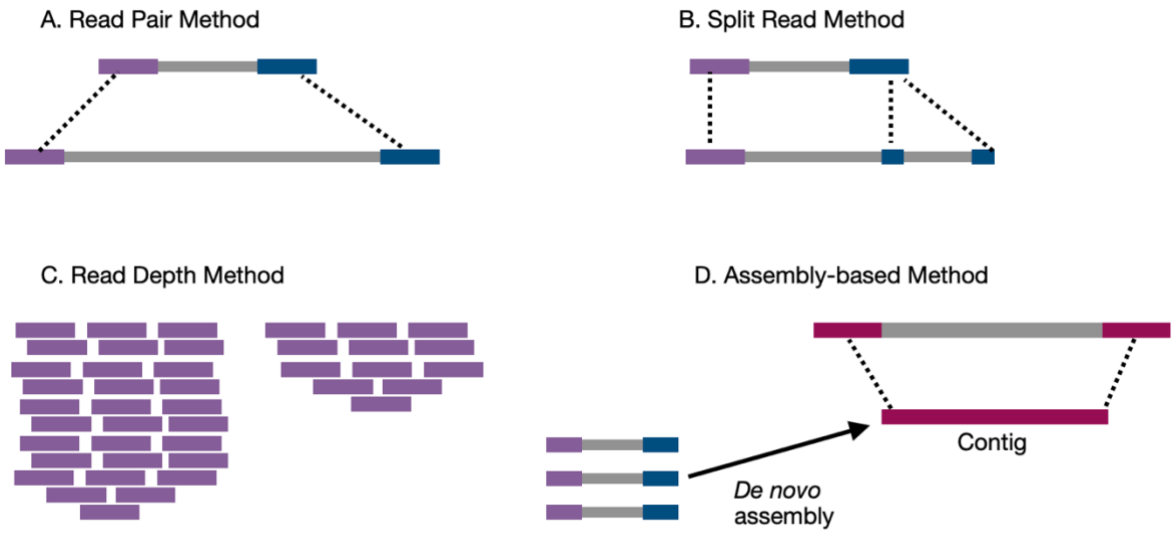


Figure 1. 3: Four basic SV detection methods from short read WGS. A. Read Pair method where a change in length of a read pair is detected as an SV site, B. Split Read Method where the splitting of a read pair indicates an SV at that site, C. Read Depth Method, where depth of reads at sites are compared to detect SV. D. Assembly-based Method, where raw reads a assembled *de novo* into contigs that are then compared to a reference to detect SV. Adapted from Zhao et al. (2013).

1.8.1 Read pair method

The read pair method sometimes called paired end method was the first method to be described for detecting SVs from WGS (Korbel *et al.*, 2007). This method uses discordantly mapped paired end reads, detecting the presence of an SV if the distance between paired reads is significantly larger or smaller than the expected distance. This method is used alone in tools such as Breakdancer (Chen *et al.*, 2009), Pindel (Ye *et al.*, 2009) and VariationHunter (Hormozdiari, 2010) and is able to detect all types of SV, although it is limited by not being able to identify variants larger than the average read pair insert size. These tools are all single sample algorithms, so outputs must be merged if one wants to analyse multiple samples. One tool, commonLAW (Hormozdiari *et al.*, 2011), is able to analyse multiple genomes in parallel, and similar to VariationHunter it compares samples to a reference genome to improve SV accuracy. The read pair method has also been used in combination with the other three commonly used methods.

1.8.2 Split read method

The split read method utilises instances where one read from a pair maps to the reference while the other fails to map, or only partially maps and is split. These partially or unmapped reads provide potential breakpoints of SVs at a single base level, so can produce accurate breakpoints of insertions or deletions. The disadvantage of this method is that it is dependent on the length of reads and is only applicable to unique regions of the genome. It is usually used in combination with one or more of the other methods (Zhao *et al.*, 2013).

1.8.3 Read depth method

Read depth methods are based on the premise that the coverage over a section of the genome is correlated to the copy number of that region. It therefore uses the depth of short reads in a region to detect changes in copy number. Read depth is calculated in predefined windows of mapped short reads, and is then normalised, and biases caused by GC content or repetitive regions corrected. The copy number is determined as a loss or gain compared to the normalised read depth, and finally these copy numbers are refined and checked, using various statistical methods. Using this methodology, more exact copy numbers can be assigned to variants compared to the capabilities of the other methods such as read pair or split read methods (Zhao *et al.*, 2013). Tools that use this method exclusively include Genome STRiP (CNVDiscovery) (Handsaker *et al.*, 2015) and CNVnator (Abyzov *et al.*, 2011).

In Genome STRiP (CNVDiscovery) this approach has been maximised to also detect multi-allelic CNVs. By utilising a distribution of read depth at a specific locus, across multiple samples, the algorithm infers the diploid read depth and then multiples of this to infer multi-allelic CNV copy numbers. Genome STRiP is one of few tools able to detect multi-allelic CNVs from population size samples.

One of the greatest limitations of this method is that is only capable of detecting unbalanced SVs, and is unable to detect any other types of SVs. Another disadvantage of this method is its sensitivity to the sequencing method employed on samples as well as overall sequence quality (Zhao *et al.*, 2013). For example, in regions which amplify poorly, there will be regions with lower coverage that may be falsely interpreted as a deletion. Additionally it is now known that PCR-based sequencing methods have a far more variable read depth coverage compared to PCR-free sequencing methods

(Collins *et al.*, 2020) which means that sequences produced using different PCR methods cannot be jointly called using a read depth based algorithm.

1.8.4 Assembly-based method

The assembly-based method differs substantially from the other three methods. In read depth, split read and read depth methods short reads are aligned to the reference as a starting point. Assembly-based method uses overlapping short reads to assemble contigs *de novo* which are then compared to the reference to identify regions of discordant copy number (Zhao *et al.*, 2013). Cortex (Iqbal *et al.*, 2012) is a tool that uses this method exclusively. Contigs from multiple samples are combined into a single graph enabling the comparison of contigs to detect SV and genotype individuals (Iqbal *et al.*, 2012). An advantage of the assembly-based method is that it does not rely on short read alignment to the reference as a starting point, which makes it comparatively unbiased or less reliant on the reference sequence. This independence from the reference sequence facilitates discovery of novel variants (Iqbal *et al.*, 2012; Zhao *et al.*, 2013).

A disadvantage of this method is that it requires high coverage sequence data to have sufficient overlapping short reads to assemble, and therefore doesn't perform well in repetitive regions. Another limitation is that the *de novo* assembly process is computationally demanding (Cameron *et al.*, 2019). Very few tools use this method exclusively, it most often is combined with one or more of the other methods.

1.8.5 Combined tools

Combined tools can be divided into those which combine two methods and those combining three methods. Tools thus far have not utilised all four methods.

The tools combining two methods all utilise the read pair method. The most common combination is the read pair and split read method combination. Common read pair and split read combination tools include DELLY (Rausch *et al.*, 2012), SoftSV (Bartenhagen and Dugas, 2016) and Wham (Kronenberg *et al.*, 2015). Fewer tools combine the read pair method with read depth and assembly-based methods. Read pair and read depth combination is used in Genome STRiP (SVDiscovery) (Handsaker *et al.*, 2011) and inGAP-sv (Qi and Zhao, 2011) and the read pair and assembly-based method combination is used in hydra-SV (Quinlan *et al.*, 2010). These combination

tools are considered to be more accurate than read pair methods alone and produce more accurate SV calls.

Triple combination tools all utilise a combination of read pair and split read methods with either read depth or the assembly-based method added. Read pair, split read and read depth combination tools include LUMPY (Layer *et al.*, 2014) and MATCHCLIP. Read pair, split read and assembly-based combination tools include GRIDDS (Cameron *et al.*, 2017) and Manta (Chen *et al.*, 2016). In Manta the algorithm first utilises paired-reads and split read to identify regions containing SV and then determines the nature of the SV using the assembly-based method. A comprehensive comparison and evaluation of SV calling algorithms found that tools that utilise assembly-based methods are very accurate SV calling algorithms (Cameron *et al.*, 2019).

1.8.6 Structural variant calling on multiple genomes

Another important difference between tools aside from the underlying detection method is whether an algorithm is designed to analyse a single genome at a time or multiple genomes in parallel. For SNV calling, it is usual to call across many thousands of genomes in parallel. In SV detection however, this has not been the case, mainly due to the immense computational power required. SV calling on a single genome is already computationally intensive, but when this is multiplied to many genomes, it becomes a significant computational challenge. This has been a large obstacle in trying to call SVs jointly.

Only a few algorithms have been designed to call SV on large sets of WGS data. Genome STRiP is one such tool, as is cn.Mops (Klambauer *et al.*, 2012). Both these tools however do not scale well beyond ~100–200 genomes becoming unreasonably computationally intensive (Collins *et al.*, 2020). A more recently described tool Popdel (Niehus *et al.*, 2021) is a tool for detection of deletions in large population scale sequencing studies. This tool has attempted to overcome the computational intensity of joint calling by having a two-step approach. In the first stage a smaller summary file is created for each WGS BAM file containing read pair information and summary information from the BAM. This smaller file is then utilised in a second stage, to avoid having to repeatedly access the BAM file, an action that greatly increases computational requirements (Niehus *et al.*, 2021). The reason for pursuing a joint

approach despite the challenges is because of the advantages of joint calling for population studies.

In order to perform population studies using algorithms designed to run on single genomes, one has to run each individual WGS through the SV discovery algorithm and then perform a merging step where all the VCF files are merged. This step is problematic, as the same variant may have been detected in multiple individuals but with slightly different coordinates and the process of deciding on the correct coordinates is complicated. Additionally, some variants may not have sufficient evidence in a single individual to be detected conclusively, where over many individuals evidence is clearer and a firm call can be made (Niehus *et al.*, 2021). Popdel is a significant step towards performing computationally efficient joint SV discovery for populations.

Eggertsson *et al.* (2017) have used a very different approach in their genotyping tool Graphtyper. Graphtyper can perform joint genotyping of large-scale population variant datasets using pangenome graphs. Using variants called by another tool, Graphtyper constructs a pangenome graph, a graphical representation of multiple genomes, including the reference and all other genomes at that specific site. Using the graph, it assigns a genotype to each genome, in context of all the other genomes in the graph. The update of this tool GraphTyper2 (Eggertsson *et al.*, 2019) has been adapted specifically for SV. GraphTyper2 takes in a set of SVs in a variant call format (VCF) file discovered by another SV discovery tool and then constructs the pangenome graph. Using the graph, it locally realigns and refines the structural variant site before assigning a genotype to each individual. The advantage of this tool is that it enables one to perform joint genotyping of variants, in large cohorts, that have been called by single sample SV discovery algorithms. An important feature of GraphTyper2 is that it considers all discovered sites, including those with low quality scores. The reason this is advantageous is that a true SV site that may have had poor evidence in one individual, may gain concordance when considered across the population. Being unable to utilise this process is a weakness of single sample discovery algorithms that GraphTyper2 has been able to overcome (Eggertsson *et al.*, 2019).

1.8.7 Application of structural variation algorithms

Since the first SV algorithm was introduced in 2007 at least 70 different tools have been described using different methods and combinations of methods. Some tools have been designed for very specific applications such as detecting somatic SVs in cancer genomes (Yau, 2013) or detecting variants within trios (Liu *et al.*, 2016).

This plethora of tools is indicative of the underlying challenges of detecting SV from WGS. Each of the four detection methods has different strengths and weaknesses and most of the newer algorithms use a combination of methods like Manta and GRIDSS that use read pair, split read and assembly-based methods. However even these combination tools often detect some types of SV better than others, and so it has been recommended to use more than one calling algorithm (Zhao *et al.*, 2013; Kosugi *et al.*, 2019).

Zhao *et al.* (2013) discuss these challenges and compare 48 different tools. They conclude that using more than one tool is the best way of ameliorating the weaknesses of a single tool. They also discuss the challenges of producing a high resolution SV detection tool and the inherent weakness that would have to be overcome by new technologies in order to produce an ideal detection tool for SV from WGS (Zhao *et al.*, 2013). A similar benchmarking exercise was performed by Kosugi *et al.* (2019) comparing 69 different SV algorithms. In this paper they do a comprehensive comparison of all these tools on both simulated and real datasets. They state that each algorithm has maximal performance for specific types and or sizes of SVs. They also perform evaluations of different pairs of algorithms. They find, that by utilising specific pairs of algorithms, precision and recall can be maximised for specific SV sizes and types. It is interesting to see the progress that has been made in the intervening 6 years between these 2 papers. Many new algorithms have been written and many of these have made significant advances in scope, accuracy and sensitivity of SV detection. One of the biggest advances for accurate SV detection has been the use of long read sequencing, which mitigates the difficulties of short read sequences. They discuss that despite its superior performance, this method still has low throughput, and is very expensive, which is limiting its wide-spread use. They compare tools that have been developed for detecting SV from long read sequences like PBHoney (English *et al.*, 2014) and Sniffles (Sedlazeck *et al.*, 2018) and found them to have similar accuracy of detection. They conclude that for both short-read and long-read sequencing the use

of multiple tools is still the best approach to maximise accuracy and sensitivity (Kosugi *et al.*, 2019).

This premise is supported by the fact that the pipelines designed for two recent population SV projects (Abel *et al.*, 2020; Collins *et al.*, 2020), which use multiple tools to produce high quality SV reference datasets. Collins *et al.* (2020) used Manta, DELLY, MELT (Gardner *et al.*, 2017) and cn.MOPS. Abel *et al.* (2020) used LUMPY, CNVnator and SpeedSeq (Chiang *et al.*, 2015).

1.9 Aims and Outline of the Study

This project had two main aims that together will assist in improving the diagnostic yield for patients with DDs in Africa:

Aim 1:

Assess the diagnostic process for DD and the characteristics of the population of patients. This was achieved by completing the following objectives:

1.1 Perform a retrospective file review of all patients presenting to the Human Genetics clinics at three Johannesburg Academic hospitals Charlotte Maxeke Johannesburg Academic Hospital, Chris Hani Baragwanath Academic Hospital and Rahima Moosa Mother and Child Hospital in 2017.
1.2 Describe the characteristics of the patient cohort presenting to the genetics clinics and assess the current diagnostic process in terms of time to diagnosis, definitive genetic aetiologies determined, overall diagnostic yield and cost of investigations.

This first aim is addressed in *Chapter Two,* where the Genetics clinic patient cohort is described, and the current diagnostic process evaluated.

Aim 2:

Perform CNV analysis on WGS of sub-Saharan African individuals to discover baseline population African CNVs. This involved completing four objectives:

2.1 Using literature, as well as expert advice from bioinformatic specialists, choose suitable algorithms for detecting and verifying CNV from multiple WGS.
2.2 Design a pipeline using the chosen algorithms to detect CNV from WGS data.

2.3 Perform CNV analysis on the dataset of African WGS

2.4 Verify variants using alternative bioinformatic tools

The second aim is addressed in *Chapter Three*. This chapter consists of the details of producing the CNV dataset from WGS from diverse African populations, and findings from the analysis of this data. This work, as mentioned in the preface was done in collaboration with the H3Africa GSK ADME study and the TrypanoGen group.

Chapter 2

Characterization of a Genetics Clinic Cohort and Evaluation of the Diagnostic Process in South Africa

---

2.1 Introduction

DD – which include NDD and congenital anomalies – affect 2-5% of children worldwide (Maulik *et al.*, 2011). The phenotypic features and genetic aetiology of DD are highly heterogeneous, making this group of disorders challenging to diagnose. For this reason guidelines have been developed over time to attempt to improve the rate of diagnosis (Mithyantha *et al.*, 2017). In 2010, CMA replaced karyotyping as the recommended first-line investigation (Miller *et al.*, 2010), improving the diagnostic yield from ~3% to about ~12%. Consequently the majority of patients with DD still remain undiagnosed. This often results in an extended period of uncertainty and ongoing testing, termed the diagnostic odyssey. This has profound psychological effects for patients and families (Carmichael *et al.*, 2015) as well as cost and medical implications.

In the past 10 years the introduction of WES has been changing this landscape rapidly, and in 2019 a consensus statement by a multi-disciplinary international group was released, recommending WES as the first-line investigation for NDD, indicating a diagnostic yield of 30–53% (Srivastava *et al.*, 2019). In 2021, the ACMG released a new guideline for the diagnosis of children with congenital anomalies, GDD and intellectual disability, recommending WES as the first-line investigation (Manickam *et al.*, 2021). It has been suggested that this approach would result in earlier diagnoses of more patients with many positive effects; it halts further diagnostic investigations, ultimately lowering investigation costs and ends the burdensome diagnostic odyssey for families. An accurate molecular diagnosis means clinicians can understand the condition better and therefore provide more accurate condition-guided management and surveillance, precision therapy where available, as well as recurrence risk assessments (Tan *et al.*, 2017).

In LMICs, like South Africa, DD are present at even higher levels than in high income countries (Maulik *et al.*, 2011; Olusanya *et al.*, 2018). Improving the diagnosis of DD is therefore of vital importance, and the large-scale implementation of exome sequencing as the first-line investigation in this setting could help to realise this objective. However, the cost of exome sequencing remains a barrier to implementation. Another significant obstacle to implementation is a lack of evidence of the need for improved diagnosis. Little has been published about patients with DD in Africa; in terms of their diagnostic outcomes or genetic aetiologies, or the diagnostic processes currently in use.

In this study we conduct a retrospective file audit of a cohort of patients who presented to one of the largest medical genetics services in sub-Saharan Africa. The aim of this audit is to identify broad groups of patient phenotypes and their resultant diagnoses and to evaluate costs of genetic tests conducted on these patients. This will enhance our understanding of this patient cohort, enabling us to predict the potential benefits of implementation of WES in this, and similar settings, and to direct testing appropriately.

2.2 Materials & Methods

2.2.1 Study Design and Case Review Selection

The study was designed as a retrospective file review of patients who attended Division of Human Genetics clinics, between 1st January 2017 – 31st December 2017. All consecutive cases were included of affected probands. Patients were excluded who attended foetal medicine and counselling clinics as these are focused on counselling for high-risk pregnancies and screening or preventative genetic testing, not the diagnosis of an affected proband. Both new patients and patients returning for follow-up visits were included. All test results for these patients were included in assessing their diagnostic test, number of tests and cost calculations, whether or not they fell within the review period.

2.2.2 The Medical Genetics Clinic

The medical genetics clinics reviewed are managed by the Division of Human Genetics, National Health Laboratory Services (NHLS) and University of the Witwatersrand, in Johannesburg, South Africa. These genetics clinics operate within the University of Witwatersrand academic hospitals and serve the public health sector. The public health sector serves 80% of the South African population. Patients are referred to the genetics clinics from hospitals and non-genetics clinics in Johannesburg, and the southern Gauteng province, as well as from clinics in neighbouring provinces, where genetic services are not available. Genetic testing is offered through the NHLS that similarly serves the public health sector. Figure 2.1 summarises the structure of these clinics and the types of referrals received. In this study we distinguish between clinical- and genetic diagnosis, with genetic confirmation of a condition referring to cytogenetic of DNA-based testing methodologies.

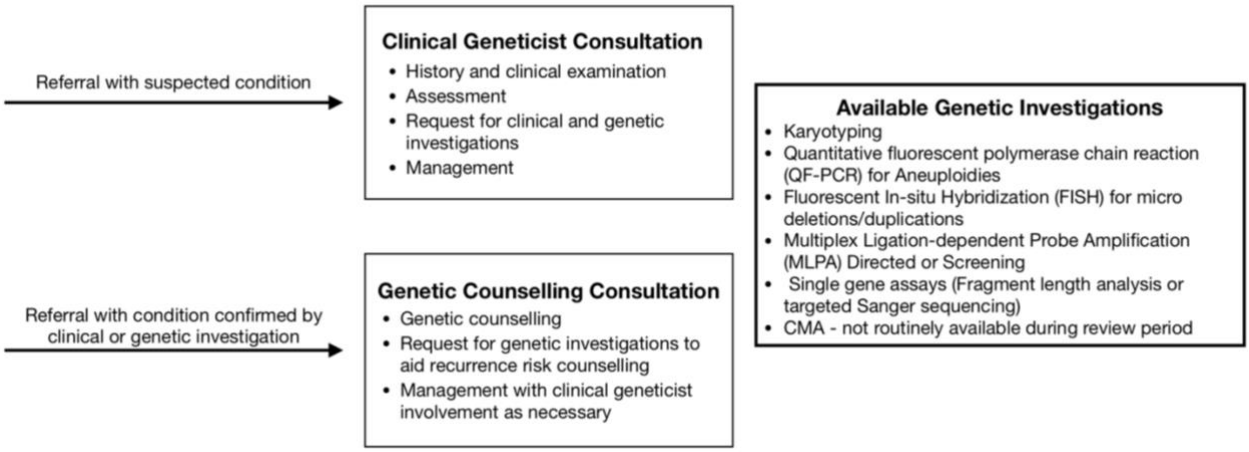


Figure 2. 1: Overview of the functional structure of the genetics clinic and the two main types of patient referrals

2.2.3 Ethics

Ethical approval to access and analyse retrospective data from clinical files in division of Human Genetics was obtained from the University of Witwatersrand Human Research Ethics Committee (HREC) clearance number M180506 (Appendix I). Ethical approval for this analysis to be included this PhD study was granted with clearance number M180885 (Appendix II).

2.2.4 Data capture & management

Genetic clinic records were used to compile a list of all patients having attended Genetics clinics in the year 2017. From these lists, patient files were retrieved from the file archive at the Division of Human Genetics. Data contained in the Human Genetics files includes their patient demographic data (including age, sex and ethnicity), clinical phenotype and assessment made by medical geneticists, summary information from other specialists where this is relevant to their genetic assessment and management. Genetic test results were accessed from the online laboratory results system NHLS LabTrak. A database was designed using REDCap to capture patient demographic data, clinical assessments and genetic investigations. Table 2.1 shows the data collected on each patient. Data were captured and managed using the REDCap electronic data capture tools hosted at the University of Witwatersrand (Harris *et al.*, 2009; Harris *et al.*, 2019).

Table 2. 1: Summary of information captured on each patient

Data type	Detail
Demographic data	*Sex, date of birth, ethnicity, country of birth*
Consultation dates	*First and most recent (in 2017)*
Main features of the patient's condition	*Patients assigned up to three phenotype categories*[a]
Genetic investigations performed	*Investigation (date, result, diagnostic yes/no)*
Diagnosis details	*Diagnosis (date of diagnosis, genetic confirmation yes/no)*
Patient summary	*Most recent clinical assessment*

[a] A list of 19 phenotypic categories was compiled to allow concise description of the main features of each patient's condition (Appendix III). As each patient was allowed more than one phenotypic category to describe them, this did not divide patients into mutually exclusive groups, rather, the data were captured to assess commonly presenting phenotypes. This was with the exception of patients who had known aneuploidies. These patients were only assigned the category of aneuploidy to allow them to be separated within the patient cohort.

2.2.5 Diagnostic costing

The cost of genetic testing was calculated by combining information on resource use (the diagnostic tests captured for each proband) and test prices extracted from the 2017 NHLS State Price list. This in-house price list presents prices agreed between the South African Department of Health and the NHLS. One diagnosis resulting from a test performed as part of a research project was included in the dataset but was not included in the costing analysis. Test cost data were summarised using means medians and ranges.

2.2.6 Data analysis

Data cleaning, analysis and visualization were performed in Stata13 (StataCorp. 2013. Stata Statistical Software: Release 13. College Station, TX: StataCorp LP.) and Apple Numbers (version 6.2.1. Apple Inc. California).

2.3 Results

2.3.1 Overview of the cohort

Files were retrieved for 88% (934/1059) of patients who attended a clinic in 2017. The resulting cohort 934 unselected patients with a male to female ratio of 1.2:1. Nationality and ethnicity were recorded for 79% (737/934) of patients and in this group most of the patients (90%) are black South Africans, with all 11 official language groups represented. Table 2.2 shows the demographic breakdown of patients. These statistics are indicative of a South African urban population, and match Johannesburg census data (StatisticsSA, 2011). This was a predominantly paediatric cohort with 95% (888/934) of the patient population were younger than 18 years of age, with 43% (403/934) of these patients presenting at clinic before the age of one year, and 75% (703/934) before the age of five years.

Table 2. 2: Demographic profile of the genetics clinic patient cohort

		Mean/Median	Ratio	N	%
Age at Presentation		4,3/1,3			
Sex	Male:Female		1.1:1		
	Male			506	54
	Female			428	46
Total				**934**	100
Nationality & Ethnicity	**South Africans**			**660**	**89,6**
	Black			568	77,2
	White			50	6,9
	Mixed Ancestry			23	3
	Indian			16	2
	Asian			3	0,5
	Non-South African			**77**	**10,4**
	Zimbabwe			32	4,3
	Democratic Republic of Congo			13	1,7
	Ethiopia			8	1
	Malawi			7	0,9
	Other Nationalities (n=7)			17	2,3
Subtotal				**737**	100
Recorded Nationality & Ethnicity				737	79
Unrecorded Nationality & Ethnicity				197	21
Total				**934**	100

2.3.2 Presentation and diagnostic outcomes

Seven hundred and forty-two (83%) patients presented with features of DD, with the three most prevalent features being global developmental delay (36%), congenital anomalies (33%) and dysmorphic features (26%). In total, 72%, 69% and 75% respectively of patients with these individual phenotypes remained undiagnosed. Thirty-two patients (3%) were deemed to have no genetic condition or presented with a phenotype within the range of normal variation, so were excluded from further analyses. With these patients removed, the cohort was 902 patients.

Figure 2.2 shows the overall diagnostic outcomes of the patient cohort. Half (473;52%) of the cohort remained undiagnosed, 48% (429) of patients who attended the clinics were provided with a diagnosis, of which 16% (145) were patients who received only a clinical diagnosis but no genetic confirmation of their diagnosis. A patient was

considered clinically diagnosed if the diagnosis was clear enough to enable condition specific management. Aneuploidies accounted for 18% (164) of all patients, and 55% of all genetically confirmed diagnoses.

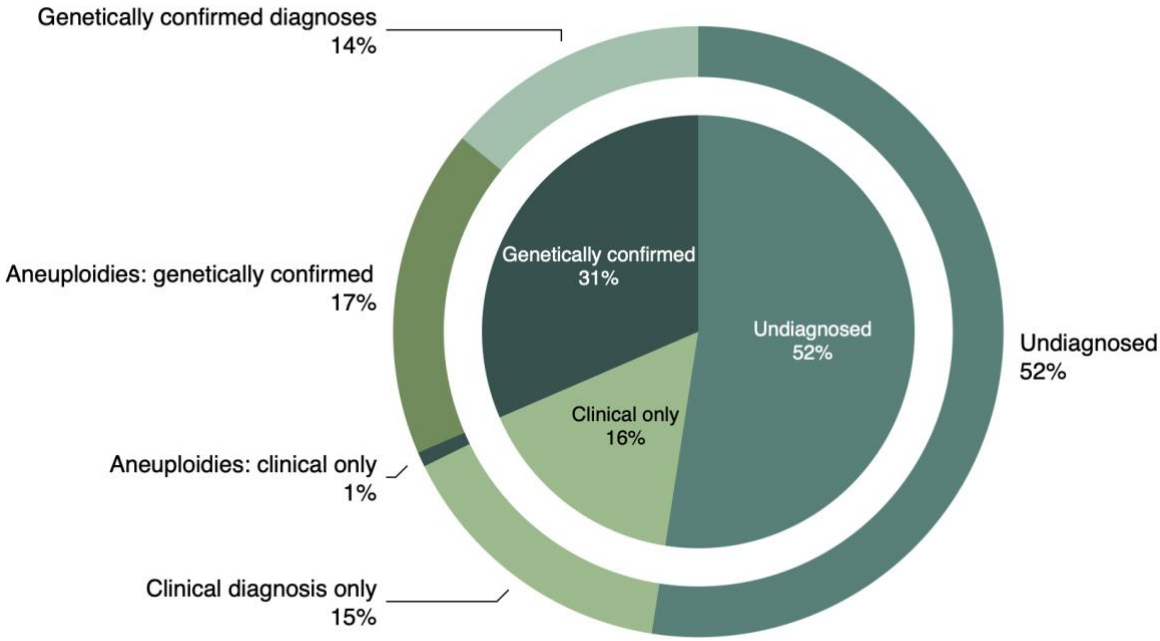


Figure 2. 2: Overall percentages of patient diagnostic outcomes. The majority of patients remained undiagnosed, and of those with genetically confirmed diagnoses the majority were diagnosed with aneuploidies. A diagnosis was considered genetically confirmed only if a positive genetic (cytogenetic or DNA based testing methodologies) result was on record for the patient.

2.3.3 Patient diagnoses

There were 130 different diagnoses made, 75 of which occurred only once in this cohort. Aneuploidies accounted for four of the top 10 diagnoses as shown in Figure 2.3. These top 10 diagnoses accounted for 54% (233) of diagnoses with Trisomy 21 being the most common diagnosis, accounting for 30% (129) of all diagnoses. In this cohort three distinct groups of patients emerged, based on their different diagnostic profiles as shown Figure 2.4 and discussed in section 2.3.4.

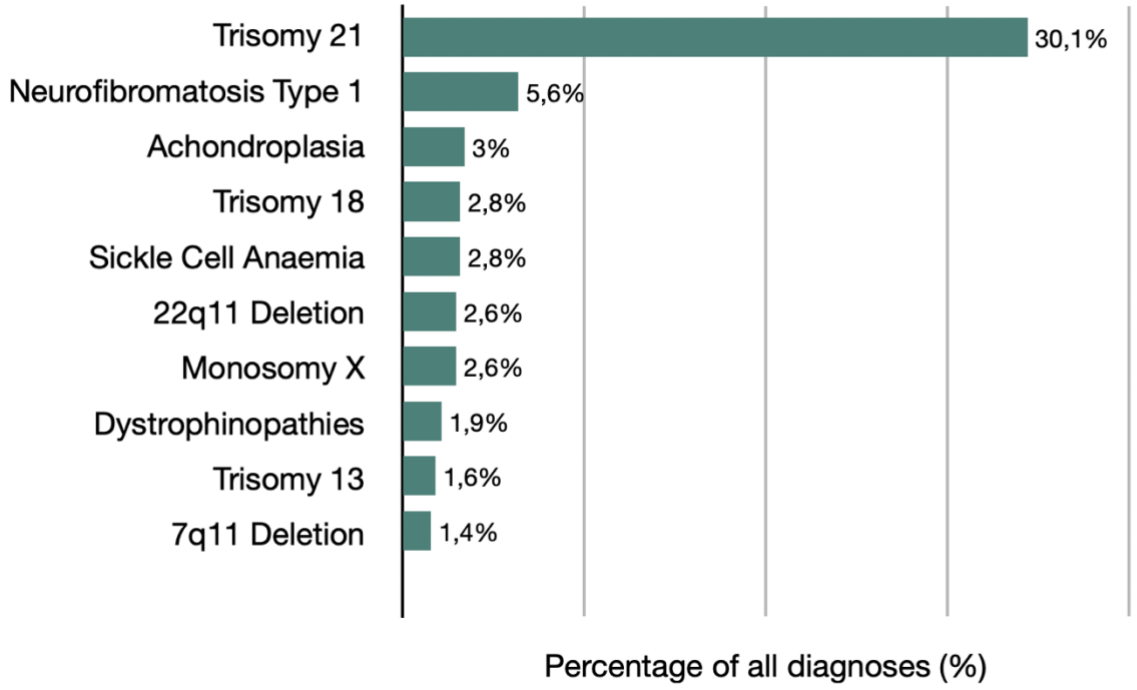


Figure 2. 3: Top 10 genetic diagnoses made in the cohort. Trisomy 21 was the most common condition diagnosed.

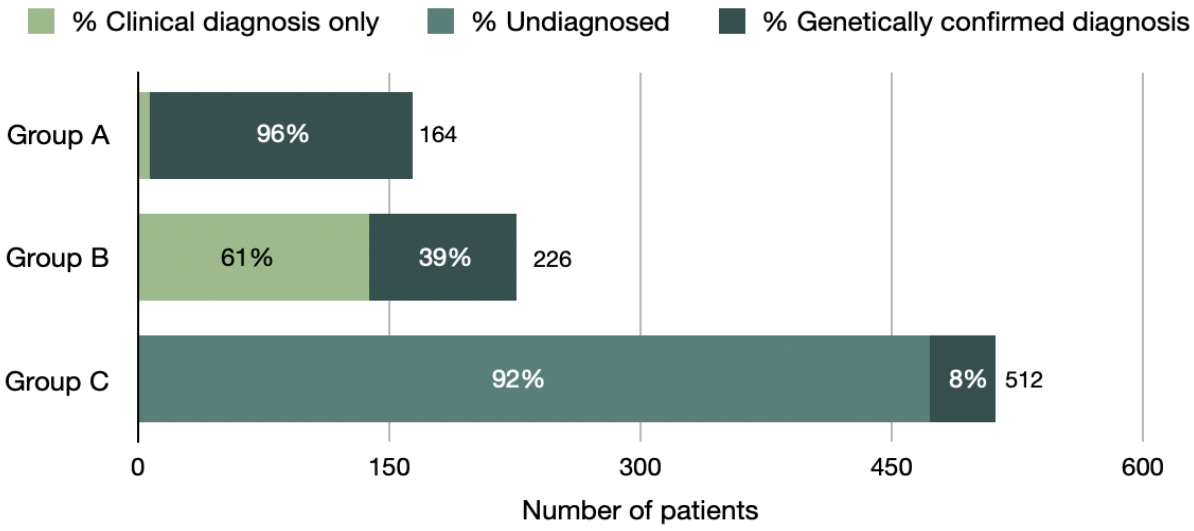


Figure 2. 4: Three main groups of patients from the file review identified with differing diagnostic profiles. Group A: Patients with aneuploidies, Group B: Patients with easily recognisable conditions, Group C: Patients with rare, less-recognisable conditions.

2.3.4 Characteristics of patient groups A, B and C

A. Group A – Aneuploidies

Group A (164/902 patients, 18%) contained patients with aneuploidies. The identifiable phenotypic features of these patients meant that referring clinicians often made the diagnosis and ordered appropriate genetic investigations. Consequently, 61% of these patients had genetically confirmed diagnoses before their first genetics clinic visit. The time to diagnosis for this group is thus negative with mean –129 days and median –29 days before their first genetic clinic visit. Aneuploidy diagnoses were genetically confirmed in 97% of cases, with either QF-PCR (84%) or karyotype (16%) and the median and mean number of genetic tests in this group was 1 and 1.1 tests respectively. The lack of genetic confirmation in the remaining patients was primarily due the unavailability of historical testing records. The mean and median costs to diagnose a patient with an aneuploidy were R1746,45 and R1707,21.

B. Group B – Easily recognisable conditions

Patients in Group B (226/902; 25%) were those diagnosed by medical geneticists with conditions with well described clinical phenotypes. In 61% of these patients no genetic confirmation could be provided. In most cases, this was because there was no test available in the system to diagnose the condition in question. NGS methodologies are not available routinely and variant or disease specific assays are limited in terms of the number of conditions covered and knowledge of African pathogenic variants.

In some cases, geneticists were sufficiently confident in a patient's clinical diagnosis that genetic confirmation was not considered essential to inform management. Two subgroups of patients were identified in Group B for whom geneticists were particularly confident in their clinical diagnosis. In the first, Group B1, (n=29), diagnostic clinical investigations were available, for instance, confirmation of sickle cell anaemia through a haematological test, and in the second, Group B2, (n=117), the conditions had a very clear distinct clinical phenotype, such as albinism. In these cases, clinicians considered the diagnoses firm enough to allow for appropriate counselling, future pregnancy risk assessment and condition specific management, without a confirmatory genetic test. The use of genetic tests in these two subgroups, and their associated costs, reflect this pattern of behaviour. The median and mean number of tests for the first group was 1 and 0,86 tests, with a median cost of R0,00 and a mean cost of R1079,25. The median

and mean number of tests for the second group was 0 and 0,63 tests, with a median cost of R0,00 and a mean cost of R1119,52.

A third subgroup (Group B3) was identified (n=80) who had phenotypes that were recognizable but not as clearly distinct as the first two groups, such as 22q11 syndrome or Noonan syndrome. These patients had suggestive diagnoses that could guide appropriate testing, but clinicians had lower confidence in clinical diagnoses, so relied more on genetic confirmation for diagnosis than in the first two subgroups of Group B. The median and mean number of tests in this group was 2 and 1,88 tests, and the median and mean costs were R3115,07 and R3226,42. The overall median and mean cost for Group B were R1467,95 and R1860,16. In this resource constrained context, patients with a greater need for testing are prioritised, and with fewer types of genetic investigations available, where there is a high degree of confidence in clinical diagnoses, genetic confirmation is not always pursued.

C. Group C – Rare, less recognizable conditions

Patients in Group C (512/902; 57%) presented with non-specific features that did not clearly point to a recognised condition. Of these patients, 90% presented with features of a DD – with developmental delay (55%), congenital anomalies (45%) and dysmorphic features (39%) most commonly observed shown in Figure 2.5.

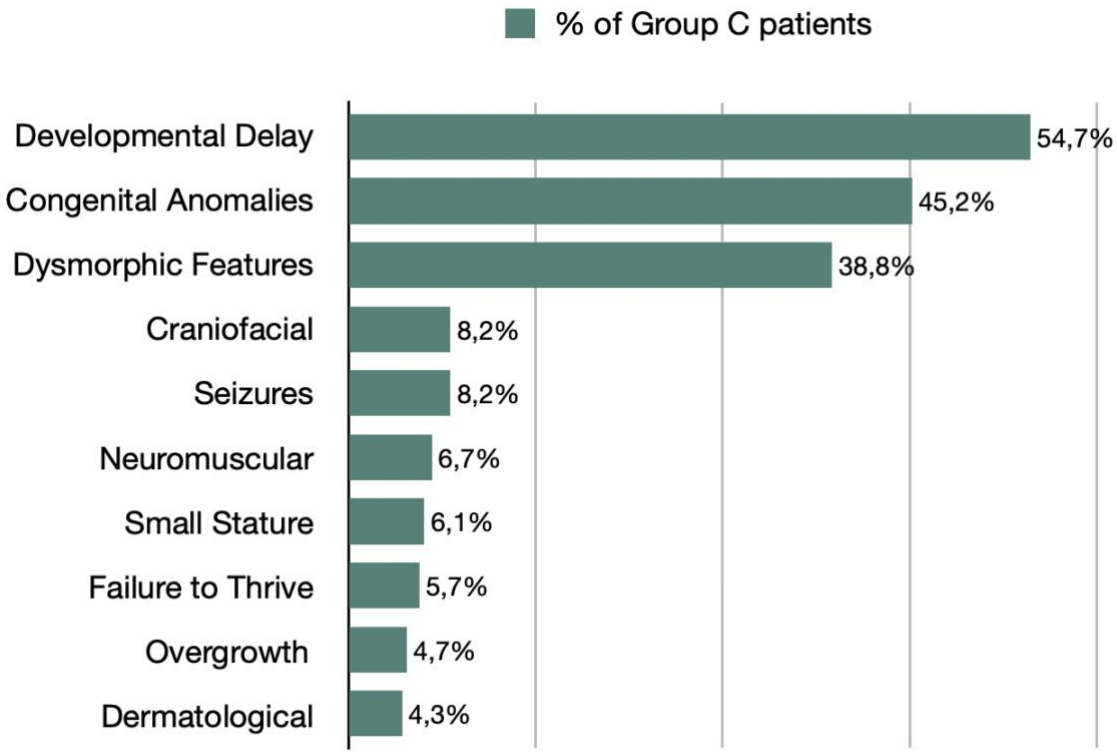


Figure 2. 5: Top ten phenotypes observed in patients in Group C. The most common phenotypes by far were developmental delay in 54.7% of patients, congenital anomalies in 45,2% and dysmorphic features in 38,8% Each patient was assigned up to 3 phenotypic categories meaning groups in each category are not mutually exclusive.

Only 8% (39/512) of patients in Group C received a diagnosis. 45 patients were considered lost to follow-up as they had not returned for a follow up visit or further testing by the end of the data capture period (~1-2 years after their first visit). These patients were removed from further analysis to prevent skewing of results.

The non-specific diagnostic tests used most commonly undertaken to provide diagnoses in this group included those able to detect chromosomal aberrations (e.g. karyotype, MLPA, for common micro-deletions and sub-telomeric deletions/duplications, and CMA) (Figure 2.6). Only 23% of these patients had single gene assays. QF-PCR aneuploidy was performed in 18% of patients, and FISH tests for specific micro-deletion and -duplication syndromes were undertaken in a minority (9%) of cases.

Many patients in Group C had numerous investigations in pursuit of a diagnosis: the median and mean number of tests per patient were 3 and 2,6 and 54% of patients had had three or more tests (Figure 2.7).

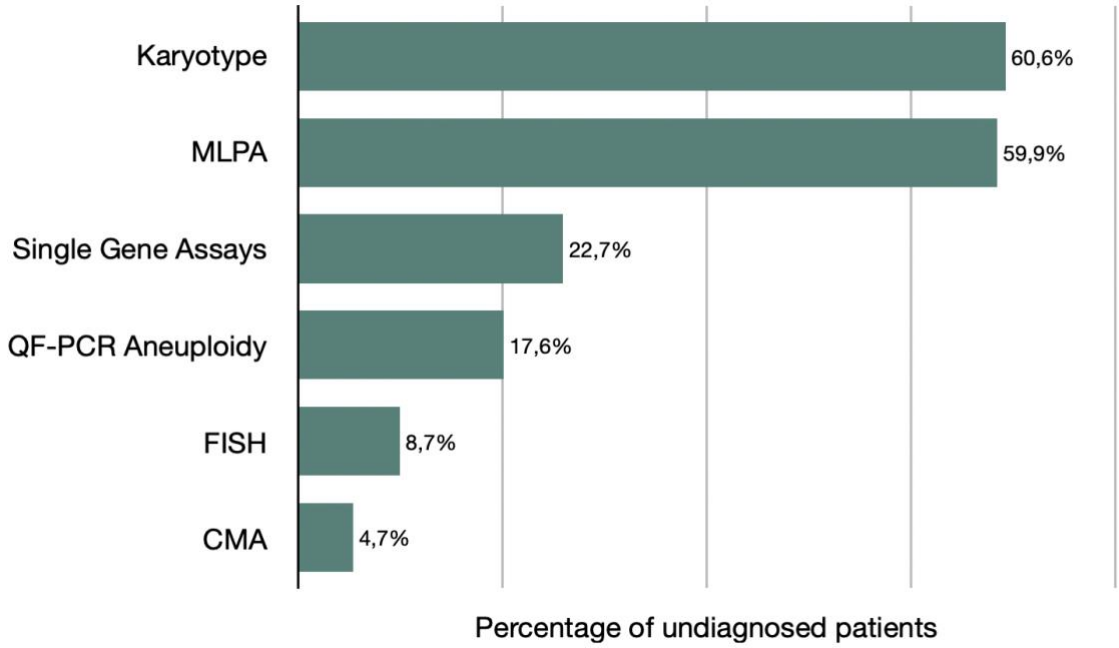


Figure 2. 6: Genetic Tests of undiagnosed patients in Group C in an attempt to reach a diagnosis. The most common tests ordered for Group C patients were Karyotypes and MLPA for detection of sub-telomeric deletions/duplications and known microdeletion/duplication syndromes. Single gene assays include Fragment analysis and Sanger sequencing.

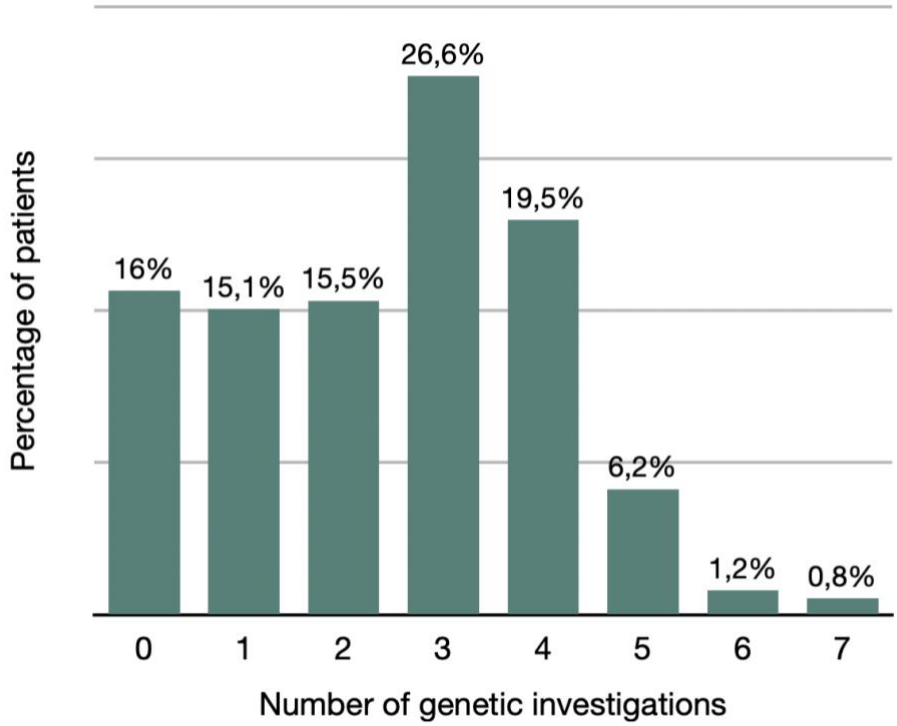


Figure 2. 7: Number of genetic investigations undergone by undiagnosed patients in Group C. The greatest percentage (26,6% of patients) had 3 tests ordered with 54.3% of patients having had 3 or more tests.

2.3.5 Cost of Genetic Investigations

Figure 2.8 shows the patients in Group C undergo multiple tests that return negative results, so remain in the clinic for several years awaiting a diagnosis. The median and mean cost of testing were R5400,87 and R4774,30 (range: R0,00 - R15 895,58) per patient. These costs were significantly higher than those incurred by patients in Groups A and B. A comparison of costs between Groups A, B and C can be found in Table 2.3. These costs do not include the cost of time with clinicians or the cost of other clinical investigations, such as radiological, haematological or metabolic testing, ordered as part of the diagnostic odyssey or for broad management principles. Consequently, the total cost per patient of seeking a diagnosis and on broad management is likely higher than these figures.

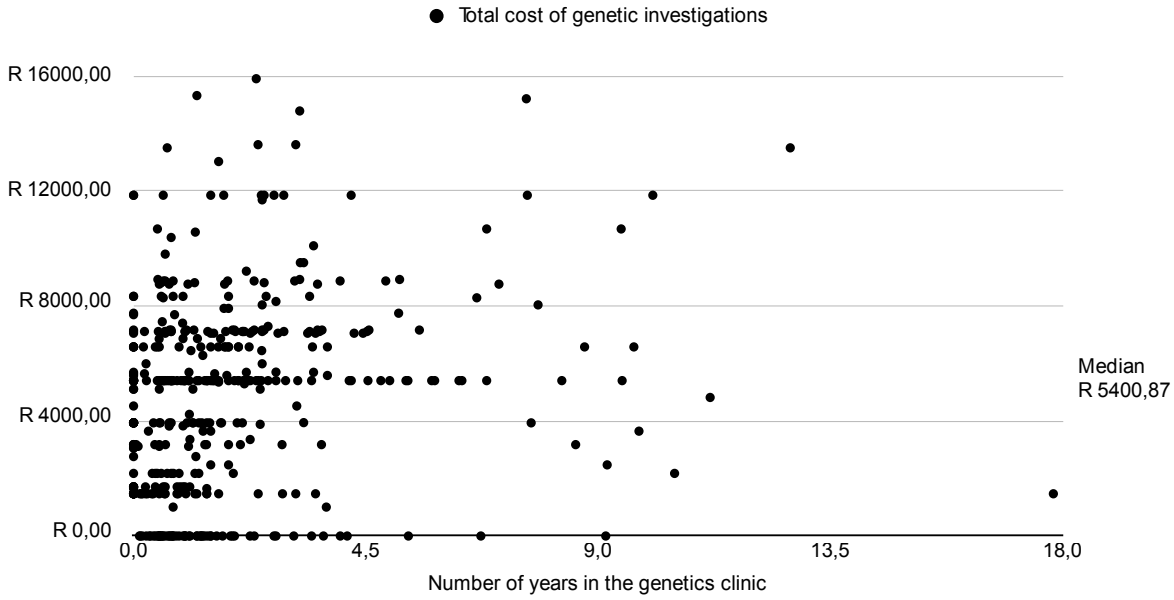


Figure 2. 8: Total cost of genetics investigations for each Group C patient against the length of time the patient has been in the clinic. Many patients have multiple investigations that with mounting costs over many years and yet remain undiagnosed. The median cost per patient was R5400,87.

Table 2. 3 Comparison of test use and costs between Groups A, B and C

Group		N (%)	Number of tests		Cost (ZAR)	
			Mean	Median	Mean	Median
Group A – aneuploidies		164 (18%)	1,1	1	1746,45	1707,21
Group B – easily recognisable conditions		226 (25%)	1,1	1	1860,16	1476,95
	Group B1	29 (3%)	0,86	1	1079,25	0
	Group B2	117 (13%)	0,63	0	1119,52	0
	Group B3	80 (9%)	1,88	2	3115,07	3226,42
Group C – rare less recognisable conditions		512 (57%)	2,6	3	4774,3	5400,87
All Patients		902 (100%)	1,9	1	3215,77	2175,42

2.4 Discussion

In this file audit of a cohort of patients presenting to a medical genetics clinic in South Africa over a one-year period, we found that the largest group of patients are those presenting with non-specific features of DD. These patients are not being effectively diagnosed by the currently available testing methodologies and may benefit from the introduction of a test with a higher yield than those currently in use, such as WES. WES is able to detect multiple types of variants across the genome and is now the recommended by the ACMG as the first-line test for children with DD (Manickam *et al.*, 2021).

The first group of patients identified (Group A) were those diagnosed with aneuploidies, with trisomy 21 being the most common diagnosis. This was expected as this is considered the most common chromosomal condition worldwide (Bull, 2020). The diagnostic process for this group of patients in South Africa differs from the process in developed countries, where most diagnoses are made through prenatal screening, increasingly by non-invasive prenatal testing (Rudolf *et al.*, 2017). In this cohort most patients were diagnosed postnatally by clinicians and then presented to the genetics clinic for genetic confirmation and/or counselling. This is most likely due to inconsistent prenatal screening for Down syndrome and other aneuploidies in antenatal care in South Africa (Urban *et al.*, 2011). Despite the fact that these patients were successfully diagnosed postnatally and do not require introduction of WES to improve diagnosis, this research highlights the fact that prenatal screening and diagnosis of aneuploidies

is currently very limited in South Africa. Testing for aneuploidies should continue, where indicated, using QF-PCR Aneuploidy.

The second group of patients (Group B) were those who could be diagnosed with confidence based on recognisable features, diagnostic clinical investigations or clinical scoring systems. This group was heterogenous, but most often diagnoses were confident enough to inform management, whether or not genetic confirmation of the diagnosis was obtained. The most decisive factor in genetic investigation choice was the availability of different tests. Given the ability of exome sequencing to detect both SNVs and CNVs throughout the genome, with a higher diagnostic yield than traditional tests, it could provide firm genetic diagnoses for many of these patients (Dillon *et al.*, 2018). Even though these patients received clinical diagnoses, confirmed genetic diagnoses from WES within a shorter timeframe would be beneficial. Such diagnoses could enable more precise management due to the impact of genotype-phenotype correlation, and in time, possible gene or variant-specific targeted therapies. For families quicker confirmed genetic diagnoses would allow more accurate prenatal risk assessment and could reduce psychosocial stress, by being certain of the genetic cause (Makela *et al.*, 2009). Additionally increasing the number of molecular diagnoses would assist in understanding the genetic epidemiology present in poorly characterized African populations.

The third and largest group of patients (Group C), were those who presented with non-specific features of DD. Only 8% of these patients received a diagnosis, on a par with the expected diagnostic yield of the traditional tests employed (Challman *et al.*, 2003). The investigation costs over time show how many of these patients remained on a diagnostic odyssey, having spent many years in genetics clinics with successive tests being done and yet remaining undiagnosed. This results in mounting costs for these patients, probably underestimated in this report, as we did not include other clinical investigations, time with clinicians or indirect costs to families. A recent study by Dragojlovic *et al.* suggests that indirect costs of patients who remain undiagnosed should be taken into account when assessing the cost-effectiveness of new testing methodologies, such as WES. Their research shows that even if initial costs are higher, having a genetic diagnosis can result in lower indirect costs compared to patients who remain undiagnosed (Dragojlovic *et al.*, 2020).

This study demonstrates that genetic diagnostic processes in use in South Africa are inadequate to diagnose patients with DDs, and new methods must be introduced if the diagnostic rate is to improve. Routine use of CMA could be a first step to improving the process as this technique is known to have a diagnostic yield for DD of 15-20% (Miller *et al.*, 2010). However, CMA is not able to detect SNVs. WES would therefore be the ideal tool to improve diagnostic yield in this context. The diagnostic yield of WES has been estimated at 30-53% for patients with NDDs (Srivastava *et al.*, 2019), and would be further enhanced if CNV analysis was also undertaken, as is increasingly possible (Gordeeva *et al.*, 2021). A further analysis of smaller subsets of different combinations of DD phenotypes in the cohort may enable further division of which patients will most benefit from first-line WES over CMA or a NGS panel of common DD genes. However as WES and WGS become more widely used, broad panels may become less valuable and the cost differential may decrease, making sub-categorisation less relevant. Studies on the value of WES and WGS has shown that doing limited testing does not benefit the majority of patients often results in patients having inappropriate testing due to the options of panels that are available. Availability of broader testing options, namely WES would ensure that the most cost-effective option for a patient can be chosen with the highest yield (Dillon *et al.*, 2018).

WES is currently only available in South Africa to those who can pay privately for international diagnostic services. At around R15 000-35 000; these costs are high and not a viable option for the majority of patients reliant on the state health system, considering the median monthly income in South Africa is R2800 per month. For exome sequencing to be viable it will have to be performed locally. The projected cost of a diagnostic exome sequence in South Africa is still relatively high due to the lack of established infrastructure, and high component and training costs (Flynn, 2020). Furthermore interpretation of the large amount of data produced by WES requires new skill sets and reliable pipelines to ensure optimal analysis. Additionally there is still a high likelihood of finding variants of unknown/uncertain significance, which may not enable firm diagnoses to inform management (Krause, 2019; Kamp *et al.*, 2021). The identification of variants of unknown significance present a specific challenge in an African context as there are few baseline population data available to reference when determining pathogenicity (Bope *et al.*, 2019). New validation and governance protocols will also be required to implement WES in clinical practice. Many of these challenges will only be overcome by capacity development through implementation

and use over time, further underscoring the importance of timely implementation of WES as a first-line diagnostic option in low- and middle-income countries.

A further challenge relates to how to make an economic case for the use of WES in this context in South Africa. There are currently no economic evaluations of the use of any form of exome or genome sequencing in an African context (Schwarze *et al.*, 2018). Such evidence is crucial given that countries across the continent have limited health budgets and many competing funding priorities, including treatments for infectious and chronic diseases that affect millions of people. The data presented in this paper on testing costs and time to diagnosis contributes to this economic evidence base but is only a first step towards generating the required evidence to support the implementation of sequencing in routine diagnostic services. Studies evaluating the cost-effectiveness of exome and genome sequencing in this setting are urgently required. These studies should go beyond a comparison of genetic testing costs to consider the costs incurred by patients before and after testing related to clinical care (both in primary care and secondary care). Importantly, such studies should also consider the full impact of possessing a genetic diagnosis on the quality of life of patients and their families. The research by Masri and Hamamy (2021) in Jordan, suggesting that WES may be cost effective in developing countries, supports further investigation of the cost effectiveness of implementing diagnostic WES in this context.

This retrospective audit of genetics clinic patients in South Africa provides insight into different groups of patients and how the current diagnostic process is serving them. In all groups there is a need to improve and upgrade testing. Although exome sequencing would not be the first-line option for a subset of patients with aneuploidies, for most patients the implementation of WES would be the best way to attain broad-based genetic diagnoses. We conclude that WES has the potential to be a worthwhile investment in a low- and middle-income country setting and will provide a much-needed leap forward in precision medicine and health improvement in these settings.

Chapter 3

An Analysis of Population Copy Number Variation in Sub-Saharan African Genomes

---

3.1 Introduction

SVs are rearrangements of genomic content at least 50bp in size, including deletions, duplications, insertions, inversions, translocations or complex SV. CNVs are a subset of SVs that cause a change in the number of copies of a genomic region (deletions and duplications). SVs are responsible for a large part of genomic variation between individuals (Sudmant *et al.*, 2015a; Alkan *et al.*, 2011) but also play a large role in the causation of diseases, having been implicated in both common diseases such as cancer (Li *et al.*, 2020), as well as rare diseases like many NDDs (Coe *et al.*, 2014). Unfortunately, however, the study of SVs has been slowed by the complexity and technical challenges of SV discovery. These challenges have resulted in a lack of high quality publicly available variant databases from large population studies which in turn has limited the growth of our understanding of the precise role of SVs in causing rare diseases, as well as their impact on normal variation and multifactorial diseases (Abel *et al.*, 2020).

An understanding of population variation informs research into genetic disorders. Early studies to map human SVs and understand their role in human disease were performed on microarray platforms (Redon *et al.*, 2006; McCarroll and Altshuler, 2007; Conrad *et al.*, 2010). These studies started to uncover the role of SVs in diseases such as NDDs (Prasad *et al.*, 2012), but lacked breakpoint accuracy and high resolution for the discovery of small SVs (Sudmant *et al.*, 2015b). With the increasing number of human genomes sequenced by NGS, methods to detect SVs from NGS WGS were developed. These SV calling methods were initially very error prone but over time have been greatly improved. Studies such as the 1000 Genomes Consortium project produced a landmark SV dataset that greatly contributed to known global SVs (Sudmant *et al.*, 2015b).

Recently a large survey of SVs was conducted on a large number of WGS to produce the gnomAD-SV database (Collins *et al.*, 2020), and this will be a valuable tool for furthering research of SVs. Current SV reference databases, such as gnomAD-SV are biased to European populations containing much smaller numbers of African genomes. Additionally, many of the African ancestry individuals included are African American individuals who have only a subset of the diversity seen in African populations (Tishkoff *et al.*, 2009; Choudhury *et al.*, 2020). This underrepresentation of African genomic data in such databases reduces their usefulness to researchers

studying individuals of African ancestry. High quality datasets of baseline population African variation, and analysis thereof, are of critical importance. They will both enrich our knowledge of the human genome (Choudhury *et al.*, 2020) as well as providing a reference of baseline African variation to aid in the interpretation of potentially pathogenic variants in the study of genetic diseases in Africa (Gurdasani *et al.*, 2015; Bope *et al.*, 2019).

Research into genetic disorders in Africa has been slowed by the lack of data from African individuals in public genetic resources (Bope *et al.*, 2019; Kamp *et al.*, 2021). A few studies investigating CNV in Africa have been done on smaller less diverse cohorts using micro-array and medium coverage WGS, but none on a large number of diverse high coverage WGS that can be incorporated into SV reference databases (Vogler *et al.*, 2010; Nyangiri *et al.*, 2020). There are a few reasons why such a dataset has not been produced to date. Firstly, there has been little availability of high coverage WGS of African individuals especially very few southern African individuals (Choudhury *et al.*, 2017). Secondly, calling CNVs from WGS involves significant bioinformatic and computational challenges that would have to be overcome if such a database is to be produced from WGS (Kosugi *et al.*, 2019; Cameron *et al.*, 2019).

Given the importance of this reference database, especially for disease research, we aimed to produce a detailed catalogue of African CNVs from 1027 African WGS from 16 sub-Saharan African countries representing east, west, central and southern African regions. From this rich dataset we aimed to uncover the diversity of the African CNV landscape and to produce a set of high-quality African reference CNVs.

3.2 Methods

3.2.1 Datasets

Figure 3.1 shows the distribution of participants across Africa whose WGS were included in this study. High coverage short read WGS from ongoing research projects as well as publicly available datasets were included in this study.

A. Cell Biology Research Laboratory HIV Study

The Cell Biology Research Laboratory HIV study selected HIV positive individuals considered to be elite controllers (individuals who maintain an undetectable viral load

for at least 12 months without treatment). These 40 individuals were all black South Africans except one who was mixed ancestry. Ethnolinguistic details were not released. The samples were sequenced by Edinburgh Genomics, Edinburgh, Scotland using the TruSeq Nano protocol (PCR-based) and high coverage sequencing (~30x) was done utilizing the Illumina SeqLab workflow system and the Illumina HiSeqX platform (Cell Biology Research Laboratory, National Institute of Communicable Diseases (NICD)/University of Witwatersrand). An application was lodged by Prof Caroline Tiemessen to the University of Witwatersrand HREC to analyse baseline variants for the research of developmental disorders from these data with clearance number M140926 (Appendix IV).

B. H3Africa-Baylor

The H3Africa data was from 347 individuals from multiple H3Africa studies and individuals were from Benin, Burkina Faso, Botswana, Cameroon, Ghana, Nigeria, Mali and Zambia. Samples were predominantly healthy controls recruited for each of the studies, with the exception of case only studies in Cameroon, Botswana, Mali and Benin. More detailed descriptions of these participants can be found in Choudhury *et al.* (2020) Supplementary Methods Table 1. None of the cases were of DDs so they were not excluded. Samples were sequenced at Human Genome Sequencing Center (HGSC) at Baylor College of Medicine, Houston, United States. Samples were prepared using the TruSeq Nano DNA Library Prep Kits (PCR-based) and underwent WGS on an Illumina TenX (150 bp) to a minimum depth of coverage of 30x (Choudhury *et al.*, 2020). Principal investigators of each H3Africa funded study gave permission for WGS to be used to call baseline population CNVs.

C. African Wits-INDEPTH Partnership for the Genomic Study (AWI-Gen) project

The AWI-Gen study is a longitudinal study with greater than 10 000 participants investigating the interplay between genetics and environmental risk factors for cardiometabolic diseases (Ramsay *et al.*, 2016). This study included WGS from 100 of these randomly selected healthy individuals who were all South-Eastern Bantu speakers living in South Africa (Ali *et al.*, 2018). Samples were sequenced at the Broad Institute, Cambridge, United States. Samples were prepared using a PCR-free method and sequenced on an Illumina HiSeq X Ten instrument with a minimum read depth of coverage of 30x. These data were accessed through Prof Scott Hazelhurst and Prof Zané Lombard who are members of the AWIGen study, and analysis of these data is

covered under the AWIGen study HREC ethics approval clearance number M170880 (Appendix V).

D. Southern African Human Genome Programme

15 samples from healthy Bantu speakers were sequenced at Illumina Service Center in San Diego, California. Samples were prepared using PCR-free methods for sequencing on the Illumina HiSeq 2000 instrument with a minimum read depth of coverage of 30x (Choudhury *et al.*, 2017). Permission to use these data to call baseline CNVs was requested from the Southern African Human Genome Programme (Appendix VI).

E. Simons Genome Diversity Project

The 25 samples in this cohort from sub-Saharan Africa were selected from the 300 individuals in the Simons Genome Diversity Project. These healthy individuals were from Botswana, Congo, Kenya, Namibia, Nigeria, Senegal, South Africa, Sudan. Samples were sequenced at an average depth of 43x at Illumina Ltd.; almost all samples were prepared using the same PCR-free library preparation (Mallick *et al.*, 2016).

F. 1000 Genomes Project Consortium

500 African samples were included from the expanded 1000 Genomes Project where 3202 samples sequenced with deep coverage. Samples were Gambian Mandinka, Mende from Sierra Leone, Yoruba from Ibadan, Nigeria, Esan from Nigeria and Luhya from Webuye, Kenya. All recruited individuals declared themselves to be healthy. Samples were prepared using Truseq DNA PCR-free (450bp) Library Preparation Kit and sequenced on an Illumina Novaseq 6000 sequencer to an average depth of 30x (Byrska-Bishop *et al.*, 2021).

3.2.2 Whole genome sequence alignment to Human Build 38

The 1000 Genomes Project Consortium data was downloaded already aligned to the human reference sequence (Genome Reference Consortium Human Build 38). All remaining data were aligned to the human reference sequence (Genome Reference Consortium Human Build 38) using the H3A-Varcall pipeline by members of H3A-Bionet (Mulder *et al.*, 2018), as these WGS are for use in other H3Africa projects.

There are many alternate contig regions in the Build 38 reference that must be appropriately processed during alignment to Build 38 by the alignment pipeline. There were some problems encountered at this step resulting in reduced coverage over these contig regions. This was detected in the results after the CNV calling pipelines had been executed, so to avoid erroneous CNV calls, these regions were removed from the analysis.

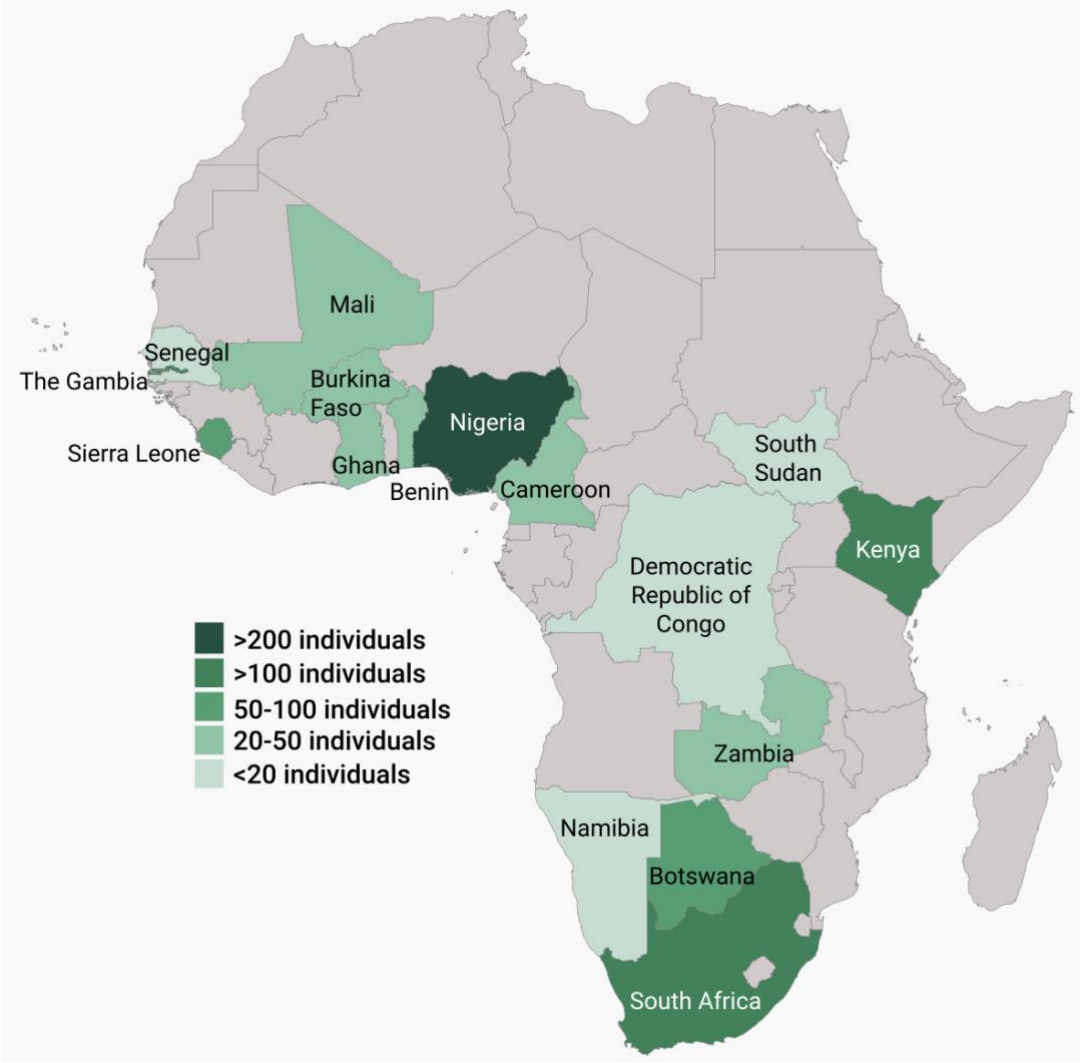


Figure 3. 1: Distribution of samples across Africa (vector map bought from https://freevectormaps.com/world-maps/africa/WRLD-AF-01-0002 and edited in figma.com)

3.2.3 Ethical clearance

Ethics for this PhD study to use these WGS to call baseline population CNVs was obtained from the University of Witwatersrand HREC with clearance number M180855 (Appendix II).

The Manta (Chen *et al.*, 2016) and Graphtyper2 (Eggertsson *et al.*, 2019) pipeline was run on all 1027 samples. The Genome STRiP pipeline (Handsaker *et al.*, 2015) was run on 919 samples. The Genome STRiP pipeline was run together with the H3A-GSK ADME Collaboration (da Rocha *et al.*, 2021) who did not have permission to use 74 of the H3A-Baylor dataset. As Genome STRiP is run jointly in all samples, these samples could not be extracted from the results after completion of the pipeline and so these samples had to be excluded from the Genome STRiP pipeline execution.

3.2.4 Copy number variant calling algorithms

Genome STRiP (Handsaker *et al.*, 2015) was chosen as it has been used widely for population scale projects, like the 1000 Genomes Project, due to its ability to detect and genotype CNVs in parallel across multiple samples. Genome STRiP is a read depth based tool able to detect deletions, duplications and multi-allelic CNV >1kb. Genome STRiP was chosen in 2018 and implemented in 2019, at which time it was the best choice given available literature. Publications in 2019 and 2020 reviewed newer tools and strongly recommended that more than one tool is used (Kosugi *et al.*, 2019), in order to overcome the weaknesses of one tool and reduce false positives. In 2020 Manta (Chen *et al.*, 2016) was chosen as a second tool because it is recommended as a reliable and sensitive tool and because it uses a different calling methodology (Kosugi *et al.*, 2019; Cameron *et al.*, 2019). Manta utilises both split reads and read pairs to identify variant sites and then uses *de novo* assembly to refine this variant (Chen *et al.*, 2016). This approach was particularly relevant to this study as African genomes are known to be highly polymorphic making *de novo* assembly preferable for detecting African variants. Additionally, Manta can detect much smaller variants (from 50bp) than Genome STRiP which would enable the characterization of a wider spectrum of CNVs. As Manta analyses samples independently from each other, Graphtyper2 (Eggertsson *et al.*, 2019) was used to genotype CNV calls identified by Manta, in parallel across the cohort. Graphtyper2 uses the sites discovered by other callers, in this case Manta, and constructs a pangenome graph. Using the graph, it locally realigns and refines the CNV site before assigning a genotype to each individual. This approach combines the advantages of the accurate single sample calling of Manta and the advantages of joint genotyping from Graphtyper2.

3.2.5 Genome STRiP

A. Implementation

Genome STRiP SVToolkit was obtained at (http://software.broadinstitute.org/software/genomestrip/download-genome-strip) and executed on the University of Witwatersrand Core Research Cluster using the DRMAA python library plug-in (Blanchard, 2018). Execution of the pipeline was performed using the queue script for SV Preprocess obtained at (http://software.broadinstitute.org/software/genomestrip/org_broadinstitute_sv_qscript_SVPreprocess.html) and CNV Discovery queue script obtained at (http://software.broadinstitute.org/software/genomestrip/org_broadinstitute_sv_qscript_CNVDiscoveryPipeline.html). Default parameters contained in these scripts were used except for memory and job thresholds which had to be increased according to the number of genomes being run. Additionally, only one chromosome at a time could be processed, due to the high memory and job requirements of the tool. The running of this pipeline was very labour intensive and computationally expensive taking several 100 000 CPU hours over several months. For this reason the pipeline could not reasonably be run multiple times with different combinations of samples or repeated after realigning of sequences.

B. Quality control of detected sites

Three problems were detected on the output of Genome STRiP. Firstly, batch effects were observed between samples that had been prepared for sequencing using PCR-based methods and those using PCR-free methods. This effect is a weakness of read depth-based tools that do multi-sample analysis and had to be overcome by running the samples in separate batches based on whether sequencing preparation was PCR-free or PCR-based. There were 99 samples for which the PCR status was uncertain and so these were run separately in a third batch to prevent creating batch effects, if they were included in the wrong PCR status group. Samples of lower quality with excessive read depth variance were observed. Read depth-based tools are particularly intolerant of this and so these samples were excluded. Thirdly we also observed splitting of CNVs into smaller overlapping or adjacent CNVs, with very similar genotypes. These limitations of Genome STRiP had been observed in two recent studies that utilised Genome STRiP (Almarri *et al.*, 2020; Jakubosky *et al.*, 2020b) and required a step to resolve these split CNVs. These quality control measures were implemented through the following pipeline.

The CNV sites detected in each batch had to be genotyped for all samples so that all discovered sites had genotypes for all samples. This was done using the Genome STRiP SVGenotyper, the queue script for which was obtained at http://software.broadinstitute.org/software/genomestrip/org_broadinstitute_sv_qscript_SVGenotyper.html. These newly genotyped sites were then annotated using Genome STRiP SVAnnotator, the queue script having been obtained at http://software.broadinstitute.org/software/genomestrip/org_broadinstitute_sv_qscript_SVAnnotator.html. These annotated VCF files from the PCR-based and PCR-free batches first had to be merged, and then a de-duplication process performed, to remove redundant sites. Where there was an exact match of breakpoints discovered in the different batches the variant with the highest quality score (GSCNQUAL) was retained and the other discarded. If there was a non-exact match the following criteria were applied to determine whether the variants were in fact the same variant detected in both batches with slightly different breakpoints. First the mode copy number for each site was calculated. To be considered likely to be the same variant, pairs of sites were required to have at least 80% reciprocal overlap, and one of the following: 1) fewer than 5% of samples with discordant non-mode copy numbers 2) have a Pearson correlation coefficient >0,95 3) Mean difference in copy number between all samples <0,2. These redundant sites were processed using a graph-based method, where sites are connected by edges, weighted according to the average percentage overlap between sites and largest GSCNQUAL score. The algorithm then iterates through the connected components in the graph and chooses the pair with the highest overlap and GSCNQUAL score. Finally, the site with the highest GSCNQUAL between that pair is chosen as the primary site and the other excluded.

For the stitching process the algorithm iterates through all pairs of adjacent variants and computes a Pearson correlation coefficient, mean difference in copy number between all samples and between non-mode samples and the discordance been non-mode copy numbers (having first calculated a mode copy number for each site). Pairs of sites had to pass the following to be considered for stitching: 1) a Pearson correlation coefficient >0,9, 2) a percent difference of <0,2% between all samples and <0,5% between all non-mode samples as well as being <30kb apart. Sites that passed these criteria were stitched together. If multiple adjacent sites pass, they were all stitched together, until a pair of sites that do not pass correlation thresholds is encountered.

These new stitched sites were re-genotyped and annotated using SVGenotyper and SVAnnotator.

Finally, a correlation was calculated between the average of the original genotypes of the constituent sites, and the genotypes of the stitched sites. This was done to establish if the new stitched sites had highly similar genotypes to the constituent sites. If the correlation between stitched sites and their constituent sites was <0.9 these stitched sites were discarded and the original individual constituent sites retained. A final VCF file was then produced from a matrix containing all sites that showed a pass or fail to indicate if the stitch site should replace its constituent site. This was done using a custom script (Available at https://github.com/emmakwiener/PhD-CNV-Analysis). Lastly variants overlapping alternate contig regions were removed due to the problems that arose during the realignment to Build 38.

3.2.6 Manta and Graphytyper2

Manta v1.6 (https://github.com/Illumina/manta) was executed in single sample mode for each of the 1027 samples using the default parameters. The individual VCF files were merged using the tool svimmer (https://github.com/DecodeGenetics/svimmer) with default parameters. This merged VCF file was then genotyped using Graphtyper v2.6.1 (https://github.com/DecodeGenetics/graphtyper). The algorithm uses two different models, i.e. 'break point' and 'coverage' to genotype as well as an 'aggregated' model that combines both approaches. The 'coverage' and 'breakpoint' models were filtered out from the VCF file, retaining only the aggregated genotypes. Only variants with the filter tag 'PASS' were retained excluding variants of low quality. The default threshold is to exclude variants with a normalised quality score (QD) <9, but Eggertsson *et al.* (2019) applied slightly stricter criteria for deletions excluding all variants with a QD<12. To ensure high quality variants we applied this same threshold. Variants >10mb were also filtered out as variants larger than this are considered of questionable accuracy by this method of detection. Insertions, inversions and translocations were excluded, and deletions and duplications were retained, as the analysis was focused on CNVs. A Hardy-Weinberg equilibrium analysis was performed on the variants including calculating excessive heterozygosity. Variants with excessive heterozygosity <1 $x10^{-15}$ were excluded. Genotypes were set to missing if their genotype quality scores GQ <20, if tagged 'FAIL1' for deletions and duplications and 'FAIL2' or 'FAIL3' for duplications. After setting missing values for low quality

genotypes, allele frequency was recalculated. Variants which now had AF=0 were then excluded. Finally, variants overlapping alternate contig regions were removed, due to the problems encountered in realigning to Build 38. The above filtering was performed using bcftools v1.12 (Danecek *et al.*, 2021), vcftools v0.1.16 (Danecek *et al.*, 2011) and some custom scripts (Custom scripts available at https://github.com/emmakwiener/PhD-CNV-Analysis).

3.2.7 Number of copy number variants per sample

The number of CNVs per sample was assessed only for samples sequenced using the PCR-free method. This was done as it has been shown that PCR-based samples are less reliable in terms of variant quality and sensitivity. For example, the gnomAD-SV study excluded PCR-based samples from variant per sample calculations, stating that PCR-based samples showed inconsistent quality and SV sensitivity (Collins *et al.*, 2020). This analysis was done for both Manta and Genome STRiP datasets using Genome STRiP SVAnnotator, Variants per Sample function.

3.2.8 Intersection of Manta and Genome STRiP datasets

BEDTools *intersect* (Quinlan and Hall, 2010) was used to find the intersection between the Manta and Genome STRiP datasets with each other. Reciprocal overlap of 80% was used to determine if the same variant sites had been detected by both tools. This percentage was also used in the Genome STRiP redundancy-collapsing pipeline to assess if the same variant had been found in the different batches of Genome STRiP. We allowed mismatch of CNV type given that Genome STRiP detects multi-allelic CNVs that may have been detected by Manta as either a deletion or duplication.

The BEDTools *intersect* output was used to determine which CNVs were found by both tools. The Manta breakpoints were kept as Manta has been shown to have more accurate breakpoint calling. Where a Manta deletion or duplication overlapped with a Genome STRiP multi-allelic CNV, the multi allelic CNV and its coordinates was kept for that variant site, to avoid losing the multi-allelic information provided by Genome STRiP. Additionally where a single variant overlapped multiple variants called by the other tool, the single variant was retained to reduce redundancy in the intersecting call set.

3.2.9 Comparison to established databases

For comparison of the variant call sets to established databases we chose the DGV because it contains all variants from the 1000 Genomes consortium (phase 1 2 and 3) and the gnomAD-SV study (Collins *et al.*, 2020) as well as many other smaller studies all aligned to Human Genome Build 38. The variant file was downloaded from (http://dgv.tcag.ca/dgv/app/downloads?ref=GRCh38/hg38). Manta and Genome STRiP variant call sets were converted to BED format and then BEDTools *intersect* (Quinlan and Hall, 2010) was used to compare them to the variant file from the DGV. A reciprocal overlap of at least 50% was used in this comparison to try and allow for the fact that variants found in older studies may have differing breakpoints but be similar variants. Additionally this percentage overlap was used in the gnomAD-SV comparison when they compared their variant call set to previously published databases (Collins *et al.*, 2020).

The intersection DD gene variant call set was also compared to the ClinVar Database (Landrum *et al.*, 2018). The latest Build 38 variant file was downloaded from (https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh38/). The variant file was processed to retain only deletions and duplications using bcftools and vcftools as well as a custom script used to extract the relevant fields into a BED format. BEDTools *intersect* was then used to compare the ClinVar and the combined DD gene variant set. A reciprocal overlap of at least 50% was used to allow for differing breakpoints of CNVs.

3.2.10 Functional analyses

To perform a location and functional impact assessment the relevant Manta and Genome STRiP VCF files were analysed with the command line version of Ensembl Variant Effect Predictor (McLaren *et al.*, 2016) software (version 103).

3.2.11 Principal component analysis

We performed a PCA on the Manta variants to assess what the most influential factor is affecting the difference of CNVs. Genome STRiP variants could not be included since they do not have individually resolved genotypes. For the purpose of this analysis an increased quality filter was performed and only sites with normalised quality score QD >15 were retained. By restricting the dataset to higher quality calls batch effects can be minimized where calls were affected by artefactual differences. This practice

was performed before principal components were analysed in the gnomAD-SV study (Collins *et al.*, 2020) and another study similar to this study (Almarri *et al.*, 2020). The PCA was performed on the higher quality dataset. PLINK and PLINK2 (Chang *et al.*, 2015) were used to convert the data to BED format, prune sites in linkage disequilibrium and finally to perform the PCA. Genesis (Buchman and Hazelhurst, 2015) was used for PCA visualization.

3.2.12 Variants overlapping developmental disorder genes

A list of 2101 genes known to contain variants that cause DD was downloaded from Gene2Phenotype https://www.ebi.ac.uk/gene2phenotype/downloads) on 14/10/2020. The coordinates of these genes were downloaded from Ensembl Biomart (version 103) (http://www.ensembl.org/biomart/martview). Using the list of coordinates CNVs that overlap the set of genes were extracted from the Manta and Genome STRiP VCF files using bcftools view -R. Files were converted to BED format and BEDtools *intersect* was used to find a set of variants common to both with a reciprocal overlap of 80%. This higher threshold was used to attain variants that were close to identical, as was used in the Genome STRiP redundancy-collapsing pipeline. Manta breakpoints were kept as Manta has been shown to have more accurate breakpoint calling but where a Manta deletion or duplication overlapped with a Genome STRiP multi-allelic CNV, the multi-allelic CNV and its coordinates was kept for that variant site to avoid losing the multi-allelic information provided by Genome STRiP.

3.2.13 Regional comparison of bi-allelic developmental disorder gene variants

The 1027 individuals were divided based on their country of origin into east, west, central or southern African regions using the United Nations region classification (United Nations Statistics Division). The bi-allelic DD gene VCF file was used to extract CNVs present in the individuals of each region. The allele frequency for each region for all CNVs was calculated using vcftools -freq. A combined matrix with frequencies for each CNV, for each of the regions, was then processed to convert the frequency of each CNV to a binary format either present =1 or absent =0 for that region. This matrix was then used in RStudio (RStudio Team, 2020) to produce an upset plot. Chi-square tests were performed using SPSS in RStudio to do sample sample-weighted comparisons.

3.2.14 Pathogenicity prediction

Variants overlapping with DD associated genes were analysed using the tool ClassifyCNV (https://github.com/Genotek/ClassifyCNV) (Gurbich and Ilinsky, 2020) to perform estimation of pathogenicity according to the latest ACMG guidelines for the interpretation and reporting of constitutional CNVs (Riggs *et al.*, 2020). ClassifyCNV can only calculate scores for deletion and duplication CNV types, so the multi-allelic CNVs (n=25) with both deletion and duplication alleles were classified as both types.

3.3 Results

3.3.1 Final number of samples included

A different number of final samples were run for the Manta and Genome STRiP pipelines. Figure 3.2 a and b shows the final numbers of samples for both pipelines as well as the reasons why, and points at which samples were excluded.

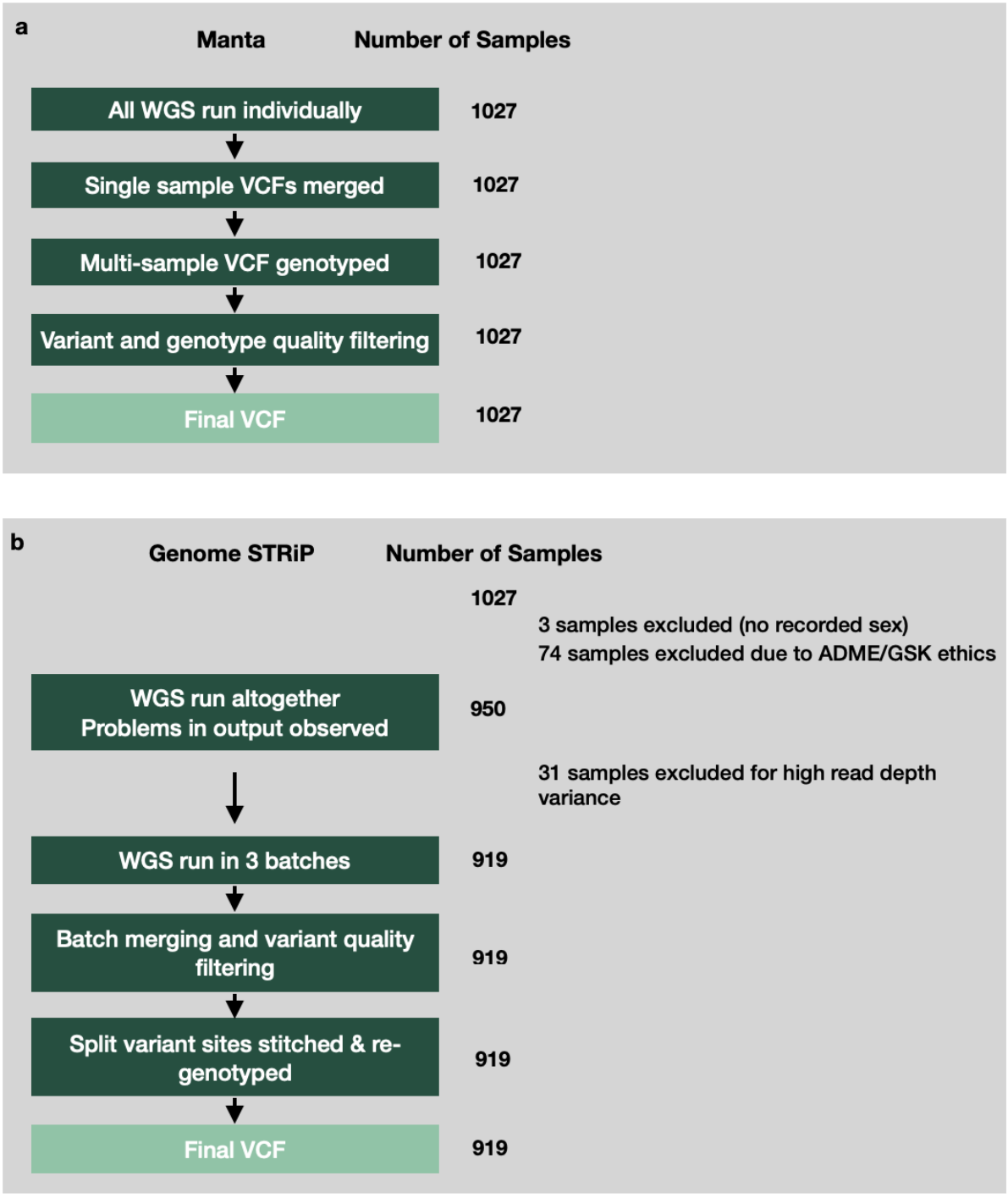


Figure 3. 2: Final number of samples included for the Manta (a) and Genome STRiP (b) pipelines

3.3.2 Overview of the results section

In the results we will focus first on the entire call sets produced by Manta and Genome STRiP to see the scope and profile of variants found by each tool. Secondly a intersecting set of variants that were called by both tools will be assessed. Lastly, we will focus on a subset of these variants that overlap genes known to be associated with

DD. The reasons for focusing on the intersecting variants and the DD gene subset will be explained at the beginning of those sections 3.3.4 and 3.3.5 respectively.

3.3.3 Entire Manta and Genome STRiP call sets

A. Variant type breakdown of each entire call set

The final Manta call set consisted of 54 810 variants, containing 43 153 deletions and 11 657 duplications. The median number of variants per sample was 11 106 (range: 10 301–12 862). This number is much higher than that seen in the gnomAD-SV where the median number for African ancestry individuals was 8 755. If just deletions and duplications are considered from the gnomAD-SV study, we see a ratio of 4:1 deletions to duplications which is similar to the ratio observed in the Manta results in this study.

The final Genome STRiP call set consisted of 15 658 variants, with 10 232 deletions, 1 810 duplications and 3 616 multi-allelic CNVs. The median number of variants per sample for the final curated Genome STRiP data set was 1 888 variants (range:1 749–2 152). This number is similar to the number to CNVs per sample reported by the 1000 Genomes project, which was 1 270 for African ancestry individuals, for which the same CNV calling pipeline was used (Auton *et al.*, 2015). Looking at the proportions of deletions, duplications and multi-allelic CNVs we see more multi-allelic CNVs than other studies that used Genome STRiP (Almarri *et al.*, 2020; Jakubosky *et al.*, 2020a). A summary of the CNVs detected by each tool is shown in Table 3.1.

Table 3. 1: Breakdown of variant types detected by each tool

CNV Type	**Manta** (n)	(%)	**Genome STRiP** (n)	(%)
Deletion	43 157	78,7	10 232	65,3
Duplication	11 657	21,3	1 810	11,6
Multi-allelic CNV	NA	NA	3 616	23,1
All	**54 814**		**15 658**	

B. Size profile of each entire call set

The Manta and Genome STRiP datasets have overlapping but different size ranges. Manta detects variants between 50bp and ~10Mb and Genome STRiP between ~1kb

to ~1Mb.This size range difference means the size profiles of the datasets from each tool differ considerably.

The median size of variant for the Manta dataset was 436bp, a number similar to the median size of variants found in the gnomAD-SV study, the recent landmark study that characterised the full size spectrum of SVs using Manta and other similar tools (Collins *et al.*, 2020). Most variants called by Manta were small; 90% of variants were <10 000 bp and 60% <1 000bp which concurs with the fact that the median size of variant was 436bp. Due to the ability of Manta to detect variants 50–100bp, 14 252 variants in this size class were detected in this study. This group of variants have not previously been well characterized, especially in African genomes, and we found that 51% of the variants detected in this size range were novel compared to only 31% of variants >100bp.

The median size of variant detected by Genome STRiP was 3 900bp. This is much greater than the median obtained from the Manta call set owing to the considerable difference in range of variant size detected by each tool. This can be seen in Figure 3.3 that shows both call sets divided into CNV size classes. 78% of the Genome STRiP variants were between 1kb–10kb in size, but the largest number of Manta variants were 100–1 000bp in size.

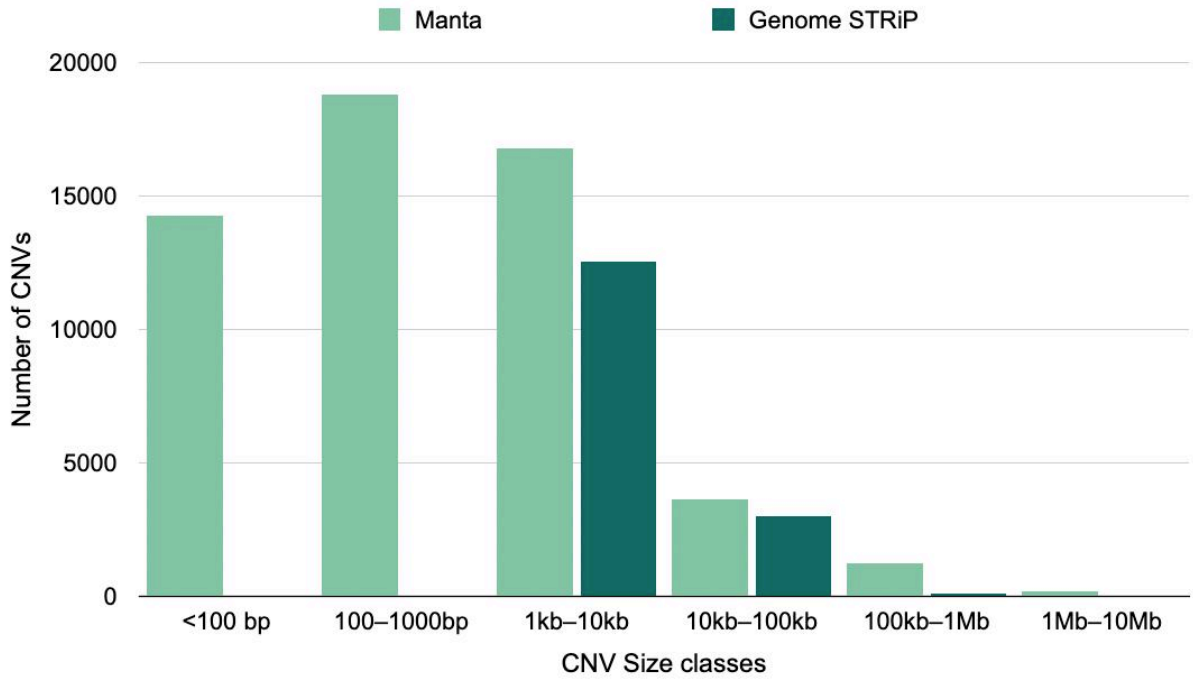


Figure 3. 3: Number of variants called by Manta and Genome STRiP in each of the size classes. Greatest number of Manta variants were between 100–1 000bp in size and the greatest number of Genome STRiP variants were 1kb–10kb in size.

The greatest number of duplications (43%) detected by Manta are <100bp while deletions have a higher number of variants between 100–1 000bp than <100bp (Figure 3.4a). Duplications detected by Genome STRiP tended to be larger than other classes of variants with the greatest number of duplications falling between 10 000bp and 100 000bp. This is seen in Figure 3.4b where the duplication plot is wider at that size range compared to the other classes of CNVs.

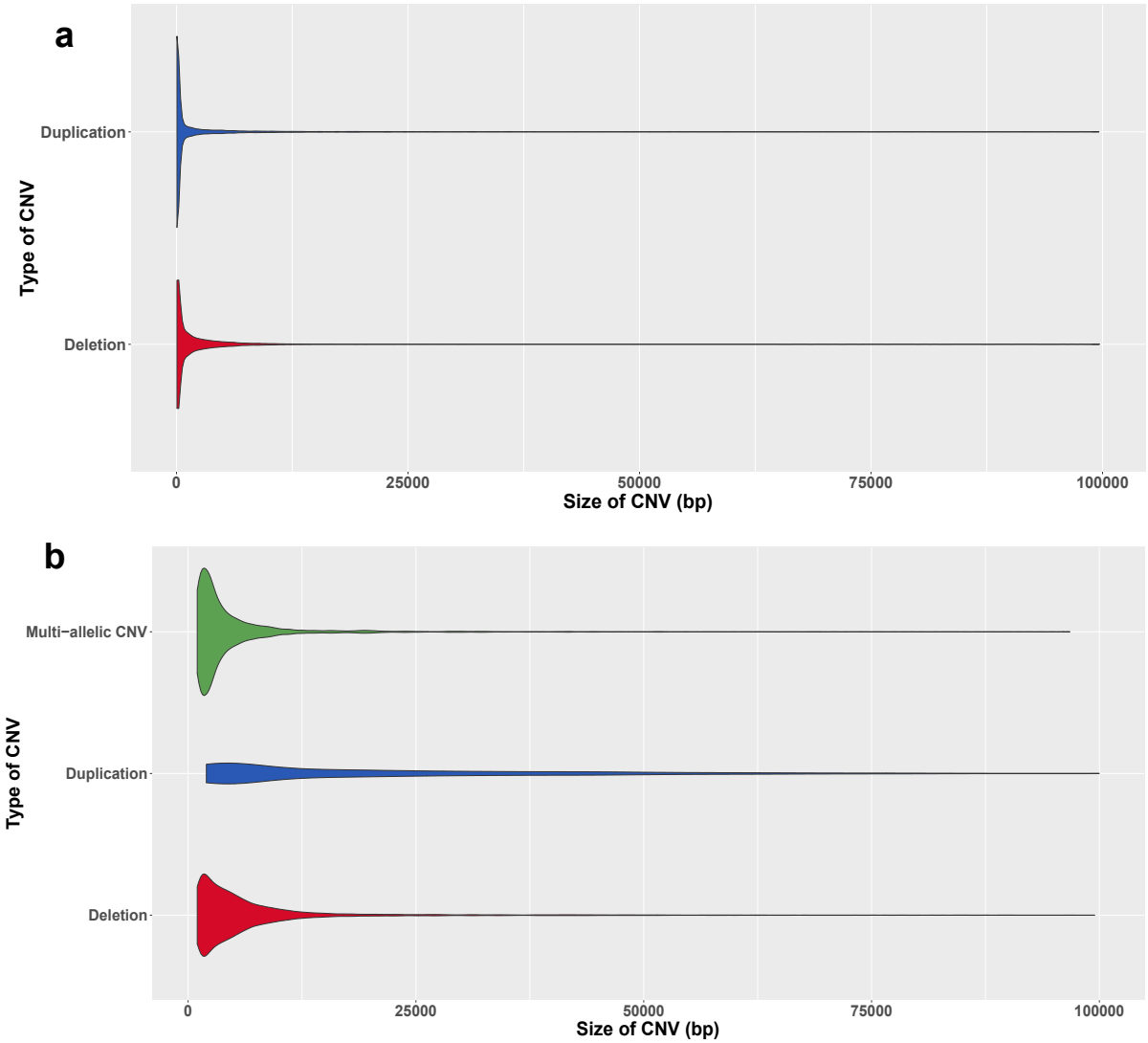


Figure 3. 4: Violin plot of CNV size distribution a) Manta variants <100 000bp according to variant type. Duplications and deletions are all bi-allelic. b) Genome STRiP variants <100 000 bp. Multi-allelic CNV are variants that range from deletion to duplication alleles with more than 3 copies. Duplication and deletion classes are bi-allelic sites with duplication or deletion alleles. For a) and b) Variants >100 000bp were excluded to enable visualization of the core set of variants.

C. Variant allele frequency profile of each entire call set

The variant allele frequency (VAF) profiles for the two tools are also quite different. The first reason for this is a difference in tool capability; Manta is able to detect lower VAF variants than Genome STRiP. Genome STRiP analyses samples in parallel, to increase the validity of sites found in multiple samples, so it only detects variants with a VAF >0,1%. Manta however calls individually on each sample and so is able to detect ultra-rare variants with VAF<0,1%. VAF can be defined in this study as the frequency of individuals with a copy number different from the reference

In the Manta call set 10 817 (19,7%) variants were singleton variants; a rate much lower than that seen by studies such as gnomAD-SV where 48,9% of variants were singletons. VAF was assessed for the Manta deletions and duplications shown in Figure 3.5. For the deletions there are more ultra-rare deletions (VAF <0,1%) than rare deletions (VAF 0,1–1%) where for duplications there are more rare duplications than ultra-rare duplications. In both variant types however, we see that the largest group of variants were common variants (VAF >5%) with 32% of deletions and 51% of duplications being common.

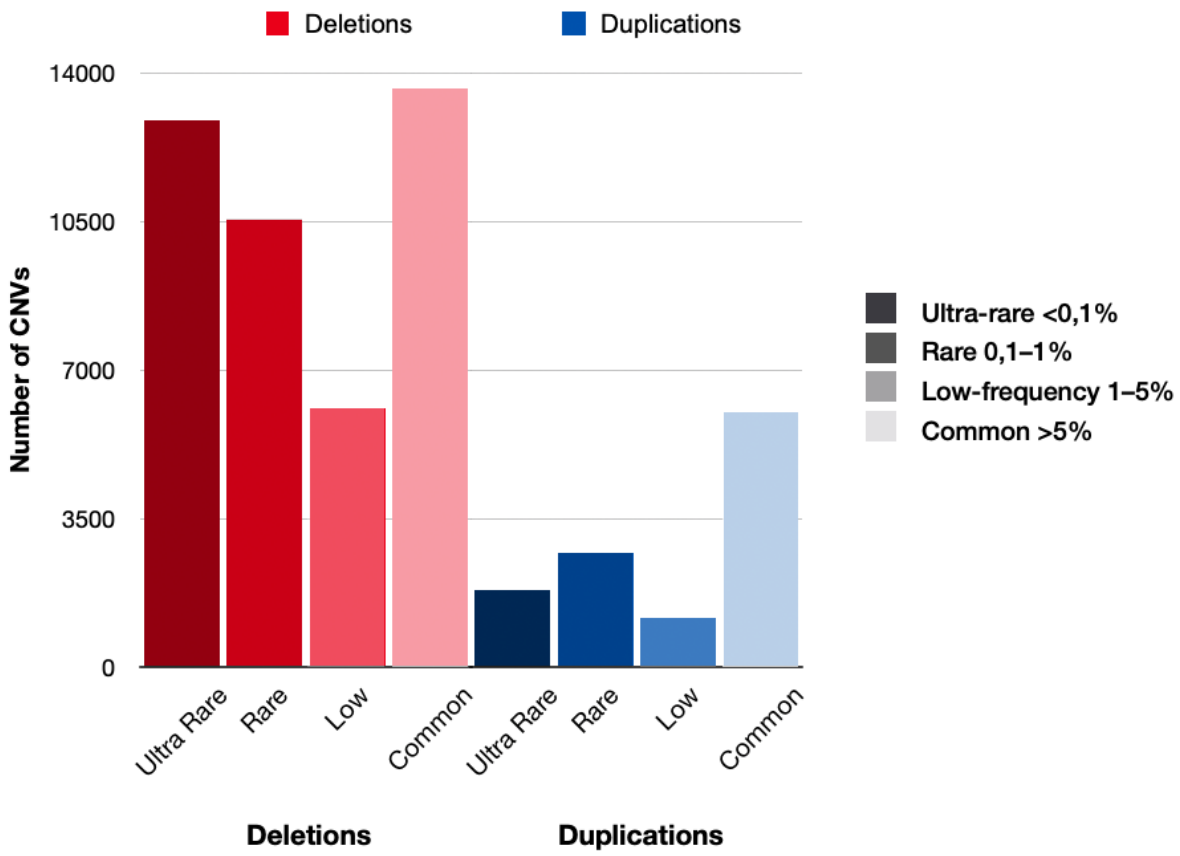


Figure 3. 5: Number of CNVs called by Manta in four VAF classes. Ultra-rare variants (<0,1%), Rare variants (0,1–1%), Low frequency variants (1–5%), Common variants (>5%). Greatest number of Manta deletions and duplications were common.

Figure 3.6 shows that Genome STRiP deletions and duplications have fewer common variants than rare and low frequency variants with a particular depletion in common duplications. Multi-allelic CNVs however have an increase in common variants compared to rare and low frequency variants.

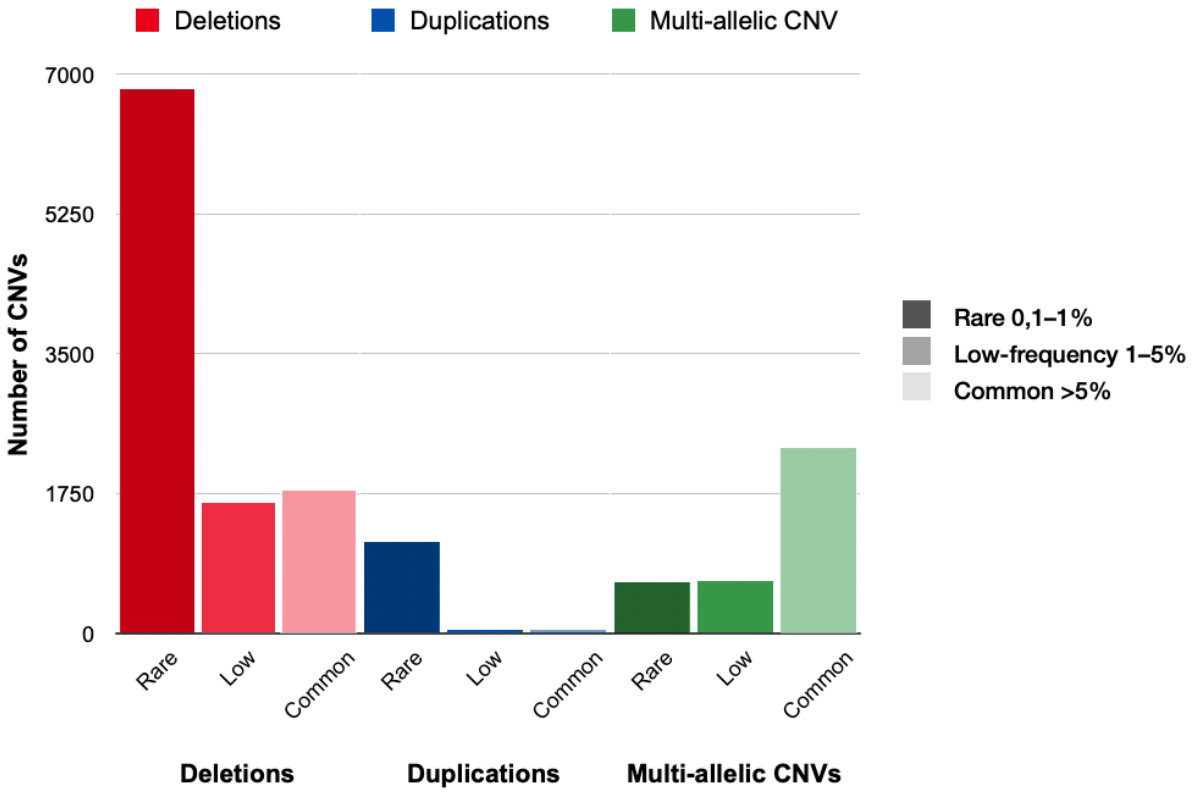


Figure 3. 6: Number of CNVs called by Genome STRiP in three VAF classes. Rare variants (0,1–1%), Low frequency variants (1–5%), Common variants (>5%).

A further breakdown of VAF in Figure 3.7 for both Manta and Genome STRiP shows a similar pattern in both call sets with most variants having frequencies <10%, and the number of variants per frequency bin decreasing to VAF of 50% and then increasing slightly per VAF bin up to 90–100%. In Figure 3.8 we see that the most variants occur at low VAF and that most variants found at high frequencies were small variants, with the exception of a few large variants >1Mb found at very high frequencies. We also see that all variants >1,25Mb are deletions.

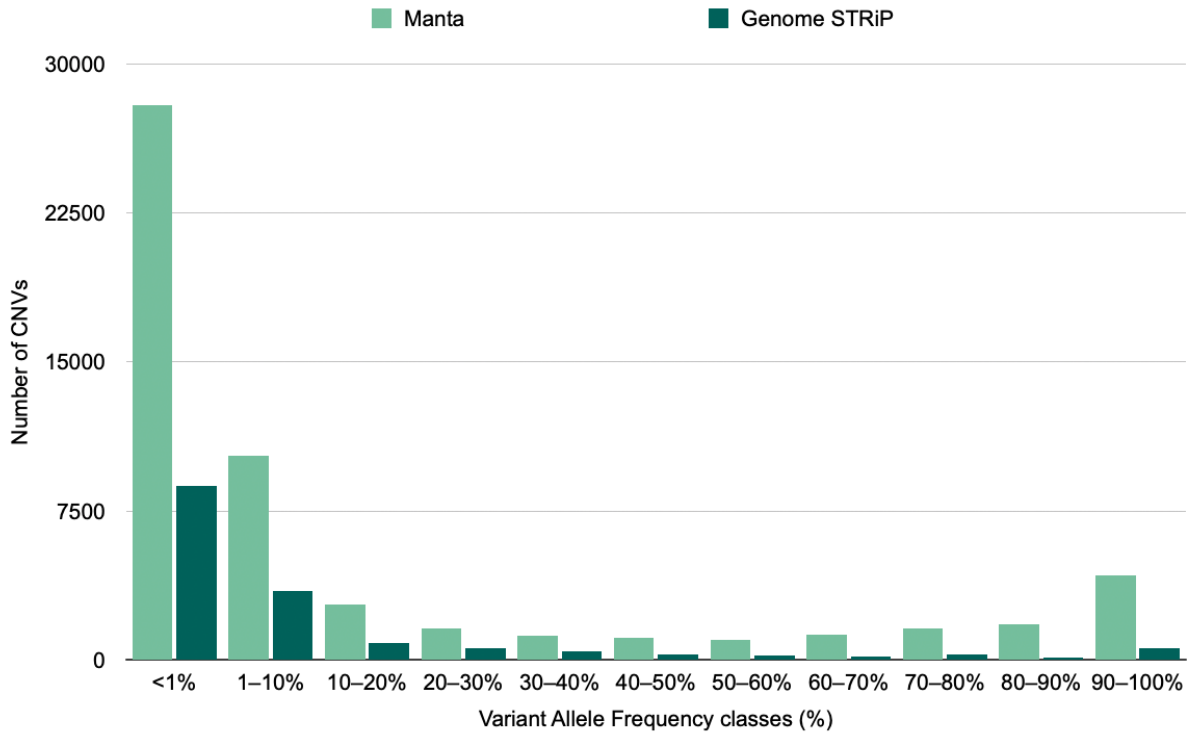


Figure 3. 7: Number of CNVs per VAF class called by Manta and Genome STRiP. The greatest number of variants in both Manta and Genome STRiP had allele frequencies <10%. Lower numbers of variants up to at an allele frequency of 50%, and then a slight increase in the number of variants at an allele frequency of 90–100%.

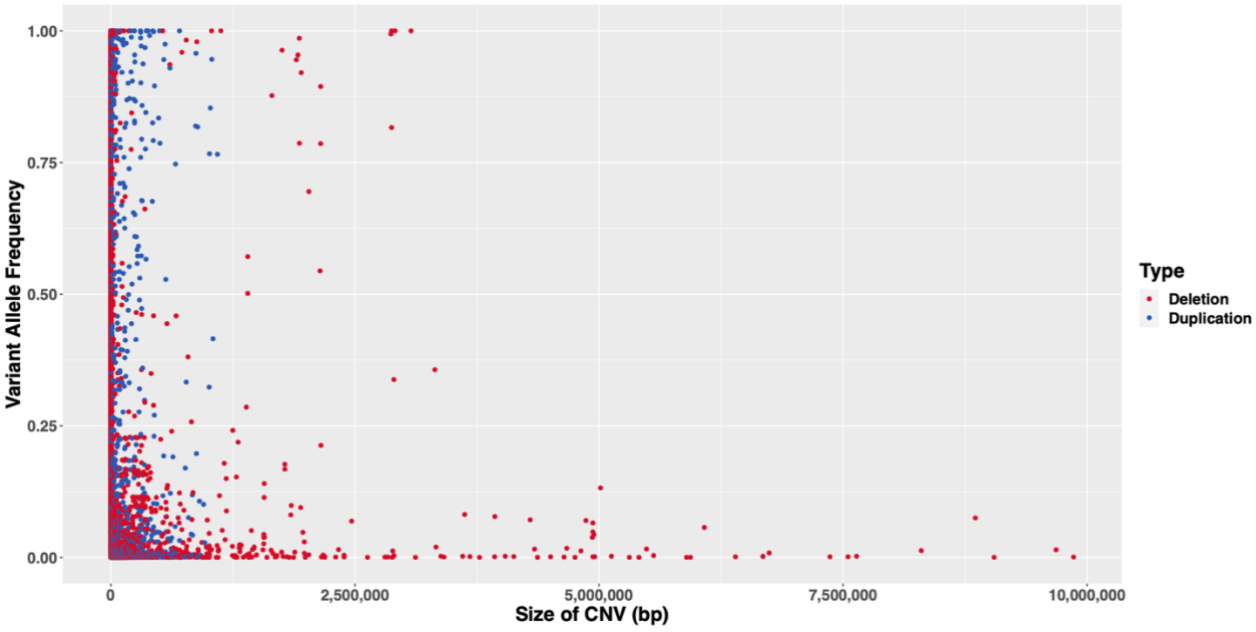


Figure 3. 8: Manta variants plotted by VAF and size (bp). The majority of variants were <1Mb and found at VAF <0,25. most variants found at higher allele frequencies were small, but a few large variants >1Mb were found at high VAF.

D. Principal component analysis of the Manta call set reveals batch effects and some regional differences

A PCA was performed on the Manta call set to detect differences in the CNVs found in all the individuals. A high-quality threshold was used on this data to reduce batch effects before performing the PCA. On visualization, samples were labelled by sequencing method, region and project to detect any batch effects that may be affecting the separation and clustering of individuals. In Figure 3.9, samples are labelled according to project, and when principal components one and two are assessed, samples cluster strongly according to which project they are from. Despite stringent quality measures there are some significant batch effects in this data that can not be removed. In Figure 3.10, where samples are labelled according to geographical region, principal components 1 and 2 yield some separation of east and west African samples and a greater degree of separation of southern African samples, especially in the H3A-Baylor samples. Given the batch effects seen in Figure 3.9 and noted separation observed in the H3A-Baylor samples in Figure 3.10, the PCA was performed for only the H3A-Baylor project data and is shown Figure 3.11. In this plot we do see differences according to African regions for east, west and southern Africa. We do not however see separation of the central African samples from other regions.

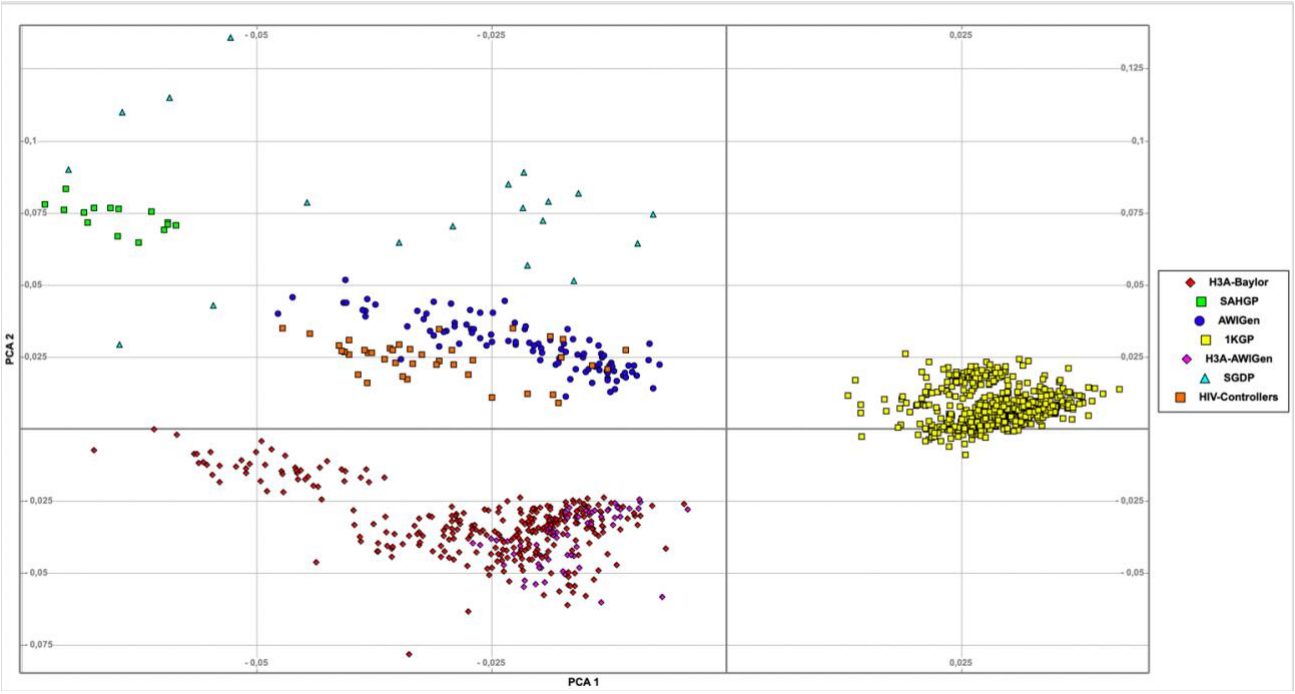


Figure 3. 9: PCA of Manta variants showing batch effects. LD pruned variants were used for PCA and PCA was visualized in Genesis. Samples colour coded according to which project each sample is from.

Projects in order Human Heredity and Health in Africa (H3A-Baylor), South African Human Genome Programme (SAHGP), African Wits-INDEPTH Partnership for the Genomic Study (AWIGen), 1000 Genomes Project (1KGP), H3A-AWIGen samples, Simons Genome Diversity Project (SGDP), HIV Elite Controller (HIV Controllers)

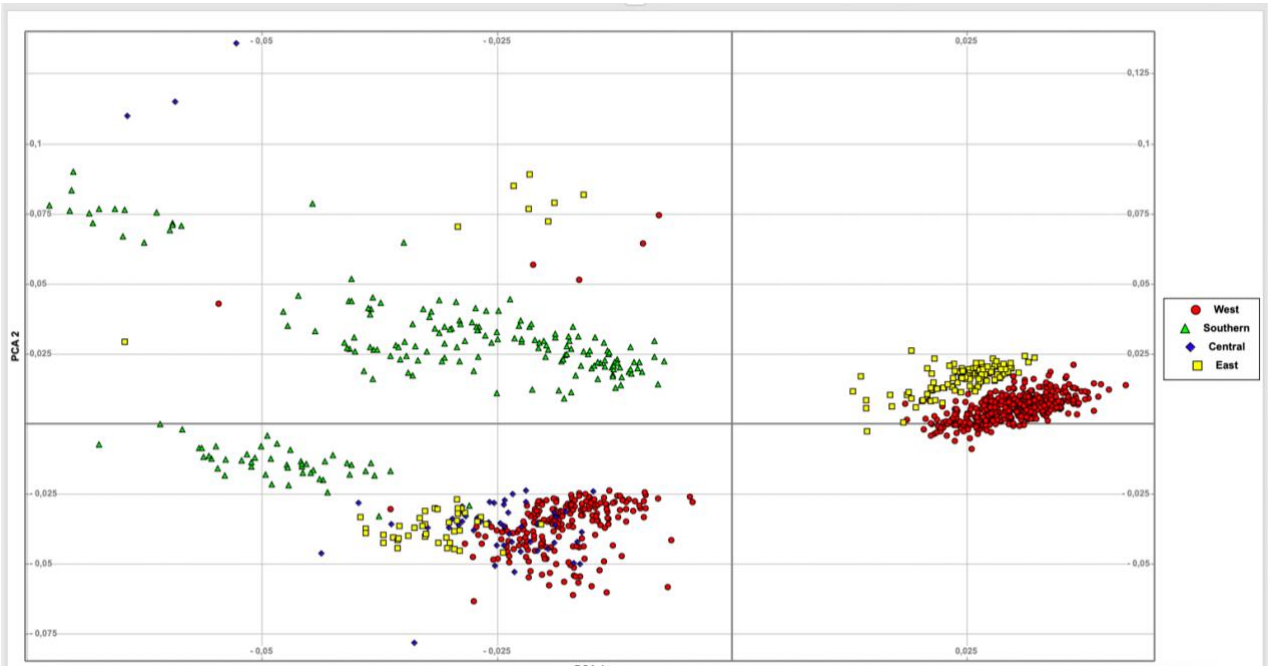


Figure 3. 10: PCA of all Manta variants showing some regional differences. LD pruned variants were used for PCA and PCA was visualized in Genesis. Samples colour coded according to region.

West Africa (West), Southern Africa (Southern), Central Africa (Central), East Africa (East).

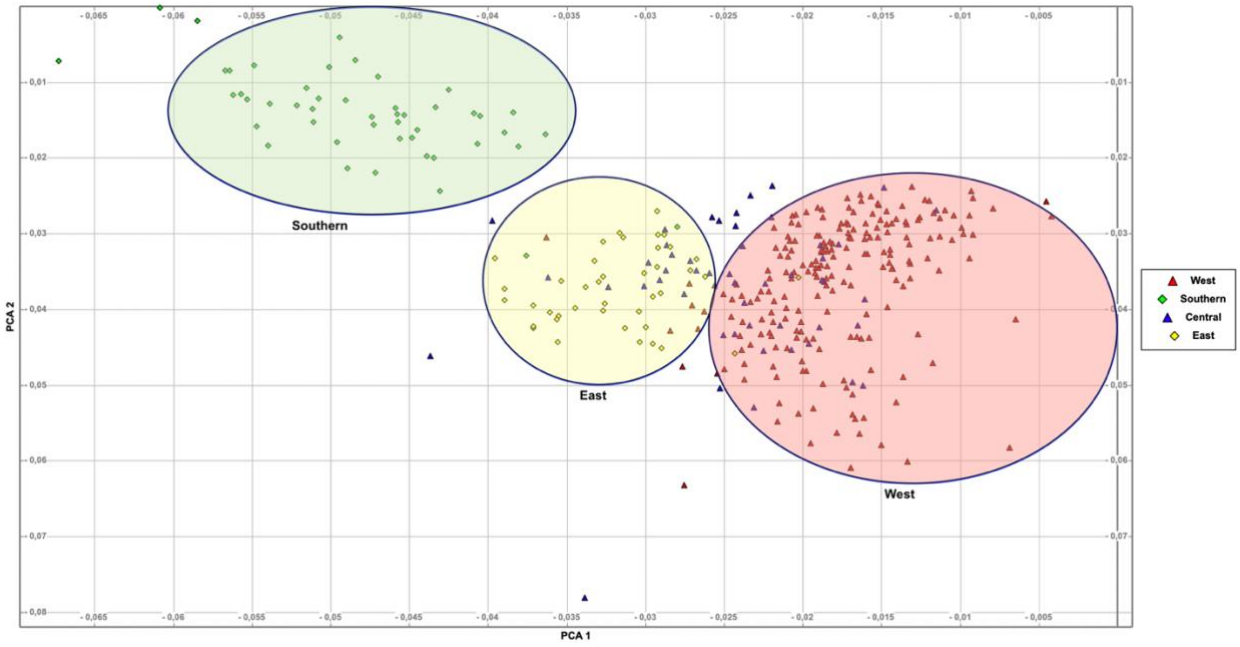


Figure 3. 11: PCA of H3Africa data showing regional differences between east, west and southern Africa. LD pruned variants were used for PCA and PCA was visualized in Genesis. Samples colour coded according to region.

3.3.4 Intersection of Manta and Genome STRiP call sets

A. Rationale

It is strongly recommended in CNV discovery that multiple CNV calling algorithms are used in combination (Kosugi *et al.*, 2019; Cameron *et al.*, 2019). Using tools with different calling methodologies allows one tool to overcome the weaknesses of the other tool. If the same variants are detected by tools utilising different methodologies it decreases the likelihood that the variant is a false positive. Additionally in this study, the analysis of the Manta dataset revealed that its size and VAF profiles did not align with expected parameters for population CNVs (Collins *et al.*, 2020). For both of these reasons we chose to focus further analyses on the variants detected by both Manta and Genome STRiP.

B. Variant type breakdown of the intersection call set

As mentioned earlier the two datasets differ considerably with regards to their range of CNV size. Of the Manta dataset 33 208 variants (those <1kb and those >1Mb) were outside of the range Genome STRiP can detect. This means only 21 606 of the Manta variants could be detected by Genome STRiP. The intersection between these comparable datasets was assessed using a reciprocal overlap of 80% to assess whether the same variant had been detected by both tools. After removing redundant and mismatching variant types, 9 001 non-redundant variants were found shared between the two datasets, as shown in Figure 3.12.

This intersection dataset is composed of 7 613 deletions, 525 duplications and 863 multi-allelic CNVs. The variant dataset is available at (https://github.com/emmakwiener/PhD-CNV-Analysis). This intersection variant dataset of CNVs identified by both tools will be used for further analyses.

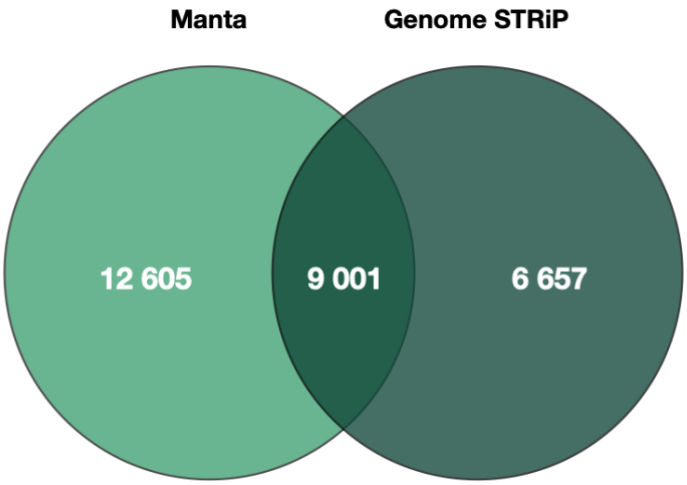


Figure 3. 12: Intersection of the comparable Manta and Genome STRiP variants. 9 001 non-redundant variants.

C. Size profile of the intersection call set

The size profile of the intersection call set in Figure 3.13 shows that the majority of variants are <15kb. The largest variants are deletions, but there are a greater number of duplications between 25-60kb than there are for multi-allelic CNVs or deletions.

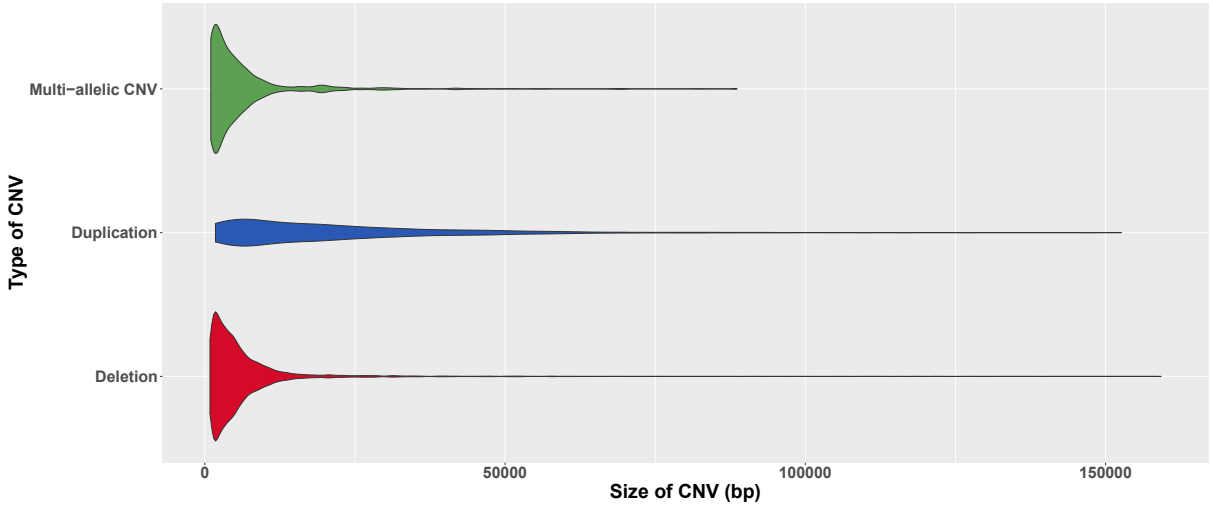


Figure 3. 13: Violin plot of CNV size distribution of intersection variant set. Multi-allelic CNVs are variants that range from deletion to duplication alleles with more than 3 copies. Duplication and deletion classes are bi-allelic sites with duplication or deletion alleles.

D. Variant allele frequency profile of the intersection call set

When the intersection variant set is divided according to VAF classes, we see that the greatest number of deletions and duplications are ultra-rare decreasing for rare, low

frequency and common variants (Figure 3.14). We see a particular depletion of common duplications and an increase in common multi-allelic CNVs. This pattern was observed in the 1000 Genomes Consortium SV study and indicates that common duplications do not remain bi-allelic but tend to become multi-allelic (Sudmant *et al.*, 2015b).

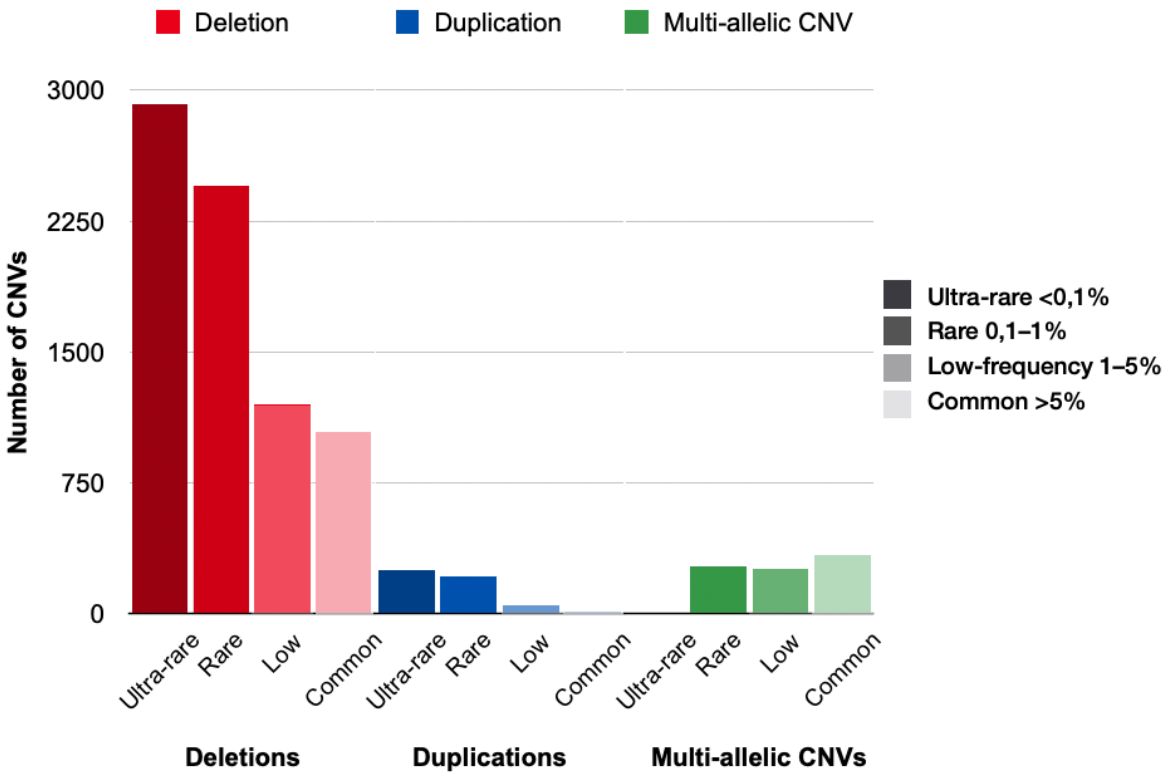


Figure 3. 14: Number of CNVs in intersection variant set in each of the four VAF classes. Ultra-rare variants (<0,1%), Rare variants (0,1–1%), Low frequency variants (1–5%), Common variants (>5%). The fewest deletions and duplications are common but the greatest number of multi-allelic CNVs are common.

E. Distribution of copy number variants across the genome

Figure 3.15 shows the chromosome visualisation of the intersection variant set. We see an increase in the concentration of CNVs in the telomeric and centromeric regions of multiple chromosomes which is expected as it is known that these regions are particularly susceptible to CNV formation. A comparison of the highest peaks to SV clusters identified in the 1000 Genomes Consortium revealed a number of matching locations of SV hotspots (Sudmant *et al.*, 2015b). We also see that there is a high concentration of multi-allelic CNVs in the p arms of chromosomes 11 and 20 and the terminal q arm of chromosome 3.

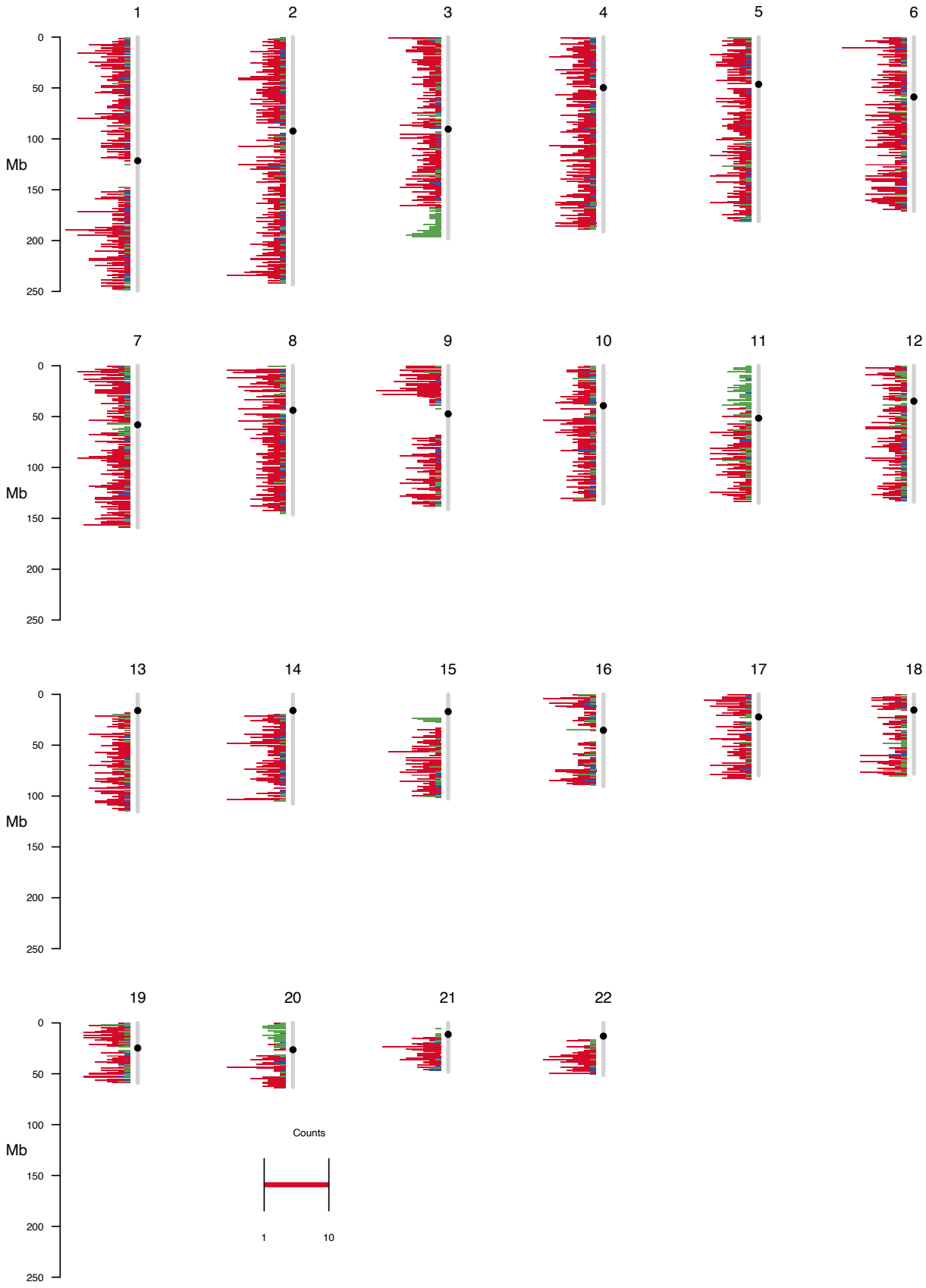


Figure 3. 15: Genomic representation of CNV density of intersection Manta and Genome STRiP variant set. Concentration of CNVs in telomeres of chromosomes 3, 7, 8, 14, 16, 18 and concentrations of multiallelic CNVs in p arms of chromosomes 11 and 20 and telomeric region of chromosome 3.

F. Functional analysis of intersection call set

Assessment of the location and functional implications of CNVs in the intersection call set is shown in Figure 3.16. The largest percentage of variants (49%) affected partial genes, 37% were intergenic, 8% near genes and 6% affected entire genes. Looking at the variants that affected partial genes 36% of all CNVs affected only introns, 8% affected regulatory sequences and 4% affected exons. In Figure 3.17 variants are divided into VAF classes, and we see the percentage of variants that overlap protein coding transcripts (whole gene duplications & deletions and exon variants) decreases with increasing VAF. Of the ultra-rare variants 18,4% of the variants overlapped a protein coding transcript, which decreases to 8,6% in rare variants, and 4,1% in low frequency variants and finally 3,3% in common variants.

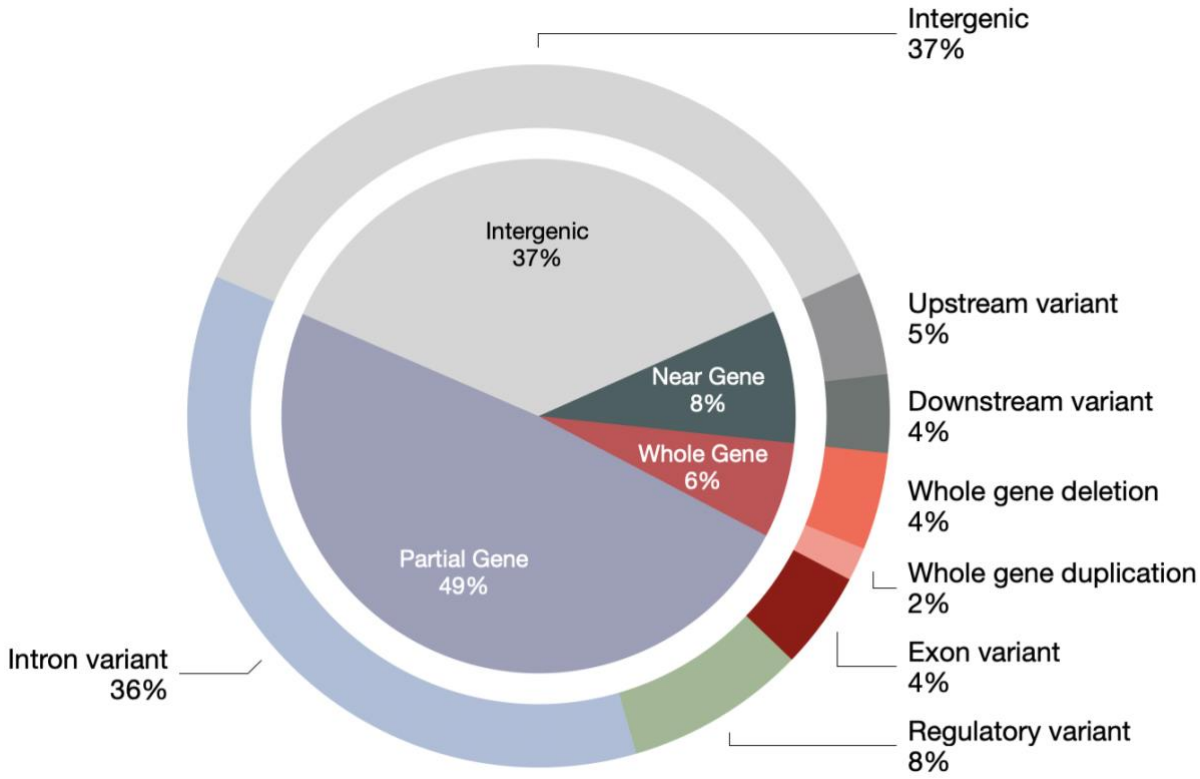


Figure 3. 16: Functional analysis of intersection variant dataset. 10% of all variants in the combined dataset overlapped exons or entire genes.

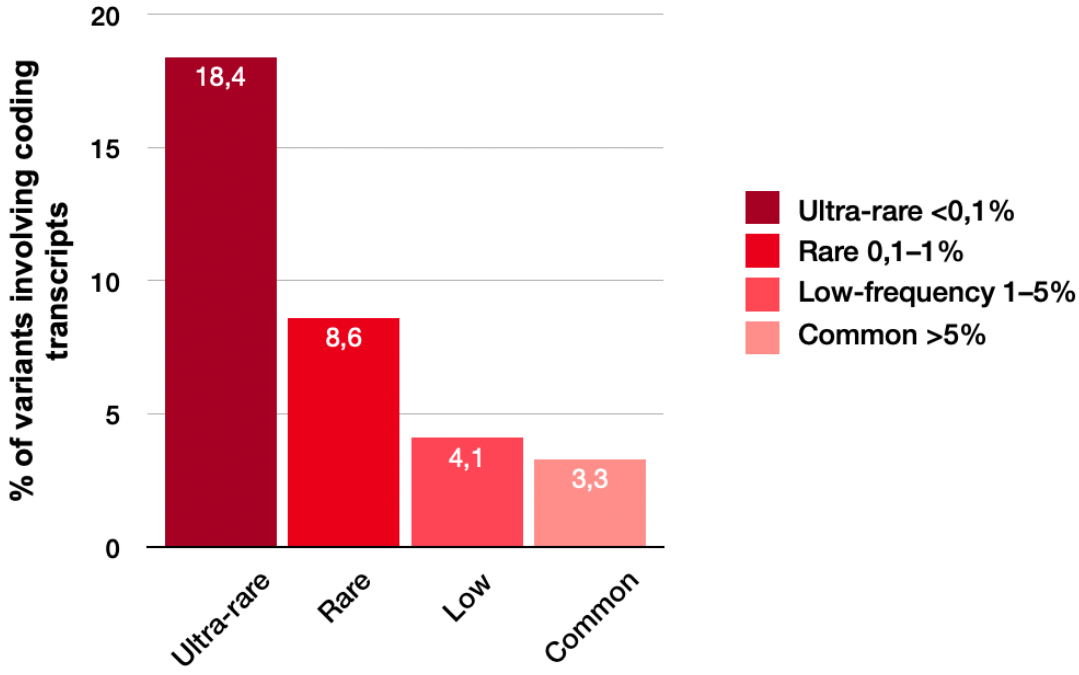


Figure 3. 17: Percentage of variants involving coding transcripts for each VAF class. Decrease in percentage of variants involving whole genes or exons as VAF increases.

G. Novel African variants

When compared to the DGV database, 1 549 (17%) novel variants were identified in the intersection variant set in this African cohort. The remaining 7 452 (83%) variants in the intersection variant set had been found by previous studies. The set of novel variants consisted of 1 356 deletions, 105 duplications and 88 multi-allelic CNVs.

Figure 3.18 shows a genomic representation of the density of these variants. In this plot we see that the novel variants are distributed across the genome with a few noted high-density areas. A large peak of deletions on the q arm of chromosome 8 near the centromere, another peak of deletions on the q arm of chromosome 1, and lastly a peak of multi-allelic CNVs at the centromere of chromosome 16.

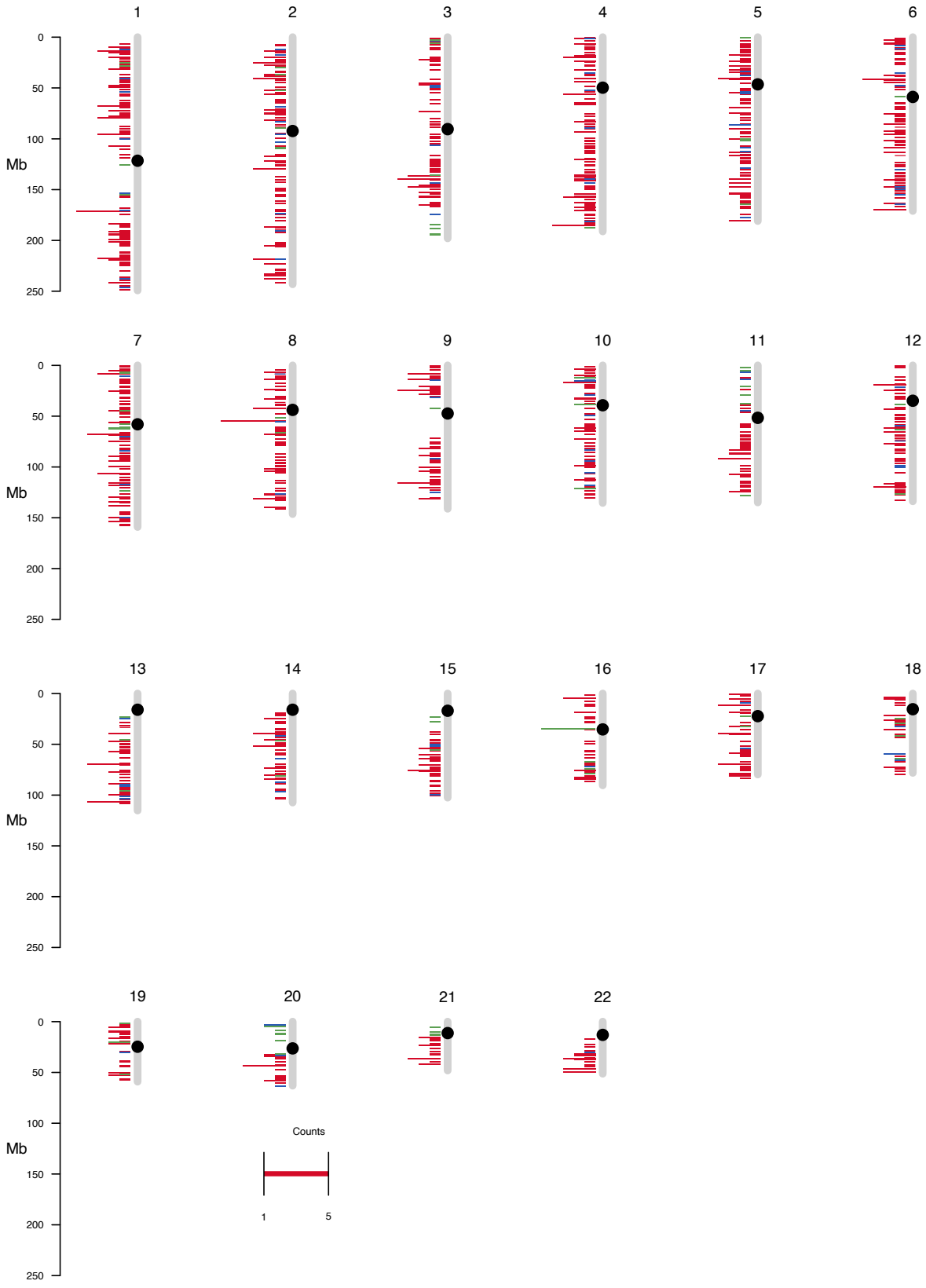


Figure 3. 18: Genomic representation of CNV density of novel variants. Novel variants distributed across the genome except for a few areas of higher density on chromosomes 1, 8 and 16.

Looking at the 88 novel multi-allelic CNVs, 64 variants were present in individuals with an absolute copy number total ranging from zero or one copies, to three copies of that variant. This means they are multi-allelic, due to having both deletion and duplication alleles for the same site, resulting in a total of three or four alleles. The remaining 24 variants showed copy number expansions and had diploid copy numbers greater than three copies. Three variants were found where individuals had more than ten copies of the variant. One variant of particular interest had a maximum total copy number for an individual of 55 copies.

The variant chr1:125175980–125177652 a 1,6kb variant is highly polymorphic with individual diploid copy number varying greatly. The distribution of copy numbers is shown in Figure 3.19. 26% of the cohort have a homozygous deletion of this region and a copy number expansion is seen, with 7 being the most common diploid copy number, extending to one individual having a total of 55 copies of this variant. The variant is intergenic and overlaps some regulatory regions. This novel variant is an occurrence of a runaway duplication and has higher total copy number in some individuals than has been previously reported (Usher and McCarroll, 2015).



Figure 3. 19: Distribution of diploid copy number for chr1:125175980–125177652. Variant copy number range from a homozygous deletion to multiple copies of the variant. The reference genome (Build 38) marked with an arrow has 2 copies and all individuals in this cohort had a variant different to the reference.

A functional analysis of all novel CNVs is shown in Figure 3.20. It shows a similar profile to the overall functional assessment of the intersection variant set, with 38% being intergenic, 7% near genes, 35% intronic variants, 7% regulatory variants with 13% of variants involving coding transcripts, either whole genes or exons.

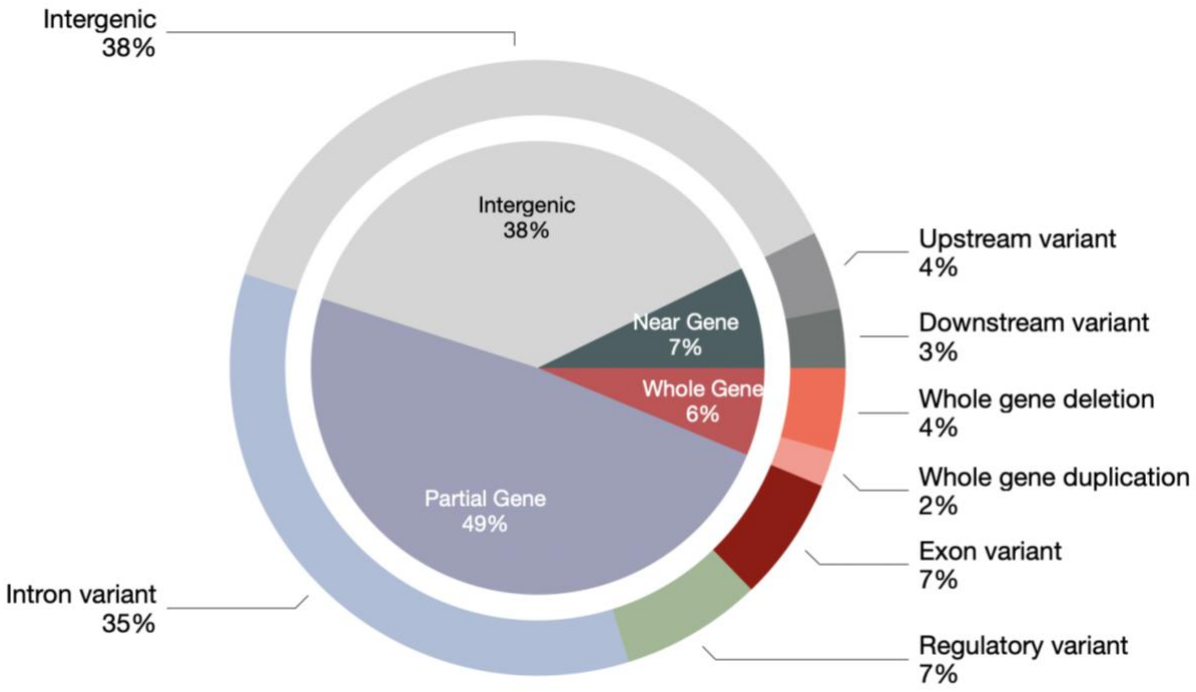


Figure 3. 20: Functional analysis of novel variants. 13% of all variants in the intersection dataset overlapped exons or entire genes.

A closer look at the novel variants that overlap exons or whole genes reveals that there are 165 variants that overlapped 200 different genes. 25 variants involved multiple genes and 139 involved one gene. Focusing on variants that overlapped at least 50% of a gene transcript, we see that all these variants are ultra-rare or rare except for one, chr7:44004036–44011050 found at an VAF of 2,5%. This variant is a 7kb multi-allelic CNV with 2,5% of individuals having diploid copy numbers between 3 and 6 copies. The variant overlaps exons 4-7/7 of the *SPDYE1*. The product of the *SPDYE1* gene is a cell cycle regulator, predicted to be a negative regulator of cyclin-dependent kinases but there is little to no evidence of disease association.

3.3.5 Copy number variants overlapping genes known to cause developmental disorders

A. Rationale

Given that the aim of this PhD is to establish baselines to help improve the diagnosis of DD, we now turn to look more closely at the population CNVs found in this study that overlap genes known to have caused a DD. These may be of particular importance in a diagnostic pipeline filtering for pathogenic variants causing DDs. 519 variants of the intersection variant set were found to overlap with a Gene2Phenotype DD gene.

B. Variant type breakdown of the developmental disorder gene call set

The intersection DD gene variant set of 519 CNVs is composed of 446 deletions, 44 duplications, 29 multi-allelic CNVs. The locations of these variants are shown in Figure 3.21. The size and VAF profiles show the same patterns as the overall intersection set of variants. Of note there are 313 CNVs with allele frequencies >0,1%. This percentage is the lower bound used in variant filtering to exclude population variants not likely to be disease causing (Pedersen *et al.*, 2021). A subset of these more common variants that overlap coding transcripts are interrogated further when the functional implications of variants are assessed.

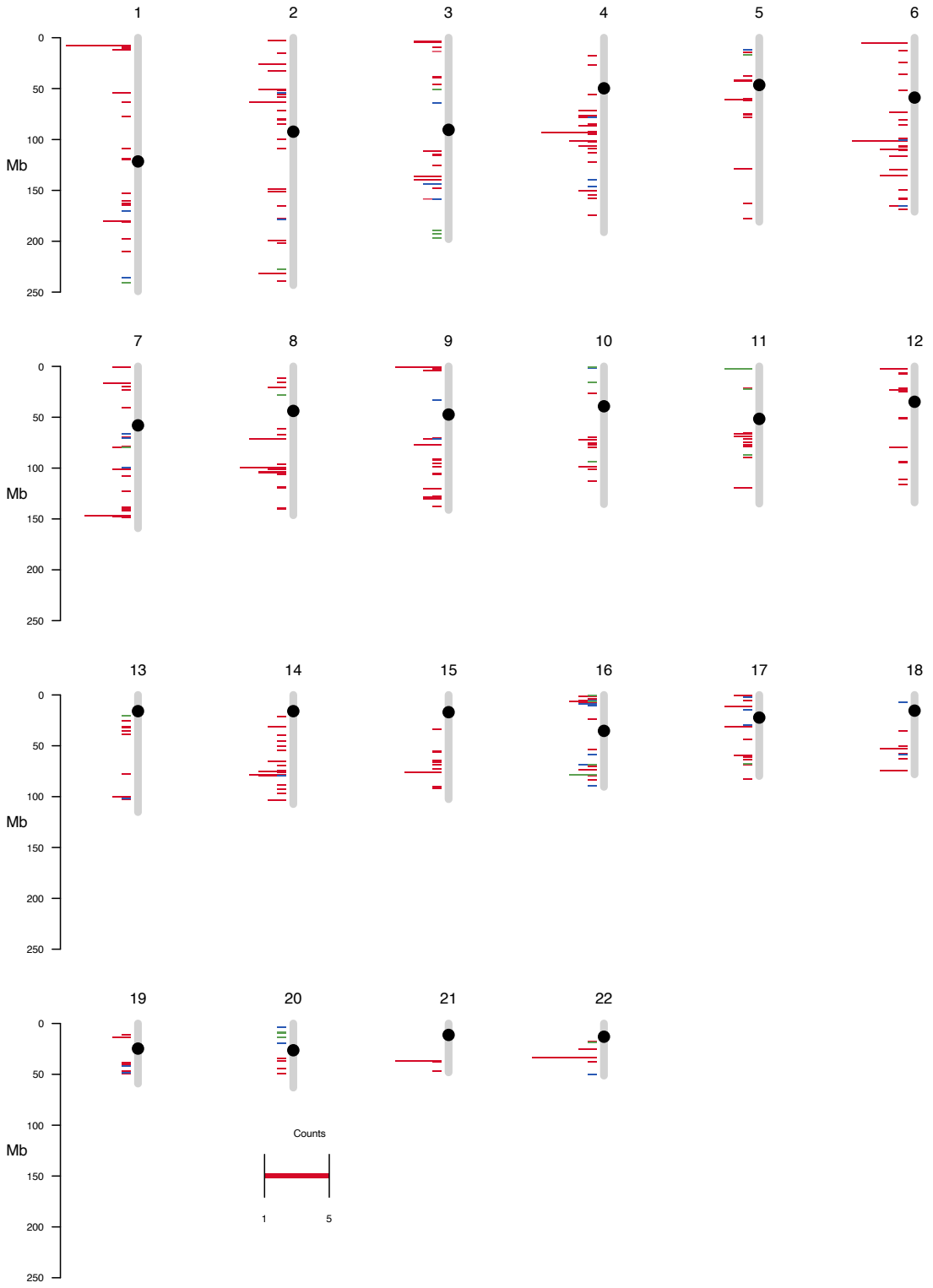


Figure 3. 21: Genomic locations of variants overlapping DD genes.

C. Novel variants that overlap developmental disorder genes

A comparison of the intersection DD gene variant set to established datasets showed that 18% of the variants are novel with 82% having been found by previous studies.

Looking at the novel variants we see that 87 are deletions, 7 are duplications and 3 are multi-allelic CNVs and in terms of their allele frequency all of them had VAF <1%. Similarly Nyangiri *et al.* (2020), who studied CNVs in a central African cohort, found that almost all novel variants were rare.

D. Differences of copy number variants in developmental disorder genes between African regions

To perform a comparison of CNVs of individuals from different African regions, all the 1027 participants were divided according to region and yielded four groups of differing sizes: 614 west African individuals, 54 central African individuals, 148 east African individuals and 211 southern African individuals. The comparison of variants between regions was performed on the 490 bi-allelic DD gene variants and is shown in an upset plot in Figure 3.22.

A chi-square test was performed to compare the proportion of CNVs seen in each African region given the number of individuals in the cohort from that region. Analysis of the total number of CNVs per region, West Africa has significantly more CNVs than expected and east and central Africa significantly less. When unique CNVs per region are considered, central Africa has significantly fewer than expected. It is expected that the CNVs unique to regions are rare or ultra-rare and so will increase with increasing sample size and that CNVs with higher VAFs will be shared by regions. This phenomenon was reported in the 1000 Genomes Consortium SV investigation (Sudmant *et al.*, 2015b). When the variants frequencies are interrogated, this trend holds true for this data.

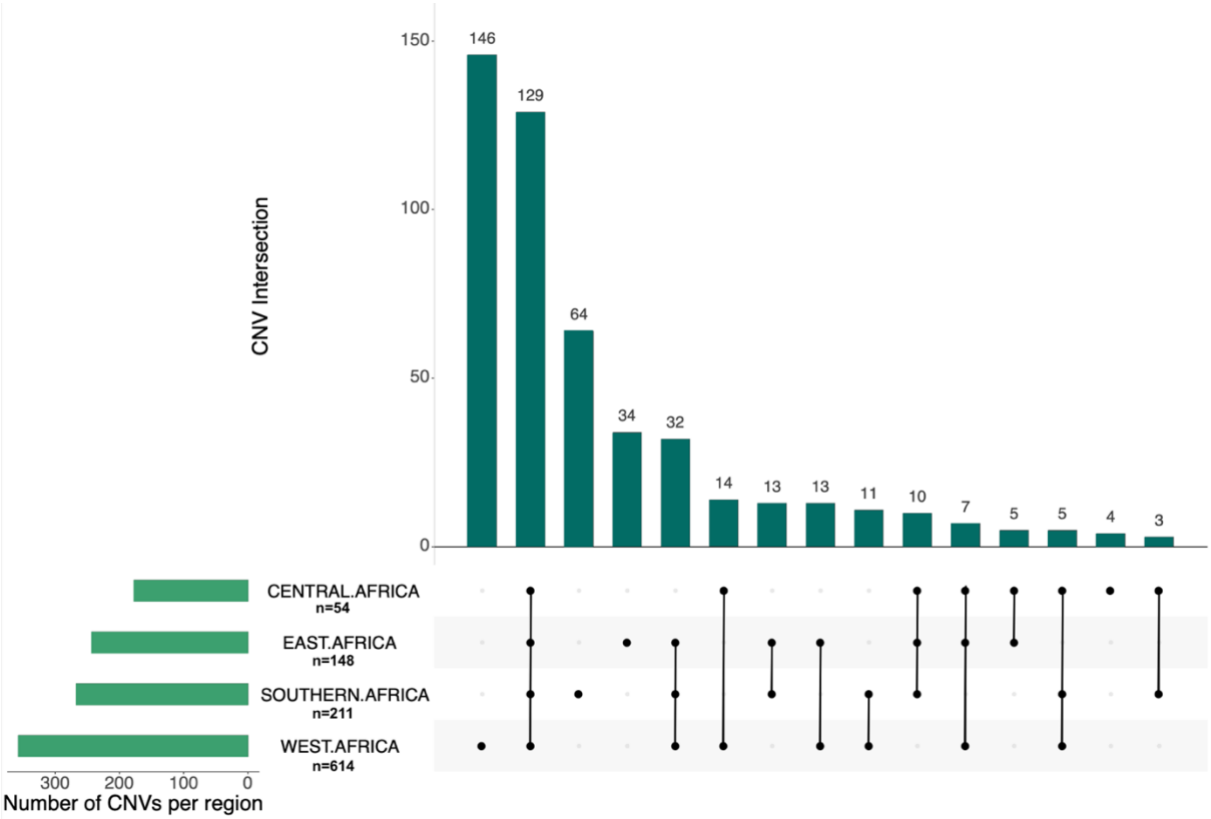


Figure 3. 22: Comparison of intersection DD gene CNVs found in individuals from each African region. 26% of variants are common to all regions as well as many unique to specific regions. The upset plot depicts the number of variants present in a region or combination of regions using the black dots to show which region or regions are being assessed. The number of individuals in each region shown by n=x.

E. Functional analysis of the developmental disorder gene variants

The functional impact of the variants in the intersection DD gene variant set is shown in Figure 3.23. 15% of variants involved coding transcripts with 7% encompassing whole genes and 8% exons. 85% of the variants involved only non-coding regions with 76% of variants affecting only introns, 8% regulatory regions and 2% intergenic.

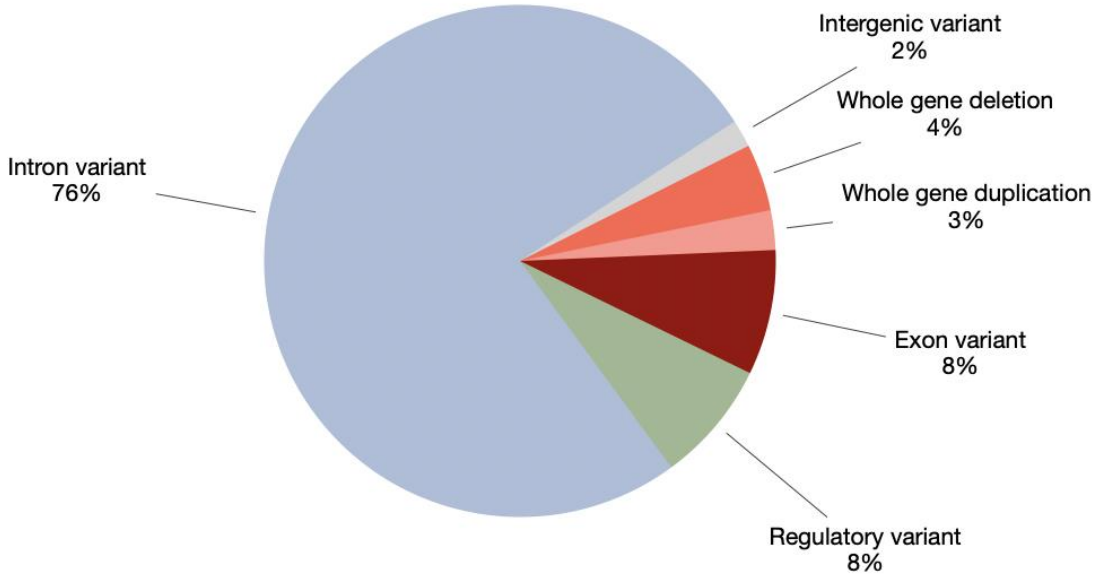


Figure 3. 23: Functional analysis of variants in intersection DD gene variant set. 76% of variants that overlapped a DD gene were intronic variants and 15% variants encompassed exons or whole genes.

Taking a closer look at the variants that contained coding sequence transcripts, we see that there were 75 variants that overlapped 98 different genes. Thirteen variants overlapped more than one gene with the remaining 62 variants overlapping a single or partial gene. Some of these CNVs overlapped a coding sequence transcript of a non-DDG2P gene. This occurred when a CNV overlapped non-coding portions of a DDG2P gene as well as coding portions of flanking non-DDG2P genes. A number of these non-DDG2P genes were small clone-based Ensembl genes.

Table 3.2 looks in more detail at the most common variants, all >1% VAF, that overlapped a whole gene or exons of genes. These 6 variants have all been previously described but we see some occurring at different VAFs in this African cohort. The variant with the highest VAF was chr5:42628337–42631029 with an allele frequency of 41,6%. This 2,7kb deletion overlapped exon 3 of the *GHR* gene that encodes growth hormone receptors. This variant has been identified in many studies and was found at a similar VAF in the 1000 Genomes Project (Sudmant *et al.*, 2015b).

The variant chr8:20224309–20225503 is a 1,2kb deletion involving exon 2 of the *ATP6V1B2* gene that encodes a subunit of vacuolar ATPase. It is a known variant and has been reported in ClinVar as benign. It does not involve the same region as the known pathogenic SNVs in this gene. This variant was seen in this cohort with a VAF

of 3,2%, a higher VAF than in the 1000 Genomes Project where it was observed at a frequency of ~1% (Sudmant *et al.*, 2015b).

The 3,6kb deletion chr1:108190709–108194629 involves the first exon of the SLC25A24 gene that encodes a calcium-binding mitochondrial carrier protein. Deletions and partial deletions of this gene are present in ClinVar as benign variants, and pathogenic SNVs were miss-sense variants in exon 5 (Landrum *et al.*, 2018). This CNV was present in this cohort at a VAF of 6,9% a number far lower than the VAF at which it was found in the 1000 Genomes project which was ~50% (Sudmant *et al.*, 2015b).

The variant chr11:71501700–71504676 is a 2,9kb deletion, found mainly in African individuals with an overall frequency of 1,2% (Collins *et al.*, 2020; Sudmant *et al.*, 2015b). This variant causes whole gene deletion of *NADSYN1*. *NADSYN1* encodes NAD synthetase 1, which catalyses the last step of the synthesis of nicotinamide adenine dinucleotide (NAD). NAD acts as a co enzyme involved in redox reactions, a substrate for post translational modifications and a precursor to cell signalling molecules. Loss of function SNVs in this gene have been shown to have autosomal recessive inheritance, with homozygous individuals having multiple congenital anomalies (Landrum *et al.*, 2018). All the individuals in this cohort were heterozygous for the variant causing this whole gene deletion. This would seem to agree with the autosomal recessive nature of the SNV cases found in this gene, where those who were homozygous for SNVs had a disease phenotype. This could by extension mean that individuals homozygous for this CNV would have a disease phenotype. It is therefore possible that this may be an unrecognized autosomal recessive condition present in Africa.

The 5,4kb deletion chr5:60911996-60917354 has been found previously with VAF 0,3%, but it was found in this cohort at a higher frequency of 1,5%. This variant involves exon 5 of *ERCC8*, where reported deletions causing DDs have involved exon 4 (Xie *et al.*, 2017).

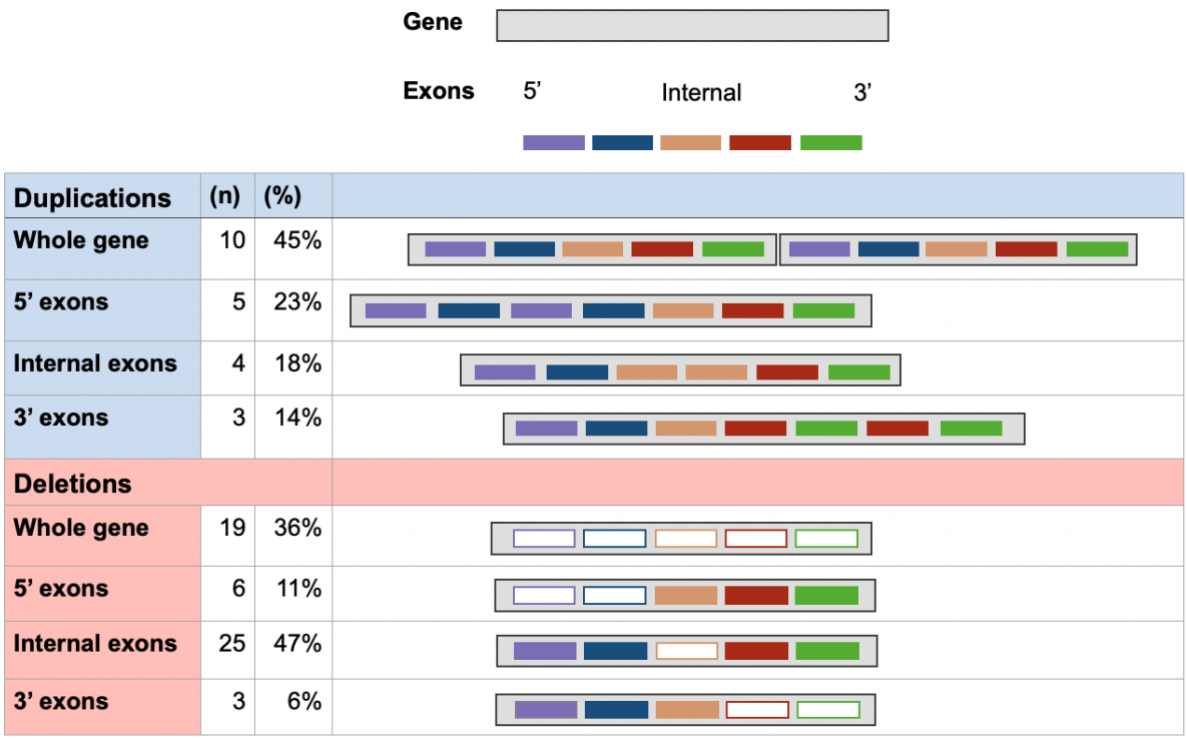
The variant involving the *RBFOX1* gene was the only one of these six CNVs that was a duplication. This 12kb duplication chr16:5465353–5478218 involves exon 2 of the *RBFOX1* gene. Pathogenic variants reported in this gene have all been deletions (Bacchelli *et al.*, 2020); however one potentially disease causing duplication has been reported involving exons 4 and 5 (Zhao, 2013). The authors did not conclusively

classify this variant as pathogenic and it involved exons 4 and 5, not exon 2 which was involved by the CNV in this study. Additionally this CNV, has been reported in other healthy cohorts, making more likely that it is a benign population CNV.

Table 3. 2: Low frequency and common population CNVs involving whole genes or exons of DD genes.

CNV	CNV Type	Gene	Exons involved	Overlap (%)	CNV Position	VAF %	OMIM Gene Number
chr5:42628337–42631029	deletion	*GHR*	3 of 10	0,91	Internal exons	41,6	600946
chr1:108190709–108194629	deletion	*SLC25A24*	1 of 10	3,64	5’ exons	6,9	608744
chr8:20224309–20225503	deletion	*ATP6V1B2*	2/4	13,83	Internal exons	3,2	606939
chr11:71501700–71504676	deletion	*NADSYN1*	-	100	Whole gene	1,7	608285
chr5:60911996–60917354	deletion	*ERCC8*	5 of 13	7,53	Internal exons	1,5	609412
chr16:5465353–5478218	duplication	*RBFOX1*	2 of 3	3,57	Internal exons	1,5	605104

We also looked at the location of all the variants that contained coding transcripts, either whole genes or exons and this is shown in Figure 3.24. This shows that the greatest number of these duplications (45%) affect whole genes and the greatest number of these deletions (47%) affect internal exons of genes.



Duplications	(n)	(%)
Whole gene	10	45%
5' exons	5	23%
Internal exons	4	18%
3' exons	3	14%
Deletions		
Whole gene	19	36%
5' exons	6	11%
Internal exons	25	47%
3' exons	3	6%

Figure 3. 24: Gene location of intersection DD gene variants that overlap coding transcripts. Deletions shown as blank exon blocks. Variants were classified as whole gene if all exons were encompassed in the variant, 5' exons if the first exon was included in the variant and 3' exons if the last exon was included in the variants.

F. Predicted pathogenicity of variants

None of the variants in the intersection DD gene variant set were pathogenic. The variants were assessed for pathogenicity according to the new ACMG recommended technical standards for interpretation CNVs. We found that 15% (83) of variants were classified as benign and remaining 85% (461) as variants of unknown significance. Pathogenic variants were not expected, given that this is a population cohort without known diseases. The exception to this would be pathogenic variants present as carrier alleles in genes with autosomal recessive inheritance. The intersection DD gene variant set was also queried against the ClinVar database where two variants matched ClinVar entries. One of the two was a CNV type mismatch; the variant in ClinVar was a deletion and the variant in this study a duplication. The other chr6:135438517-135442182 was classified in ClinVar as 'Likely pathogenic'. The variant classification was partly based on the variant not having been reported as a benign population variant (Carss *et al.*, 2017); however it has subsequently been reported in gnomAD-SV at a frequency of 0,07% in African populations, but not in other populations. It was

present in this cohort at a slightly lower frequency but its presence in another healthy cohort adds evidence to it being benign.

3.4 Discussion

3.4.1 Overview

In this study we produced a data set of CNVs from a large cohort of African individuals using two different CNV calling pipelines. This dataset will serve as a valuable resource for African genetic disease research and has uncovered some of the richness of the African CNV landscape.

3.4.2 Profiles of separate datasets

The analysis of each call set separately was a valuable way of understanding the capabilities, strengths and limitations of each approach. Manta has a wide size spectrum of CNV detection and detected variants from 50bp to 10 Mb where Genome STRiP has a narrower range of size detection than Manta being between 1kb and 1Mb. A large number of variants <100bp were detected by Manta, and as this size range of variants has not been previously well characterised, this made them an interesting set of variants to study in these understudied African populations (Almarri *et al.*, 2020). These small variants have not been studied in as large an African cohort as this study and so the variants detected in this size range were expected to be previously unreported. This was confirmed by the fact that >50% of these were novel when compared the DGV — 20% higher than the percentage of novel variants in the Manta variants >100bp. As these small variants are largely novel and previously uncharacterised, it is therefore important they are validated. This size of variant cannot be detected by Genome STRiP, so these variants still need to be validated through another bioinformatic approach or by laboratory confirmation.

The variants >1Mb detected by Manta were also of interest as this size of variant is usually rare. Collins *et al.* (2020) found that 3,8% of the population have at least one SV larger than 1Mb and also stated that CNV size is a key determinant influencing selection against variants, keeping their population frequency low. Analysis of these large variants in the Manta data showed a number at VAF far greater than expected. Figure 3.8 shows a number of variants >1Mb with VAF between 75–100%. These

variants are therefore highly likely to be false positives and cannot be reported with confidence unless they can be validated by another technology or other CNV calling algorithms. It is not completely clear why these were observed, but it has been shown that Manta is less accurate on large deletions that on medium or small ones (Kosugi *et al.*, 2019). Cameron *et al.* (2019) showed that most SV callers, including Manta, reach peak accuracy at 300–500bp, and gradually become less accurate with increasing size.

The VAF profile of the Genome STRiP variants aligned to what was seen in the 1000 Genomes project, and we saw the same tendency of common duplications to be multi-allelic not bi-allelic (Handsaker *et al.*, 2015; Sudmant *et al.*, 2015b). The VAF profile of the Manta variants in Figure 3.5 showed that a general trend of large numbers of common variants, is again far higher than expected. The gnomAD-SV study reported <10% of variants for all SV types were found at VAF >1% (Collins *et al.*, 2020) and the 1000 Genomes Project SV paper reported that 65% of the variants had VAF <0,2% (Sudmant *et al.*, 2015b). The reason for Manta detecting so many variants with high VAF is uncertain. This added further evidence to support focusing further analyses on the variants that had been detected by both tools.

3.4.3 The intersection variant set

There is always a tension between finding as many CNVs as possible and ensuring the accuracy of the CNVs found. In this study we chose to use Manta and Genome STRiP, two algorithms that use different calling approaches and have different capabilities, to detect the full scope of African CNVs. Manta can assess a wide size spectrum and Genome STRiP can detect multi-allelic CNVs, a class of variants not well studied in African individuals (Almarri *et al.*, 2020). This strategy maximised sensitivity, utilising tools that employ different calling methods and that detect a wide range of variants to ensure a discovery of a wide scope of African CNV. The outputs from these two tools could then be combined using the union of these datasets, keeping all variants called by one or other tool. This approach means that CNVs called by just one algorithm are used in the final dataset; however, it is highly recommended that consensus variants called by more than one algorithm should be prioritised in CNV discovery, to reduce false positives (Cameron *et al.*, 2019; Kosugi *et al.*, 2019). In this second approach only the intersection of the datasets is retained, excluding all variants only called by one tool.

This meant there was a tension between producing a CNV dataset with wider scope or one with greater accuracy. If one chose to retain only variants called by both tools, accuracy would be increased but variants that cannot be called by both tools would be lost. The difference in size range of CNV detection, between Manta and Genome STRiP, meant that the small variants <1kb and the large variants >1Mb, that could only be detected by Manta, would be lost. Despite the loss of these variants, the factor that weighed towards choosing to retain only variants called by both tools, was the large number of high VAF CNVs called by Manta. These high VAF CNVs, especially those >1Mb, did not conform to expected population CNV parameters (Collins *et al.*, 2020; Abel *et al.*, 2020; Sudmant *et al.*, 2015b) and so were believed possibly to be false positives. CNV calling algorithms are known to lose accuracy with increasing size and these are at the edge of the detection ability of such tools (Cameron *et al.*, 2019). Additionally it is shown in Sudmant *et al.* (2015b) and Collins *et al.* (2020) that size is the greatest factor influencing selection against variants decreasing the likelihood that large variants will be found at high VAF. If these variants are in fact truly so common in Africa they will require validation by another tool or technology. We therefore decided to go forward with the intersecting set of variants detected by both Manta and Genome STRiP. This set may have a smaller size spectrum but there is a higher degree of certainty of the accuracy of these CNVs and so we chose to focus further analyses on this set of variants. The limited size of the variant dataset produced in this study necessitates further work to produce a variant dataset from this diverse set African WGS that cover the whole size spectrum. This will be achieved by adding additional CNV algorithms to the pipeline for this project.

The size and VAF and functional profiles of the intersection variant set matched expected parameters for population CNVs as described in the 1000 Genomes Project and gnomAD SV study (Sudmant *et al.*, 2015b; Collins *et al.*, 2020). The distribution of CNVs over the genome in Figure 3.13 showed mainly expected density hotspots over telomeres and known hotspot regions, but the large concentration of multi-allelic CNVs cannot be fully explained. There are some multi-allelic CNV hotspots in the 3q region, centromeric region of chr15 and 20p region, reported by the 1000 Genomes consortium, however not to the extent that was seen here. Some of the multi-allelic CNV hotspots were novel when compared to the DGV, so it appears these regions may harbour unique African multi-allelic CNVs, like those described by Almarri *et al.* (2020).

3.4.4 Novel African variants

Comparison of the intersection variant set to DGV showed that 17% (1 549) variants were novel. This is a lower percentage than the percentage of novel variants found by the 1000 Genomes Project (60%) and the gnomAD-SV study (86%). A breakdown of the number of novel CNVs (deletions, duplications and multi-allelic CNVS) is not given in the gnomAD-SV study but the percentage of novel CNVs found by the 1000 Genomes study was 59% still a lot higher than the percentage seen in this study. This is most likely because both these projects were landmark studies, using new methods and larger cohorts than previous studies, that enabled them to detect far more SVs than had been previously described. The lower percentage of novel variants in this study is not unexpected, as this study was on a smaller cohort and only looked at CNVs, excluding other types of SV that contributed substantially to the overall number of SV discovered in the gnomAD-SV study. Another reason for this is probably because of the conservative approach that was taken to only assess the intersecting variants, that significantly reduced the size range of variants. We saw that >50% of the Manta variants <100bp were novel, showing that this size range of variants has not been well studied in African individuals. This highlights the importance or continuing this work to produce a wider size range dataset from multiple CNV calling algorithms. The novel African CNVs detected in this study show some of the richness of African CNVs and further study should uncover more novel African CNVs and other types of SVs for inclusion into SV reference databases.

A number of high copy number multi-allelic CNVs have been reported in African populations (Handsaker *et al.*, 2015; Almarri *et al.*, 2020), so the novel multi-allelic CNVs found were assessed closely. 24 of the 88 novel multi-allelic CNVs had copy number expansions >3 copies. A number of these were in centromeric regions, known to be prone to multi-allelic expansions (Collins *et al.*, 2020). The runaway duplication of interest with diploid copy number reaching 55 copies is higher than previously described (Usher and McCarroll, 2015). This variant site was identified by both Manta and Genome STRiP, but the VAF of 100% is concerning for a novel variant, especially given that many of the individuals included in this study were also assessed in the 1000 Genomes project for multi-allelic CNVs. However, the sequences included from the 1000 Genomes Project, were recently sequenced high coverage sequences, not the low-medium coverage sequences of the individuals used for the 1000 Genomes project analysis in 2015. It is therefore possible that this variant is a true novel African

runaway duplication with copy numbers higher than previously described, that could not be detected on the low coverage sequences analysed in 2015 (Usher and McCarroll, 2015).

3.4.5 African regional differences

The PCA performed on the entire Manta dataset showed significant batch effects despite increasing quality threshold to overcome this. This is an unfortunate consequence of having combined WGS datasets from multiple projects, but the value gained by having a more diverse and the larger dataset was more important for African CNV discovery.

Despite the batch effects, there were some differences being observed, especially in the H3A-Baylor data, and so we chose to examine this project's data alone. In the H3A-Baylor PCA we do see separation of the east west and southern African regions, indicating that there are differences in the CNV landscape between these African regions. The DD gene regional comparison, performed on the entire cohort, in the upset plot also showed there are differences in the variants involving these disease genes, although given the unequal sample size we cannot make a statistically significant assertion of these differences. Nyangiri *et al.* (2020) in their analysis of CNV in three African ethnolinguistic groups, including individuals from central, east and west Africa, were unable to find differences between these groups. In this study we had a greater number of individuals and a more diverse cohort, and so we were are able to see some differences between regions, but still not at the same resolution possible with SNVs (Choudhury *et al.*, 2020). In order to improve upon this regional comparison a cohort with regional groups of similar size, preferably from a single project, would be needed.

3.4.6 Baseline population variants involving developmental disorder genes

The baseline CNVs that overlapped DD genes were an important focus in this study. CNVs involving these genes are likely to be found in DD research in Africa and so knowledge of population CNVs present in healthy African individuals is valuable information. 313 of the 519 CNVs involving DD genes were present in this cohort at VAF >0,1%, the lower cut off used in variant prioritising pipelines to isolate potentially pathogenic variant (Pedersen *et al.*, 2021). 150 of these had VAF >1%, a frequency at which variants are usually considered to be a probably benign common population

variant (Collins *et al.*, 2020). The novel CNVs identified in this subset of variants were all ultra-rare or rare, so although they may not be common benign variants, many of the rare variants are present at frequencies >0,1% and will therefore be a useful reference for research into DDs in Africa.

The functional analysis of the variants involving DD genes showed that the majority (85%) only involved non-coding portions of these genes. This is an expected finding given that these genes are known disease genes and disruption of coding transcripts is likely to result in disease. When the 15% that do involve coding regions were examined, we saw that duplications were more likely to involve whole genes while deletions were more likely to involve only internal exons. This pattern was also observed by Truty *et al.* (2019) in their analysis of intragenic CNVs in Mendelian disease genes. They discuss that baseline CNVs in clinically relevant genes are either benign due to not affecting the structural integrity of a gene or they are pathogenic but found in autosomal recessive genes as carrier alleles. This explanation aligns with what was seen in the variants involving coding regions that were present at VAF >1%. The 6 variants described in this set of variants demonstrate both these types of baseline CNVs. The *GHR* gene variant is a good example of the first mechanism described. This 2,6kb CNV, was found as a very common variant, that deleted an internal exon of the GHR gene. In this case alternative splicing of exon three is a known phenomenon and so a deletion of this exon is well tolerated and does not disrupt the structural integrity of the gene (Stallings-Mann *et al.*, 1996). The variant resulting in a whole gene deletion of the *NADSYN1* gene shows the second type of baseline CNV described by Truty *et al.* (2019) of being a pathogenic variant in an autosomal recessive gene seen as a carrier allele. We observed a VAF of 1,7%, but all individuals with the variant were heterozygous, suggesting that this variant may be showing autosomal recessive inheritance like the loss of function SNVs reported for this gene (Szot *et al.*, 2020).

In these 6 examples of common baseline CNVs in DD genes, different VAFs for some of these known variants were observed in this African cohort compared to the VAFs observed in previous less diverse cohorts. Knowledge of the frequency at which such variants are seen in Africa contributes to the overall knowledge of genetic variation in DD genes in Africa.

The pathogenicity prediction of the DD gene variants, using ClassifyCNV (Gurbich and Ilinsky, 2020), a program based on the new ACMG CNV interpretation guidelines (Riggs *et al.*, 2020), showed that none of the variants were pathogenic, with 85% being variants of uncertain significance and 15% being benign. This is not an unexpected finding for baseline CNVs in a healthy cohort. Additionally, the ACMG guidelines are aimed at diagnostic use, and so include patient and literature specific information (Riggs *et al.*, 2020) that is not part of the programmatic prediction of pathogenicity in ClassifyCNV (Gurbich and Ilinsky, 2020). The comparison to ClinVar showed one variant that matched a “likely pathogenic” ClinVar variant. Closer examination showed that this variant has subsequently been reported in gnomAD-SV at a frequency of 0,07% in African populations, but not in other populations. The presence of this variant in another healthy cohort adds evidence to it being benign.

3.4.7 Conclusion

In conclusion the variant call set produced through this study will be a valuable tool for African genetic disease research and has been an important first step to producing a comprehensive high confidence reference of African CNVs across the size spectrum. As discussed earlier this dataset is limited in size range and requires further work using extra CNV calling algorithms to produce a comprehensive CNV variant set across the size spectrum. The limitations of this work and further work to be pursued will be discussed in detail in Chapter four. The production of such a reference database is of critical importance and so further work on this, will be prioritised.

Chapter 4

Conclusion

4.1 Improvement Required in the Diagnosis of Developmental Disorders

The aim of this PhD was to address two baseline gaps required to help improve the current diagnostic process for DDs in Africa. The first gap identified was a lack of research describing the population of patients who present with DDs and the yield of the current diagnostic process for these patients. The second baseline gap identified was the fact that CNVs from diverse African populations have not been well characterised. These reference CNVs are an important tool for genetic disease research in Africa.

DDs are a group of heterogenous disorders including NDDs and congenital anomalies. They have been increasing in prevalence in LMICs like South Africa (Olusanya *et al.*, 2018). Lack of research and data on childhood disabilities such as DDs has resulted in services to diagnose and treat such conditions, being poorly funded and prioritised in South Africa (Kamp *et al.*, 2021; Malherbe *et al.*, 2021). The DDD-Africa study aims to improve the understanding of DDs in Africa, in terms of prevalence, frequency and aetiologies, by utilising WES to detect genetic variants causing DDs in patients in Africa. It also aims to produce a practical plan for the implementation of WES in an African setting from the experience and capabilities gained through the study. The baselines gaps addressed in this PhD are important parts of ensuring the aims of DDD-Africa can be achieved.

The first baseline gap was addressed in this thesis by the first aim (*Chapter Two*) in the file review, characterising the population of patients with DDs and evaluating the existing diagnostic process. The second baseline gap was addressed in aim 2 (*Chapter Three*) where a catalogue of population baseline CNVs was produced from CNV analysis of WGS of a diverse African cohort.

4.2 Diagnostic Test with Higher Yield Needed for Patients with Developmental Disorders

In the retrospective file review of patients from the Johannesburg Human Genetics clinics, we saw that patients who present common features of DDs, such as developmental delay, congenital anomalies or dysmorphic features, but no clear gestalt, comprise 50% of the cases seen in these clinics. Almost all, 92% of these patients remained undiagnosed despite having undergone numerous different

traditional tests, with mounting costs, in an attempt to reach a diagnosis. Without a clear phenotype to guide targeted testing, the most commonly utilised tests in this setting were a karyotype and MLPA (that allow screening for microdeletions and duplications). These two tests were utilised mainly because they were the only available options for broad non-targeted investigation of these individuals, despite the fact they do not have very high yields for diagnosis of patients with DDs. This approach was shown to be inadequate to diagnose the majority of patients with non-specific DDs.

Our study clearly showed that there is a great need for improving diagnostic practices for patients with DDs in this setting. WES has been shown to have a high diagnostic yield of 30-53%, much higher than traditional tests and CMA (Srivastava *et al.*, 2019) and is now the ACMG recommended first-line test for patients with DD (Manickam *et al.*, 2021). It has been shown that first-line WES results in earlier diagnoses of more patients with many positive effects. Firstly, it halts further diagnostic investigations, ultimately lowering investigation costs and ends the burdensome diagnostic odyssey for families (Carmichael *et al.*, 2015; Dillon *et al.*, 2018; Dragojlovic *et al.*, 2020). An accurate molecular diagnosis means clinicians can understand the condition better, and therefore provide more accurate condition-guided management and surveillance, precision therapy where available (Tan *et al.*, 2017), as well as recurrence risk assessment. We therefore suggest that the introduction of WES is the best way forward to improve diagnostic yield for patients with non-specific DD, although a thorough cost effectiveness analysis is still required.

Assessing the cost effectiveness of first-line WES from a public health point of view has been challenging due to the complexity of measuring the full cost incurred by a patient remaining undiagnosed (Tan *et al.*, 2017; Schwarze *et al.*, 2018). These costs encompass genetic test costs, costs incurred by the patient's family over time as well as important quality of life implications (Dragojlovic *et al.*, 2020). The review by Schwarze *et al.* (2018) assessing WES cost effectiveness studies concludes that the cost effectiveness of WES is highly dependent on the clinical context, patient population and health system factors. They highly recommend context specific analyses be initiated assessing costs as comprehensively as possible.

Through the patient file audit, we were able to provide motivation for the approach being taken in the DDD-Africa study, to use WES to investigate DD in Africa and to further WES implementation. The audit showed how much there still is to uncover

about the aetiologies of DDs in Africa, which the utilisation of WES in the DDD-Africa study will begin to uncover. DDD-Africa will provide evidence of the diagnostic yield of WES in this setting which is an important next piece of evidence needed to motivate for the widespread implementation of WES. The evaluation of the current testing methods and their diagnostic yield, in the file review, will serve as good baseline to compare the outcomes and diagnostic yield obtained from utilising WES in the DDD-Africa Study. The DDD-Africa study will help build a practical plan for WES implementation in this setting, as well as building capabilities and skills in WES analysis and interpretation, to facilitate the much-needed introduction of WES.

4.3 A Catalogue of African Population Copy Number Variants

CNVs are an important cause of DDs, accounting for ~15% of NDDs (Kaminsky *et al.*, 2011), and so the analysis of CNVs will be an important part of WES implementation to improve the diagnosis of DDs. Part of the process of diagnostic WES is the process of variant filtering and prioritization. As described previously an important part of this process is filtering out common variants present in populations. Current reference databases used as a source of population variants for filtering, such as gnomAD-SV, have low representation of African populations, limiting their use in African genetic disease research. Baseline population CNVs from diverse African populations have not been well characterised and so this was the second critical gap in baseline data addressed in this study.

In order to address the second aim, a CNV discovery pipeline was produced to call CNVs from 1027 WGS. This pipeline utilised two different CNV calling algorithms Manta & Genome STRiP. These tools utilise different calling methodologies that enabled us to detect a wide range of CNVs. The use of WGS to produce this baseline CNV dataset gives it significant future value for genetic diagnostics, as WGS is quickly becoming the diagnostic test of choice for many genetic conditions like DD (Lee *et al.*, 2021).

Two separate variant sets were produced using these tools and each was analysed to assess the scope of CNVs detected. Manta detected a wide size range of deletions and duplications, especially a large set of small variants <100bp. Genome STRiP has a narrower size range of detection but can detect multi-allelic CNVs that Manta cannot detect.

It is strongly recommended that multiple tools are used to discover variants, to reduce false positive rates. Unfortunately, the different size ranges of Manta and Genome STRiP meant that retaining only variants called by both tools would result in the loss of Manta variants (<1kb and >1Mb) outside the range of Genome STRiP. We had however been concerned about the high number of high VAF variants called by Manta, and so chose to conservatively retain only variants called by both tools, to ensure high quality variants. This high confidence intersection variant set, although limited in size range (1kb–1Mb), conformed to expected parameters for population CNVs in terms of VAF, size and functional assessment, which increased confidence in the variant set validity.

A third of this intersection high confidence variant set were found at population VAF > 1% making them common baseline African population CNVs. A subset of these variants were novel African variants not previously reported, and these will be valuable to studies in genetics in Africa. The exclusion of the variants only called by one tool, resulted in the exclusion of the small Manta variants (<100bp), a large number of which were novel. This indicates there is still an important mine of novel African CNVs that could be uncovered by further work with additional tools. Once this expanded work is published, we will aim for the addition of this set of African variants to current reference databases. This addition will increase the utility of such databases for research into genetic diseases in African individuals

Given that the focus of this project is to produce baselines, to aid improvement of the diagnosis of DD, we focused separately on the variants that involved DD associated genes. This analysis uncovered a 519 CNVs overlapping DD genes, which adds to our knowledge of genetic variation in these genes in Africa. For a number of variants, VAF in this African cohort differed from VAF for other populations. The African VAFs from this study will be able to inform interpretation of variants detected in other studies. The subset of these variants that involved protein coding transcripts showed known mechanisms for baseline CNVs in disease genes; variants that did not disrupt the structural integrity of gene products, like in the *GHR* gene, and those that are possibly pathogenic variants present in autosomal recessive genes like the *NADSYN1* gene (Truty *et al.*, 2019).

Through this research we saw a number of novel multi-allelic hotspots and also three variants with copy number >10, of which there are not a great number known (Usher

and McCarroll, 2015; Almarri *et al.*, 2020). One such variant was the multi-allelic CNV with >50 copies. Copy number expansions with such a high total copy number have not been previously reported, and warrant further study to validate such findings (Usher and McCarroll, 2015). This is likely just a fraction of novel African SVs, given the fact that we assessed only CNVs (deletions and duplications) of a limited size range. Further work into CNV discovery from this set of high coverage WGS will be of great value and would serve to enrich reference databases with African CNVs, that are important for genetic disease research on African ancestry individuals.

4.4 Challenges and Study Limitations

A number of limitations can be identified in both parts of this study. The main limitation in the retrospective file audit was the data source, which in this case were paper based clinical files. Firstly, not all patient files could be retrieved; a number of case files had been removed from the file archive and had to be located elsewhere and in some cases files had been lost completely. Secondly, case information was not uniformly recorded, and in many cases was incomplete. Missing data was greatest for non-genetic investigations, which meant these investigations could not be included in the cost analysis. This limited the scope of diagnostic costing to genetic investigations only, reducing our ability to accurately assess the total cost involved in reaching a diagnosis. Lastly all information captured had to be interpreted from hand-written consultation notes, which has inevitable limitations based on clarity of handwriting. These limitations highlighted the value of electronic health records. Electronic health records would have prevented loss or incomplete case files, reduced potential errors and enabled many more analyses, such as more comprehensive costing. As mentioned in Chapter Two although this study described approximate diagnostic costs this does not constitute and cannot replace a thorough cost effectiveness analysis. The lack of cost effectiveness evidence is a limitation to the conclusion that WES is a cost-effective test in this diagnostic context.

The CNV analysis involved some very different challenges and limitations. We encountered the multiple reasons why SV and CNV research has lagged behind SNV research. Implementation and running of the variant calling algorithms is challenging, partially due to complexity of such programs and partially due to the high computational demands (Kosugi *et al.*, 2019). Significant challenges were faced during

implementation and analysis, for example the Genome STRiP pipeline, as mentioned earlier, taking several months to execute. These challenges exposed some of the pitfalls of specific tools chosen and the practical aspects that have to be considered when choosing and combining tools. Due to the tool combination, the high confidence variant set produced is limited in CNV size range (1kb-1Mb). These tools are also very sensitive to sequencing methods and inter-sample differences, which can result in erroneous calling (Zhao *et al.*, 2013). In the Manta results we saw likely false positive variants >1Mb at excessively high VAF. CNV calling algorithms are known to have high rate of false positive CNV detection if a single tool is used alone, so this finding is not truly surprising and is why multiple tools should be used in combination. We also found that standard formats and analysis tools are not yet well suited to CNV analysis, and standards for comparisons have not yet been set, which makes the analysis challenging.

There are some inherent limitations to the detection of CNVs from short read NGS technology. Short read NGS performs poorly on low complexity regions, and these regions are known to be enriched for CNVs caused by NAHR. The weaknesses in the sequencing of such regions with short read NGS results in errors of CNV calling in these regions, which in turn produce errors of both false positives calls and missed calls. Additionally, the small size of reads means that SVs such as complex rearrangements, nested SVs or large insertions may be lost in the assembly of short reads (Mahmoud *et al.*, 2019). Long read sequencing technologies have enabled the detection of more accurate SVs for a few reasons. Long reads can span the length of small SVs and long stretches of repetitive DNA, enabling more accurate assembly and thus better detection of SVs. The disadvantages of long read sequencing are its high cost and a high sequencing error rate. Despite these disadvantages long read sequenced genomes have been included in many SV discovery studies and have been shown to increases the scope of SV detection substantially (Mahmoud *et al.*, 2019; Chaisson *et al.*, 2019).

Another limitation mentioned previously were the batch effects that resulted from combining multiple projects WGS, and the resultant different sized regional cohorts. This limitation occurred because of the scarcity of high coverage WGS from African individuals. All available high coverage African WGS had to be included to achieve a large sample size. Related to this limitation is the lack of phenotype information on the individuals included. In this study the cohorts from each project were recruited cross-

sectionally from their respective populations, aside from the individuals with known HIV infection, which is why this cohort has been used to determine baseline population CNVs. Without accompanying phenotype information for individuals, however, it is possible that mild undiagnosed conditions such as mild intellectual disability or congenital anomalies may have been present in these individuals.

4.5 Future Research

Future work related to the first aim is mainly around showing the utility and cost effectiveness of WES in this setting. An important part of this is research into the specific barriers that exist and will need to be overcome to successfully implement WES. Additionally, a clinical trial should be done to demonstrate the clinical utility and diagnostic yield of WES, as well as qualitative work into the perceived patient benefit. A cost effectiveness analysis for WES in this context is also a crucial piece of research required. This work ideally needs to be undertaken by a health economist, to provide sufficient evidence to inform public health policy.

For the analysis of African CNVs in the second aim, the next steps in this research are to produce a variant set of high quality, across the CNV size spectrum, from the current dataset of WGS. This work will then be deposited into a public CNV database like DGV to increase the representation of African genomes and thus the utility of such databases in an African research setting. In order to achieve this additional tools need to be added to the pipeline. Tools that use a combination of approaches, especially including the assembly-based method, perform better than those that use a single approach (Cameron *et al.*, 2019). For this reason the following tools have been considered as good options to add to the Manta pipeline: GRIDDS, a tool similar to Manta, that uses split read, read pair and assembly-based methods (Cameron *et al.*, 2017), DELLY that uses split read and read pair information (Rausch *et al.*, 2012) and LUMPY that uses split read, read pair and read depth approaches (Layer *et al.*, 2014). There is evidence in the evaluations by Kosugi *et al.* (2019) and Cameron *et al.* (2019) that these tools are high quality callers that work well together. LUMPY has slightly higher accuracy for deletions than Manta and they all have a similar size range to Manta, meaning one could use variant sites detected by two or three tools to produce a high confidence set of CNVs across the size spectrum.

Another aspect that will be required before implementing this new pipeline is the realigning of some of the samples from Human reference genome Build 37 to Build38. The problems that were encountered at this step mean that the sample realignment to Build 38 needs to be redone, taking into account these alternate contig regions, before a new pipeline is run on the data.

Using a variety of tools as suggested above is currently the most accepted way to ensure high confidence CNVs from short read WGS. However, another possibility that should be considered is using additional technologies such as long read WGS. Long read sequencing of some of these individuals has already been done and so will be a good way to validate the findings from the short read WGS data. The limitation of this is that only a few individuals have been sequenced using long read technology and so one cannot expect to find all variants from in the large cohort that may not be present in those individuals. In the future if the number of long read sequenced individuals can be increased this would be a valuable way to validate CNVs found in this dataset.

Given the findings in our PCA, showing the difference between African regions, a larger more diverse cohort from each sub-Saharan region would be beneficial to characterise African CNVs more broadly. Additionally considering the batch effect problems that we encountered in this dataset, it would ideal if individuals could be recruited and sequenced together to minimise batch effects that may mask true differences of CNVs observed in African regions.

The multi-allelic CNVs found in this study had far greater maximum copy numbers than have been previously described, and we also saw hotspots of novel multi-allelic CNVs. This points to the fact that this group of variants warrants further study using additional tools or technologies, to better understand the landscape of multi-allelic CNVs in Africa. cn.MOPS is another read depth based tool like Genome STRiP that was used in the gnomAD-SV study to call multi-allelic CNVs. This tool would be a good additional tool to use for the calling of multi-allelic CNVs to compare to the Genome STRiP results for congruency. However, given the complex nature of these variants, they may require the design of a laboratory validation protocol to ensure that the findings of variants are true.

4.6 Concluding Remarks

Through the research undertaken in this PhD, the patient cohort with DDs was characterised and the diagnostic process evaluated. Baseline population CNVs from diverse African populations catalogued for use in the interpretation of WES. This work has provided much-needed baseline data to help improve the diagnosis of DDs in Africa.

References

Abel, H.J., Larson, D.E., Regier, A.A., *et al.* (2020) Mapping and characterization of structural variation in 17,795 human genomes, *Nature,* 583(7814), pp. 83-89.

Abyzov, A., Urban, A.E., Snyder, M., *et al.* (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing, *Genome Res,* 21(6), pp. 974-984.

Ali, S.A., Soo, C., Agongo, G., *et al.* (2018) Genomic and environmental risk factors for cardiometabolic diseases in Africa: methods used for Phase 1 of the AWI-Gen population cross-sectional study, *Glob Health Action,* 11(sup2), pp. 1507133.

Alkan, C., Coe, B.P. and Eichler, E.E. (2011) Genome structural variation discovery and genotyping, *Nat Rev Genet,* 12(5), pp. 363-376.

Almarri, M.A., Bergstrom, A., Prado-Martinez, J., *et al.* (2020) Population Structure, Stratification, and Introgression of Human Structural Variation, *Cell,* 182(1), pp. 189-199 e115.

American Psychiatric Association (2013) *Diagnostic and statistical manual of mental disorders.* 5th Edition edn. Washington DC: American Psychiatric Association.

Auton, A., Brooks, L.D., Durbin, R.M., *et al.* (2015) A global reference for human genetic variation, *Nature,* 526(7571), pp. 68-74.

Bacchelli, E., Cameli, C., Viggiano, M., *et al.* (2020) An integrated analysis of rare CNV and exome variation in Autism Spectrum Disorder using the Infinium PsychArray, *Sci Rep,* 10(1), pp. 3198.

Backenroth, D., Homsy, J., Murillo, L.R., *et al.* (2014) CANOES: detecting rare copy number variants from whole exome sequencing data, *Nucleic Acids Res,* 42(12), pp. e97.

Bartenhagen, C. and Dugas, M. (2016) Robust and exact structural variation detection with paired-end and soft-clipped alignments: SoftSV compared with eight algorithms, *Brief Bioinform,* 17(1), pp. 51-62.

Blanchard, D. (2018) *DRMAA Python Library*. Available at: https://pypi.org/project/drmaa/.

Bope, C.D., Chimusa, E.R., Nembaware, V., *et al.* (2019) Dissecting in silico Mutation Prediction of Variants in African Genomes: Challenges and Perspectives, *Front Genet,* 10, pp. 601.

Buchman, R.W. and Hazelhurst, S. (2015) *Genesis* (Version 0.3.0).

Bull, M.J. (2020) Down Syndrome, *N Engl J Med,* 382(24), pp. 2344-2352.

Byrska-Bishop, M.E., Uday, S., Zhao, X., *et al.* 2021. High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. bioRXiv.

Cameron, D.L., Di Stefano, L. and Papenfuss, A.T. (2019) Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software, *Nat Commun,* 10(1), pp. 3240.

Cameron, D.L., Schroder, J., Penington, J.S.*, et al.* (2017) GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly, *Genome Res,* 27(12), pp. 2050-2060.

Carmichael, N., Tsipis, J., Windmueller, G.*, et al.* (2015) "Is it going to hurt?": the impact of the diagnostic odyssey on children and their families, *J Genet Couns,* 24(2), pp. 325-335.

Carss, K.J., Arno, G., Erwood, M.*, et al.* (2017) Comprehensive Rare Variant Analysis via Whole-Genome Sequencing to Determine the Molecular Pathology of Inherited Retinal Disease, *Am J Hum Genet,* 100(1), pp. 75-90.

Chaisson, M.J.P., Sanders, A.D., Zhao, X.*, et al.* (2019) Multi-platform discovery of haplotype-resolved structural variation in human genomes, *Nat Commun,* 10(1), pp. 1784.

Challman, T.D., Barbaresi, W.J., Katusic, S.K.*, et al.* (2003) The yield of the medical evaluation of children with pervasive developmental disorders, *J Autism Dev Disord,* 33(2), pp. 187-192.

Chang, C.C., Chow, C.C., Tellier, L.C.*, et al.* (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets, *Gigascience,* 4, pp. 7.

Charng, W.L., Karaca, E., Coban Akdemir, Z.*, et al.* (2016) Exome sequencing in mostly consanguineous Arab families with neurologic disease provides a high potential molecular diagnosis rate, *BMC Med Genomics,* 9(1), pp. 42.

Chen, K., Wallis, J.W., McLellan, M.D.*, et al.* (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation, *Nat Methods,* 6(9), pp. 677-681.

Chen, X., Schulz-Trieglaff, O., Shaw, R.*, et al.* (2016) Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications, *Bioinformatics,* 32(8), pp. 1220-1222.

Chiang, C., Layer, R.M., Faust, G.G.*, et al.* (2015) SpeedSeq: ultra-fast personal genome analysis and interpretation, *Nat Methods,* 12(10), pp. 966-968.

Choudhury, A., Aron, S., Botigue, L.R.*, et al.* (2020) High-depth African genomes inform human migration and health, *Nature,* 586(7831), pp. 741-748.

Choudhury, A., Ramsay, M., Hazelhurst, S.*, et al.* (2017) 'Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans', *Nat Commun*, 8(2062). DOI: 10.1038/s41467-017-00663-9 (Accessed 27 March 2018).

Christianson, A.L., Zwane, M.E., Manga, P.*, et al.* (2002) Children with intellectual disability in rural South Africa: prevalence and associated disability, *J Intellect Disabil Res,* 46(Pt 2), pp. 179-186.

Coe, B.P., Witherspoon, K., Rosenfeld, J.A.*, et al.* (2014) Refining analyses of copy number variation identifies specific genes associated with developmental delay, *Nat Genet,* 46(10), pp. 1063-1071.

Collins, R.L., Brand, H., Karczewski, K.J.*, et al.* (2020) A structural variation reference for medical and population genetics, *Nature,* 581(7809), pp. 444-451.

Conrad, D.F., Andrews, T.D., Carter, N.P.*, et al.* (2006) A high-resolution survey of deletion polymorphism in the human genome, *Nat Genet,* 38(1), pp. 75-81.

Conrad, D.F., Pinto, D., Redon, R.*, et al.* (2010) Origins and functional impact of copy number variation in the human genome, *Nature,* 464(7289), pp. 704-712.

Couper, J. (2002) Prevalence of childhood disability in rural KwaZulu-Natal, *S Afr Med J,* 92(7), pp. 549-552.

Cui, C., Shu, W. and Li, P. (2016) 'Fluorescence In situ Hybridization: Cell-Based Genetic Diagnostic and Research Applications', *Front Cell Dev Biol*, 4. DOI: 10.3389/fcell.2016.00089 (Accessed 9 March 2018).

D'Aurizio, R., Pippucci, T., Tattini, L.*, et al.* (2016) Enhanced copy number variants detection from whole-exome sequencing data using EXCAVATOR2, *Nucleic Acids Res,* 44(20), pp. e154.

D'Haene, E. and Vergult, S. (2021) Interpreting the impact of noncoding structural variation in neurodevelopmental disorders, *Genet Med,* 23(1), pp. 34-46.

da Rocha, J.E.B., Othman, H., Botha, G.*, et al.* (2021) The Extent and Impact of Variation in ADME Genes in Sub-Saharan African Populations, *Front Pharmacol,* 12, pp. 634016.

Danecek, P., Auton, A., Abecasis, G.*, et al.* (2011) The variant call format and VCFtools, *Bioinformatics,* 27(15), pp. 2156-2158.

Danecek, P., Bonfield, J.K., Liddle, J.*, et al.* (2021) Twelve years of SAMtools and BCFtools, *Gigascience,* 10(2).

De Coster, W. and Van Broeckhoven, C. (2019) Newest Methods for Detecting Structural Variations, *Trends Biotechnol,* 37(9), pp. 973-982.

Deciphering Developmental Disorders, S. (2015) Large-scale discovery of novel genetic causes of developmental disorders, *Nature,* 519(7542), pp. 223-228.

Dillon, O.J., Lunke, S., Stark, Z.*, et al.* (2018) Exome sequencing has higher diagnostic yield compared to simulated disease-specific panels in children with suspected monogenic disorders, *Eur J Hum Genet,* 26(5), pp. 644-651.

Dragojlovic, N., van Karnebeek, C.D.M., Ghani, A.*, et al.* (2020) The cost trajectory of the diagnostic care pathway for children with suspected genetic disorders, *Genet Med,* 22(2), pp. 292-300.

Durkin, M. (2002) The epidemiology of developmental disabilities in low-income countries, *Ment Retard Dev Disabil Res Rev,* 8(3), pp. 206-211.

Eggertsson, H.P., Jonsson, H., Kristmundsdottir, S., *et al.* (2017) Graphtyper enables population-scale genotyping using pangenome graphs, *Nat Genet,* 49(11), pp. 1654-1660.

Eggertsson, H.P., Kristmundsdottir, S., Beyter, D., *et al.* (2019) GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs, *Nat Commun,* 10(1), pp. 5402.

English, A.C., Salerno, W.J. and Reid, J.G. (2014) PBHoney: identifying genomic variants via long-read discordance and interrupted mapping, *BMC Bioinformatics,* 15, pp. 180.

Feuk, L., Carson, A.R. and Scherer, S.W. (2006) Structural variation in the human genome, *Nat Rev Genet,* 7(2), pp. 85-97.

Fieggen, K.J., Lambie, L.A. and Donald, K.A. (2019) Investigating developmental delay in South Africa: A pragmatic approach, *S Afr Med J,* 109(4), pp. 210-213.

Firth, H.V., Richards, S.M., Bevan, A.P., *et al.* (2009) DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources, *Am J Hum Genet,* 84(4), pp. 524-533.

Firth, H.V., Wright, C.F. and Study, D. (2011) The Deciphering Developmental Disorders (DDD) study, *Dev Med Child Neurol,* 53(8), pp. 702-703.

Flore, L.A. and Milunsky, J.M. (2012) Updates in the genetic evaluation of the child with global developmental delay or intellectual disability, *Semin Pediatr Neurol,* 19(4), pp. 173-180.

Flynn, K.A. (2020) *Evaluating whole exome sequencing on the Ion Torrent S5™ as a potential diagnostic tool for developmental disorders.* Masters Unpublished Degree type thesis or dissertation, University of Witwatersrand, Johannesburg [Online] Available at: https://hdl.handle.net/10539/31944 (Accessed.

Gardner, E.J., Lam, V.K., Harris, D.N., *et al.* (2017) The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology, *Genome Res,* 27(11), pp. 1916-1929.

Geiersbach, K.B., Gardiner, A.E., Wilson, A., *et al.* (2014) Subjectivity in chromosome band-level estimation: a multicenter study, *Genet Med,* 16(2), pp. 170-175.

Gordeeva, V., Sharova, E., Babalyan, K., *et al.* (2021) Benchmarking germline CNV calling tools from exome sequencing data, *Sci Rep,* 11(1), pp. 14416.

Grantham-McGregor, S., Cheung, Y.B., Cueto, S., *et al.* (2007) Developmental potential in the first 5 years for children in developing countries, *Lancet,* 369(9555), pp. 60-70.

Gurbich, T.A. and Ilinsky, V.V. (2020) ClassifyCNV: a tool for clinical annotation of copy-number variants, *Sci Rep,* 10(1), pp. 20375.

Gurdasani, D., Carstensen, T., Tekola-Ayele, F., *et al.* (2015) The African Genome Variation Project shapes medical genetics in Africa, *Nature,* 517(7534), pp. 327-332.

Handsaker, R.E., Korn, J.M., Nemesh, J.*, et al.* (2011) Discovery and genotyping of genome structural polymorphism by sequencing on a population scale, *Nat Genet,* 43(3), pp. 269-276.

Handsaker, R.E., Van Doren, V., Berman, J.R.*, et al.* (2015) Large multiallelic copy number variations in humans, *Nat Genet,* 47(3), pp. 296-303.

Haraksingh, R.R., Abyzov, A. and Urban, A.E. (2017) Comprehensive performance comparison of high-resolution array platforms for genome-wide Copy Number Variation (CNV) analysis in humans, *BMC Genomics,* 18(1), pp. 321.

Harris, P.A., Taylor, R., Minor, B.L.*, et al.* (2019) The REDCap consortium: Building an international community of software platform partners, *J Biomed Inform,* 95, pp. 103208.

Harris, P.A., Taylor, R., Thielke, R.*, et al.* (2009) Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support, *J Biomed Inform,* 42(2), pp. 377-381.

Hastings, P.J., Lupski, J.R., Rosenberg, S.M.*, et al.* (2009) Mechanisms of change in gene copy number, *Nat Rev Genet,* 10(8), pp. 551-564.

Hormozdiari, F., Hajirasouliha, I., McPherson, A.*, et al.* (2011) Simultaneous structural variation discovery among multiple paired-end sequenced genomes, *Genome Res,* 21(12), pp. 2203-2212.

Hormozdiari, F.H., Iman; Dao, Phuong; Hach, Faraz; Yorukoglu, Deniz; Alkan, Can; Eichler, Evan E.; Sahinalp, S. Cenk (2010) Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery, *Bioinformatics,* 26(12), pp. i350-i357.

International Schizophrenia, C. (2008) Rare chromosomal deletions and duplications increase risk of schizophrenia, *Nature,* 455(7210), pp. 237-241.

Iqbal, Z., Caccamo, M., Turner, I.*, et al.* (2012) De novo assembly and genotyping of variants using colored de Bruijn graphs, *Nat Genet,* 44(2), pp. 226-232.

Jakubosky, D., D'Antonio, M., Bonder, M.J.*, et al.* (2020a) Properties of structural variants and short tandem repeats associated with gene expression and complex traits, *Nat Commun,* 11(1), pp. 2927.

Jakubosky, D., Smith, E.N., D'Antonio, M.*, et al.* (2020b) Discovery and quality analysis of a comprehensive set of structural variants and short tandem repeats, *Nat Commun,* 11(1), pp. 2928.

Jehee, F.S., Takamori, J.T., Medeiros, P.F.*, et al.* (2011) Using a combination of MLPA kits to detect chromosomal imbalances in patients with multiple congenital anomalies and mental retardation is a valuable choice for developing countries, *Eur J Med Genet,* 54(4), pp. e425-432.

Kaminsky, E.B., Kaul, V., Paschall, J.*, et al.* (2011) An evidence-based approach to establish the functional and clinical significance of copy number variants in intellectual and developmental disabilities, *Genet Med,* 13(9), pp. 777-784.

Kamp, M., Krause, A. and Ramsay, M. (2021) Has translational genomics come of age in Africa?, *Hum Mol Genet,* 30(20), pp. R164-R173.

Karczewski, K.J., Francioli, L.C., Tiao, G.*, et al.* (2020) The mutational constraint spectrum quantified from variation in 141,456 humans, *Nature,* 581(7809), pp. 434-443.

Kaufman, L., Ayub, M. and Vincent, J.B. (2010) The genetic basis of non-syndromic intellectual disability: a review, *J Neurodev Disord,* 2(4), pp. 182-209.

Klambauer, G., Schwarzbauer, K., Mayr, A.*, et al.* (2012) cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate, *Nucleic Acids Res,* 40(9), pp. e69.

Korbel, J.O., Urban, A.E., Affourtit, J.P.*, et al.* (2007) Paired-end mapping reveals extensive structural variation in the human genome, *Science,* 318(5849), pp. 420-426.

Kosugi, S., Momozawa, Y., Liu, X.*, et al.* (2019) Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing, *Genome Biol,* 20(1), pp. 117.

Krause, A. (2019) New genetic testing technologies: Advantages and limitations, *S Afr Med J* 109(4), pp. 207-209.

Kromberg, J.G., Sizer, E.B. and Christianson, A.L. (2013) Genetic services and testing in South Africa, *J Community Genet,* 4(3), pp. 413-423.

Kronenberg, Z.N., Osborne, E.J., Cone, K.R.*, et al.* (2015) Wham: Identifying Structural Variants of Biological Consequence, *PLOS Comput Biol,* 11(12).

LaFramboise, T. (2009) Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances, *Nucleic Acids Res,* 37(13), pp. 4181-4193.

Landrum, M.J., Lee, J.M., Benson, M.*, et al.* (2018) ClinVar: improving access to variant interpretations and supporting evidence, *Nucleic Acids Res,* 46(D1), pp. D1062-D1067.

Lappalainen, I., Lopez, J., Skipper, L.*, et al.* (2013) DbVar and DGVa: public archives for genomic structural variation, *Nucleic Acids Res,* 41(Database issue), pp. D936-941.

Layer, R.M., Chiang, C., Quinlan, A.R.*, et al.* (2014) LUMPY: a probabilistic framework for structural variant discovery, *Genome Biol,* 15(6), pp. R84.

Ledergerber, C. and Dessimoz, C. (2011) Base-calling for next-generation sequencing platforms, *Brief Bioinform,* 12(5), pp. 489-497.

Lee, H.F., Chi, C.S. and Tsai, C.R. (2021) Diagnostic yield and treatment impact of whole-genome sequencing in paediatric neurological disorders, *Dev Med Child Neurol,* 63(8), pp. 934-938.

Levsky, J.M. and Singer, R.H. (2003) Fluorescence in situ hybridization: past, present and future, *J Cell Sci,* 116(Pt 14), pp. 2833-2838.

Li, Y., Roberts, N.D., Wala, J.A.*, et al.* (2020) Patterns of somatic structural variation in human cancer genomes, *Nature,* 578(7793), pp. 112-121.

Love, M.I., Mysickova, A., Sun, R.*, et al.* (2011) Modeling read counts for CNV detection in exome sequencing data, *Stat Appl Genet Mol Biol,* 10(1).

MacDonald, J.R., Ziman, R., Yuen, R.K.*, et al.* (2014) The Database of Genomic Variants: a curated collection of structural variation in the human genome, *Nucleic Acids Res,* 42(Database issue), pp. D986-992.

Mahmoud, M., Gobet, N., Cruz-Davalos, D.I.*, et al.* (2019) Structural variant calling: the long and the short of it, *Genome Biol,* 20(1), pp. 246.

Makela, N.L., Birch, P.H., Friedman, J.M.*, et al.* (2009) Parental perceived value of a diagnosis for intellectual disability (ID): a qualitative comparison of families with and without a diagnosis for their child's ID, *Am J Med Genet A,* 149A(11), pp. 2393-2402.

Malherbe, H.L., Aldous, C., Christianson, A.L.*, et al.* (2021) Modelled epidemiological data for selected congenital disorders in South Africa, *J Community Genet,* 12(3), pp. 357-376.

Mallick, S., Li, H., Lipson, M.*, et al.* (2016) The Simons Genome Diversity Project: 300 genomes from 142 diverse populations, *Nature,* 538(7624), pp. 201-206.

Manickam, K., McClain, M.R., Demmer, L.A.*, et al.* (2021) Exome and genome sequencing for pediatric patients with congenital anomalies or intellectual disability: an evidence-based clinical guideline of the American College of Medical Genetics and Genomics (ACMG), *Genet Med,* 23(11), pp. 2029-2037.

Manning, M., Hudgins, L. and Committee, P.P.a.G. (2010) Array-based technology and recommendations for utilization in medical genetics practice for detection of chromosomal abnormalities, *Genet Med,* 12(11), pp. 742-745.

Masri, A. and Hamamy, H. (2021) Cost Effectiveness of Whole Exome Sequencing for Children with Developmental Delay in a Developing Country: A Study from Jordan, *Journal of Paediatric Neurology*, pp. s-0040-1722265

Maulik, P.K., Mascarenhas, M.N., Mathers, C.D.*, et al.* (2011) Prevalence of intellectual disability: a meta-analysis of population-based studies, *Res Dev Disabil,* 32(2), pp. 419-436.

McCarroll, S.A. and Altshuler, D.M. (2007) Copy-number variation and association studies of human disease, *Nat Genet,* 39(7 Suppl), pp. S37-42.

McCarroll, S.A., Hadnott, T.N., Perry, G.H.*, et al.* (2006) Common deletion polymorphisms in the human genome, *Nat Genet,* 38(1), pp. 86-92.

McLaren, W., Gil, L., Hunt, S.E.*, et al.* (2016) The Ensembl Variant Effect Predictor, *Genome Biol,* 17(1), pp. 122.

Miller, D.T., Adam, M.P., Aradhya, S.*, et al.* (2010) Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies, *Am J Hum Genet,* 86(5), pp. 749-764.

Mithyantha, R., Kneen, R., McCann, E.*, et al.* (2017) Current evidence-based recommendations on investigating children with global developmental delay, *Arch Dis Child,* 102(11), pp. 1071-1076.

Mulder, N., Schwartz, R., Brazas, M.D.*, et al.* (2018) The development and application of bioinformatics core competencies to improve bioinformatics training and education, *PLOS Comput Biol,* 14(2), pp. e1005772.

Niehus, S., Jonsson, H., Schonberger, J.*, et al.* (2021) PopDel identifies medium-size deletions simultaneously in tens of thousands of genomes, *Nat Commun,* 12(1), pp. 730.

Nyangiri, O.A., Noyes, H., Mulindwa, J.*, et al.* (2020) Copy number variation in human genomes from three major ethno-linguistic groups in Africa, *BMC Genomics,* 21(1), pp. 289.

Olivier, L., Curfs, L.M. and Viljoen, D.L. (2016) Fetal alcohol spectrum disorders: Prevalence rates in South Africa, *S Afr Med J,* 106(6 Suppl 1), pp. S103-106.

Olusanya, B.O., Davis, A.C., Wertlieb, D.*, et al.* (2018) Developmental disabilities among children younger than 5 years in 195 countries and territories, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016, *The Lancet Global Health,* 6(10), pp. e1100-e1121.

Packer, J.S., Maxwell, E.K., O'Dushlaine, C.*, et al.* (2016) CLAMMS: a scalable algorithm for calling common and rare copy number variants from exome sequencing data, *Bioinformatics,* 32(1), pp. 133-135.

Parenti, I., Rabaneda, L.G., Schoen, H.*, et al.* (2020) Neurodevelopmental Disorders: From Genetics to Functional Pathways, *Trends in Neurosciences,* 43(8), pp. 608-621.

Pedersen, B.S., Brown, J.M., Dashnow, H.*, et al.* (2021) Effective variant filtering and expected candidate variant yield in studies of rare human disease, *NPJ Genom Med,* 6(1), pp. 60.

Pfundt, R., Del Rosario, M., Vissers, L.*, et al.* (2017) Detection of clinically relevant copy-number variants by exome sequencing in a large cohort of genetic disorders, *Genet Med,* 19(6), pp. 667-675.

Plagnol, V., Curtis, J., Epstein, M.*, et al.* (2012) A robust model for read count data in exome sequencing experiments and implications for copy number variant calling, *Bioinformatics,* 28(21), pp. 2747-2754.

Pos, O., Radvanszky, J., Buglyo, G.*, et al.* (2021) DNA copy number variation: Main characteristics, evolutionary significance, and pathological aspects, *Biomed J,* 44(5), pp. 548-559.

Prasad, A., Merico, D., Thiruvahindrapuram, B.*, et al.* (2012) A discovery resource of rare copy number variations in individuals with autism spectrum disorder, *G3 (Bethesda),* 2(12), pp. 1665-1685.

Qi, J. and Zhao, F. (2011) inGAP-sv: a novel scheme to identify and visualize structural variation from paired end mapping data, *Nucleic Acids Res,* 39(Web Server issue), pp. W567-575.

Quinlan, A.R., Clark, R.A., Sokolova, S., *et al.* (2010) Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome, *Genome Res,* 20(5), pp. 623-635.

Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features, *Bioinformatics,* 26(6), pp. 841-842.

Ramsay, M., Crowther, N., Tambo, E., *et al.* (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, *Glob Health Epidemiol Genom,* 1, pp. e20.

Rausch, T., Zichner, T., Schlattl, A., *et al.* (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis, *Bioinformatics,* 28(18), pp. i333-i339.

Redon, R., Ishikawa, S., Fitch, K.R., *et al.* (2006) Global variation in copy number in the human genome, *Nature,* 444(7118), pp. 444-454.

Riggs, E.R., Andersen, E.F., Cherry, A.M., *et al.* (2020) Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen), *Genet Med,* 22(2), pp. 245-257.

Ronaghi, M., Karamohamed, S., Pettersson, B., *et al.* (1996) Real-time DNA sequencing using detection of pyrophosphate release, *Anal Biochem,* 242(1), pp. 84-89.

Rothberg, J.M., Hinz, W., Rearick, T.M., *et al.* (2011) An integrated semiconductor device enabling non-optical genome sequencing, *Nature,* 475(7356), pp. 348-352.

RStudio Team (2020) *RStudio: Integrated Development for R* (Version 1.4.1717). Available at: https://www.rstudio.com/.

Rudolf, G., Tul, N., Verdenik, I., *et al.* (2017) Impact of prenatal screening on the prevalence of Down syndrome in Slovenia, *PLoS One,* 12(6), pp. e0180348.

Schuster, S.C., Miller, W., Ratan, A., *et al.* (2010) Complete Khoisan and Bantu genomes from southern Africa, *Nature,* 463(7283), pp. 943-947.

Schwarze, K., Buchanan, J., Taylor, J.C., *et al.* (2018) Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature, *Genet Med,* 20(10), pp. 1122-1130.

Sedlazeck, F.J., Rescheneder, P., Smolka, M., *et al.* (2018) Accurate detection of complex structural variations using single-molecule sequencing, *Nat Methods,* 15(6), pp. 461-468.

Sherry, S.T., Ward, M.H., Kholodov, M., *et al.* (2001) dbSNP: the NCBI database of genetic variation, *Nucleic Acids Res,* 29(1), pp. 308-311.

Shevell, M., Ashwal, S., Donley, D., *et al.* (2003) Practice parameter: evaluation of the child with global developmental delay: report of the Quality Standards Subcommittee of the American Academy of Neurology and The Practice Committee of the Child Neurology Society, *Neurology,* 60(3), pp. 367-380.

Short, P.J., McRae, J.F., Gallone, G.*, et al.* (2018) De novo mutations in regulatory elements in neurodevelopmental disorders, *Nature,* 555(7698), pp. 611-616.

Singleton, A.B. (2011) Exome sequencing: a transformative technology, *Lancet Neurol,* 10(10), pp. 942-946.

Smedley, D., Jacobsen, J.O., Jäger, M.*, et al.* (2015) Next-generation diagnostics and disease-gene discovery with the Exomiser, *Nat Protoc,* 10(12), pp. 2004-2015.

Smedley, D., Köhler, S., Czeschik, J.C.*, et al.* (2014) Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases, *Bioinformatics,* 30(22), pp. 3215-3222.

Srivastava, S., Love-Nichols, J.A., Dies, K.A.*, et al.* (2019) Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders, *Genet Med,* 21(11), pp. 2413-2421.

Srour, M. and Shevell, M. (2014) Genetics and the investigation of developmental delay/intellectual disability, *Arch Dis Child,* 99(4), pp. 386-389.

Stallings-Mann, M.L., Ludwiczak, R.L., Klinger, K.W.*, et al.* (1996) Alternative splicing of exon 3 of the human growth hormone receptor is the result of an unusual genetic polymorphism, *Proc Natl Acad Sci U S A,* 93(22), pp. 12394-12399.

StatisticsSA 2011. General Household Survey. Pretoria: Statistics South Africa.

Stavropoulos, D.J., Merico, D., Jobling, R.*, et al.* (2016) Whole Genome Sequencing Expands Diagnostic Utility and Improves Clinical Management in Pediatric Medicine, *NPJ Genom Med,* 1.

Strauss, K.A., Gonzaga-Jauregui, C., Brigatti, K.W.*, et al.* (2018) Genomic diagnostics within a medically underserved population: efficacy and implications, *Genet Med,* 20(1), pp. 31-41.

Strømme, P. (2000) Aetiology in severe and mild mental retardation: a population-based study of Norwegian children, *Dev Med Child Neurol,* 42(2), pp. 76-86.

Stuppia, L., Antonucci, I., Palka, G.*, et al.* (2012) Use of the MLPA assay in the molecular diagnosis of gene copy number alterations in human genetic diseases, *Int J Mol Sci,* 13(3), pp. 3245-3276.

Sudmant, P.H., Mallick, S., Nelson, B.J.*, et al.* (2015a) 'Global diversity, population stratification, and selection of human copy-number variation', *Science*, 349(6253), pp. 3761. DOI: 10.1126/science.aab3761 (Accessed 14 February 2018).

Sudmant, P.H., Rausch, T., Gardner, E.J.*, et al.* (2015b) An integrated map of structural variation in 2,504 human genomes, *Nature,* 526(7571), pp. 75-81.

Szot, J.O., Campagnolo, C., Cao, Y.*, et al.* (2020) Bi-allelic Mutations in NADSYN1 Cause Multiple Organ Defects and Expand the Genotypic Spectrum of Congenital NAD Deficiency Disorders, *Am J Hum Genet,* 106(1), pp. 129-136.

Talevich, E., Shain, A.H., Botton, T., *et al.* (2016) CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing, *PLOS Comput Biol,* 12(4), pp. e1004873.

Tan, T.Y., Dillon, O.J., Stark, Z., *et al.* (2017) Diagnostic Impact and Cost-effectiveness of Whole-Exome Sequencing for Ambulant Children With Suspected Monogenic Conditions, *JAMA Pediatr,* 171(9), pp. 855-862.

Teer, J.K. and Mullikin, J.C. (2010) Exome sequencing: the sweet spot before whole genomes, *Hum Mol Genet,* 19(2), pp. 145-151.

Tishkoff, S.A., Reed, F.A., Friedlaender, F.R., *et al.* (2009) The genetic structure and history of Africans and African Americans, *Science,* 324(5930), pp. 1035-1044.

Tran, N.Q.V. and Miyake, K. (2017) Neurodevelopmental Disorders and Environmental Toxicants: Epigenetics as an Underlying Mechanism, *Int J Genomics,* 2017, pp. 7526592.

Truty, R., Paul, J., Kennemer, M., *et al.* (2019) Prevalence and properties of intragenic copy-number variation in Mendelian disease genes, *Genet Med,* 21(1), pp. 114-123.

Turnbull, C., Scott, R.H., Thomas, E., *et al.* (2018) The 100 000 Genomes Project: bringing whole genome sequencing to the NHS, *BMJ,* 361, pp. k1687.

Tuzun, E., Sharp, A.J., Bailey, J.A., *et al.* (2005) Fine-scale structural variation of the human genome, *Nat Genet,* 37(7), pp. 727-732.

Uddin, M., Pellecchia, G., Thiruvahindrapuram, B., *et al.* (2016) Indexing Effects of Copy Number Variation on Genes Involved in Developmental Delay, *Sci Rep,* 6, pp. 28663.

United Nations Statistics Division *Standard country or area codes for statistical use (M49)*. New York: United Nations. Available at: https://unstats.un.org/unsd/methodology/m49/overview/ 2021).

Urban, M.F., Stewart, C., Ruppelt, T., *et al.* (2011) Effectiveness of prenatal screening for Down syndrome on the basis of maternal age in Cape Town, *S Afr Med J,* 101(1), pp. 45-48.

Usher, C.L. and McCarroll, S.A. (2015) Complex and multi-allelic copy number variation in human disease, *Brief Funct Genomics,* 14(5), pp. 329-338.

van Toorn, R., Laughton, B. and van Zyl, N. (2007) Aetiology of cerebral palsy in children presenting at Tygerberg Hospital, *S Afr J Child Health,* 1(2), pp. 74-77.

Vissers, L., van Nimwegen, K.J.M., Schieving, J.H., *et al.* (2017) A clinical utility study of exome sequencing versus conventional genetic testing in pediatric neurology, *Genet Med,* 19(9), pp. 1055-1063.

Vogler, C., Gschwind, L., Rothlisberger, B., *et al.* (2010) Microarray-based maps of copy-number variant regions in European and sub-Saharan populations, *PLoS One,* 5(12), pp. e15246.

Xie, H., Li, X., Peng, J.*, et al.* (2017) A complex intragenic rearrangement of ERCC8 in Chinese siblings with Cockayne syndrome, *Sci Rep,* 7, pp. 44271.

Yau, C. (2013) OncoSNP-SEQ: a statistical approach for the identification of somatic copy number alterations from next-generation sequencing of cancer genomes, *Bioinformatics,* 29(19), pp. 2482-2484.

Ye, K., Schulz, M.H., Long, Q.*, et al.* (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads, *Bioinformatics,* 25(21), pp. 2865-2871.

Zhang, F., Gu, W., Hurles, M.E.*, et al.* (2009) Copy number variation in human health, disease, and evolution, *Annu Rev Genomics Hum Genet,* 10, pp. 451-481.

Zhao, L., Liu, H., Yuan, X.*, et al.* (2020) Comparative study of whole exome sequencing-based copy number variation detection tools, *BMC Bioinformatics,* 21(1), pp. 97.

Zhao, M., Wang, Q., Wang, Q.*, et al.* (2013) Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives, *BMC Bioinformatics,* 14 Suppl 11, pp. S1.

Zhao, W.W. (2013) Intragenic deletion of RBFOX1 associated with neurodevelopmental/neuropsychiatric disorders and possibly other clinical presentations, *Mol Cytogenet,* 6(1), pp. 26.

Zhao, X., Li, C., Paez, J.G.*, et al.* (2004) An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays, *Cancer Res,* 64(9), pp. 3060-3071.

Appendix I

Human Genetics Retrospective Data Ethics Certificate

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

R14/49 Prof Amanda Krause et al

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M180506

NAME: (Principal Investigator)	Prof Amanda Krause et al
DEPARTMENT:	Human Genetics National Health Laboratory Service
PROJECT TITLE:	Using anonymised residual diagnostic samples and -file data for audit, research and development
DATE CONSIDERED:	25/05/2018
DECISION:	Approved unconditionally
CONDITIONS:	
SUPERVISOR:	
APPROVED BY:	Dr CB Penny, Chairperson, HREC (Medical)
DATE OF APPROVAL:	23/08/2018

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary on the Third Floor, Faculty of Health Sciences, Phillip Tobias Building, 29 Princess of Wales Terrace, Parktown, 2193, University of the Witwatersrand. 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 resubmit the application to the Committee. **I agree to submit a yearly progress report**. The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. In this case, the study was initially reviewed in **May** and will therefore be due in the month of **May** each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature

Date 05/10/2018

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

Appendix II

Emma Wiener PhD Study Ethics Certificate

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

R14/49 Drs E Wiener and N Carstens

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
CLEARANCE CERTIFICATE NO. M180885

NAME: (Principal Investigator)	Drs E Wiener and N Carstens
DEPARTMENT:	School of Patholgy Department of Human Genetics National Health Laboratory Service
PROJECT TITLE:	Establishing a baseline for developmental disorder diagnosis by evaluating current processes and mapping commmon benign copy number variation in Africa
DATE CONSIDERED:	Ad hoc
DECISION:	Approved unconditionally
CONDITIONS:	Sub-study under M160830
SUPERVISOR:	Professor A Krause and Dr Z Lombard
APPROVED BY:	Dr CB Penny, Chairperson, HREC (Medical)
DATE OF APPROVAL:	14/08/2018

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary on 3rd floor, Phillip V Tobias Building, Parktown, University of the Witwatersrand, Johannesburg.

I/We fully understand the conditions under which I am/we are authorised 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 resubmit 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 of the meeting when the study was initially reviewed. In this case, the study was initially reviewed in **2018/08/01** and will therefore reports and re-certification will be due early in the month of **2018/08/01** each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature

15/08/2018
Date

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

Appendix III

Phenotypic Category List

Phenotypic Categories

Aneuploidy

Autism

Congenital Anomalies

Connective Tissue

Craniofacial

Dermatological

Disorders of Sexual Differentiation

Dysmorphism

Failure to Thrive

Global Developmental Delay

Haematological

Metabolic

Neuromuscular

Other

Overgrowth

Seizures

Skeletal Dysplasia

Small Stature

Tumours

Vascular

Appendix IV

CBRL HIV Study Ethics Certificate

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

25 July 2018

Professor CT Tiemessen
SARCHI Chair of HIV Vaccine and Translational Research
NICD
NHLS
Sent by e-mail to: CarolineT@nicd.ac.za

Dear Professor Tiemessen

Re: **Protocol Ref No:** M140926
Protocol Title: *HIV-1 Positive South African elite and long term controllers: viral and host targets for functional cure strategies – paediatric and adult*
Principal Investigator: Professor CT Tiemessen

I refer to your e-mails of 3 and 15 July to the Chairperson of the HREC (Medical).

We note that three new projects will make use of the whole genome sequencing data that you have collected in the course of the study originally approved as M140926. This is approved. For the record, these projects are:

1. Design of a new genotyping array suitable for African populations (chip evaluation and imputation project)
2. Investigation of variation in the ADME genes in African populations
3. Control data for development disorders in African populations

We further note and approve of your proposal that details of any further projects which come along between now and 16 October 2019, when the current ethics clearance expires (or such extended date as may be agreed from time to time), which use the same genomic data, under the existing provisions of anonymity and confidentiality which you have described, need only be sent to us for purposes of notification, rather than approval *de novo*.

Thank you for keeping us informed and updated.

Yours Sincerely

I Burns

.....
Mr I Burns
For the Human Research Ethics Committee (Medical)

Research Office Secretariat:

Physical address: Phillip Tobias Building, 3rd Floor, Office 302, Corner York Road and Princess of Wales Terrace, Parktown, Johannesburg 2193.
Postal address: Private Bag 3, Wits 2050
Tel Nos. +27 (0)11-717-1234/2656/2700/1252
Office E-mail: HREC-Medical ResearchOffice@wits.ac.za
Website:http://www.wits.ac.za/research/about-our-research/ethics-and-research-integrity/

Appendix V

AWIGen Ethics Certificate

R14/49 Prof Michele Ramsay

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M170880

NAME: (Principal Investigator)	Prof Michele Ramsay
DEPARTMENT:	School of Pathology/Division of Human Genetics National Health Laboratory Services
PROJECT TITLE:	Wits: INDEPTH Partnership - Genomic and Enviromental Risk Factors for Cardiometabolic Disease in African Populations
DATE CONSIDERED:	03/10/2014(Initial Approval 24/08/2017)
DECISION:	Approved unconditionally
CONDITIONS:	Renewal for 5 Years Valid for the Period 01 September 2017 - 30 September 2022 (Previously M121029)
SUPERVISOR:	
APPROVED BY:	Professor CB. Penny Co-Chairperson, HREC (Medical)
DATE OF APPROVAL:	12/09/2017

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary in Room 10004,10th floor, Senate House/3rd floor, Phillip Tobias Building, Parktown, University of the Witwatersrand. I/We fully understand the conditions under which I am/we are authorised 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 resubmit to the Committee. **I agree to submit a yearly progress report**. The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. in this case, the study was initially review September and will therefore be due in the month of September each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

M. Ramsay
Principal Investigator Signature

Date 18 September 2017

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

Appendix VI

SAHGP Data Permission

Dr Z Lombard
NHLS & University of the Witwatersrand
South Africa

10 January 2019

Dear Dr Lombard,

Re: Request to access the SAHGP data – whole genome sequence data on 24 South African individuals – Request code: SAHGP015

Project title: Deciphering Developmental Disorders in Africa (DDD-Africa) - Evaluating Clinical Exome Sequencing in an African Setting

Thank you for your application and request.

The SAHGP Data Access Committee has reached the following conclusions:

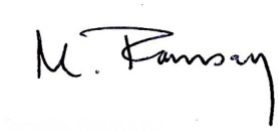
Your request is approved.

To gain access to the data, the next step is to complete the Data Access Agreement and have it signed by the relevant individuals. Please find the form attached to the email.

Once we receive the completed and signed agreement we will send a message to the European Genome Phenome Archive to contact you with regard to data transfer. Alternatively we can make the data available through the Wits cluster.

Please contact the undersigned should you require further clarification.

Yours sincerely,



Michele Ramsay

On behalf of the SAHGP Data Access Committee
Michele.ramsay@wits.ac.za

Appendix VII

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 **Emma Karin Wiener** (Student number: **603862**) am a student registered for the degree of **Doctor of Philosophy (Med)** in the academic year **2022**.

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: **25/02/2022**

Appendix VIII

Turnitin Report

Emma Wiener 603862 PhD Thesis - 25.02.22.docx

ORIGINALITY REPORT

9%	8%	6%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.ncbi.nlm.nih.gov Internet Source	1%
2	www.biorxiv.org Internet Source	1%
3	hdl.handle.net Internet Source	<1%
4	www.frontiersin.org Internet Source	<1%
5	www.nature.com Internet Source	<1%
6	link.springer.com Internet Source	<1%
7	Submitted to University of Witwatersrand Student Paper	<1%
8	academic.oup.com Internet Source	<1%
9	learning.ashg.org Internet Source	<1%