

https://dspace.org/

The Role of Copy Number Variants in the aetiology of developmental disorders in South Africa – a whole exome sequencing study

Louw, Nadja

2025

University of the Witwatersrand, Johannesburg

https://hdl.handle.net/10539/48453

Louw, Nadja. (2025). The Role of Copy Number Variants in the aetiology of developmental disorders in South Africa – a whole exome sequencing study [PhD thesis, University of the Witwatersrand, Johannesburg]. WIReDSpace. https://hdl.handle.net/10539/48453

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

The Role of Copy Number Variants in the aetiology of developmental disorders in South Africa – a whole exome sequencing study

Nadja Louw

2153705



Supervisors: Prof Zané Lombard

Dr Nadia Carstens

Declaration

I, Nadja Louw (2153705), declare that this Thesis is my own, unaided work, unless stated in 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 any other University.



Signature

12 May 2025

Date

Publications and presentations

- **Louw N**, Carstens N, Lombard Z; for DDD-Africa as members of the H3Africa Consortium. Incorporating CNV analysis improves the yield of exome sequencing for rare monogenic disorders-an important consideration for resource- constrained settings. Front Genet. 2023 Dec 14;14:1277784. doi: 10.3389/fgene.2023.1277784. PMID: 38155715; PMCID: PMC10753787.
- **The 4th DS-I Africa Consortium Meeting, Mauritius, 19 November 2024**, Oral presentation: CNV detection from exome sequencing data: Outcomes from the DDD- Africa cohort
- **South African Society of Huma Genetics (SASHG) Biennial Congress, South Africa, 29 October 2024,** Oral presentation: CNV detection from exome sequencing data: Outcomes from the DDD- Africa cohort
- **Wits Faculty of Health Sciences Research Day, 5 September 2024**, Oral presentation: CNV detection from exome sequencing data: Outcomes from the DDD- Africa cohort
- **Genomics of Rare Disease Conference, United Kingdom, 25 April 2023,** Poster and speed presentation: Improving detection of developmental disorders in resource constrained countries by incorporating copy number variant analysis
- **International Congress od Human Genetics (ICHG), South Africa, 24 February 2023**, Oral presentation: Copy number variant identification from exome sequencing data – a possible approach for African developmental disorder datasets?
- **Wits Faculty of Health Sciences Research Day 15 September 2022**, Poster presentation: The role of copy number variants in the aetiology of developmental disorders in South Africa – a whole exome sequencing study
- **19th H3Africa Consortium Meeting, Nigeria, 29 May 2022**, Oral presentation: The role of copy number variants in the aetiology of developmental disorders in South Africa - a whole exome sequencing study

Abstract

Developmental disorders are rare conditions, causing a mental or physical impairment. Genetic heterogeneity and highly variable clinical manifestations poses great challenges in the diagnosis of developmental disorders. New sequencing technologies have become an important and cost-effective tool in the diagnosis of rare diseases and results can impact clinical practice significantly. Copy number variants (CNVs) play a major role in the pathogenesis of developmental disorders and thus it is important to investigate the presence of CNVs within these patient cohorts. The introduction of exome sequencing (ES) has allowed the detection of CNVs and single nucleotide variants (SNVs) exome wide with a single test. This is a valuable approach to implement especially in a limited resource setting like South Africa. Despite SNVs being well studied, the incorporation of CNV bioinformatics tools into variant calling pipelines has not been implemented routinely as there is no current gold standard for CNV detection from exome data. Very limited clinical CNV studies have been carried out on next generation sequencing (NGS) data in patients of African ancestry. This study thus aimed to identify the most appropriate bioinformatics approach to detect CNVs from exome data. Subsequently, it was implemented in a developmental disorder variant analysis pipeline for ES data generated by the Deciphering Developmental Disorders in Africa (DDD-Africa) study, to establish the role that CNVs play in this African cohort. The DDD-Africa study recruited and performed detailed clinical phenotyping and ES on 500 African patients with an undiagnosed developmental disorder. Four different bioinformatics tools (CANOES, CLAMMS, XHMM and InDelible) have been applied to a set of samples (N=100) with known CNVs which served as a truth set. Functional equivalence evaluation was carried out in order to identify the most effective CNV calling tool or combination of tools to implement on the DDD-Africa exome dataset. Implementing the chosen tools onto the first batch of ES data consisting of 287 participants, yielded a ~7% diagnostic yield. Integrating CNV detection tools into a standard variant analysis pipeline from ES data can improve diagnostic yield while also promoting an improved cost benefit for ES. These results could not only end the diagnostic odyssey, but may lead to better care and management for families with developmental disorders in Africa.

Acknowledgements

I would like to thank my supervisors for continued support, assistance and mentorship. It has been a great learning curve and always helpful to have both of you as an example of great researchers and leaders in the field.

Thank you to Dr Phelelani T Mpangase for the help and patience with bioinformatic analysis during the entire project. To Dr David Twesigomwe for also assisting with coding and helping with the CNV overlapping regions.

To the entire DDD-UK team and especially Dr Eugene Gardner with assistance and helping interpret InDelible output and Dr Petr Danecek for assistance with the CNV truth set and data pertaining to this.

To the Faculty of Health Sciences Biostatistics department at the University of the Witwatersrand, a word of thanks for assistance with the functional equivalence evaluation.

To my family, especially my husband for all the love and support and for believing in me even in times that I did not. I will always be grateful for the opportunities which allowed me to be able to complete my PhD.

Research reported in this publication was supported by the National Institute of Mental Health of the National Institutes of Health under Award Number U01MH115483.

Table of Contents

Chapter 1: Introduction and literature review *1*

1.1 Deciphering Developmental Disorders **2**

1.2 Developmental disorders within Africa and other low- and middle income countries **3**

1.3 Diagnosing developmental disorders **4**

1.4 Genetic causes of developmental disorders **6**

1.5 Genetic testing for diagnosing DD caused by copy number variants **8**

1.5.1 Chromosomal Microarray for developmental disorder diagnosis 9

1.5.2 Next generation sequencing for developmental disorder diagnosis 11

1.5.3 Exome sequencing for combined detection of SNVs and CNVs 11

1.6 Bioinformatics approaches for CNV detection **13**

1.6.1 Read depth-based approach 16

1.6.2 Split read-based approach 17

1.6.3 Paired-end mapping approach 17

1.6.4 Assembly-based approach 18

1.6.5 The ensemble approach 18

1.7 Best approaches for CNV calling from ES data **18**

1.8 Value of CNV calling from ES in resource constrained countries **20**

1.9 CNV interpretation and classification **23**

1.10 Aim and objectives **24**

Chapter 2: Bioinformatics CNV detection and functional equivalence evaluation *25*

2.1. Introduction **26**

2.2. Methods **27**

2.2.1 Participants 28

2.2.2 Data generation and download 29

2.2.3 CNV calling from ES data 29

2.2.4 Calculating functional equivalence of CNV tools 38

2.3 Results **41**

2.3.1 Identification of the optimal method for CNV identification through functional equivalence evaluation 41

2.4 Discussion **45**

Chapter 3: Screening DDD-Africa exome data for pathogenic CNVs 48

3.1. Introduction 49

3.2. Methods 52

3.2.1 DDD-Africa participants 52

3.2.2 DDD-Africa data generation 53

3.2.3 Identifying shared CNVs from the different bioinformatics tools 54

3.2.4 Data analysis 54

3.2.5 Classifying CNVs according to ACMG and ClinGen guidelines 55

3.2.6 Comparison of manual classifications with CNV-ClinViewer outcome 58

3.2.7 ClinTAD tool for additional pathogenicity evidence 60

3.2.8 Copy number variant validation using Array CGH 62

3.3. Results 62

3.3.1 Applying chosen CNV calling tools to the DDD-Africa dataset 62

3.3.2 Classification of combined CNVs from all three tools 70

3.3.3 Additional analyses of shortlisted CNVs not meeting manual quality filtering parameters 76

3.4 Discussion 78

3.4.1 Comparison of CNVs identified from the DDD-Africa dataset 78

3.4.2 CNV classification and interpretation 81

Chapter 4: Study Conclusion 89

4 Conclusion 90

5 Challenges and study limitations 91

6 Future studies 93

7 References 95

8 Appendices 116

Appendix I: EGA data access agreement 117

Appendix II: DDD-Africa Ethics certificates 125

Appendix III: Truth set CNV's summary 128

Appendix IV: InDelible SV output files 130

Appendix V: XHMM CNV output file 134

Appendix VI: CANOES CNV output file 136

Appendix VII: CLAMMS CNV output file 138

Appendix VIII: Script created to identify overlapping CNV's between CNV tools 140

Appendix IX: Additional information for patients with LP/P CNVs identified 144

Appendix X: Website Links 148
Appendix XI: Plagiarism declaration 150
Appendix XII: Turnitin report 152

List of Tables:

Table 1.1: Comparison of common DD diagnostic methods 10

Table 1.2: Summary of bioinformatics tools developed for CNV detection from NGS data 15

Table 2.1: Advantages of the four CNV calling tools applied in this study 30

Table 2.2: Description of the variables used for sensitivity and specificity calculations 40

Table 2.3: Sensitivity and specificity values from the functional equivalence evaluation for the three CNV tools, CANOES, XHMM and CLAMSS..... 44

Table 3.1: Summary of CNVs called from the DDD-Africa dataset with each tool. 63

Table 3.2: Likely disease-causing CNVs identified with the three different bioinformatics CNV tools..... 72

List of Figures:

Figure 1.1: DDD-UK Study workflow (Wright et al., 2015)..... 5

Figure 1.2: Outline of structural variation including CNVs as large deletions and duplications. 6

Figure 1.3: Improvement and enhancement of CNV detection methods over time 7

Figure 1.4: Grey (60%)de novo mutations involved in severe DD..... 8

Figure 1.5: Illustration of the four main approaches used to detect CNVs from short-read NGS data. 14

Figure 1.6: Average molecular diagnostic yield of exome SNVs and combined SNV/CNVs. 22

Figure 2.1: Outline of methodology of this study 28

Figure 2.2: Implementation of default InDelible filtering parameters 32

Figure 2.3: Flowchart of CNV calling with XHMM 35

Figure 2.4: Overview of CANOES computational steps 37

Figure 2.5: Venn diagram showing overlapping and unique CNVs called by the three CNV tools. 41

Figure 2.6: Overlapping CNVs between different CNV tools for CNVs called from all 90 truth set samples (A) and from the 59 CMA validated truth set samples only (B). 42

Figure 2.7: Receiver Operating Characteristic (ROC) curves for CLAMMS, XHMM and CANOES 44

Figure 3.1: (A)Manual and (B)Web-based classification of CNVs..... 55

Figure 3.2: Criteria of the online scoring rubric from the CNV pathogenicity calculator for a CNV loss (A) and CNV gain (B)..... 57

Figure 3.3: Overview of CNV-ClinViewer with all features and output.. 59

Figure 3.4: Example cases on ClinTAD showing topologically associated domains 61

Figure 3.5: Proportion of deletions (DEL) and duplications (DUP) identified using the three CNV tools..... 63

Figure 3.6: Overlapping CNVs between the three CNV tools when implemented on DDD-Africa data. 65

Figure 3.7: (A) Average number of CNVs called per sample for each CNV tool and (B) Average size of the CNVs called for each tool. 65

Figure 3.8: Average size of CNVs identified by CLAMMS, CANOES and XHMM 66

Figure 3.9: Box and whisker plot of the number of CNVs called by the three different CNV tools for each sample. 67

Figure 3.10: Number of patients with specific number of CNV calls (ranging from 0, 1-10, 11-20, 21-50 and >50 number of CNVs)..... 67

Figure 3.11: Percentage CPU usage from each task of the CNV tools 69

Figure 3.12: Venn diagram showing all LP/P CNVs identified from the probands only as classified by CNV-ClinViewer 70

Figure 3.13: Number of CNVs classified as likely pathogenic/pathogenic, benign and VUS from all three CNV calling tools (CANOES, CLAMMS and XHMM) for all samples after classification with CNV-ClinViewer. 76

Abbreviations

ACMG – American College of Medical Genetics and Genomics

Array CGH – Array Comparative Genomic Hybridisation

AUC – Area Under the Curve

BAM – Binary Alignment Map

bp – base pairs

ClinGen – Clinical Genome Resource

CMA – Chromosomal Microarray

CNV – Copy Number Variant

CPU – Central Processing Unit

CRAM – Compressed Reference-oriented Alignment Map

DD – Developmental Disorder

DDD-Africa - Deciphering Developmental Disorders in Africa

DDD-UK – Deciphering Developmental Disorders in the United Kingdom

DDG2P – DD Genotype to Phenotype

DGV – Database of Genomic Variants

DNA – Deoxyribonucleic Acid

EGA – European Genome-phenome Archive

ES – Exome Sequencing

FISH – Fluorescent *In Situ* Hybridization

FN – False Negative

FP – False Positive

GB – Gigabyte

HI – Haploinsufficient

HMM – Hidden Markov Model

HPO – Human Phenotype Ontology

Kb – kilobase

LMIC – Low-Middle Income Country

LOEUF - loss-of-function observed/expected upper bound fraction

Mb – Megabase

MLPA – Multiplex Ligations-dependent Probe Amplification

NGS – Next Generation Sequencing

OMIM - Online Mendelian Inheritance in Man

pLI – Loss Intolerance probability

QC – Quality Control

RAM – Random Access Memory

ROC – Receiver Operating Characteristic

SNV – Single Nucleotide Variant

SV – Structural Variation

TAD – Topologically Associated Domain

TN – True Negative

TNR – True Negative Rate

TP – True Positive

TPR – True Positive Rate

TS – Triplosensitive

TSV – Tab-separated Values

VUS – Variant of Uncertain Significance

WES – Whole Exome Sequencing

WGS – Whole Genome Sequencing

Chapter 1: Introduction and literature review

Parts of this chapter were published as a mini review article in Frontiers in Genetics:

Louw, N., Carstens, N., Lombard, Z., & for DDD-Africa as members of the H3Africa Consortium (2023). Incorporating CNV analysis improves the yield of exome sequencing for rare monogenic disorders-an important consideration for resource-constrained settings. *Frontiers in genetics*, *14*, 1277784. https://doi.org/10.3389/fgene.2023.1277784

1.1 Deciphering Developmental Disorders

Developmental disorders (DD) are rare conditions, causing a mental or physical impairment. Children with DD experience a lag in fine- or gross motor, language, social, and cognitive developmental domains. The term global developmental delay is used where children below five years of age experience a delay in two or more of these developmental domains. It is estimated that DD affect 2-3% of the global population and over 50% of these cases are due to genetic causes (Moeschler and Shevell, 2014, Mithyantha et al., 2017). These disorders include Rett - (Hagberg et al., 1983), Angelman - (Angelman, 1965) , CHARGE - (Hall, 1979) and Noonan syndrome (Noonan, 1968) to name a few. DD are understudied and becoming more recognised in low- and middle income countries (LMICs) which can be partly attributed to the prevention and treatment of infectious diseases leading to a decrease in childhood mortality (Bitta et al., 2017). This leads to more children with DD being part of the community and thus seemingly more prevalent.

The genetic aetiology of DD include structural variations, single gene defects as well as chromosomal abnormalities and disruptions in the epigenetic process of methylation. DNA methylation does not alter the DNA sequence itself but rather regulates gene expression (Holliday and Pugh, 1975). DD are genetically heterogeneous, mutations at multiple distinct loci thus result in a similar phenotype. This heterogeneity together with highly variable clinical manifestations makes DD notoriously difficult to diagnose. A genetic diagnosis is of immense importance, given these challenges, not only for parents and caregivers, but also for doctors treating children with DD. This could have immediate implications for recurrence risk, clinical management and avoidance of unnecessary testing. Classic diagnostic testing for DD has a relatively poor yield and newer approaches can lead to a more precise diagnosis and cost less than the cascade of genetic tests required at present in order to diagnose DD.

To address this issue the Deciphering Developmental Disorders UK (DDD-UK) study was very successfully carried out in the United Kingdom where ~14000 patients with DD were recruited and ~41% received a genetic diagnosis (Firth et al., 2011).

Similarly, the DDD-Africa study (Appendix X) was established, modelled on the DDD-UK study, to provide a genetic diagnosis for children with DD in an African setting. The main aim of the DDD-Africa study is to improve the scope of genetic services in Africa by incorporating next generation sequencing (NGS) in order to diagnose DD more accurately. To achieve this objective, 500 African patients with DD as well as their parents were recruited and detailed clinical phenotyping carried out followed by exome sequencing (ES). No genetic diagnosis has been confirmed in these patients prior to recruitment.

1.2 Developmental disorders within Africa and other low- and middle income countries

It has been reported that about 250 million children within LMICs are at risk of DD (Lu et al., 2016). In sub-Saharan Africa the number of children younger than five years of age with DD rose from 8.6 million in 1990 to almost double (14.6 million) in 2016 (Olusanya et al., 2018). A study carried out in Nigeria has also seen a steep increase in cases of DD seen in clinic from 153 cases in 2017 to 236 cases in 2019 (Adeniyi and Adeniyi, 2020). There is an overall decline in the under-five childhood mortality rate; however, over 50% of these deaths occurred within sub-Saharan Africa. As mentioned in the above introduction, a large factor contributing to the decrease in childhood mortality rates is better control and management of infectious disease, especially within LMICs. This has resulted in an increased prevalence of DD in these countries, placing additional strain on already fragile healthcare system. In South Africa, about four million people are affected by rare disease (Moosa et al., 2022) but limited large-scale investigations into the aetiology and prevalence of genetically-based DD have been carried out in the country. The NeuroDev study shared data from a trio pilot study where exome data was analysed for children with DD in South Africa and Kenya (Kipkemoi et al., 2023). It was found that there is a need for a more collaborative effort to investigate DD within Africa (Lim et al., 2023). As introduced in section 1.1, the DDD-Africa study has been established to address these issues and to make diagnosing these conditions easier and more attainable for LMICs.

1.3 Diagnosing developmental disorders

Due to the heterogenous nature of DD and the fact that many phenotypes are non-specific, the genotype-first approach is recommended and has had successful outcomes for other studies (Wilczewski et al., 2023, Wright et al., 2023). Relatively increased accessibility and decreased cost of NGS have also led to more diagnoses for rare disease and new gene discoveries (Lohmann and Klein, 2014, Satam et al., 2023, Yang et al., 2023). First-line investigations for DD have been established; however, have changed over time as technologies and costs improve. A clinical evaluation of the patient is also very important in order to establish in-depth clinical phenotypes which will assist with diagnosis after the genotype is known. Metabolic and biochemical blood and urine testing should also be considered for children with DD (Merino Elia and Coghill, 2021) especially if blood testing at birth was not performed. This would not be attainable in LMICs where many of these tests are not readily available and costly. Detailed medical and family history is also important to confirm a specific diagnosis. An outline of the genotype-first approach from the DDD-UK study can be seen in Figure 1.1. This highlights the importance of detailed family history in variant prioritisation and in-depth phenotypes for manual review of the genotype and matching this with the phenotype. The genotype driven approach also allows for identification of subsets of patients with similar disorders. A genotype-phenotype database is key in prioritising novel variants and establishing a baseline of rare variants and gene-disease associations (Wright et al., 2015). Identification of more individuals with a specific variant in the same gene, together with deep phenotyping, gives greater insight into the range of phenotypes associated with a particular gene. This reverse-phenotyping approach is invaluable to identify novel disease genes and determine a diagnosis (de Goede et al., 2016).

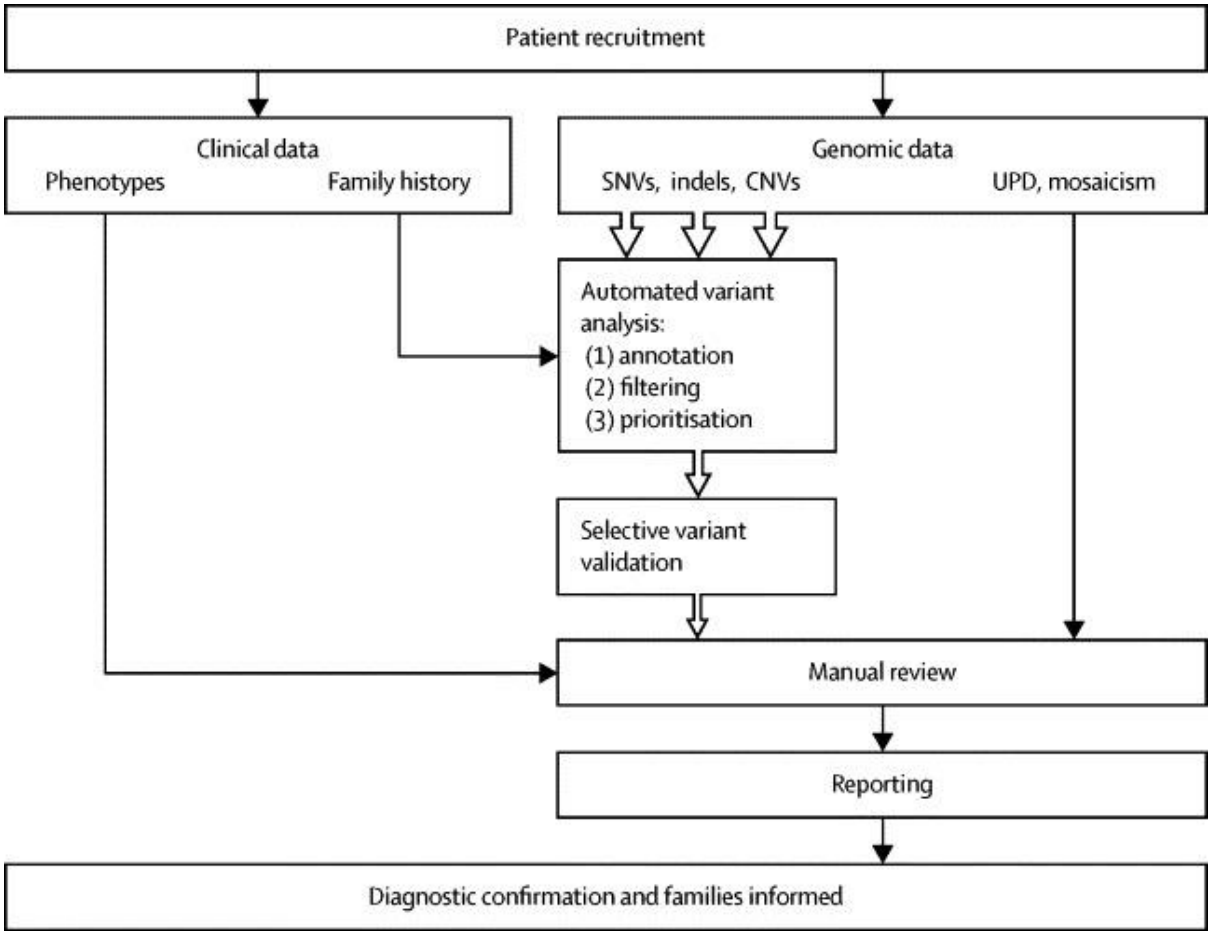


Figure 1.1: DDD-UK Study workflow (Wright et al., 2015)

SNV=single nucleotide variant. Indel=insertion or deletion. CNV=copy number variant. UPD=uniparental disomy.

A molecular diagnosis is important to not only reduce the burden on patients and families but may also lead to changes in the management of a patient. This also opens up the possibility for a patient with a rare disorder and their families to become part of support groups and other networks relating to their condition (Manickam et al., 2021). Identification of the appropriate clinical investigations and genetic testing for DD is thus of immense importance to make an accurate diagnosis and end the diagnostic odyssey.

1.4 Genetic causes of developmental disorders

Genetic variants associated with DD vary in size from a single nucleotide variant (SNV) to large duplications and deletions (up to several Mb) as well as complex rearrangements. SNVs are responsible for a large proportion of DD and have been quite well characterised in the DD context (Deciphering Developmental Disorders, 2015). These variants cause protein alteration due to either non-synonymous -, missense – or nonsense variants. Many tools and pipelines have been developed to identify and annotate SNVs and many functional studies have also been carried out on these variant types. It has been shown that copy number variants (CNVs) explain the pathogenesis of a large proportion of children with intellectual disability and ~15% attributed specifically to large CNVs (>100kb) (Testard et al., 2021). In line with this, previous reports have shown that most pathogenic CNVs identified in various patient cohorts are >100kb (Vermeesch et al., 2007, Cooper et al., 2011, Moosmann et al., 2015). CNVs also tend to have a more severe consequence on phenotype when compared to SNVs due to their large size and deletion or duplication of entire coding regions (Park et al., 2019). It is thus important to further investigate CNVs as a contributor to the pathogenicity of DD. CNVs are part of structural variation (Figure 1.2) which includes large deletions and duplications, spanning a region of over 50bp (previously over 1kb) and up to several mega base pairs (Alkan et al., 2011, MacDonald et al., 2014, Zarrei et al., 2019).

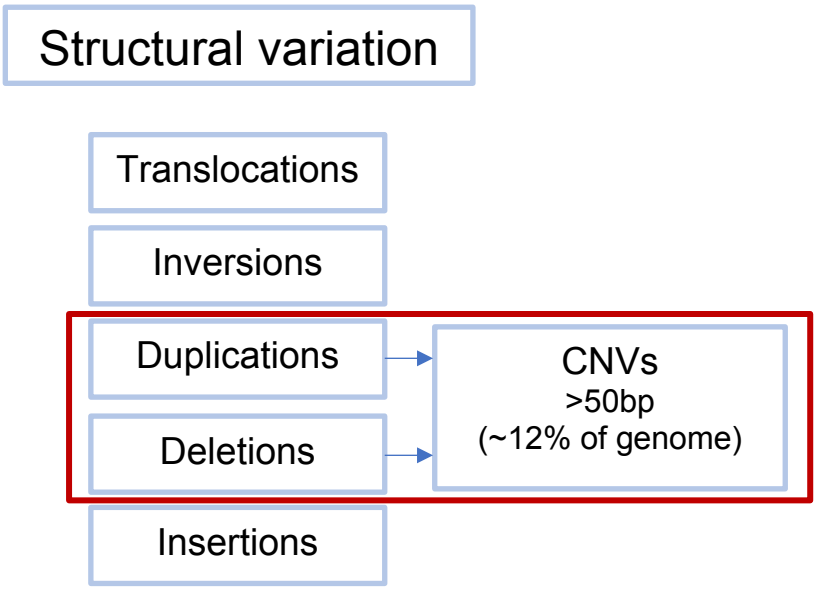


Figure 1.2: Outline of structural variation including CNVs as large deletions and duplications.

Large, unbalanced rearrangements change the diploid status of the locus. The gene dosage is thus altered which may also lead to a disease phenotype. It is very important to accurately identify CNVs as they are responsible for the largest portion (~1.5%) of per base genetic variation within the genome (Fromer et al., 2012, Tan et al., 2014). Over the years, CNV detection methods have improved due to increased resolution and detection capacity (Figure 1.3). This figure visually presents how detection methods have improved to detect a variety of variant lengths to finally allowing for genome-wide CNV detection with high-resolution since the 21st century. The hallmarks of CNV detection can be seen which ultimately allowed for development of CNV detection from NGS including ES, Whole Genome Sequencing (WGS) and long-read sequencing.

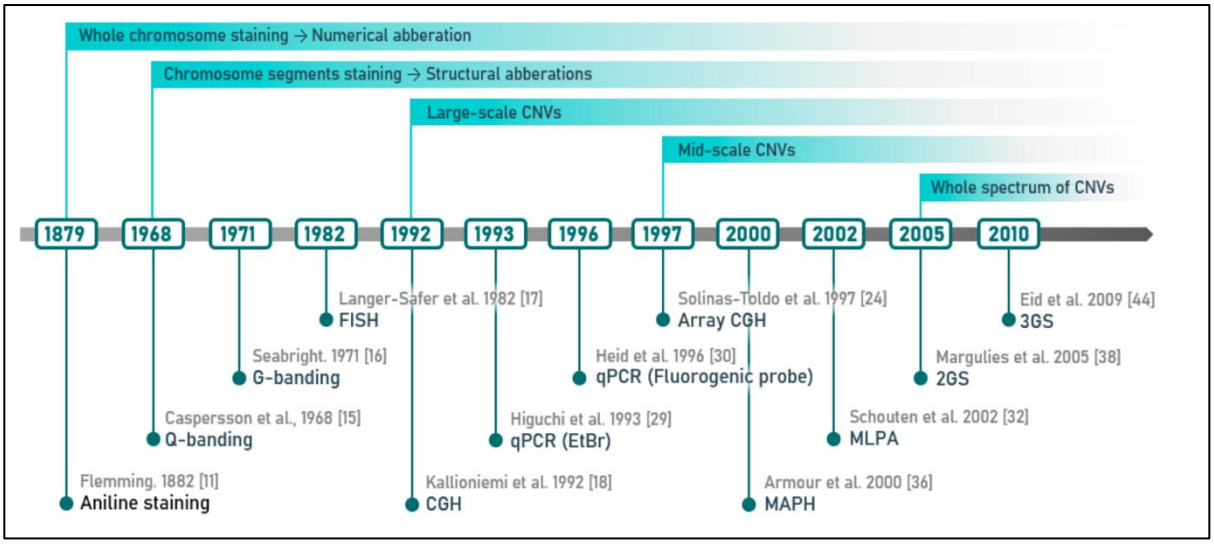


Figure 1.3: Improvement and enhancement of CNV detection methods over time. Figure from (Pös et al., 2021)

The frequency of CNVs in the human genome is estimated at approximately 12% (Zare et al., 2017). This includes recurrent variation which contribute to population diversity as well as pathogenic variants associated with the development of certain disease phenotypes (Redon et al., 2006, Zarrei et al., 2015, Pös et al., 2021). The majority of CNVs are thought to be inherited and less than 500kb in size (Zare et al., 2017). Previous studies (Gilissen et al., 2014) have reported that rare and *de novo* CNVs play a more notable role especially in neurological disorders (Figure 1.4). In the DDD-UK study, a total of 76% of the variants identified were indeed *de novo* and ~7% attributed to *de novo* CNVs (Wright et al., 2023, Danecek et al., 2024). This number is lower than previously reported as 85% of the DDD-UK cohort did receive prior array CGH testing.

There is thus a higher probability for larger CNVs to be pathogenic and the causal variant in patients with DD. Identification of clinically relevant CNVs are being prioritised in the hope of characterising phenotypic consequences thereof. Increasing the number of CNVs identified and reported ultimately improves our understanding of the disease aetiology of these variants.

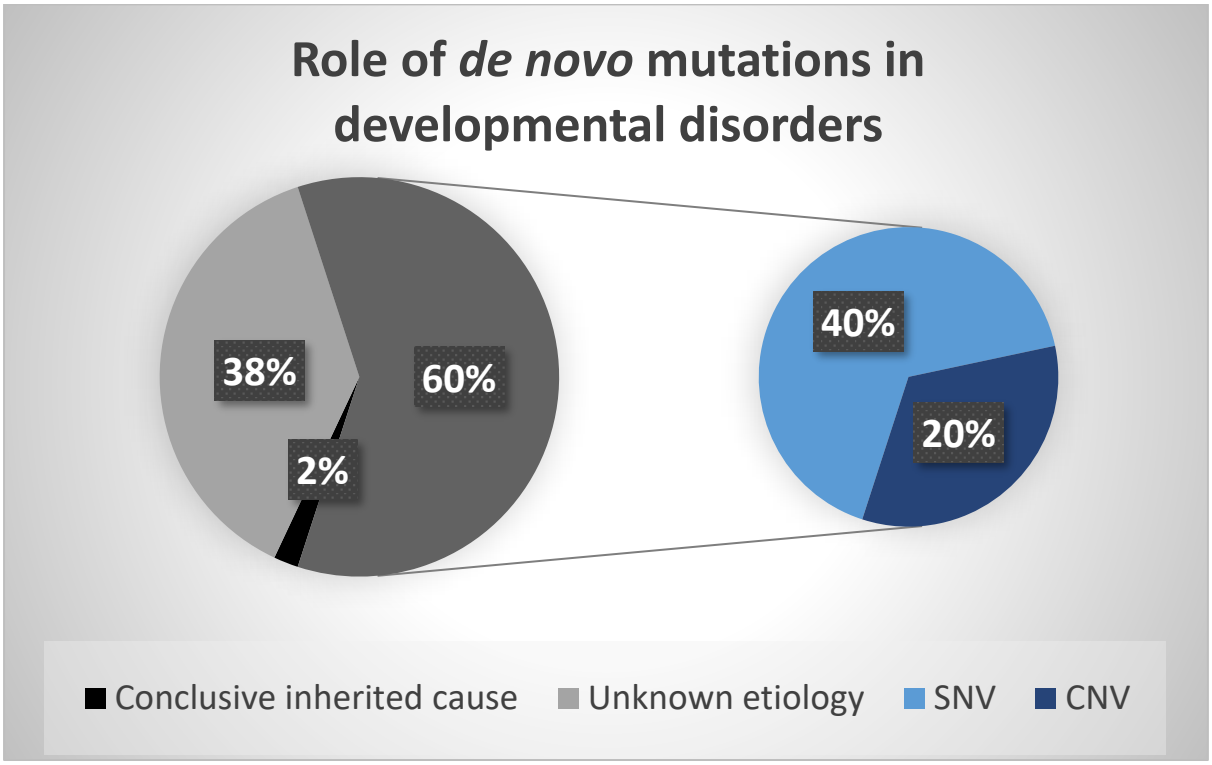


Figure 1.4: Grey (60%)*de novo* mutations involved in severe DD. *de novo* mutations are responsible for ~60% of severe DD/ID phenotype of which around 20% are predicted to be CNVs (Figure adapted from (Gilissen et al., 2014)).

The detection of pathogenic CNVs allows for an improved prognosis, better clinical management as well as more accurate genetic counselling for the patient and their family.

1.5 Genetic testing for diagnosing DD caused by copy number variants

Previously, more traditional testing methods have been used for genetic diagnosis of DD (Figure 1.3). These include cytogenetic testing, namely karyotyping as well as single-gene targeted Sanger sequencing and multiplex ligation-dependant probe amplification (MLPA).

These tests do not have a very high individual diagnostic yield and can be costly to do since a cascade of testing is usually carried out. Identifying SNVs has come a long way since implementation of NGS platforms.

The most widely used approach was Sanger sequencing which is quite laborious and where only one region of a limited size can be sequenced at a time. Newer technologies such as targeted NGS panels and exome/genome sequencing have allowed for a much better resolution in multiple genes or even genome-wide with a single test. Additionally, NGS allows for detection of SNVs and CNVs without additional assays and by using the same data. As discussed in section 1.4, a large proportion of the aetiology of DD can be attributed to CNVs and thus at present the recommended first-line investigation for the identification of genetic causes of DD is chromosomal microarray (CMA).

1.5.1 Chromosomal Microarray for developmental disorder diagnosis

Chromosomal microarray testing includes Array Comparative Genomic Hybridisation (Array CGH) (Kallioniemi et al., 1992) and single nucleotide polymorphism (SNP) arrays (LaFramboise, 2009) and has been the recommended first-tier test for DD diagnoses as it showed a much higher yield than previous methods (Miller et al., 2010, Jang et al., 2019). This increased yield can mainly be attributed to better resolution thus improving the sensitivity for identifying sub microscopic deletions and duplications. CMA can detect losses or gains genome-wide using a single chip, including sub microscopic changes which are too small to be detected with previous methods like fluorescent in-situ hybridisation (FISH). It can also characterise breakpoints of CNVs more accurately than MLPA. The diagnostic yield of CMA for DD is 15-20% (Table 1.1) which is a significant increase from a conventional G-banded karyotype offering ~3% diagnostic yield (excluding Down syndrome and other recognisable disorders) (Miller et al., 2010).

Although CMA is still being used in many laboratories as standard practice, newer sequencing technologies specifically NGS offer technical advantages thus being more appropriate and cost-effective when compared to CMA.

Table 1.1: Comparison of common DD diagnostic methods

Method	Advantages and limitations	Diagnostic yield
Karyotyping	Large rearrangements (3-10Mb), low throughput, not able to detect small intragenic rearrangements.	~3% (Miller et al., 2010)
FISH	detects balanced rearrangements and mosaicism, cannot detect small rearrangements (<100kb).	2.5-6% (Biesecker, 2002, Ravnan et al., 2006)
MLPA	Detects small rearrangements, limited number of targets (~40), high throughput, cannot detect copy neutral loss, only known targets – no discovery.	~5% (Miclea et al., 2021)
CMA	Small and large CNV detection, only reporting CNVs >200kb, probe locations predefined, difficulty detecting novel and rare CNVs.	15-20% (Miller et al., 2010, D'Arrigo et al., 2016)
Exome sequencing	Detect SNVs and CNVs within protein coding exonic regions, cannot determine exact breakpoints, higher depth of coverage than WGS.	25-53% (Dharmadhikari et al., 2019, Srivastava et al., 2019, Dong et al., 2020, Wright et al., 2023)
Whole genome sequencing	Detect SNVs and CNVs genome-wide, costly, high resolution, low coverage.	30-50% (Gilissen et al., 2014, van der Sanden et al., 2023)

1.5.2 Next generation sequencing for developmental disorder diagnosis

Instead of focusing on single genes or regions, with NGS, multiple genes or even the entire genome can be studied to identify disease-causing variants. This allows the detection of different types of variation with one test instead of ordering a cascade of testing which is costly and requires more resources. Incorporating this approach for diagnosing DD has been particularly valuable due to the heterogenous nature both molecularly and phenotypically. It is thus vital to introduce a test which can detect a wider range of variation for diagnosing DD. Targeted gene panels are effective tools for the diagnosis of DD. Although limited to only a number of specific single genes, these panels have good sensitivity and are more cost-effective than exome – or genome sequencing (Mellone et al., 2022). The introduction of exome – and genome sequencing has; however, allowed combining the detection of CNVs and SNVs with a single test. ES has shown an increased coverage as well as an additional 30% genes being characterised and novel gene findings, demonstrating major benefits for ES to diagnose DD specifically (LaDuca et al., 2017). To this end, ES has become more widely used and is recommended as first-tier for diagnostic purposes, especially for children with suspected monogenic disorders (Monroe et al., 2016, Stark et al., 2016, Srivastava et al., 2019, Hu et al., 2020).

1.5.3 Exome sequencing for combined detection of SNVs and CNVs

Exome sequencing is a widely used molecular approach and is recommended as a first-tier test for diagnostic purposes for rare monogenic disorders (Stark et al., 2016, Hu et al., 2018, Srivastava et al., 2019). Although whole exome sequencing (WES) is the more widely used term, it only includes protein coding exonic regions, targeting <25% of all exons. The technically correct term is therefore ES instead of WES (Aspden et al., 2023). ES identifies variants within coding exonic regions and is predominantly centered around SNV discovery. Although ES only identifies variants in the protein coding exonic regions, it is estimated that over 85% of mutations identified in Mendelian disorders are within the exome (Guo et al., 2012). It is important to note that Fragile X syndrome, which is the most common inherited disorder linked to intellectual disability, cannot be detected by ES. It has been stated that Fragile X should be a second-tier test after NGS analysis (Zhang et al., 2022).

The overall diagnostic rate of ES is estimated at ~25% (Yang et al., 2013, Lee et al., 2014, Fung et al., 2020) and yields as high as 36% (Srivastava et al., 2019) and 41% (Dong et al., 2020, Wright et al., 2023) have been reported in patients with rare monogenic developmental disorders (Table 1.1). In studies with consanguineous patients involved, a yield of up to 86% has been reported (Hiz Kurul et al., 2022).

Due to recent computational advances it is possible to incorporate exome-based CNV detection, making it more practical and cost effective, especially for disorders where both SNVs and CNVs are involved in disease aetiology. CNVs are responsible for the pathogenesis of up to 15% of rare monogenic cases (Truty et al., 2019, Testard et al., 2022) and tend to have a more severe consequence on phenotype when compared to SNVs due to their large size and effect on entire coding regions (Park et al., 2019). ES has proven to be a cost-effective first-tier test in developed countries, predicting a cost saving of between $1,484 and $3,242 per diagnosis (Schwarze et al., 2018). The costs of implementing a diagnostic exome within LMICs is still thought to be higher than in developed countries due to the lack of established infrastructure, high cost of reagents and the need for personnel training (Wiener et al., 2023). This being said, studies show that traditional genetic testing and pre-ES investigations can cost up to six times more than local ES costs (Cordoba et al., 2018, Masri and Hamamy, 2021).

Progress has been made regarding joint SNV and CNV investigations in LMICs; however, the gold standard for CNV detection remains CMA despite its inability to detect SNVs or smaller insertions and deletions. Despite the advances of ES, it is still not routinely used, especially in countries where genetic testing is limited. ES as a first-line investigation would be beneficial for many patients and a worthwhile investment in resource limited setting (Wiener et al., 2023). It is thus important to explore ways to incorporate CNV analysis into ES pipelines. As previously noted, CMA is frequently employed, leading to additional testing costs, and is limited in its ability to report CNVs of smaller sizes.

On its own CMA cannot detect SNVs or smaller insertion and deletions, leading to additional genetic testing to rule out any other plausible disease-causing variants. This will, in turn, inflate the costs of obtaining a possible diagnosis.

It is thus impractical to apply microarray only especially in resource limited settings and for diseases where both SNVs and CNVs play an important role in the pathogenesis (Park et al., 2019).

When comparing diagnostic rates of WGS, an increased yield of only 1-2% have been shown (Wright et al., 2018, Barbitoff et al., 2020). Considering the cost implication as well as the additional bioinformatics analysis required, the feasibility of WGS has not been proven. Implementing ES can overcome some of these issues as it allows for detection of both SNVs and CNVs simultaneously when applying exome data CNV calling. It can also detect large and small CNVs thus eliminating the need for tests like MLPA and CMA (Royer-Bertrand et al., 2021). Implementing CNV calling from ES data requires incorporating bioinformatics CNV calling tools, developed specifically for this type of data. Several tools have been developed to this end and are being used successfully to compliment the SNV calling pipeline.

Taking the above into account, a critical assessment of literature reflecting current standard approaches for CNV analysis and incorporation of bioinformatics CNV tools into existing NGS pipelines is thus necessary. This is even more crucial within a LMIC as a proof of principle due to low resource settings facing even more challenges to implement genetic testing. Keeping costs to a minimum while maximising diagnoses is crucial within these settings. Implementing ES as a first-tier test while incorporating CNV analysis could be a fundamental step in furthering genetic testing within low resource settings.

1.6 Bioinformatics approaches for CNV detection

Many bioinformatics tools have been developed for the identification of CNVs from genome and ES data. The four main approaches to detect CNVs (Figure 1.5) are read depth, paired-end mapping, split read, assembly approach (Zhao et al., 2013). A combination or ensemble approach is also commonly used as none of the methods alone detect all CNVs with high specificity and sensitivity.

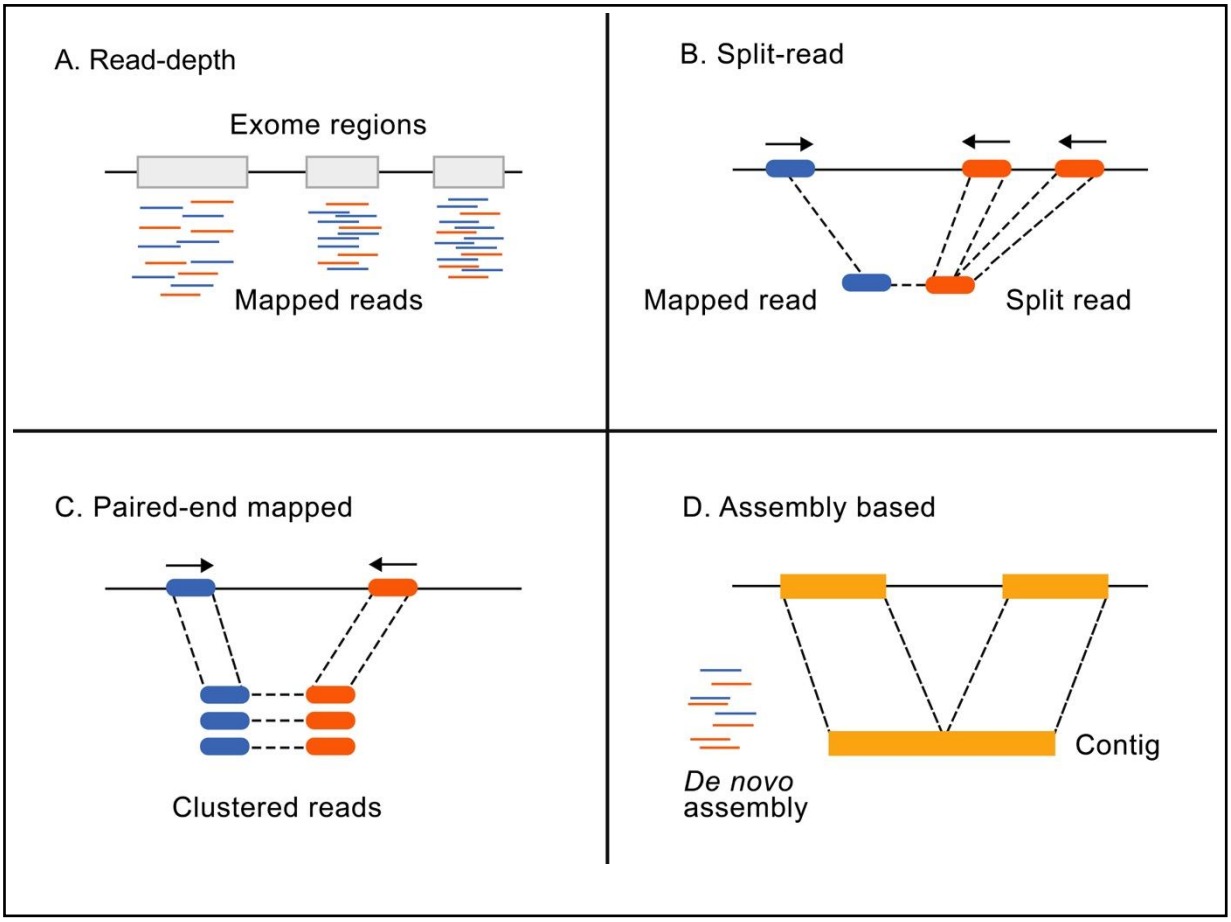


Figure 1.5: Illustration of the four main approaches used to detect CNVs from short-read NGS data. Adapted from (Zhao et al., 2013)

In this section, the focus is on the most recent and most widely used CNV tools (Gabrielaite et al., 2021) divided into categories according to these different detection approaches (Table 1.2).

Table 1.2: Summary of bioinformatics tools developed for CNV detection from NGS data

Tool	Method	Advantages/Attributes	Reference
Tools designed for exome sequencing			
CANOES	Read depth	Rare CNVs from ES data	Backenroth et al., 2014
CLAMMS	Read depth	Large population ES studies, easy integration into automated variant-calling pipeline	Packer et al., 2016
CN-Learn	Read depth	Small sample size needed, good precision and high-confidence	Pounraja et al., 2019
CODEX	Read depth	Do not require matched normal controls for normalisation	Jiang et al., 2015
CoNIFER	Read depth	More insertions identified, low memory required.	Krumm et al.,2012
EXCAVATOR2	Read depth	Detects CNVs with genome-wide resolution	D'Aurizio et al., 2016
ExomeDepth	Read depth	Good control of technical variability between samples, effective across wider range of exome datasets. Small and heterozygous deletions	Plagnol et al., 2012
HMZDelFinder/ HMZDelFinder _opt	Read depth	Rare, intragenic homozygous and hemizygous deletions	Gambin et al., 2017
XHMM	Read depth	Explore novel classes of CNVs	Fromer et al., 2012
InDelible	Split read	Smaller SVs (21-500bps) missed by conventional methods	Gardner et al., 2020
Tools designed for whole genome sequencing			
Gustaf	Split read	Identifies size and location of dispersed duplications and translocations; 30-100bp and >500bp	Trappe et al., 2014
PINDEL	Split read	Large deletions and medium sized insertions	Ye et al., 2018
PRISM	Split read	SVs and precise breakpoints	Jiang et al., 2012
SVseq2	Split read	INDEL from low coverage WGS data, exact breakpoints	Zhang et al., 2012
BreakDancer	Paired-end	Variety of SVs including indels, inversions, and translocations	Chen et al., 2009
HYDRA	Paired-end	Diverse SVs, includes transposons and segmental duplications	Quinlan et al., 2010
PEMer	Paired-end	~3 kilobases or larger SVs	Korbel et al., 2009
Ulysses	Paired-end	High specificity over complete spectrum of variants	Gillet-Markowska et al., 2015

Tool	Method	Advantages/Attributes	Reference
Magnolya	Assembly based	No mapping of reads to reference genome, *de novo* CNV detection	Nijkamp et al., 2012
CNVer	Ensemble approach	Better mitigate the sequencing biases causing uneven local coverage and accurately predict CNVs	Medvedev et al., 2010
DELLY	Ensemble approach	Full spectrum of genomic rearrangements, including complex events.	Raesch et al., 2012
GenomeSTRIP	Ensemble approach	Whole genome discovery and genotyping of deletions	Handsaker et al., 2011
LUMPY	Ensemble approach	Increased sensitivity of SV detection	Layer et al., 2014
Tools designed for exome or whole genome sequencing			
Manta	Ensemble approach	Less computational time/space, intense large-scale SVs, medium-sized indels and large insertions	Chen et al., 2016
Cn.MOPS	Read depth	Larger CNVs identified. Fast average running time.	Klambauer et al., 2012
CNVkit	Read depth	<100kb CNVs, more effective for deletions	Eric et al., 2016
GATKgCNV	Read depth	Rare CNVs, Determine copy number biases and CNVs	Babadi et al., 2023

*Whole Genome Sequencing (WGS); *Structural variation (SV)

1.6.1 Read depth-based approach

This approach relies on the depth of coverage to estimate copy number from the corresponding genomic region. Higher depth of coverage at a specific region indicates a gain whereas a lower depth of coverage indicates a loss of copy number. Since ES has higher depth-of-coverage than WGS it is ideal for using this approach. Most tools developed to date for the identification of CNVs from ES data are thus based on this approach. CLAMMS (Packer et al., 2016), CoNIFER (Krumm et al., 2012), ExomeDepth (Plagnol et al., 2012), XHMM (Fromer et al., 2012), cn.MOPS (Klambauer et al., 2012) and GATKgCNV (Babadi et al., 2022) are amongst the most recent and often used read depth-based tools. As only the exonic regions are sequenced, some considerations need to be addressed for these tools to function optimally.

When using read depth-based CNV detection one should take into consideration that most tools require the use of a reference panel of samples. An assumption of the read depth approach is that reads are distributed uniformly across the genome, unfortunately this is not the case for ES. Reference samples are thus used to control for these biases created by regions of variable depth across exons by establishing a baseline for CNV calling which ensures that CNVs are called accurately. These samples should ideally be matched in terms of preparation and sequencing platform and even sequencing batch if possible to limit technical biases which might hinder CNV detection. Several tools require matched case-control samples as input; however, many tools use multiple test samples as a cohort to serve as reference samples for the analysis. The number of samples to be used ranges from less than ten to hundreds of samples, for instance cn.MOPS requires a minimum of six samples whereas XHMM a minimum of 50 samples. Several read depth-based tools have been developed and implemented on ES data (Krumm et al., 2012, Tan et al., 2014, Pfundt et al., 2017, Zhao et al., 2020, Gordeeva et al., 2021).

1.6.2 Split read-based approach

This approach detects unmatched read pairs, thus one read aligns to the reference genome while the other read fails to map or aligns only partially to the genome. This potentially identifies the breakpoints for CNVs. A few recent tools developed with this approach are PINDEL (Ye et al., 2018), PRISM (Jiang et al., 2012), SVseq2 (Zhang et al., 2012) and Gustaf (Trappe et al., 2014). One tool developed specifically for ES data is InDelible (Gardner et al., 2021) which was designed to target smaller structural variation (21 – 500bp) mostly missed with other CNV calling tools.

1.6.3 Paired-end mapping approach

This approach was the first to put forth the possibility of using NGS data to detect CNVs (Tuzun et al., 2005, Korbel et al., 2009). It relies on the insert size from the library preparation process and identifies any decreased insert size or swapped read directions between read pairs to identify a CNV or mobile element, insertions, inversions and tandem duplications.

In regions of low complexity containing segmental duplications, this approach seems to be limited. A number of tools have been developed with BreakDancer (Fan et al., 2014), HYDRA (Quinlan et al., 2010), PEMer (Korbel et al., 2009) and Ulysses (Gillet-Markowska et al., 2015) being the most widely used (Zhao et al., 2013, Gabrielaite et al., 2021).

1.6.4 Assembly-based approach

The assembly-based approach assembles reads *de novo* and does not align to a reference genome. Overlapping reads are assembled and these contigs are then compared to the reference genome, identifying regions with contradictory copy numbers. A minimum read coverage is required for tools based on this approach to be used successfully. The specific threshold for coverage is not well defined but can be inferred from read depth-based methods as it usually requires higher coverage than these tools. The most commonly used assembly based tool is Magnolya (Nijkamp et al., 2012).

1.6.5 The ensemble approach

None of the above-mentioned methods alone detects the full spectrum of CNVs with high sensitivity and specificity and thus it is recommended to use an ensemble approach (Coutelier et al., 2022). In this regard several tools have been developed to integrate multiple approaches and increase performance. These include DELLY (Rausch et al., 2012), LUMPY (Layer et al., 2014), Manta (Chen et al., 2016), CNVer (Medvedev et al., 2010) and GenomeSTRIP (Handsaker et al., 2011). Although this approach is recommended, there is still no gold standard for CNV detection especially from ES data.

1.7 Best approaches for CNV calling from ES data

As there are many tools available for CNV detection from ES data (Table 1.2), recommendations have been made focusing on the use of these tools for optimal results. In a recent comparative analysis of ES-focused CNV tools (Zhao et al., 2020) the recommendations for obtaining the best results were related to the specific dataset.

In terms of accuracy, it was recommended to use CNVkit (Talevich et al., 2016) if CNV size is small (<100kb), if CNV size is larger using cn.MOPS seems to be optimal. If the dataset presents with more insertions, using CoNIFER is recommended and CNVkit is seemingly the best for identifying deletions compared to duplications. If there is no prior knowledge on the dataset then using cn.MOPS and CoNIFER together is recommended (Guo et al., 2013, Zhao et al., 2020).

Different tools have been designed to obtain optimal sensitivity and specificity focused on rare or common CNVs as well as population size. Previous limitations, for instance only being able to identify CNVs spanning at least two or more exons, GC content or mapability biases as well as sequencing noise have been addressed (Pounraja et al., 2019, Bigio et al., 2020, Filer et al., 2021, Babadi et al., 2022). CLAMMS was developed to be more suitable for large population studies (Packer et al., 2016) and integrate more easily into an automated variant-calling pipeline. In order to more accurately identify rare and intragenic homozygous and hemizygous deletions, HMZDelFinder (Gambin et al., 2017) was developed and the newer HMZDelFinder_opt (Bigio et al., 2020) outperforms the older version in terms of accuracy and specifically identification of partial exon deletions. ExomeDepth has also been widely used (Marchuk et al., 2018, Rajagopalan et al., 2020, Zhai et al., 2021) and was designed to control for technical variability between samples. CANOES (Backenroth et al., 2014) is complementary to methods like XHMM and CoNIFER and accuracy can be improved when using CANOES in combination with one of these. CN-Learn identifies true CNVs with higher precision and recall rates without compromising performance even with as few as 30 samples (Pounraja et al., 2019). This tool uses CNVs predicted by four different CNV callers (CANOES, CODEX, XHMM and CLAMMS) which was found to enhance performance instead of using the tools as standalone methods.

Another study also merged results from CANOES and HMZDelFinder after each tool was applied separately (Dong et al., 2020). It was also suggested to combine GATKgCNV, LUMPY, DELLY and cn.MOPS which had the best recall and captures different CNVs (Gabrielaite et al., 2021). As LUMPY and DELLY have been developed for WGS data, GATKgCNV and cn.MOPS should be used with ES data.

In a recent study, CNVkit, XHMM, EXCAVATOR2, and ExomeDepth were used for ES based CNV calling in order to maximize the sensitivity and make ES a more powerful tool to diagnose neurodevelopmental disorders (Zhai et al., 2021).

While no additional sequencing costs are involved in the exome-CNV analysis, computational costs relating to additional data analysis should be considered. Comparing computational costs of exome CNV tools, the average expected central processing unit usage was 5.68GHz and an average of 267,55Mb of space was used for a 11.2Mb series of datasets with 100x coverage (Zhao et al., 2020). Implementing CANOES on 285 samples took ~6 minutes per sample using a 2.3 GHz central processing unit core (Backenroth et al., 2014) and for CLAMMS (Packer et al., 2016) an estimate of ~50MB random-access memory is required per process.

The ensemble approach clearly yields optimal results, while increasing the sensitivity and specificity of CNV detection (Välipakka et al., 2020) but might not always be feasible. Factors such as computational costs and resource availability should also be taken into account when deciding on a method of CNV detection. Individual implementation of CNV calling tools could still be helpful and lead to increased diagnostic yields but are largely influenced by the available computing infrastructure in specific environments as well as adequate representation of the different calling tools. CNV calling from ES data should be particularly attractive in resource constrained settings with reduced capital expenditure and infrastructure required.

1.8 Value of CNV calling from ES in resource constrained countries

In a previous study (Dong et al., 2020), an overall yield of 41.4% was reported by simultaneous analysis of SNV and CNV of which 12% is attributed to CNVs. In this study, both CANOES and HMZDelFinder were applied separately for CNV detection. Another study found SNV and exome-based CNV calling yielded an overall diagnostic rate of 58.8% of which diagnostic CNVs accounted for 17.6% (Xiang et al., 2021). A comprehensive method was used for CNV identification which included combining XHMM and principal component analysis with CNVKit.

Similarly, it was found that incorporating exome-based CNV detection into conventional SNV analysis for a single trio-ES test significantly improved the diagnostic rate (Zhai et al., 2021). When combining SNV and CNV analysis, an overall diagnostic yield of 54% was obtained which included 18.9% from CNV analysis alone. CNVs in this study were detected by CNVkit, XHMM, EXCAVATOR2, and ExomeDepth, the detected CNVs were all retained and annotated thereafter.

In an effort to identify the cause of congenital heart disease in 96 child participants from Nigeria, a combined approach was taken by making use of ES from patient and parents (where available) and performing XHMM CNV analysis on the data (Ekure et al., 2021). Ten percent of the cohort presented with pathogenic and likely pathogenic variants of which one was a large CNV contributing to the phenotype.

Assessing the genomic aetiology of autism spectrum disorder in India, ES yielded a diagnosis for 29,7% of individuals in total of which CNVs contributed 3%, CMA analysis carried out on the same cohort yielded 2.9% diagnoses (Sheth et al., 2023). Combined CNV and SNV analysis from ES data thus significantly increased the diagnostic yield compared to only using CMA (29.7% vs 2.9%). Detection of CNVs for children with DD in South Africa and Kenya was carried out in a pilot study (Kipkemoi et al., 2023) using GATK-gCNV (Babadi et al., 2022) and seqr (Pais et al., 2022). This yielded a diagnosis for ~22% of children of which ~10% were due to CNVs. The combined SNV and CNV analysis from the discussed literature has been shown to increase diagnostic yield by as much as 18% (Figure 1.6) which is an additional diagnosis for ~2 out of every 100 individuals. The average increased yield attributed to CNVs from the discussed research is 10.6% without additional testing costs involved.

Implementing ES as a first-tier for diagnosis, especially when incorporating CNV analysis, has proven to be efficient, cost-effective and can end the diagnostic odyssey for patients who would not have otherwise necessarily received a molecular diagnosis.

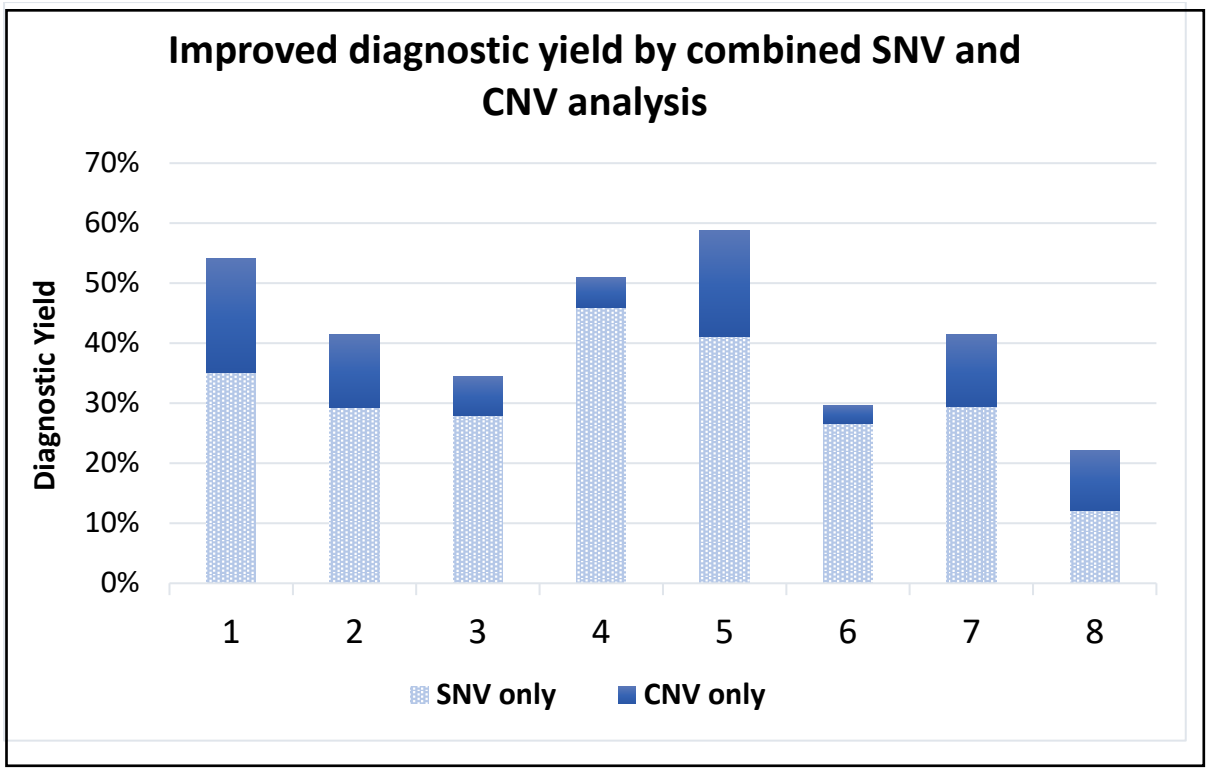


Figure 1.6: Average molecular diagnostic yield of exome SNVs and combined SNV/CNVs from 1. Zhai et al., 2021, 2. Truty et al., 2019, 3. Pranav et al., 2023, 4. Moosa et al., 2022, 5.Xiang et al., 2021, 6. Sheth et al., 2023, 7. Dong et al. 2020, 8. Kipkemoi et al., 2023.

As is the case for most resource limited settings, the cost of sequencing a trio and availability of both parents is always a limiting factor.

A study carried out in India (Pranav Chand et al., 2023), on children with neurodevelopmental delay found that a proband-only ES approach yielded a molecular diagnosis for 31.5% of these children. Addition of parental samples increased this yield by only 3% and CNVs contributed to 6.5% of these diagnoses. Another study conducted in China had an overall molecular diagnostic rate of 28.8% after analysing 1323 paediatric patients only which proved to be a relatively efficient and cost-effective approach in a developing country (Hu et al., 2018). A South African study (Moosa et al., 2022) found that proband-only ES is a very valuable tool for diagnosis, especially if CNV analysis is included. A molecular diagnostic yield of 51% was obtained with 46% of patients presenting with SNVs and 5% with CNVs. Even though trio-ES has been shown to have the best outcome for a positive molecular diagnosis (Wright et al., 2023), proband-only exome analysis has proven to be a feasible option for diagnosis in settings with limited resources or difficulty obtaining parental samples.

1.9 CNV interpretation and classification

Due to the large size of CNVs (>50bp) and many protein coding regions involved, classifying and interpreting CNVs are even more difficult compared to changes disrupting a single gene or base pair. The lack of representation in population frequency databases has made clinical interpretation and classification of CNVs more challenging especially in LMICs. An advantage of identifying CNVs in underrepresented populations is the expansion of variant representation in predominantly European focused public data repositories. Recent progress has been made to contribute CNVs from African population groups to variant databases (Nyangiri et al., 2020, Romdhane et al., 2021, Yilmaz et al., 2021) as the lack of diversity of high-quality genomic data, specifically from Africa, hampers the implementation of appropriate genetic services and exacerbates healthcare inequalities (Baine-Savanhu et al., 2023). Standardised CNV reporting is possible by using specific ACMG and Clinical Genome Resource (ClinGen) guidelines for CNV classification (Riggs et al., 2020) but careful evaluation of CNVs is encouraged to ensure that only likely disease-causing CNVs matching the patient phenotype are reported. This not only requires trained staff for CNV classification but also clinicians to carry out genotype-phenotype correlation. This process often required further input from a multi-disciplinary team which is not always available in resource constrained environments. Resolving VUSs remains challenging if population representation is inadequate. Furthermore, additional investigations including validation and functional experiments are often not available in LMIC laboratories. To alleviate some of these difficulties with CNV interpretation and classification, several tools have been developed enable a more convenient CNV annotation and interpretation. These tools provide support for annotation and/ or classification of CNVs and many are web-based, easy to use and freely available. A recent review has summed up these tools comprehensively to make it easier for the clinician, laboratory scientist and genetic counsellors to make a decision as to which tool would work best in their setting (Pös et al., 2021).

This study investigated different bioinformatics tools to identify the optimal approach in an African setting and evaluate the utility of this approach within a LMIC, the study aims and objectives are outlined below in more detail.

1.10 Aim and objectives

This study aimed to evaluate the optimal bioinformatics approach to identify CNVs from exome data for DD cohorts in a limited resource setting. Subsequently, this approach will be implemented in DD variant analysis pipeline for ES data generated by the DDD-Africa study.

Objectives:

1. Evaluate and compare available bioinformatics tools used for detecting CNVs from ES data for functional equivalence by comparing the sensitivity and specificity of each tool

2. Implementation of the chosen bioinformatics tool(s) in the DDD-Africa specific patient ES dataset in order to identify putative disease-causing CNVs

3. Annotation and interpretation of identified CNVs to assess pathogenicity using the ClinGen/ACMG guidelines

Chapter 2: Bioinformatics CNV detection and functional equivalence evaluation

2.1. Introduction

CNV detection from ES data relies mainly on depth of coverage data where the coverage is compared between samples and CNV predictions are made according to the increase or decrease in depth of coverage compared to reference samples. Biases can be introduced leading to false positive and false negative results specifically relating to depth of coverage methods. Difference in GC content of genomic regions, PCR amplification and sequencing errors can introduce biases which can affect the analysis and thus the outcome of results. In order to try and compensate for these biases, specific steps are in place for each tool to ensure the effects are minimised. Normalisation of the read depth data is a crucial step in each of these analysis tools to ensure the copy number is correctly interpreted.

As discussed in chapter 1, the recommendation is to use a combination of these tools (ensemble approach), as none of the methods alone calls CNVs with high sensitivity and specificity (Välipakka et al., 2020). A classical approach for optimised CNV detection from NGS data is to use multiple tools in parallel then compiling the data, keeping only overlapping CNVs from multiple tools (Collins et al., 2019). In this study, this was evaluated to determine whether using different tools really is the best approach for exome CNV calling on this dataset and whether using one tool versus different tools together presents with variable outcomes. To this end, the functional equivalence of these tools has been assessed. In this study, functional equivalence was defined as the shared property of two different tools, run on the same dataset, producing almost identical results. The sensitivity and specificity of the tools were assessed in order to measure functional equivalence and the number of true CNVs correctly identified by each tool were also evaluated.

Three CNV calling tools (CANOES (Backenroth et al., 2014), CLAMMS (Packer et al., 2016) and XHMM (Fromer et al., 2012)) and one structural variation (SV) calling tool (InDelible) (Gardner et al., 2021) were implemented and compared in this study. The same tools were also implemented in the DDD-UK study together with CoNVex (Deciphering Developmental Disorders, 2015) which was designed in-house and not available for use.

These calling tools are specifically designed for exome data and CANOES, CLAMMS as well as XHMM use the read depth method to call CNVs while InDelible was designed using the split read method. These tools all make use of different statistical models for CNV detection even though three tools use the read depth method for CNV detection. As mentioned previously, the read depth method is best for detection of CNVs from exome data. Each of the tools also has their own advantages which are discussed in more detail in this chapter.

Chapter 2 outlines the methods and results of the functional equivalence evaluation after these tools were applied to identify known CNVs from the truth set exome data and the decision on which tool or tools to implement further on the DDD-Africa data.

2.2. Methods

The outline of this study's methods can be seen below (Figure 2.1), with two different datasets, namely the truth set (phase 1) together with the functional equivalence calculation (phase 2) as well as the DDD-Africa dataset (phase 3). Phases 1-2 will be discussed in this chapter followed by a detailed discussion of phase 3 in chapter 3. During phase 1, all four CNV tools were applied to the truth set downloaded from the European Genome-phenome Archive (EGA) (Lappalainen et al., 2015) database to detect previously identified CNVs validated by more than one tool and/or CMA. Functional equivalence evaluation from the truth set results was carried out during phase 2 and the most appropriate tool or combination of tools identified. The truth set data consists of 100 samples with known CNVs, thus the desired outcome and the accuracy of the CNV calling tools could be predicted. Although several CNVs were identified in each sample, only one was included for the functional equivalence evaluation. This CNV was the only one validated as a true CNV by the DDD-UK study. All CNV calling tools were downloaded and installed using default settings and applied to the truth set in order to establish functional equivalence. The tool or tools presenting with the best sensitivity, specificity and identification of the known CNVs was applied during phase 3. The chosen approach was applied to the first batch of DDD-Africa exome data (287 samples) to identify new putative disease-causing CNVs.

These CNVs were classified using the ACMG/ClinGen guidelines (Riggs et al., 2020) as either pathogenic, likely pathogenic, VUS, likely benign or benign. Positive results (LP/P CNVs) were returned to families by a genetic counsellor on the DDD-Africa team. The prediction of the known CNVs from the truth set of data has thus informed the decision on which tool or combination of tools to implement on the DDD-Africa dataset.

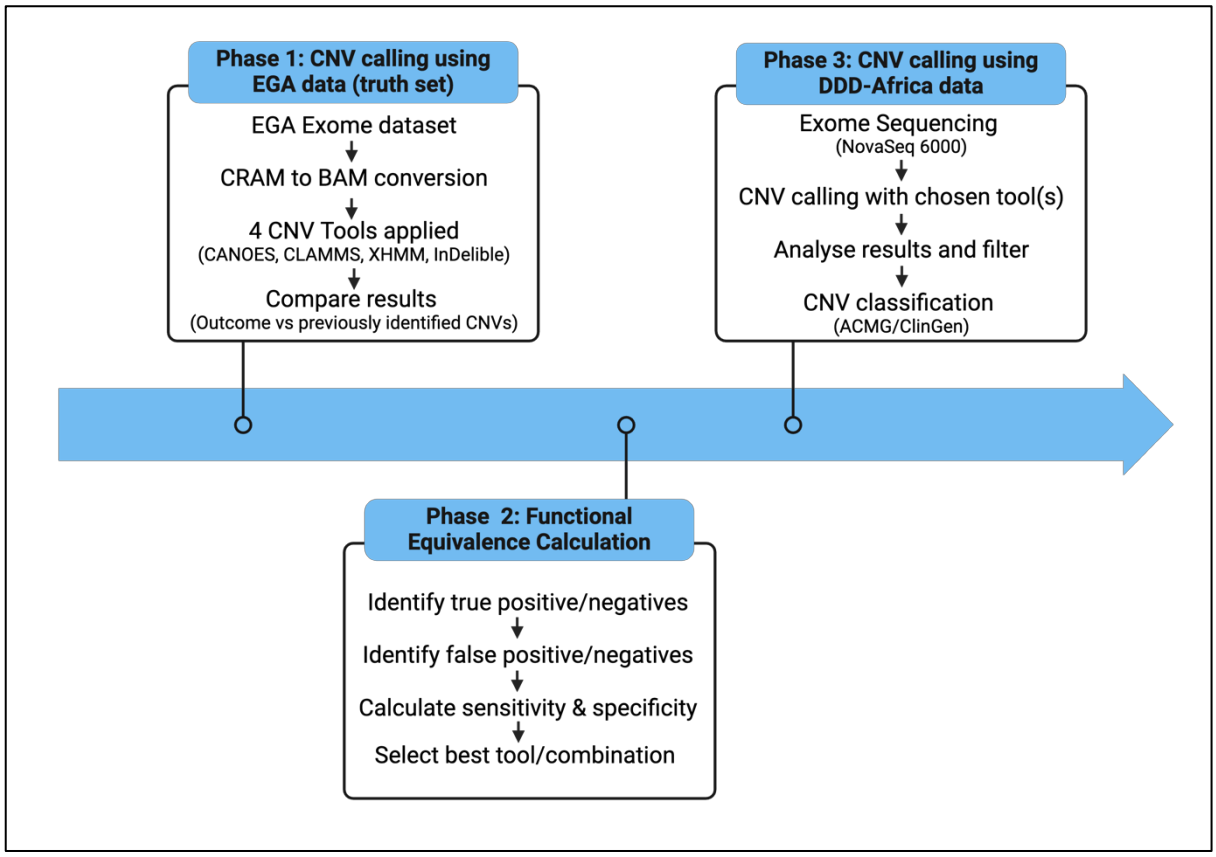


Figure 2.1: Outline of methodology of this study. Two datasets were used, namely the truth set downloaded from EGA and the DDD-Africa dataset which have undergone ES as part of the DDD-Africa project. Created with BioRender.com

2.2.1 Participants

Data from ES of one hundred DDD-UK samples (truth set) were obtained from the EGA database after specific permissions were obtained. Permission was obtained after successful application, to download 100 samples from the DDD-UK dataset (EGAD00001002748) consisting of 4293 trio samples (Appendix I). A sample size of 100 was chosen to include a similar number of affected individuals as the DDD-Africa dataset (117).

No personal or identifying information was obtained and the sample IDs consisted of only an EGA code. One hundred patient samples were downloaded for the purposes of a truth set to evaluate and compare the specificity and sensitivity of the four different bioinformatics tools. At the time of obtaining the data, only the codes of the probands with identified CNVs were given within a linked dataset as part of the larger DDD-UK study. The parental samples were not linked to these codes, thus only patient samples were downloaded to be used as part of the truth set. These samples all have CNVs identified through the DDD-UK study, thus serving as a truth set. One CNV per individual was included which were classified as pathogenic, likely pathogenic, likely benign or benign and were detected with more than one tool. These truth set CNVs were identified with at least two or more of the CNV calling tools applied in the DDD-UK study, which includes CLAMMS, CANOES, CoNVex and XHMM. A total of 59 (68.8%) of the CNVs were validated by CMA.

2.2.2 Data generation and download

The data for the truth set was generated by the DDD-UK study through 75-base paired-end sequencing (Illumina HiSeq) (Deciphering Developmental Disorders, 2015). One hundred Compressed Reference-oriented Alignment Map (CRAM) files, aligned to the older reference genome (hg19), were downloaded from the EGA database. These files were first converted to FASTQ files and were re-aligned to reference genome GRCh38 before converting to Binary Alignment Map (BAM) files using Samtools (Li et al., 2009). To obtain functional equivalence using the truth set, only CNVs located within autosomes were included as CANOES does not call CNVs on sex chromosomes. This led to the exclusion of ten of the 100 samples and thus comparison was done only with the remaining 90 truth set samples (90 CNVs). A summary of the 90 CNVs can be seen in Appendix III.

2.2.3 CNV calling from ES data

The four tools used during phase 1 have different advantages which are summarised in Table 2.1. In this section, these tools are discussed in further detail, specifically outlining the steps followed in order to call CNVs using exome data and the expected output from each tool.

Table 2.1: Advantages of the four CNV calling tools applied in this study

Tool	Advantages/Attributes	Reference
InDelible	Smaller SVs (21-500bp), specifically designed for DD data	Gardner et al., 2020
CANOES	Rare CNVs from ES data	Backenroth et al., 2014
CLAMMS	Large population ES studies, easy integration into automated variant-calling pipeline	Packer et al., 2016
XHMM	Explore novel classes of CNVs from ES data	Fromer et al., 2012

The different attributes of these tools, the different statistical methods implemented by each tool, and the previous application thereof in the DDD-UK study motivated the inclusion in this study. It is of interest to identify whether this would lead to different results for each tool or whether the tools show functional equivalence. Implementing a combination of these tools could also be complementary and lead to a higher sensitivity and thus an increased yield. Implementation of sophisticated bioinformatics tools is challenging and even more so within Africa and other LMICs (Mulder et al., 2016, Shaffer et al., 2019, Tibiri et al., 2025). This study would indicate whether one tool could be implemented instead of multiple tools if functional equivalence is proven.

A Nextflow workflow (Appendix X) was created for CNV calling using the four CNV calling tools. The CNV calling tools were containerised using Docker (Merkel, 2014).

- InDelible

InDelible was implemented as outlined on the GitHub repository (Appendix X). It was originally designed for the ascertainment of insertion-deletions (InDels) (>20bp) and SVs from ES data for which most other approaches have proven insufficient. A “target gene list” was also used by InDelible containing genes of interest for the end user which reduces search space. Variants identified are limited to the Developmental Disorders Genotype to Phenotype (DDG2P) gene list (Thormann et al., 2019).

The DDG2P list is a database of all known causative DD loci (Appendix X). Only variants with at least one breakpoint within the coding sequence of a gene on this list are called by InDelible. The variant calling process involves six different steps, namely Fetch, Aggregate, Score, Database, Annotate and *de novo*:

- Fetch: BAM files are inspected for reads where only part of the aligned sequence match the reference genome and others do not (split reads).
- Aggregate: Reads are clustered based on chromosome and position to identify positions with multiple split reads.
- Score: A random forest model is used to score reads using a truth set.
- Blast: The likely breakpoints of split reads are determined.
- Annotate: Split read clusters are annotated with population allele frequencies and whether they intersect with protein-coding genes.
- De novo: Inheritance status and likely *de novo* variants are determined by assessing clusters for presence or absence in parental samples.

During the first step (Fetch) soft-clipped reads (areas where split reads are identified) are extracted from the BAM files and then aggregated to identify locations with multiple clipped reads (Aggregate). A random forest model is then used to score positions according to the number/quality of clipped reads as well as sequence context. In the database step, sites across individuals are merged, breakpoint frequencies assigned as well as type of variant and breakpoints determined where possible. These putative SVs are then annotated and lastly *de novo* events are called where parental samples are available.

The output file from these steps is in the form of a tab-separated values file (TSV) with details of each SV identified by InDelible (Appendix IV). After the initial TSV output, additional filtering is applied in order to reduce variants and ensure only highly likely variants, specific to the genes of interest are called (Figure 2.2). The default filters were first applied as follows, minor allele frequency (maf) lower than 0,4 and average mapping quality (avg_mapq) of at least 20. The number of reads with both 5' and 3' split reads (pct_double_split) as well as the SV type (svtype) filters were applied as outlined below. This was done to ensure the type of SV is known (deletion or duplication) and is not a segmental duplication or translocation.

The breakpoints should intersect only genes included on the DDG2P gene list (Thormann et al., 2019) and the variant must be in an exonic region. Lastly, specific filters for parental structural rearrangements should also be applied, where available. These samples were not available for the truth set data and thus step 4 was omitted.

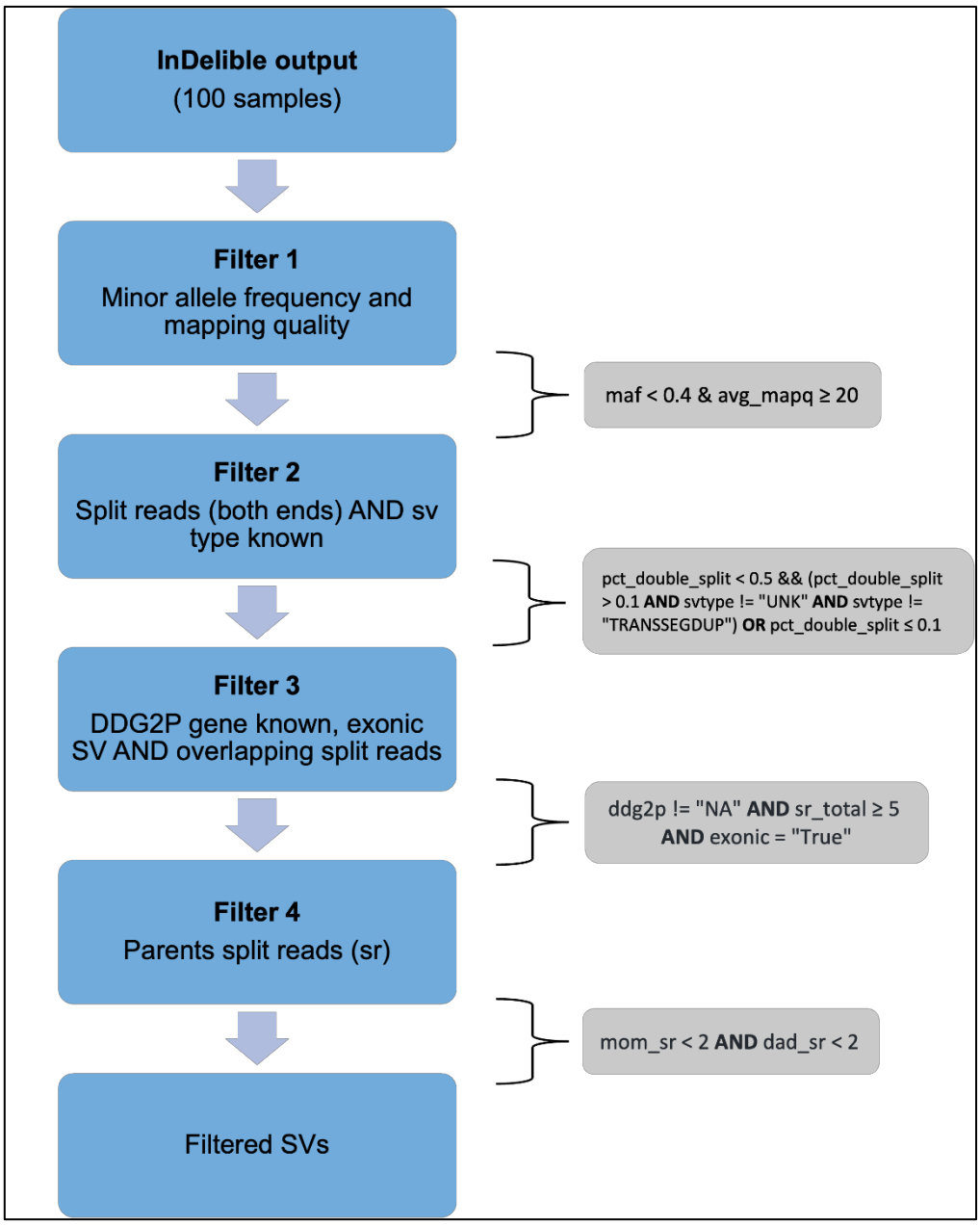


Figure 2.2: Implementation of default InDelible filtering parameters. Filter 1 was to include only SVs with minor allele frequencies (maf) less than 0,04 and average mapping quality >= 20 (avg_mapq). For filter 2 the SV type has to be known (not to be unknown or segmental duplication or translocation) and number of reads with both 5' and 3' split reads (pct_double_split) <0.5. Filter 3 ensured that only SVs in genes from DDG2P are included and only exonic SVs with total number of split reads in reads overlapping this breakpoint less than 5. Filter 4, not included in truth set, was to have mother and father samples' split reads less than 2 (number of split reads in the bam/cram provided to mother and father with the same 'position').

- XHMM

Exome Hidden Markov Model (XHMM) software was downloaded and all steps followed as on GitHub (Appendix X) and published (Fromer and Purcell, 2014). This tool is specifically focused, but not limited to, large cohorts and calling novel, rare CNVs (<5%) over 200kb in size. A reference FASTA file of the genome and a list of exome targets was needed for XHMM analysis. This tool was run with default settings and the outline of these steps can be seen in Figure 2.3. The most important steps are running depth of coverage calculations (steps 5-6) and removing of extreme targets and samples (steps 7-13), data normalisation (excluding samples with extreme variability in normalised depth – step 14-15), CNV calling per sample (using a hidden Markov model – step 18), and statistical genotyping (step 19) (Fromer et al., 2012). The 287 BAM files from the first batch of DDD-Africa data were divided into 57 groups for the coverage calculations to optimise computational resources by doing these in parallel. To ensure that targets with extreme GC content or read depth, low complexity and highly variable targets are excluded, the following filters are implemented. Targets were excluded if they were smaller than 10bp or larger than 10kb, if they had a mean depth <10 reads or >500 reads. Samples were also excluded if their mean depth over all targets were <24 or >200 reads, if they had extremely high variance and if mapping quality was <20. The output after running these steps with all quality scores is in a DATA.xcnv file format (Appendix V).

In order to further filter out false positive results, specific filtering parameters were put in place for the DDD-Africa dataset. These were not required for the truth set as the CNVs to be identified were known. Specific quality scores are given for each CNV which gives an indication of the reliability of the output and whether the CNV is a true event and not due to an artifact or false positive (Appendix V). The Q_some score is a phred-scaled quality score of some CNV within the interval and the Q_non_dip is a phred-scaled quality score of the sample not being diploid in the interval, thus having a deletion or duplication event. The quality scores (Q_some and Q_non_dip) were set to be >=60 and only include probands and affected siblings to identify disease-causing CNVs within the affected individuals. These filters ensure that most CNVs identified due to sequencing- or other technical biases would be filtered out.

If a particular CNV was present in more than five individuals of which they were not all affected individuals or parents with a similar phenotype, these were also excluded from further analyses. Lastly, CNVs over 100kb in size were prioritised as these are more likely to be pathogenic compared to smaller CNVs. CNVs over 100kb will thus be referred to as large whereas CNVs <100kb will be referred to as small. XHMM does not differentiate between homozygous or heterozygous deletions or copy numbers of three and higher and thus a single prediction value is made per sample. The coverage and quality of the regions spanning the CNV was assessed using the integrative genomic viewer (IGV) (Robinson et al., 2011) in specific cases where quality parameters were not met.

The depth of coverage calculation for each sample was found to take 5–6 hours for 100 samples and the subsequent commands required 1-2 hours in total. The maximum memory for a single process to run was 3-4GB (Fromer et al., 2012).

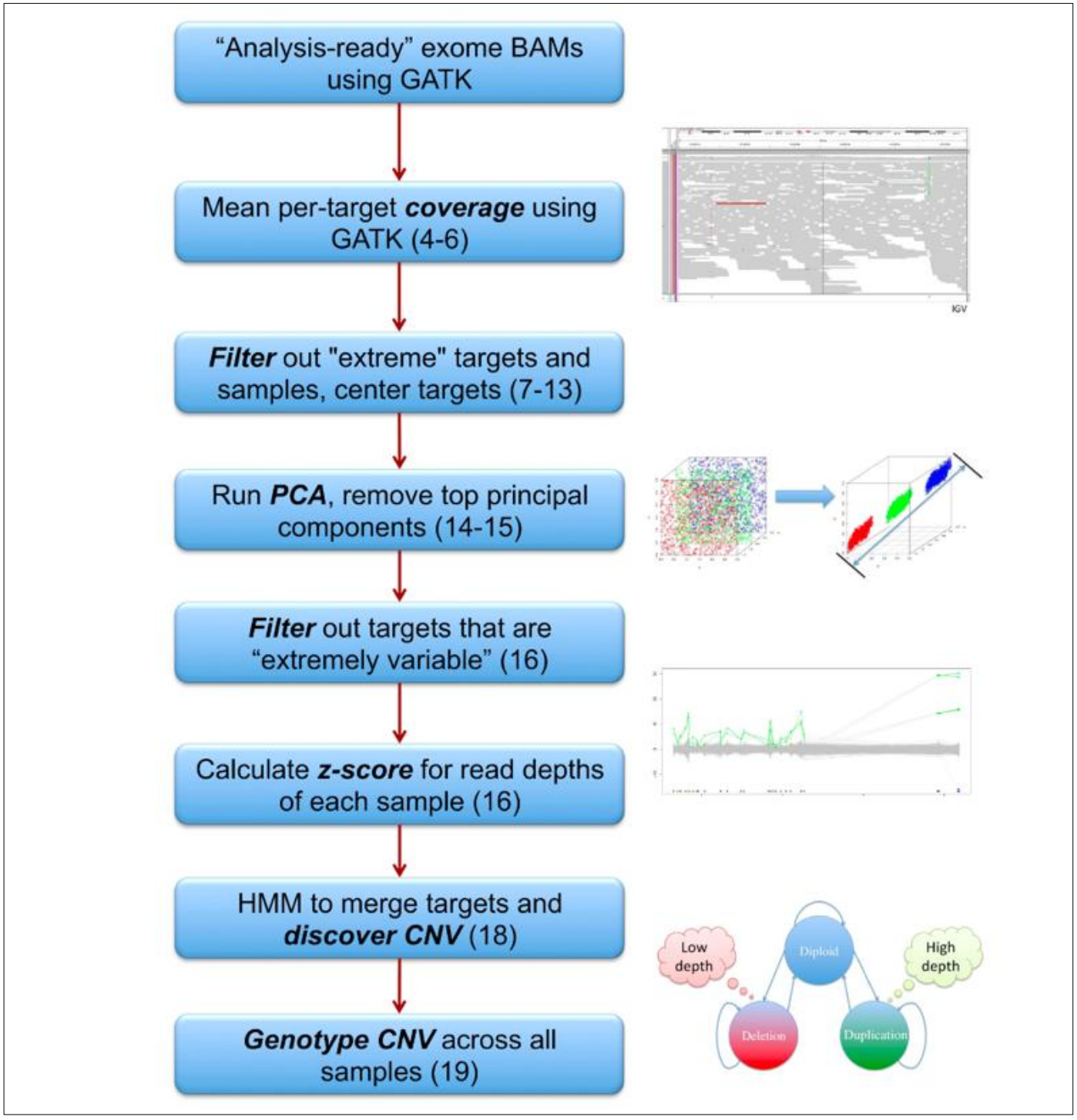


Figure 2.3: Flowchart of CNV calling with XHMM. The main steps in XHMM workflow are listed with corresponding step numbers as outlined above. Key steps are depicted graphically on the right. Figure from (Fromer and Purcell, 2014).

- CANOES

CNVs with an Arbitrary Number Of Exome Samples (CANOES) software was downloaded (Appendix X) and applied using default settings (Backenroth et al., 2014). It uses negative binomial distribution to call CNVs, which has been found to be a good model for over dispersed sequence depth data and is focused on calling rare CNVs 100kb-10Mb. This method is complementary to Gaussian approach tools like XHMM, which has been shown to be less accurate for exome data and could lead to oversight especially for smaller deletions (Plagnol et al., 2012). More accurate calls will be possible when using CANOES in combination with one of these methods which was demonstrated by a reduced and more realistic *de novo* rate from trio data. This tool requires a data frame with the coordinates, GC content and read count per sample for each exome capture region.

As mentioned above, CANOES uses a negative binomial distribution to model read counts and estimates variance of the read counts using a regression-based approach based on selected reference samples in a given dataset. A pooling strategy is adopted by CANOES to build its reference model. Samples with the closest mean and variance are used to independently define each sample's reference. There are a number of computational steps taken by CANOES (Figure 2.4), starting with BAM files of interest and ending with a segmentation of the exome targets into normal, deletion and duplication regions. This is carried out by using a Hidden Markov Model (HMM) and assignment of most likely copy numbers using maximum likelihood. For the DDD-Africa sample set, there were a total of 287 samples of which 171 were unaffected samples.

The unique set of filtering parameters used for the CNVs identified from CANOES are as follows, Q_SOME: a Phred-scaled quality score for the CNV is calculated and all samples <80 were filtered out. Results were further filtered to select for only probands with a CNV of at least 100kb in size. Samples with more than 50 CNVs called (N_{max}=50) are automatically flagged as the Q_SOME score is set to reflect as not applicable (NA). Analyses are restricted to the autosomes due to complications in calling CNVs on the sex chromosomes. The output file (.csv) from CANOES can be seen in Appendix VI.

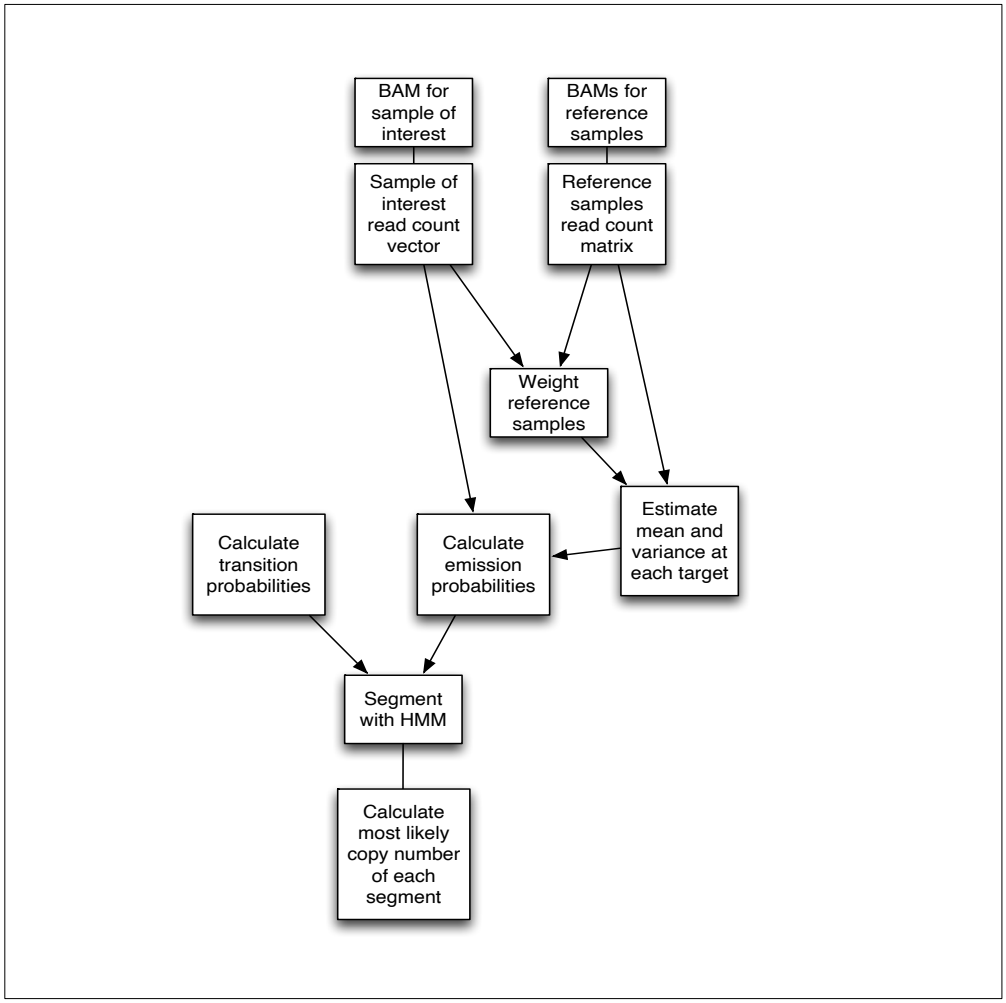


Figure 2.4: Overview of CANOES computational steps. A HMM is used to segment exome targets into normal, deletion and duplicated regions. The most likely copy number of each segment is identified using maximum likelihood. Figure from (Backenroth et al., 2014).

- CLAMMS

Copy number estimation using Lattice-Aligned Mixture Models (CLAMMS) software was installed (Appendix X) and run with default parameters (Packer et al., 2016). This tool was developed to address the problem that most CNV detection tools are not suitable for large population studies. CLAMMS is scalable due to its probabilistic framework. The speed of analysis is also fast, making it easier to analyse multiple samples together. CLAMMS is also quite robust and able to handle samples of different quality and coverage. It is best at calling smaller CNVs (<100kb); however, still calls larger CNVs with good precision. Additionally, it can integrate more easily into an automated variant-calling pipeline and also recognise more common variants.

The CLAMMS algorithm can be divided into three steps:

- Normalisation of coverage values for individual samples to correct average depth of coverage and GC amplification biases.

- A panel of reference samples is used to model the expected distribution of coverage across samples to identify a particular CNV state. These reference samples are selected based on seven Quality Control (QC) metrics and using a k-d tree data structure for each sample. A finite mixture model fit for each exome capture region, based on the reference panel, is used. CNV numbers 0-3 are considered exome-wide and in regions of known duplication numbers 4-6 are also considered.

- A HMM is used to call each individual CNV from each sample's normalised coverage values for each region.

A minimum mapping quality of 30 is used for a read to be counted. A sample.cnv.bed file is the final output (18 columns representing different scores) with the CNVs identified for each sample (Appendix VII). CNVs were filtered according to the Q_some score as per output file and only those which scored above or equal to 500, were included in the final shortlist. Only samples with positive Q_exact scores (Q_exact not negative) were included, this is a non-Phred-scaled quality score measuring whether the called CNV state and breakpoints match the coverage profile. A score lower than zero is questionable and was thus excluded. Lastly, the largest CNVs were prioritised (>100kb). In terms of computational time, one hundred samples from the original publication took a total of 7.5 minutes (45 seconds per sample).

2.2.4 Calculating functional equivalence of CNV tools

In order to more accurately determine the functional equivalence only truth set samples validated by CMA (59/90) was included in the sensitivity and specificity calculation. The computational tools applied in this study have been validated by comparing their sensitivity and specificity to measure functional equivalence.

The concept of functional equivalence has been studied using WGS data with different pipelines, Regier et al., 2018 found that results from different pipelines using the same dataset is virtually indistinguishable. As previously mentioned, functional equivalence in this study is defined as the shared property of two different tools, run on the same dataset, producing essentially the same results, regardless of the different tools' respective breakpoints for a specific CNV. Sensitivity is the proportion of positive results that are truly identified as positive and specificity is the proportion of negative results that are correctly identified as such. The sensitivity was calculated via the top equation and is defined as the true positive rate (TPR). The specificity was calculated via the bottom equation and is defined as the true negative rate (TNR).

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN}$$

$$TNR = \frac{TN}{N} = \frac{TN}{TN + FP}$$

In these equations, there are six values: Positive (P), Negative (N), True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) (Zhao et al., 2013). In order to obtain these values, the total number of targets (probes) located within the CNVs overlapping with true CNVs for each tool were calculated. The probes (targets) used are from the ES analysis showing the coordinates where each probe is located within the exome. This is also the file included in the analysis steps of the CNV calling tools in order to identify the exome capture regions. Each CNV identified with each tool has the number of targets already included within the output file for these tools (Appendix IV-VII). In order to obtain the number of targets within the CNVs coordinates of the truth set, a script was created to identify the number of targets within each CNV (interval). These values are thus defined as (Table2.2):

Table 2.2: Description of the variables used for sensitivity and specificity calculations

Abbreviation	Definition
P (Positive)	Number of targets which are within the real CNV
N (Negative)	Number of targets which are outside of the real CNV
TP (True Positive)	Number of targets which are within both the detected CNV and the real CNV
TN (True Negative)	Number of targets which are out of both the detected CNV and the real CNV
FP (False Positive)	Number of targets which are within the detected CNV and out of the real CNV
FN (False Negative)	Number of targets which are out of the detected CNV and within the real CNV

The number of overlapping CNVs between the tools were also identified in order to accurately determine functional equivalence. This process was aimed at determining the consistency of these tools' results. A high rate of overlapping CNVs between the tools would indicate that the results have high consistency and are trustworthy. This was done using a Venn diagram where the quantitative value was defined as N_(overlap) and N (CNV tool) (Figure 2.5). The results of overlaps between CNVs detected using all truth set samples (90) and only those validated by CMA (59) will be compared. Only three of the tools are represented here as mentioned above, the results for InDelible was not sufficient to be included. The value for N is the total number of CNVs that are uniquely detected by a CNV tool and N_(overlap) is the number that overlapped with those CNVs also detected by the other tool.

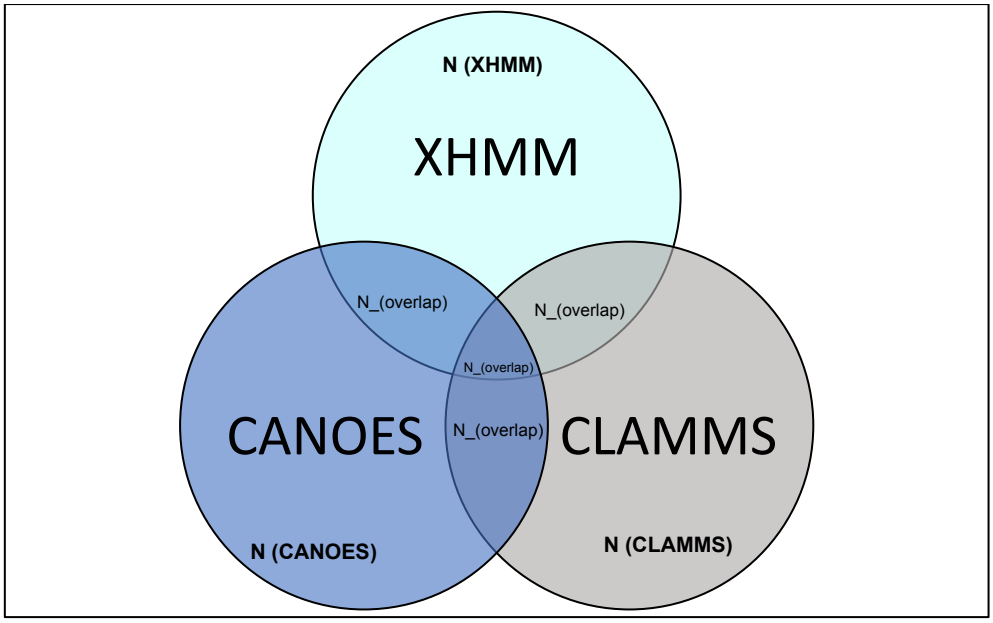


Figure 2.5: Venn diagram showing overlapping and unique CNVs called by the three CNV tools.

2.3 Results

2.3.1 Identification of the optimal method for CNV identification through functional equivalence evaluation

2.3.1.1 True CNVs identified between CNV calling tools

Before filtering, over 2 million SVs were identified by InDelible which is an average of ~20,000 SVs per sample. Only three of the 90 known CNVs overlapped with SVs identified by InDelible; however, the advantage of using InDelible might not have been apparent as there were not enough truth set CNVs within the scope of InDelible for a fair evaluation. Only eight of the CNVs were between 20-500bp in size, which is the optimal size for which InDelible was designed. A total of 40 CNVs were below 100kb in size and 17 of these CNVs under 10kb. This tool was developed for smaller SVs and as this study is trying to identify the best approach to use for rare, disease-causing CNV analysis, InDelible was not used for the main dataset and was left out of further comparison and analyses. This tool was initially included to try and close the gap where most CNV tools identify only larger CNVs and where the smaller CNVs (one exon and less) are missed by the majority of these tools. The CNV calling tools did identify CNVs as small as 500bp in some instances but mostly missed the CNVs below 10kb.

After analysis of the full truth set data (90 CNVs), the bioinformatics tools developed to identify CNVs specifically and not all SVs, were more accurate in detecting the true CNVs. For each of the three other tools (CANOES, CLAMMS and XHMM), the overlap of detected true CNVs between the tools has been illustrated by a Venn diagram (Figure 2.6A). For CANOES, 68% (61/90) of the expected CNVs could be identified with the truth set calls of which 46 were deletions and 15 duplications. XHMM identified 61% (55/90) including 40 deletions and 15 duplications and CLAMMS identified 57% (51/90) of the known CNVs, including 36 deletions and 15 duplications. The 90 truth set CNVs included 13 benign, 34 likely benign, eight VUS and 35 likely pathogenic classified CNVs. A total of 25 CNVs were missed by all three tools of which nine were below 10kb in size, with the majority below 100kb (17/25). Thirteen of these CNVs were deletions and 12 duplications.

A total of 45 CNVs overlapped between all three tools; however, 65/90 (72%) of the known CNVs were detected by adding the individual results of all three tools together. Out of the 45 CNVs overlapping all tools, 39 have been validated with CMA. The majority of these CNVs were classified as likely pathogenic or pathogenic (LP/P) (23/45) whereas 18 CNVs were classified as likely benign and four CNVs were benign. Although this detection rate is quite low, upon further investigation it was seen that the CNVs were quite small as 37 CNVs were below 100kb and most of these tools are optimised for detection of larger CNVs. The individual tool which identified most of the known CNVs out of the four was CANOES, calling 61 of the true CNVs.

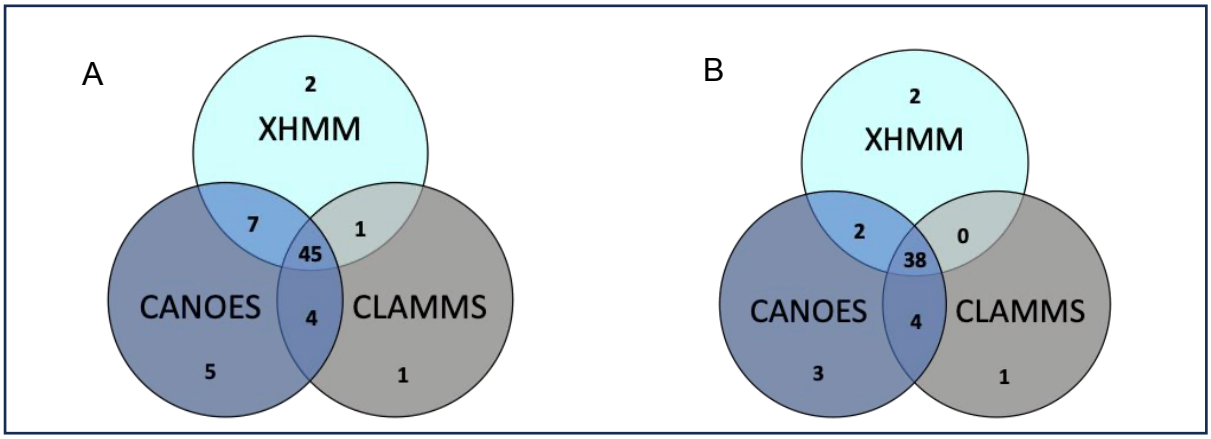


Figure 2.6: Overlapping CNVs between different CNV tools for CNVs called from all 90 truth set samples (A) and from the 59 CMA validated truth set samples only (B). The majority of the CNVs were identified by combining all three tools' results together 65/90.

Analysis of the CMA validated truth set CNVs only (Figure 2.6B), showed improved detection by the three tools. A total of 47 CNVs were detected by CANOES (79.7%), 43 by CLAMMS (72.9%) and 43 by XHMM (71.2%). In total the three tools together detected 50 (84.7%) of the CNVs of which 36 were deletions and 14 duplications. The nine CNVs not detected by any of the tools were small with an average of 84.83kb of which the largest was 470.39kb (duplication) and the smallest 0.51kb (deletion). Five CNVs were deletions and the remaining four duplications. Thirty eight CNVs were detected by all three tools of which the majority were deletions (28/38) and the average CNV size was ~2Mb.

2.3.1.2 Calculating sensitivity and specificity of the three CNV tools

As evident from the previous section (2.3.1.1) the overlapping consistency was not ideal and 28% of CNVs were missed; however, the sensitivity and specificity of the tools seem to be consistent when considering the number of probes which did overlap with true CNVs from the truth set. When comparing these result to those when only the CMA validated CNVs were included, overlapping consistency increased to ~85%. The CNVs identified by the three tools were smaller than the truth set CNVs. Even though these tools detected the correct CNV, the full size of the CNVs were different, underestimating the CNV size rather than overestimating. This also led to more false negative probes and less false positive probes in the identified CNV regions and thus the true negative rate (specificity) was higher than the true positive rate (sensitivity) for each tool (Table 2.3). When comparing the results from the CMA confirmed CNVs to the CNVs which was only bioinformatically confirmed, an increased sensitivity is observed. These tools thus show a better ability to prioritise true positive CNVs (confirmed with CMA). The sensitivity and specificity of CANOES was the best followed by XHMM and then CLAMMS. The area under the curve (AUC) for each tool is depicted in Figure 2.7 showing the sensitivity (y-axis) and 1 minus specificity on x-axis. The sensitivity and specificity of these tools are approaching one and thus the AUC shows that an almost perfect prediction capability is achieved by each tool. These tools thus distinguish well between the true positive and true negative CNVs.

Table 2.3: Sensitivity and specificity values from the functional equivalence evaluation for the three CNV tools, CANOES, XHMM and CLAMSS.

			CNVs confirmed with CMA	**CNVs confirmed bioinformatically**
CANOES	Specificity	TNR	0.99	0.99
	Sensitivity	TPR	0.84	0.62
XHMM	Specificity	TNR	0.99	0.99
	Sensitivity	TPR	0.73	0.60
CLAMMS	Specificity	TNR	0.99	0.99
	Sensitivity	TPR	0.67	0.51

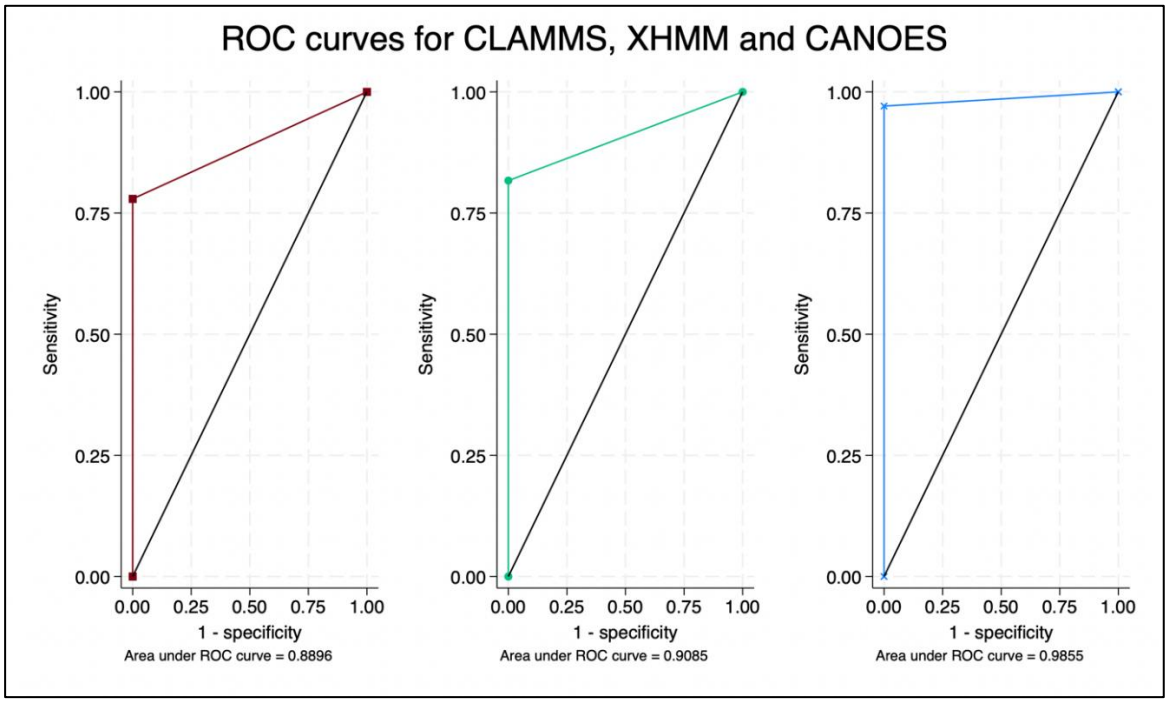


Figure 2.7: Receiver Operating Characteristic (ROC) curves for CLAMMS, XHMM and CANOES illustrating sensitivity on the y-axis and 1-specificity as well as the area under the curve (AUC) for each tool (maximum of one). All three figures curve close to the top left of the x-axis toward one on the y-axis and thus represent good sensitivity and specificity.

All three methods together presented with the best result as 72% of the CNVs were detected and this increased to 85% when only including the CMA validated CNVs. Even though on its own the tools did not reach this percentage of known CNVs identified, it is seen that the different tools do show functional equivalence when inspecting the sensitivity and specificity. Implementing a single tool would thus give a similar outcome when measuring sensitivity and specificity of tools to detect true positive CNVs.

CANOES did detect the largest number of CNVs (68%) second to the ensemble approach although a number of CNVs would have been missed if only one tool was applied. The ensemble approach showed superior results which is especially apparent when inspecting the overlapping CNVs between the tools (Figure 2.6). InDelible was not used for the following analyses, but all three other CNV calling tools (CANOES, CLAMMS and XHMM) were applied to the DDD-Africa dataset.

2.4 Discussion

The main aim addressed in this chapter was to calculate the functional equivalence of the four CNV calling tools and subsequently identify the best tool or combination of tools to implement onto the DDD-Africa exome dataset. It is evident from the truth set results that three of the four tools (CANOES, CLAMMS and XHMM) present with similar results; however, all three tools together delivered superior results. InDelible identified only three of the known CNVs correctly although the majority of CNVs were not within this tool's scope. This tool was not implemented on the main DDD-Africa dataset. Not one tool identified all the true CNVs, but CANOES had the best calling rate followed by XHMM and lastly CLAMMS. The ensemble approach was superior as a total of 72% of CNVs were identified when combining the individual results from all three CNV tools. This informed the decision to implement all three the CNV tools for further analysis using the batch 1 samples from the DDD-Africa dataset.

InDelible was initially chosen to be incorporated as it uses split read method and paired with one or more of the read depth CNV tools might lead to better performance and higher molecular diagnostic yield. There might be several reasons for InDelible not performing optimally as it is trained to identify smaller structural variants (21-500bp). As mentioned in section 2.3.1.1, only eight of the CNVs within the truth set were within the optimal size range for InDelible. The SVs identified must intercept DD associated genes on the DDG2P panel specifically which might also have led to the exclusion of some of these CNVs. Analysis of the truth set did not include parental samples and thus a full trio dataset was not available and consequentially the final *de novo* step of InDelible could not be carried out. InDelible is focused on highly penetrant dominant *de novo* variants as this is the primary cause of DD (Kaplanis et al., 2020).

This tool would be useful to incorporate into the CNV analysis pipeline for DD patients where a full trio is available as this tool is focused on *de novo* variants. This could identify smaller SVs (one exon or less) which might be missed by the CNV calling tools available at present. In a previous study, InDelible identified 59 additional SVs, not previously identified by other exome CNV calling tools for patients enrolled in the DDD-UK study (Gardner et al., 2021). A total of 29 of these were plausible pathogenic variants and increased the putative diagnostic variants (21-500bp in size) with 42%. An increase of 2-3% of total molecular diagnostic yield was reported from incorporating InDelible on ES data. A future recommendation should thus be to incorporate this tool when a dataset with full trios is available.

A large number of the known CNVs were not identified with any of the three CNV tools and the main reason for this shortcoming could be the fact that 40 of the CNVs in the truth set were below 100kb. The majority of the CNVs missed with these tools are smaller than 10kb even though CNVs as small as 500bp were indeed identified with these tools. The fact that the majority of the 45 overlapping CNVs from the truth set were classified as LP/P shows that although not all the CNVs were correctly called, the more pathogenic CNVs seem to be selected for. The average size of the 45 CNVs identified was also large (~1600kb) with only twelve below 100kb in size. Not all of the true CNVs met the filtering parameters applied to the batch 1 DDD-Africa dataset even though most of the CNVs not meeting these parameters were classified as likely benign.

All three tools presented with good specificity, but with varying sensitivity. CANOES presented with the best sensitivity followed by XHMM and lastly CLAMMS. This thus signifies the tools' ability to detect a true positive CNV (true positive rate). Specificity in this study seems to be higher than previously predicted and thus the true negative rate might be overestimated due to the large number of ES probe targets not within the CNVs, artificially increasing the negative predictive value. The ROC curves (Figure 2.7) show that all three tools' area under the curve approaches 1 which is optimal, with CANOES showing the best outcome when compared to XHMM and CLAMMS. The sensitivity of the tools was 84% (CANOES), 73% (XHMM) and 67% (CLAMMS) respectively.

This seems to be mostly in keeping with literature where sensitivity of the tools ranged between 70-80% (Fromer et al., 2012, Backenroth et al., 2014) although CLAMMS did not present with optimal sensitivity in a 2022 study where it was below 50% (O'Fallon et al., 2022). In another study, CANOES specifically presented with very good sensitivity, specificity and positive predictive value (Quenez et al., 2021). Overall, these statistics (functional equivalence evaluation) are consistent with the expected performance for a screening tool.

Although the sensitivity and specificity calculations showed a low likelihood of false positive results, some CNVs were still missed by using this specific ensemble approach. These tools are thus not meant to be implemented as a comprehensive CNV typing solution, but rather in screening pipelines. The individual tool identifying most (68%) of the true CNVs was CANOES which also showed the most optimal sensitivity and specificity. Even though CANOES did present with the best sensitivity and specificity, all three CNV tools were incorporated into the DDD-Africa dataset as together 72% of the true CNVs were detected. The advantage of using all three is more evident when studying the outcome of the specific tools' ability to identify the true CNVs (Figure 2.6). A large proportion of the true CNVs were identified by all three tools, but there were still true CNVs identified by only specific tools. This indicated the need to investigate all the CNVs identified by the tools individually and not only those overlapping between the tools. Starting with interpretation of the overlapping CNVs between tools did, however, speed up the curation process of the CNVs.

Overall the tools show functional equivalence from sensitivity and specificity analyse; however, it was seen that the ensemble approach lead to the best combined detection rate (72%). It would still be recommended to use more than one tool if possible as these tools have different strengths due to the algorithmic differences. Discussed in chapter 1 are recommendations from other studies to apply additional tools, not covered in this study. Many of these studies have also found the use of an ensemble approach to work best as presented in section 1.7. It is evident that all three tools have different outcomes and strengths when applied to exome sequencing data. Applying the three CNV tools (CANOES, CLAMMS and XHMM) together and interpreting the results individually will be carried out for the DDD-Africa data and discussed in chapter 3.

Chapter 3: Screening DDD-Africa exome data for pathogenic CNVs

3.1. Introduction

It is essential to be able to correctly determine the pathogenicity of variants for patients to not only receive an accurate molecular diagnosis, but also appropriate medical care. The decrease in cost and widespread use of NGS together with an enhanced quality of genome-wide analyses led to increased detection of variants. This has created a higher demand for improved understand of these variants and for the correct interpretation of their impact on human health. Interpretation of NGS data in a clinical setting is complicated and often leads to ambiguous results between different laboratories. The use of different technologies, incomplete penetrance, variable expressivity, and unreported clinical symptoms are often factors challenging variant interpretation (Lincoln et al., 2021, Vaseghi et al., 2023). As the impact of SNVs is better understood in comparison to other types of variation, great effort led to the development of many tools for pathogenicity prediction specifically for these variants (Liu et al., 2020, Garcia et al., 2022). To further alleviate difficulties with SNV interpretation, specific guidelines by the American College of Medical Genetics and Genomics (ACMG) have also been developed to ensure consistent and standardised variant classification and reporting (Richards et al., 2015). This includes a five-tier terminology system using the terms pathogenic, likely pathogenic, uncertain significance, likely benign, and benign. Standardised variant classification and reporting helps to keep results consistent and up to a professional standard across laboratories and institutions. Inconsistent results between laboratories create confusion for scientists, clinicians and patients and makes interpretation and feedback of results very difficult.

Although some of the basic principles of genomic variability interpretation are consistent, these rules are unfortunately not directly translatable to CNV interpretation. More recently, ACMG and Clinical Genome (ClinGen) Resource have published guidelines for CNV classification (Riggs et al., 2020) to enable standardised CNV reporting. The aim is to accurately interpret CNVs with consistent methods and standards which are readily available and updated regularly.

The same five tier terminology system is used although the criteria for classification of CNVs differs from SNVs due to the nature of CNVs being larger, more complex and encompassing many protein coding regions. This is a semiquantitative point-based scoring metric with separate criteria for CNV losses and CNV gains. A ClinGen CNV pathogenicity calculator (Appendix X) has been developed and is based on CNV scoring metrics according to ACMG technical standards and streamlines the classification of CNVs (Patel et al., 2017).

Databases and tools have been developed to standardise CNV reporting and evaluation for instance the DECIPHER (Bragin et al., 2014) database which is a web-based platform facilitating secure deposition, analysis, and sharing of genomic variation together with the associated phenotypes for patients with rare disorders (Firth et al., 2009, Bragin et al., 2014). It includes an interactive genome browser to visualise and interpret SNVs and CNVs against datasets with other pathogenic variants reported as well as population variation. Information regarding recurrent CNVs is becoming more readily available; however, the majority of CNVs are still unique which cause challenges for interpretation. The Database of Genomic Variants (DGV) has been developed to provide insight into specific genetic variants (MacDonald et al., 2014). This database contains structural variations identified in healthy individuals which is helpful for correlating genomic variation with phenotype data. Another freely available public archive of human genetic variation is ClinVar which provides interpretations of clinical significance to disease with supporting evidence and confidence levels (Landrum et al., 2017). The Genome Aggregation Database (gnomAD) summarises data from a wide variety of large-scale exome and genome sequencing projects, making it available to the wider scientific community (Karczewski et al., 2020).

Many web-based tools for CNV interpretation and classification have also been created, using the Riggs et al., 2020 published standards, making it more attainable and realistic for a broader audience to use and understand. Although these published guidelines (Riggs et al., 2020) make it easier to interpret and standardise CNV reporting, implementation on a large scale remains a challenge. Each CNV requires the approval and consideration of a clinician before the final score can be given.

CNV calling pipelines and online classification resources and tools are constantly being improved to try and overcome some of the difficulties faced with variant classification, specifically related to CNVs (O'Fallon et al., 2022, Lemire et al., 2024).

Annotation of CNVs is also notoriously difficult as the breakpoints for the exact same CNV may differ between methods of identification and clinical interpretation is complex as it often spans across many genes. Multiple functional genomic elements are affected by CNVs which may or may not result in a phenotype. CNVs can also span non-coding regions which could have an effect on gene expression, but the functional consequences of these variations often require further studies. Other features such as variable expressivity and incomplete penetrance can also play a role in whether or not a CNV will result in expression of a phenotype.

CNVs involving genes sensitive to a change in dosage may lead to development of specific phenotypes related to these genes. Determining whether specific genes are indeed dosage sensitive is another challenge as not all genes have been curated or are well understood. Correlation of CNVs to clinical details of the patient is often insufficient especially the rare CNVs, which might have never before been reported**.** Many CNVs are recurrent, although it is difficult to establish recurrent CNVs without extensive data on specific population groups to confirm these CNVs. Data availability is an important factor to consider when interpreting variants as the availability of more data and previously reported variants improves the interpretation of these variants. This is a limitation, especially within underrepresented populations; however, with an increase in NGS use and additional data availability, more variants are being deposited into genomics databases. This has also enabled various CNVs located in different regions of the genome to be detected, not only in affected individuals, but also in the general population. This being said, there still remains a lack of data especially within African cohorts as many CNVs are not well-characterised or submitted to available databases.

In this chapter, methods and results from applying CANOES, CLAMMS and XHMM to the first batch of the DDD-Africa exome data will be discussed. Furthermore, classification of the exome CNVs identified by the three chosen CNV tools is discussed. Classification was completed both manually and with a web-based tool.

The outcomes of the manual and artificial intelligence-based classification were compared to ensure no CNVs were classified incorrectly with either method, overlooking CNVs which could be pathogenic or which should be classified as VUS or benign. All CNVs classified as LP/P were further investigated and evaluated together with patient clinical details. Results were summarised to measure the final molecular diagnostic yield attributed to CNVs in this patient cohort.

3.2. Methods

3.2.1 DDD-Africa participants

Data from 287 participants were included as the DDD-Africa study batch 1 samples, consisting of 117 affected individuals (62 male, 55 female), 108 mothers and 62 fathers. Included in this dataset is a total of 60 trios, 46 duos, three extended families with more than one affected individual and six singletons. The majority (90%) of individuals were of black African origin. The age range of the affected individuals are 1-14 years. All participants were recruited through the DDD-Africa study from Charlotte Maxeke Hospital, Chris Hani Baragwanath Hospital, Nelson Mandela Hospital and Rahima Moosa Mother and Child Hospital in Johannesburg. Ethics was obtained for this study through the Human Research Ethics Committee of the University of the Witwatersrand (M180678 and M230567) part of the DDD-Africa study (Appendix II). The affected participants (patients) were included into the study based on the below inclusion criteria. Inclusion A and B involving developmental delay are mutually exclusive as well as inclusion B and C involving major and minor malformations. It was thus not possible for patients to be included within both these categories together. The majority of this patient cohort had severe developmental delay (40%), microcephaly (40%), hypotonia (40%) and delayed speech (60%). Seizures and vision impairment were present in ~20% and 50% presented with a structural abnormality.

Inclusion criteria:

A. **Patient is developmentally delayed; (moderate to profound)**
   Over and above the main inclusion criteria of developmental delay, there may be dysmorphic features present, but additional clinical features are not necessary for inclusion in this category.
B. **Multiple major malformations* in TWO or more different organ systems**
   Developmental delay and other additional clinical features are not necessary for inclusion in this category.
C. **Only one major malformation, and patient also has clinically relevant minor anomalies and/or dysmorphic features**
   Developmental delay and other additional clinical features are not necessary for inclusion in this category.
D. **Mild developmental delay, and patient also has clinically relevant minor anomalies and/or dysmorphic features**
   It is essential that a group of medical geneticists reach clinical consensus about the minor anomalies and dysmorphic features for inclusion in this category.

Patients were excluded from the study if there was a suspicion of the condition not being due to a monogenic cause. This included evidence of an acquired brain lesion with predominant neurological manifestations, or a suspected multifactorial or environmental cause. Patients with mild developmental delay only were also excluded. All participants signed an informed consent/assent form and gave consent/assent for their data to be used and stored in a public domain database without identifiers. Parents or caregivers signed on behalf of minor patients and those unable to give assent on their own. Deidentification of participant samples and data was done by using a code and numbering system. The majority of the patients did receive genetic screening (including karyotyping, MLPA and CMA) prior to inclusion.

3.2.2 DDD-Africa data generation

Exome sequencing of the DDD-Africa samples was completed at Wellcome Sanger Institute using ISC-Twist library preparation (Twist Biosciences, San Francisco, CA, USA) with samples run in pools of 96-plex using Illumina paired-end sequencing (Illumina, San Diego, CA, USA) on the Illumina NovaSeq 6000 platform at an average depth of ~40x. This is lower than the average depth of ES (~100x); however, ES with Twist has been shown to perform well at lower sequence coverage compared to other exome capture techniques (Yaldiz et al., 2023). This technique was also implemented for DDD-UK (Firth et al., 2011).

The three chosen bioinformatics CNV tools (CANOES, CLAMMS and XHMM) were applied to the DDD-Africa data as detailed in section 2.2.3, generating data files with all detected CNVs per sample for each tool. Candidate CNVs were viewed on IGV to assess quality and coverage of the regions spanning the CNVs.

3.2.3 Identifying shared CNVs from the different bioinformatics tools

A custom script was created in order to identify CNVs overlapping between the CNV tools (Appendix VIII) using Python 3.0 (Rossum and Drake, 2009). This was specifically created for the DDD-Africa dataset as the detected CNVs from this amount of samples are too many to manually analyse possible overlaps. Although four tools were applied, only results from three were sufficient to be used. The script specifically includes CNVs overlapping from the different tools (different for the length of each CNV), the larger the CNV, the larger the overlapping region would need to be. In literature, a reciprocal overlap of 50% has been used to consider two CNVs as similar (Pang et al., 2010, Castellani et al., 2014). This script allowed for up to 25% deviation from each breakpoint, thus for CNVs of 1000bp, a difference in size of 250bp on either side of the breakpoint was taken into account from the primary and distal breakpoints. Most of these CNVs overlapped a large core region with only a few base pair differences from each of the tools' calling breakpoints. A Venn diagram (Figure 2.5) was used to illustrate the common CNVs between the three tools as well as the unique CNVs called by each tool.

3.2.4 Data analysis

Output files containing all CNVs identified using the three CNV calling algorithms (CANOES, CLAMMS, XHMM) were used for an initial step of annotation and classification. Each tool's CNV output was analysed individually first before results were compared between the tools. Manual – as well as semi-automated CNV classification was carried out (Figure 3.1). As manual classification is more time consuming and more vulnerable to errors, an online classification tool was also incorporated in parallel to ensure consistent results. Manual classification was performed with the online dosage sensitivity calculator (Patel et al., 2017) and secondly CNV-ClinViewer (Macnee et al., 2022) was also used to classify all CNVs obtained by these three CNV calling tools.

The results from manual classification and CNV-ClinViewer were compared and interpreted together to identify the LP/P CNVs.

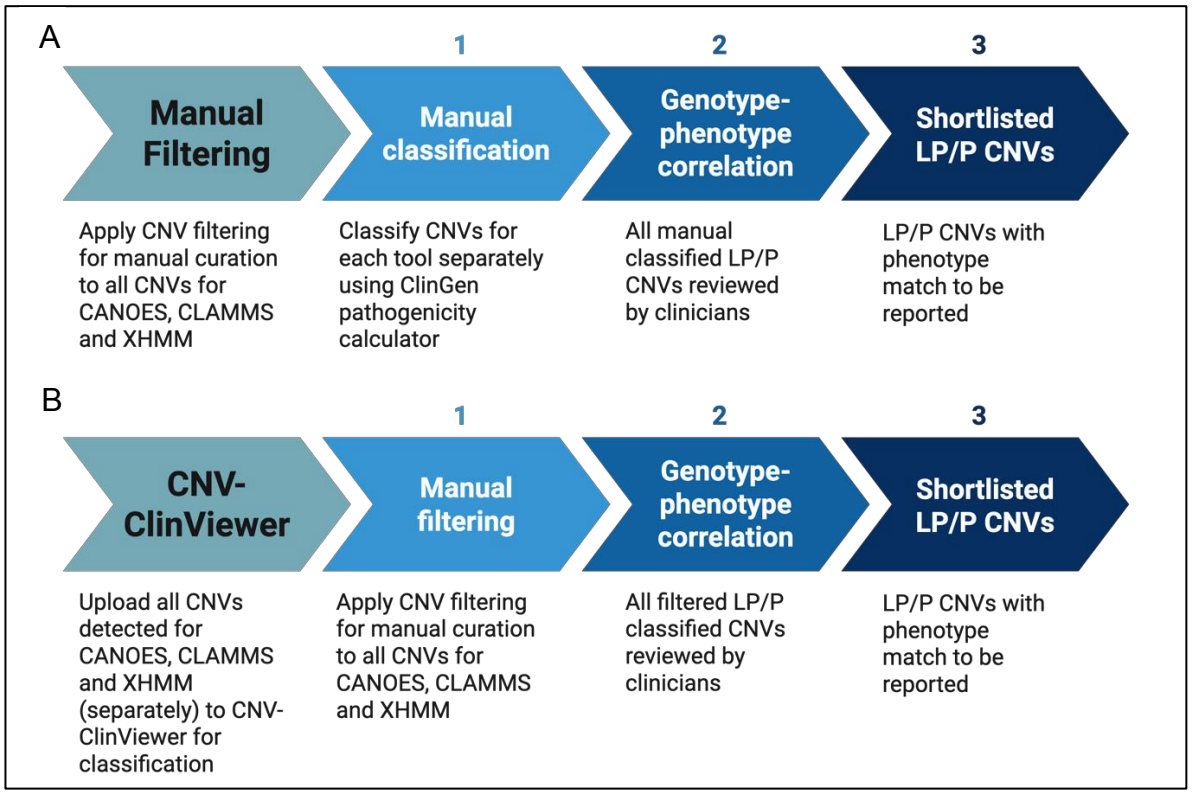


Figure 3.1: (A) Manual and (B) Web-based classification of CNVs. For manual classification CNVs were first filtered before classifying with ACMG/ClinGen guidelines in order to reduce the number of CNVs.

3.2.5 Classifying CNVs according to ACMG and ClinGen guidelines

CNVs were prioritised as described in chapter 2 (2.2.4) with applied filters for each tool as a screening strategy. This allowed analyses of the CNVs with enough evidence to be classified as LP/P. These CNVs were then further curated to ensure consistency in classification and to receive clinical input before being reported. This was thus a screening strategy rather than a comprehensive CNV analysis, leaving only the most likely disease-causing CNVs to be interpreted. These CNVs were prioritised according to size (largest to smallest) as well as whether the particular CNV was *de novo*, inherited or present in many other healthy individuals in the cohort. These prioritised CNVs were then evaluated on the ClinGen dosage sensitivity curation website, where the given coordinates of each CNV was entered in the search bar (GRCh38 region) (Riggs et al., 2018). This established whether there were any haploinsufficient (HI) or triplosensitive (TS) genes within the given CNV, which is a requirement in order to classify these CNVs appropriately.

The HI genes are intolerant of deletion and do not generate enough protein when one of the alleles is deleted, meaning it may lead to a disease phenotype. TS genes on the other hand are intolerant of duplication thus an additional copy of these genes may cause a disease phenotype. Details regarding known population regions and Online Mendelian Inheritance in Man (OMIM) genes (Hamosh et al., 2002), which are established disease-causing genes, are also given. This curation site also displays the DECIPHER HI index, the gnomAD probability of being loss-of-function intolerant (pLI) score and the gnomAD prediction loss of function (LOEUF) which highlights whether a gene is more intolerant to loss of function variation. Lastly, it is indicated whether the gene/region has been curated by the ClinGen curation team which plays a crucial role in the decision-making process of CNV pathogenicity prediction.

Recommended steps were then followed to score these CNVs according to ACMG and ClinGen guidelines (Riggs et al., 2020) using the ClinGen CNV pathogenicity calculator. There are separate calculations for CNV loss and CNV gain, and points are applied according to individual evidence categories for each given CNV (Figure 3.1). Section one relates to the content of the CNV and whether there are any protein coding elements involved. Section two is used to establish whether the CNV overlaps with any HI or TS regions. Section three evaluates the number of genes within the CNV while section four is a more detailed evaluation of the genomic content by comparing to cases in literature, public databases or internal lab data. Lastly, section five evaluates the inheritance mode of the CNV. The complete detailed sections can be found in the original publication (Riggs et al., 2020). Each CNV was then classified as benign (<-0.99), likely benign (-0.90 to -0.98), variant of unknown significance (VUS) (0 to -0.89), likely pathogenic (0.90 to 0.98) or pathogenic (>0.99), according to the final score. CNV classification scores are calculated from each piece of evidence either in support of (positive value) or refuting pathogenicity (negative value).

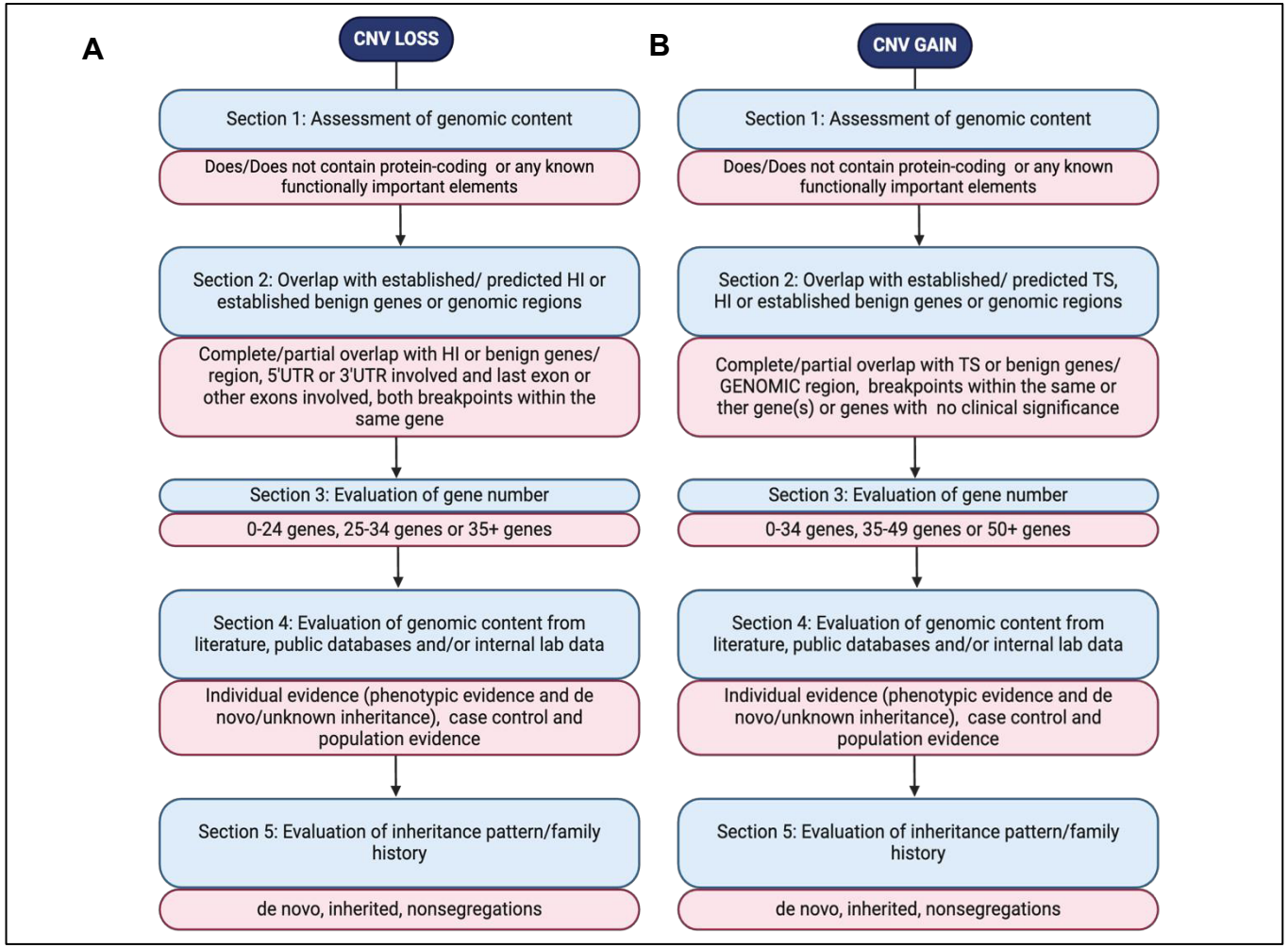


Figure 3.2: Criteria of the online scoring rubric from the CNV pathogenicity calculator for a CNV loss (A) and CNV gain (B).Each CNV was scored accordingly and an output for each was obtained with the estimated pathogenicity score after all criteria was evaluated. Figure from (Riggs et al., 2020).

After initial classification and review of patient clinical information, shortlisted LP/P CNVs were then uploaded onto the DECIPHER database where CNVs were compared to other similar reported variants. This is very useful as patient variants and phenotypes from around the world are submitted and can be compared to the patient of interest's variant and phenotype. This also highlights the more common variants and thus it is very useful for variant classification purposes. All CNVs scored as LP/P were investigated further to confirm whether the phenotype matched the genotype and that it explains the developmental disorder of the specific patient. Although these CNVs were shortlisted by classification, the final outcome (LP/P or VUS) was determined by confirming whether the CNV is likely to be disease-causing. This final step in deciding if a variant is likely linked to the patient's phenotype was made at a multidisciplinary review meeting, consisting of medical scientists, medical geneticists and – registrars, paediatricians and scientists in the field of human genetics from the DDD-Africa project.

3.2.6 Comparison of manual classifications with CNV-ClinViewer outcome

CNV-ClinViewer (Macnee et al., 2022), was also utilised in order to compare to the results from manual curation. The shortlisted CNVs were thus reviewed using the two strategies to strengthen the pipeline and improve the discovery power. As some filtering was done before manual classification, the smaller CNVs (<100kb) or CNVs not meeting filtering criteria might have been overlooked and thus this web-application classified all CNVs from each tools' output. This open-source web-application enables evaluation and classification of CNVs on a visual interface and allows for comparison with previously reported CNVs. This web-application uses the output from the different CNV calling tools (bed, text or excel files) and classifies the CNV according to ACMG/ClinGen guidelines using ClassifyCNV (Gurbich and Ilinsky, 2020). It is designed to guide researchers and clinicians in their decision-making process. Curated CNVs from ClinVar (Landrum et al., 2017), UK-biobank (Sudlow et al., 2015) and gnomAD (Karczewski et al., 2020) are used during the classification process. It thus combines analysis, annotation, classification and clinical evaluation (Figure 3.3).

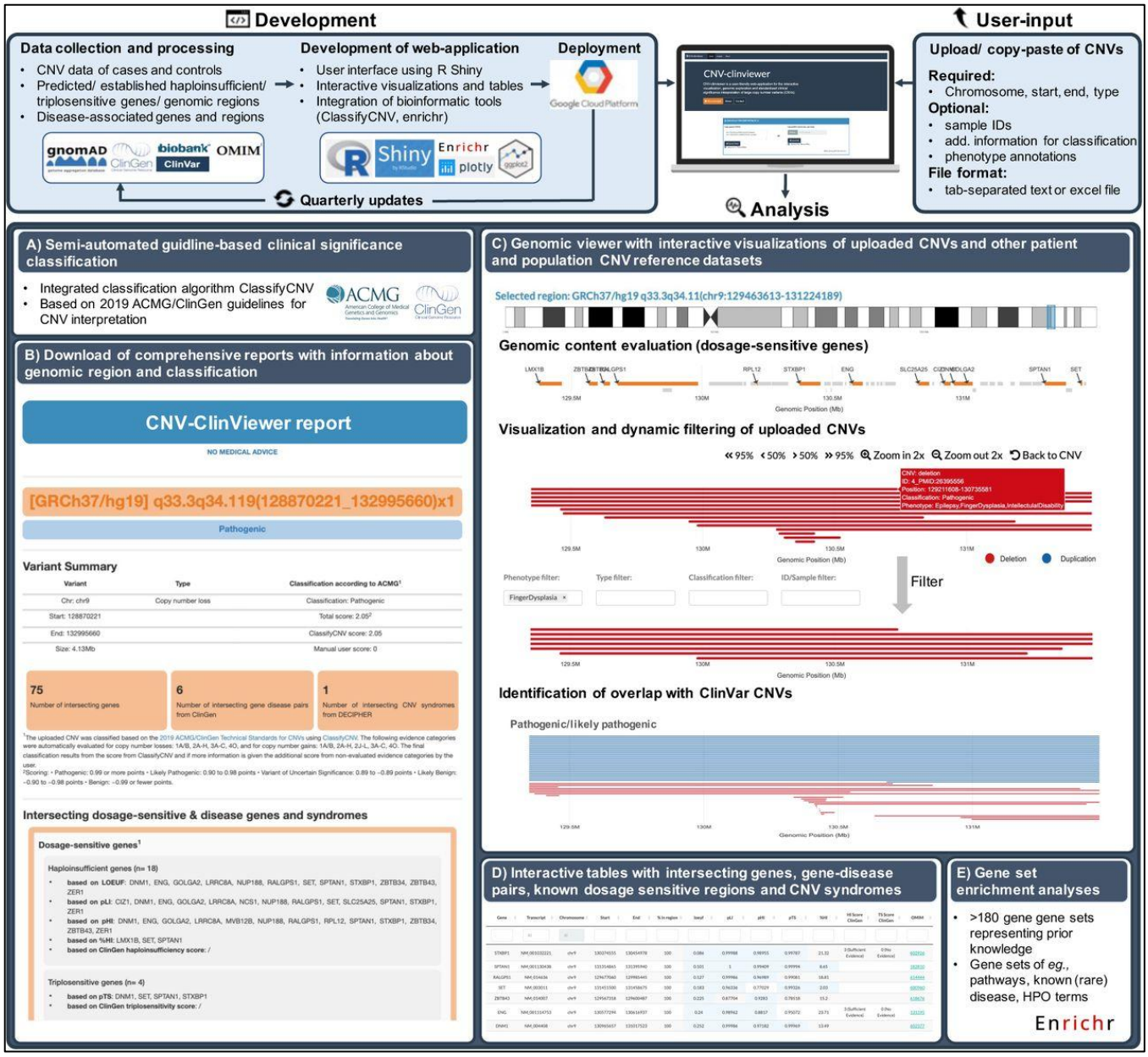


Figure 3.3: Overview of CNV-ClinViewer with all features and output. A) Intergrated semi-automated classification of CNVs is based on 2019 ACMG/ClinGen Technical Standards for CNVs by ClassifyCNV. B) comprehensive downloadable report on individual CNVs. C) Interactive visualisation of uploaded CNVs with the genomic viewer also showing other pathogenic and general population CNVs reported. D) Information on intersecting genes, gene-disease associations, known dosage sensitive regions and known CNV syndromes can be viewed and downloaded on this interactive table. E) Gene-set enrichment analyses (GSEA) can be performed on genes in a specific region or CNV. Figure from (Macnee et al., 2022) under license https://creativecommons.org/licenses/by-nc-nd/4.0/.

3.2.7 ClinTAD tool for additional pathogenicity evidence

Where CNVs were difficult to interpret and there was insufficient annotation to classify, the ClinTAD tool was used to further evaluate pathogenicity (Spector and Wiita, 2019). This tool evaluates the genomic impact of CNVs in the context of topologically associated domains (TAD). These domains have an influence on epigenetic modifications such as histone methylation and regulates gene-enhancer interactions. The disruption of these structural domains by a CNV can also have a pathogenic effect and lead to alternate regulation of transcription. A previous study investigating deletion CNVs on DECIPHER found that up to 11.80% of these pathogenic CNVs may involve disruption of TADs (Ibn-Salem et al., 2014). This tool can thus also be used to evaluate VUS classified variants to gain additional evidence for pathogenicity in cases where there is a phenotype match.

The ClinTAD tool displays the TADs (Figure 3.3) which is originally obtained from human induced pluripotent stem cells and published by Dixon et al. (2012), also displays the genes in the region (annotations obtained from Ensembl (Harrison et al., 2023)), VISTA enhancers in the region and CNVs from the DGV. This assists with interpretation to determine whether there are genes in the region matching the phenotype, enhancers which might influence genes within the TAD regions, but not necessarily within the CNV region and to identify other similar CNVs reported which might be pathogenic. Overall, this tool gives researchers some insight into the three dimensional organisation of the genome, how it interacts and what implications this may have on the development of disease. Figure 3.3 highlights three different scenarios where TADs are interrupted by CNVs (A, B, C and D) and one scenario where the CNV has no effect on TAD boundaries (D).

There is also a statistics tab which shows gene and clinical descriptor matches to the variant being investigated across 500 random locations within the genome. Clinical descriptors are given in the form of Human Phenotype Ontology (HPO) terms. The score for gene match is related to how many genes in the potentially affected TAD could be related to the patient's phenotype.

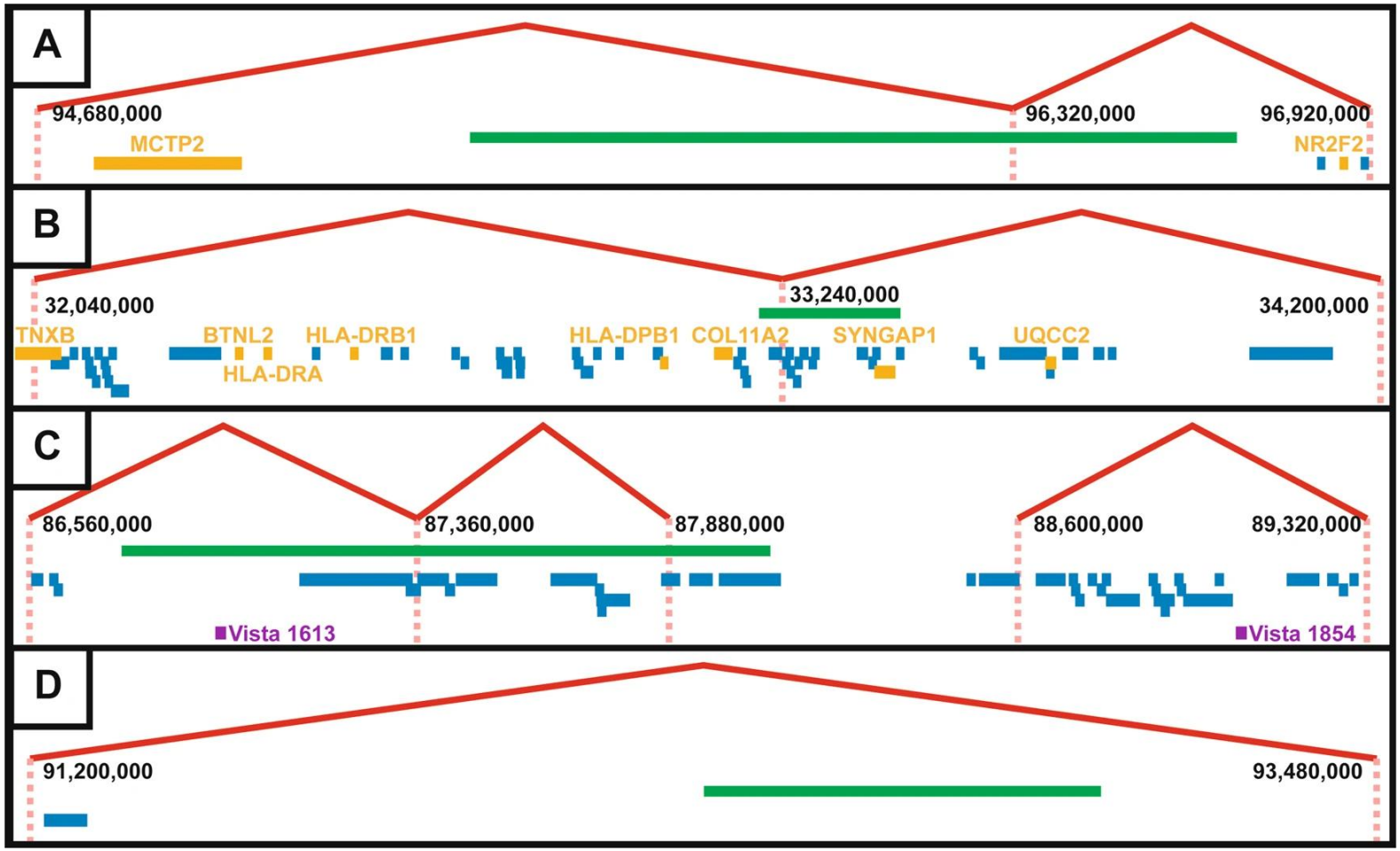


Figure 3.4: Example cases on ClinTAD showing topologically associated domains (red) and the boundaries (dashed), the copy number variant (green), genes with no phenotype matches (blue), genes with phenotype matches (orange), VISTA enhancers (purple), blue lines represent genes. Genes are associated with phenotypes that match to the patient (orange). CNVs which overlap a TAD boundary are represented in picture A and B as well as phenotype matches in adjacent TADs showing that TAD interruption could have contributed to the patients' phenotype. In picture C there were overlapping TAD boundaries, but no phenotype matches and D showed no overlapping TADs nor phenotype matches, lowering chances of being clinically relevant CNVs. Picture from Spector & Wiita, 2019, *Reproduced with permission from Springer Nature*

The weighted HPO matches score counts the number of phenotypes in nearby genes matching to the patient and considers the frequency in which each HPO occurs. This can then be used to compare the CNV to the randomly generated CNVs. A rare phenotype has a higher score compared to common phenotypes with a lower score. If the score is much higher at the actual location compared to other random locations, the likelihood of pathogenicity is higher and thus has a high probability to be disease-causing.

3.2.8 Copy number variant validation using Array CGH

Seven of the CNVs identified were validated by CMA using Agilent SurePrint G3 ISCA v2 CGH 8x60K microarray platform as per manufacturer's instructions (www.genomics.agilent.com). Agilent CytoGenomics 5.3 was used for data analysis with the relevant tracks from DGV, International Standards for Cytogenomic Arrays, OMIM and in house tracks. This analysis was carried out by the National Health Laboratory Service (NHLS), Braamfontein in parallel to the study as part of routine medical care.

3.3. Results

3.3.1 Applying chosen CNV calling tools to the DDD-Africa dataset

A total of 27,467 CNVs were identified between the three tools when applied to the DDD-Africa dataset. The number of CNVs (4703) identified using CANOES (Table 3.1) represents an average of ~16 CNVs per sample. After filtering parameters this was reduced to an average of ~3 CNVs per sample (773 CNVs). A total of 6539 CNVs were identified after XHMM was applied to the data from 287 individuals. These CNVs represent only 185 individuals as 102 were filtered out after GC content analysis and normalisation. The average CNV count per sample for XHMM was thus ~35. After filtering steps were complete, a total of 226 CNVs (>100kb) were identified from 70 probands, detecting an average of ~3 CNVs per patient. A total of 16,225 CNVs were identified using CLAMMS (Table 3.5), which is by far the tool detecting the most CNVs overall. An average of ~57 CNVs were called per individual before filtering parameters were in place and after filtering parameters were applied, 99 CNVs were identified from 67 affected individuals (~1.5 CNVs per individual).

Table 3.1: Summary of CNVs called from the DDD-Africa dataset with each tool.

	CANOES	CLAMMS	XHMM
Total CNVs	4703	16225	6539
Total DEL	3551	4241	3101
Total DUP	1152	11984	3438
Average size	152,66kb	37,41kb	77,71kb
Largest CNV	9081,04kb	12093,77kb	8664,53kb
Smallest CNV	0,424kb	0,29kb	0,29kb

The majority of the CNVs identified by CANOES were deletions as opposed to duplications (Figure 3.5). In contrast to CANOES, duplications represented the majority of CNVs from the CLAMMS output. A close to 50/50 ratio of deletions to duplications were identified by XHMM; however, there were slightly more duplications than deletions.

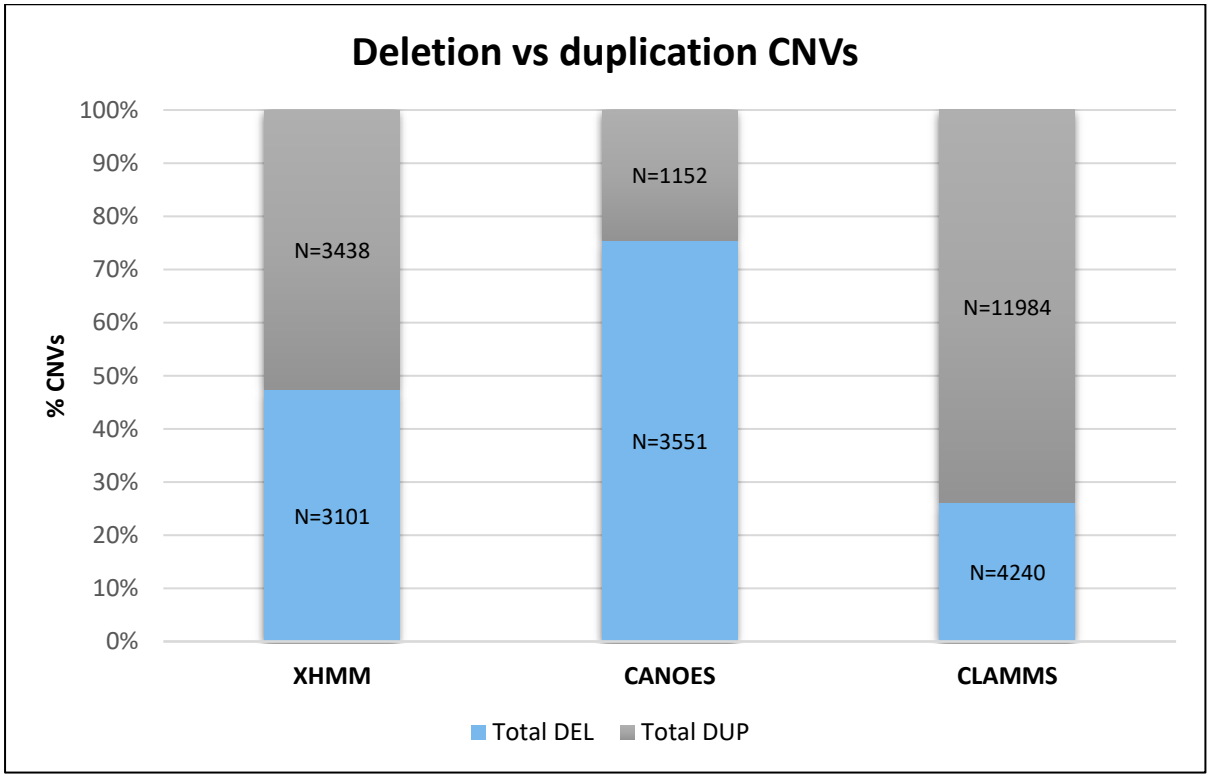


Figure 3.5: Proportion of deletions (DEL) and duplications (DUP) identified using the three CNV tools.

3.3.1.1 Comparison of CNVs identified from the DDD-Africa dataset

The number of CNVs identified with each tool was counted as well as the number of CNVs overlapping between the tools (Figure 3.6). As the breakpoints differ slightly from one tool to the next, the CNVs were not necessarily identical, but presented with at least a 50% overlap when comparing the CNVs identified by the different tools. These overlapping CNVs were also identified in the same individual in each instance. There was a total of 952 overlapping CNVs between CANOES and CLAMMS [575 deletions 377 duplications] of which 319 CNVs overlapped with XHMM as well. A total of 299 of these overlapping CNVs were within 104 different affected individuals [160 deletions, 139 duplications]. A total of 815 CNVs overlapped between CANOES and XHMM, of which 313 CNVs were from 70 affected individuals. Between CLAMMS and XHMM, a total of 707 CNVs overlapped and 280 CNVs within 70 different probands. There were a total of 319 CNVs which overlapped between all three tools of which 164 were deletions and 155 duplications. The size of these CNVs ranged from 0.38kb up to ~9Mb and the majority of these are recurring CNVs identified from a number of samples. Only 100 of the 319 CNVs were distinct within only one sample.

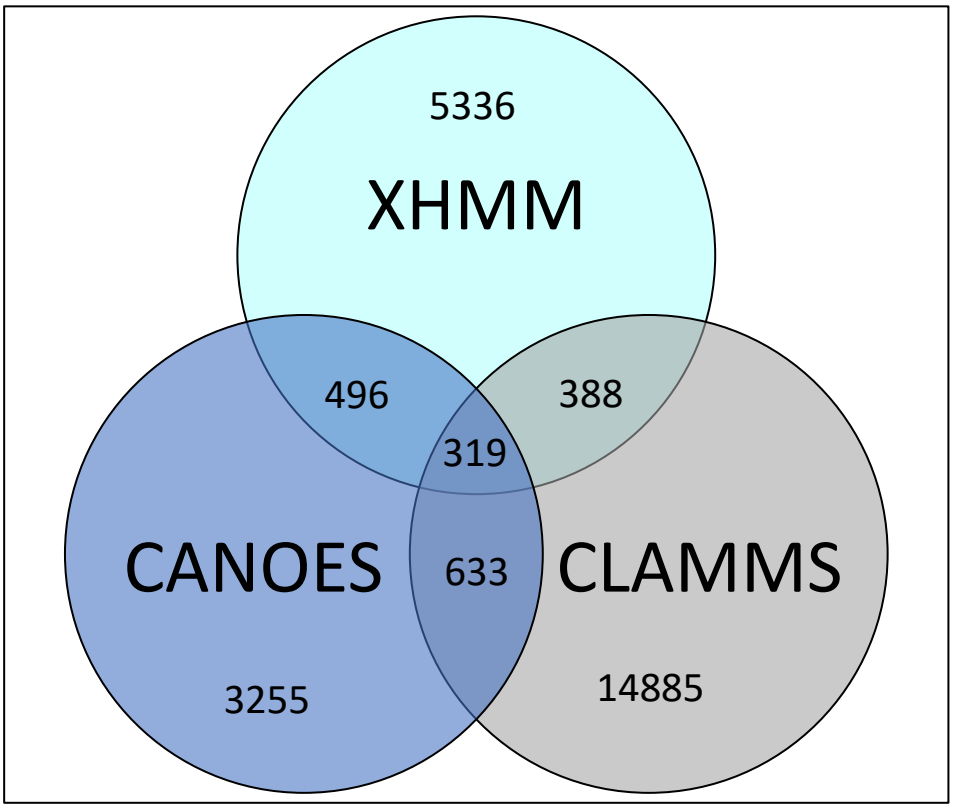
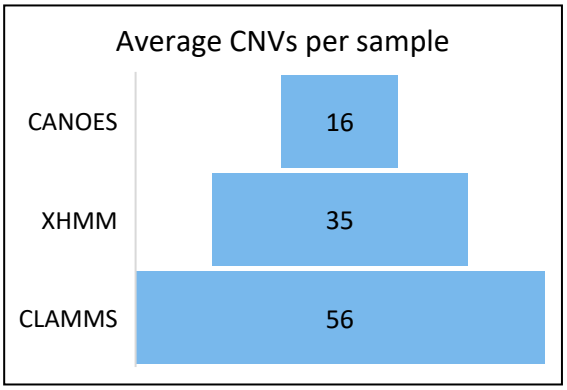


Figure 3.6: Overlapping CNVs between the three CNV tools when implemented on DDD-Africa data.

As seen in Figure 3.7 A and B below the more CNVs called by a tool, the smaller the size of those CNVs and on the contrary, if the average CNV size called by a tool was larger, fewer CNVs were called on average per sample.

A.



B.

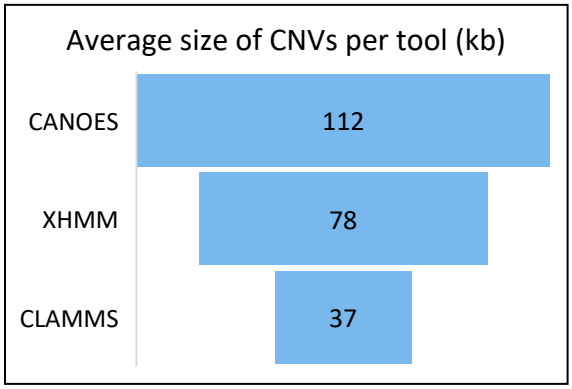


Figure 3.7: (A) Average number of CNVs called per sample for each CNV tool and (B) Average size of the CNVs called for each tool.

This study shows that these bioinformatics tools can identify CNVs of a wide range of sizes ranging from <100kb up to several megabases and shows a superior overall range of CNVs identified compared to most reported CNVs from microarray (Figure 3.8).

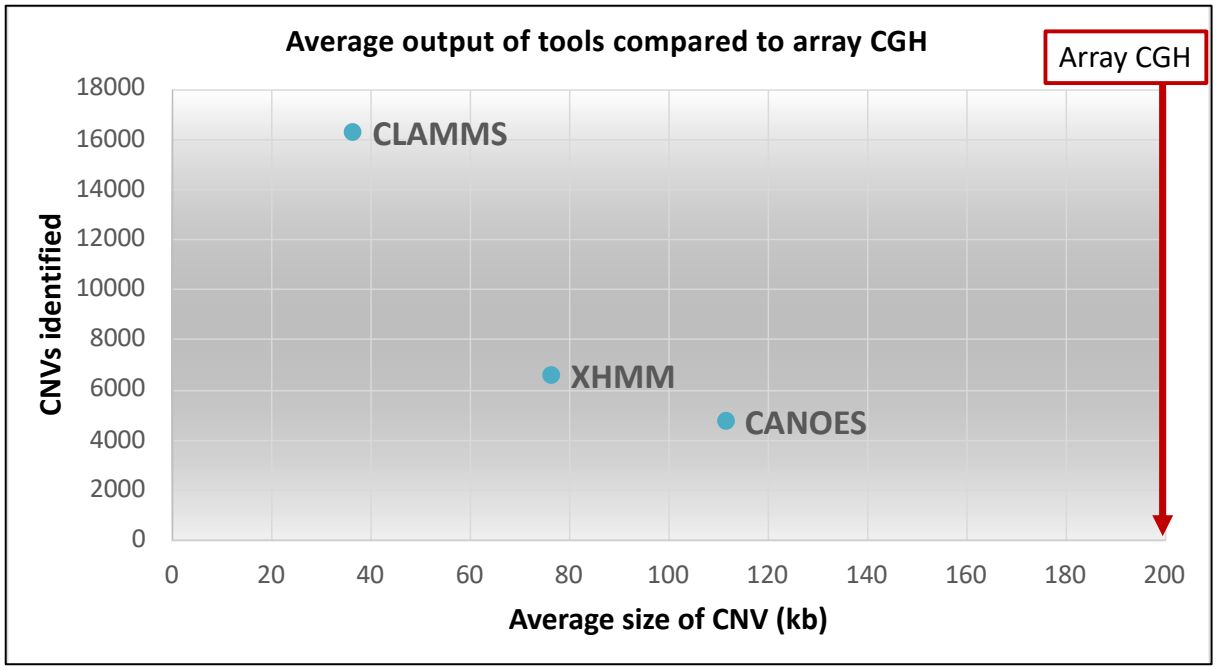


Figure 3.8: Average size of CNVs identified by CLAMMS, CANOES and XHMM. The average size of the CNVs is significantly smaller than the minimum size reported from Array CGH. Even though the CNV tools can also identify much larger CNVs (up to several Mb) the range of identified CNVs seems to be wide.

The total number of CNVs called per sample also varies between the tools and as mentioned within the filtering parameters for CANOES, samples with more than 50 CNVs called were excluded from further analyses, thus the below boxplot (Figure 3.9) depicts these samples (<= 50 CNVs) for each of the three tools. Even though CLAMMS called the most CNVs per sample on average when considering the total samples, the number of CNVs called per sample in this subset is lower. The output from XHMM was the least variable with fewer outliers than CANOES and CLAMMS; however, a larger number of CNVs were called per sample for this subset.

The tools also varied in the range of CNVs called per sample and in the below Figure 3.10 it can be seen that XHMM presented with 102 samples with no CNVs calls. These samples were removed after GC content calculation as they were highly variable. This might be due to the statistical method used including principal component analysis which effectively removes GC content bias.

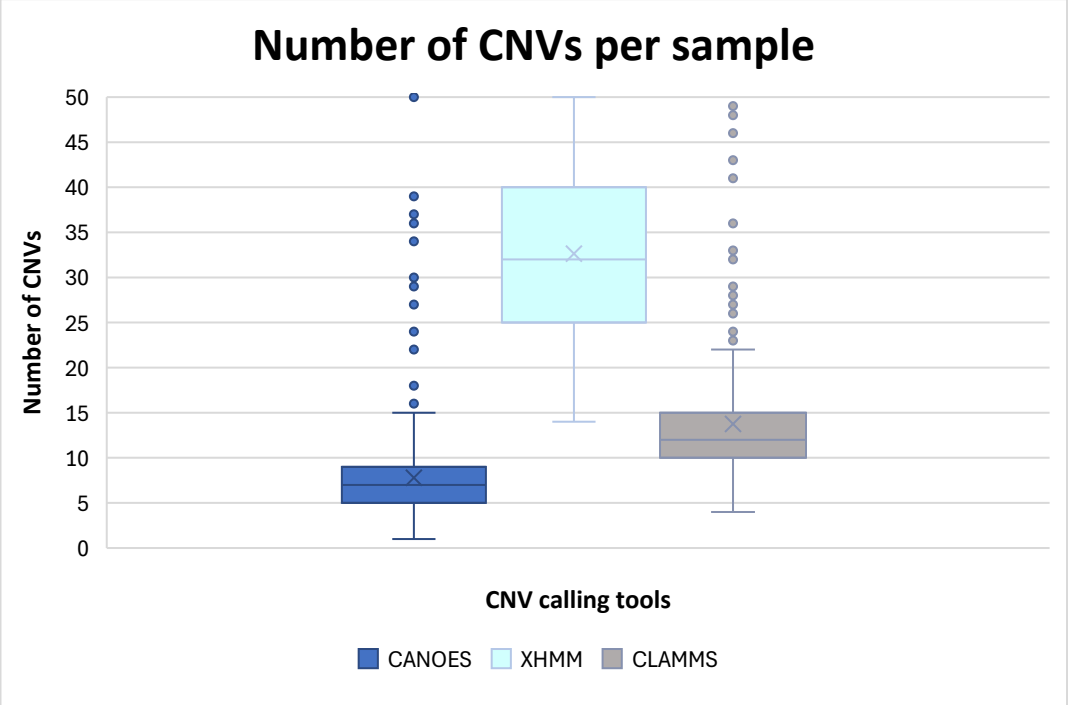


Figure 3.9: Box and whisker plot of the number of CNVs called by the three different CNV tools for each sample. Only samples with 50 or less CNVs called in total were included for this graph. CANOES (N=270), XHMM (N=163), CLAMMS (N=254).

The majority of samples had a range of between 1–10 CNVs called by CANOES whereas for XHMM the majority of samples had 21–50 CNVs identified. The majority of samples had 11-20 CNVs called by CLAMMS in this subset.

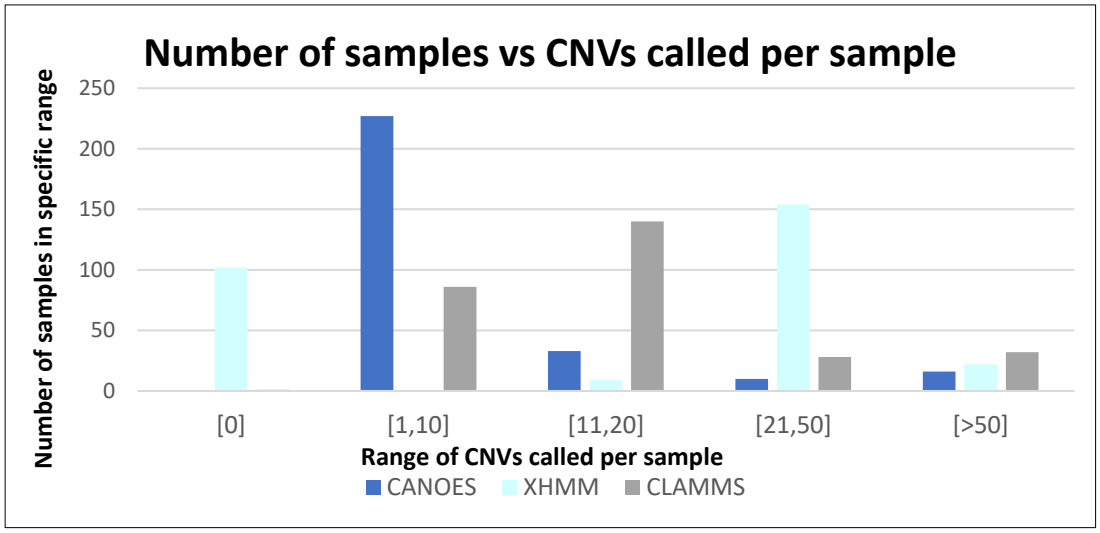


Figure 3.10: Number of patients with specific number of CNV calls (ranging from 0, 1-10, 11-20, 21-50 and >50 number of CNVs)

3.3.1.2 Computational cost of implementing CNV calling tools

CANOES was the most computationally costly as a total of ~34.1GB random access memory (RAM) was used in the steps to call CNVs. XHMM was second most computationally costly (3.3GB) followed by CLAMMS (531.5MB) which was the least computationally intensive in terms of memory usage. The below graph (Figure 3.11) depicts the percentage central processing unit (CPU) usage for each process of the three tools and as illustrated, CANOES required the most percentage CPU followed by XHMM and lastly CLAMMS. The CPU percentage is measured in terms of a single core and thus a CPU of over 100% indicates that the machine used multiple cores which is not an option for all machines. The most computationally intense task was the genReadcounts step of CANOES where read counts were generated for each sample using bedtools (Quinlan and Hall, 2010), taking an average of 1879.7 minutes and 27.5GB of memory. The total time per sample for CANOES was 11.8 minutes while XHMM and CLAMMS took ~1.9 minutes and 0.4 minutes respectively per sample to complete all tasks. The Amazon computational costs of $0.3469 per hour (https://aws.amazon.com/ec2/pricing/on-demand/) and monthly storage costs of $0.1047 (https://aws.amazon.com/ebs/pricing/) can be used to estimate physical costs incurred. Using these costs, it is estimated that CANOES had cost ~$20 for the 287 samples, XHMM cost ~$3 and CLAMMS ~$1 (Mpangase et al., 2021). The application of these tools to ES data is relatively cost-effective, which constitutes a critical factor for their implementation in LMICs.

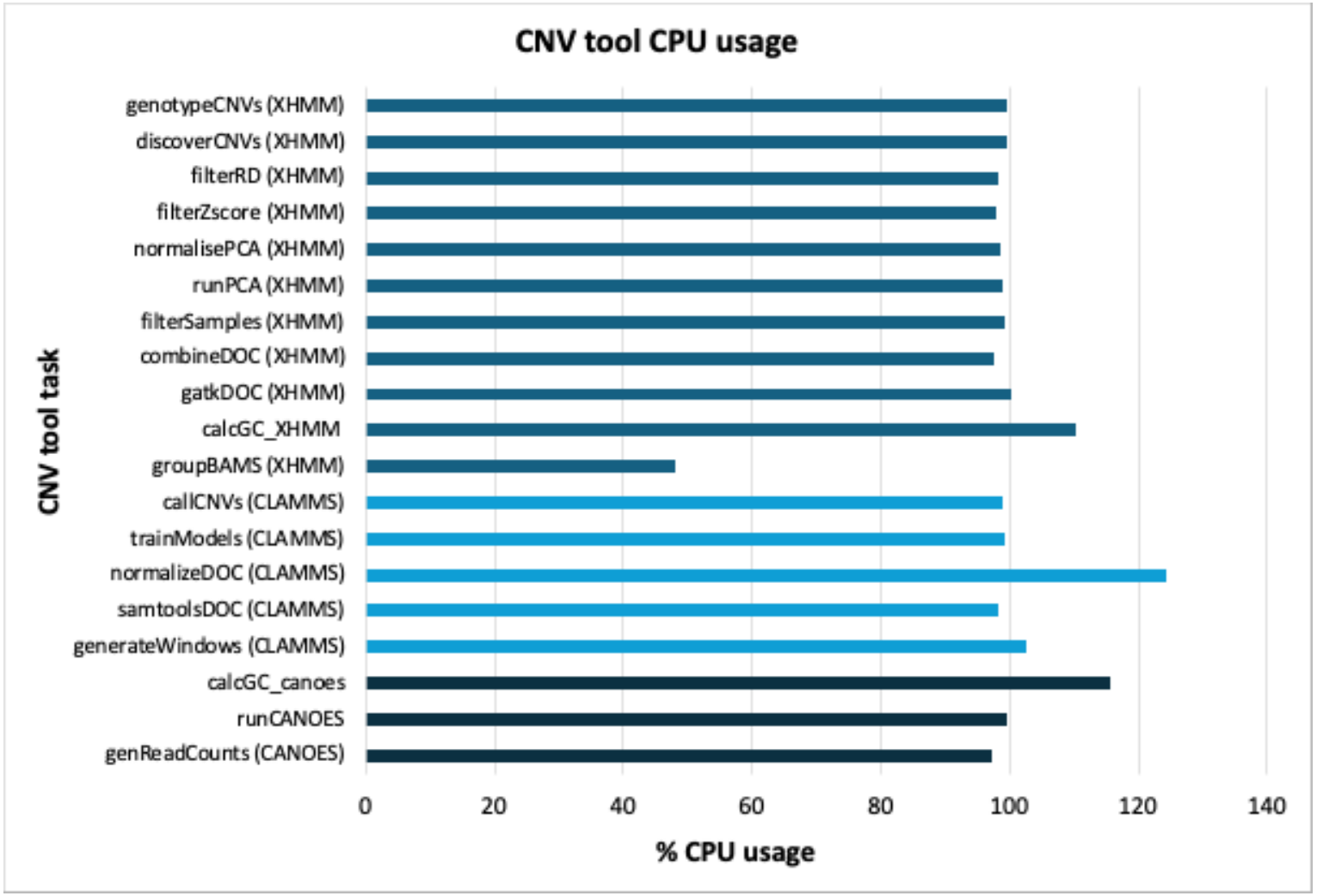


Figure 3.11: Percentage CPU usage from each task of the CNV tools. All tasks from XHMM at the top (teal), CLAMMS in the middle (light blue) and CANOES at the bottom (dark blue) are separately indicated, depicting median CPU usage for each of these tasks.

3.3.2 Classification of combined CNVs from all three tools

As mentioned in section 3.2.1, the DDD-Africa batch 1 dataset consists of 287 samples of which 117 are affected individuals and the remaining are parental samples. A total of 35 LP/P CNVs were identified within the proband samples; however, only six of these overlapped all three tools (Figure 3.12). These numbers do not include any CNVs classified as LP/P from unaffected individuals. An additional two of the three CNVs detected by CANOES and CLAMMS only were also shortlisted taking the total to eight. The remaining 27 CNVs were excluded due to not meeting all quality and filtering parameters as set out in figure 3.1. These eight CNVs (Table 3.1), identified from eight different affected individuals, were the final shortlist after analysis and filtering.

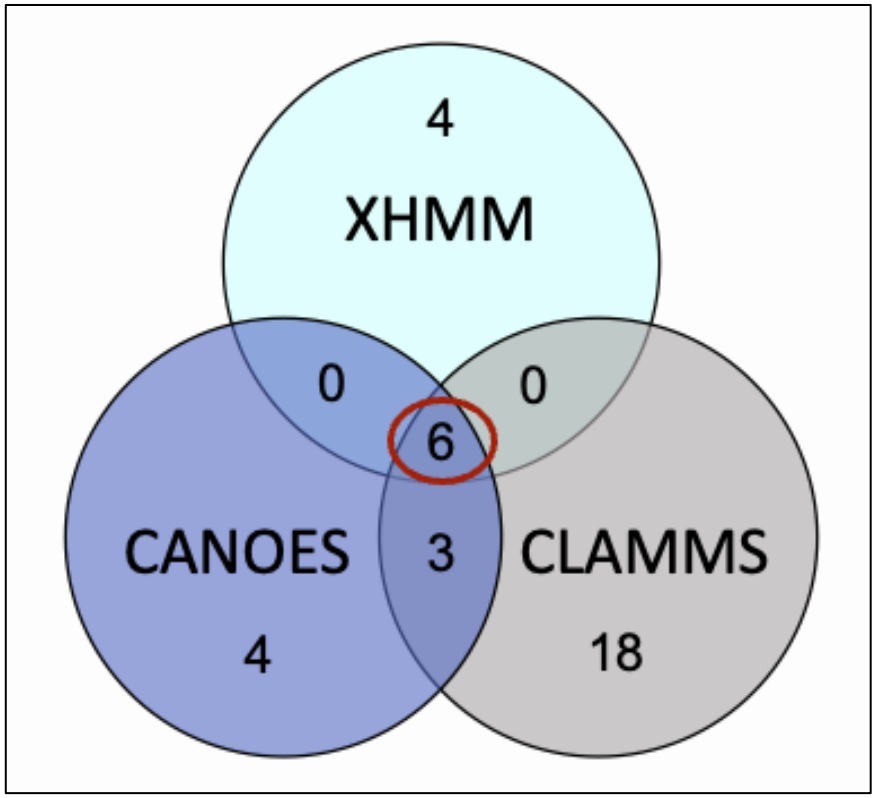


Figure 3.12: Venn diagram showing all LP/P CNVs identified from the probands only as classified by CNV-ClinViewer. Only six CNVs were shared between all three tools (encircled in red).

Seven of these CNVs were confirmed with Array CGH and the CNV from one sample could not be confirmed due to technical challenges. The DNA quality was not of sufficient quality and was depleted thereafter.

Parental data was inspected to confirm the inheritance of these CNVs although other LP/P CNVs identified in parental samples only were not analysed further in this study. These results were compared to each patient's phenotype and disease presentation in order to confirm whether these variants do indeed contribute to the disease phenotype. The average size of the eight CNVs is ~6.3Mb and all were deletions apart from one large duplication (~13Mb) which was identified on chromosome two. All eight patients were included within category A for moderate to profound ID/DD and four patients also had major and minor malformations (inclusion category C). The main clinical features of these patients can be seen in Table 3.2. Additional information regarding specific patient phenotypes is included in Appendix IX.

The number of CNVs which was shortlisted as LP/P from the affected individuals only were similar between the manual classification of CNVs and making use of the automated web-based classification tool (CNV-ClinViewer). There were indeed several CNVs which were excluded before manual classification due to low quality parameters which will be discussed in the following section (3.3.3). The remaining 27 CNVs not shortlisted were either too small (<100kb) or did not meet the filtering parameters, specifically the quality scores and were likely false positive CNVs. These were classified as LP/P with CNV-ClinViewer as this tool does not take quality scores into account before classification. After filtering, there were no additional CNVs added to the shortlist or classified as LP/P which were not also identified by manual classification.

Table 3.2: Likely disease-causing CNVs identified with the three different bioinformatics CNV tools.

Sample ID	CNV interval (GRCh38)	Chromosome position	HI/TS Genes	CNV type	CNV Size (kb)	Tools	ACMG/ClinGen classification	Pathogenicity score	Patient HPO terms
D3S_0009_01_1	**16:69245485-72960252**	**16q22.1-22.3**	*AP1G1, CALB2, NFAT5, SF3B3, TAT, ZFHX3*	DEL	3715	CANOES, CLAMMS and XHMM	Pathogenic	1.05	HP:0000750:Delayed speech and language development HP:0011343:Moderate global developmental delay HP:0001643: Patent ductus arteriosus
D3S_0040_01_1	**6:153282150-157897057**	**6q25.3-25.1**	*ARID1B, ZDHHC14*	DEL	4615	CANOES, CLAMMS and XHMM	Pathogenic	1.00	HP:0000750: Delayed speech and language development HP:0011344: Severe global developmental delay
D3S_055_01_1	**8:26338775-31173210**	**8p21.2-p12**	*EXTL3, HMBOX1, PBK, RBPMS*	DEL	4834	CANOES, CLAMMS and XHMM	Pathogenic	1.05	HP:0000750: Delayed speech and language development HP:0002283: Global brain atrophy HP:0000252: Microcephaly HP:0011343: Moderate global developmental delay

Sample ID	CNV interval (GRCh38)	Chromosome position	HI/TS Genes	CNV type	CNV Size (kb)	Tools	ACMG/ClinGen classification	Pathogenicity score	Patient HPO terms
D3S_0070_01_1	**11:24497007-33142168**	**11p14.3-p13**	*CSTF3, EIF3M, ELP4, IMMP1L, MPPED2, PAX6, RCN1, WT1*	DEL	8645	CANOES, CLAMMS and XHMM	Pathogenic	1.15	HP:0000526: Aniridia HP:0000750: Delayed speech and language development HP:0025502: Overweight HP:0011344: Severe global developmental delay
D3S_0085_01_1	**12:1262908-5647947**	**12p13.33-p13.31**	*CACNA1C, CCND2, PRMT8*	DEL	4385	CANOES, CLAMMS and XHMM	Pathogenic	1.05	HP:0000750: Delayed speech and language development HP:0000252: Microcephaly HP:0011344: Severe global developmental delay
D3S_0101_01_1	**17:53822906-58756569**	**17q22-q23.1**	*ANKFN1, APPBP2, CLTC, DYNLL2, MPO, MSI2, NOG, RAD51C, RPS6KB1, SRSF1, SUPT4H1, TRIM37, VEZF1, VMP1, YPEL2*	DEL	4934	CANOES, CLAMMS and XHMM	Pathogenic	2.05	HP:0001344: Absent speech HP:0000252: Microcephaly HP:0011344: Severe global developmental delay

Sample ID	CNV interval (GRCh38)	Chromosome position	HI/TS Genes	CNV type	CNV Size (kb)	Tools	ACMG/ClinGen classification	Pathogenicity score	Patient HPO terms
D3S_0109_01_1	**2:41947349-55322795**	**2p22.1-p16.3**	*FBXO11*	DUP	13375	CANOES and CLAMMS	Likely Pathogenic	0.90	HP:0001344: Absent speech HP:0002059: Cerebral atrophy HP:0011344: Severe global developmental delay
D3S_0114_01_1	**12:23534153-29783848**	**12p12.1-p11.21**	*KRAS, PTHLH, SOX5*	DEL	6250	CANOES and CLAMMS	Pathogenic	1.35	HP:0000750: Delayed speech and language development HP:0001508: Failure to thrive HP:0000252: Microcephaly HP:0011343: Moderate global developmental delay

*Array CGH was performed on grey shaded samples for validation of the CNVs; HI/TS genes based on %HI (Haploinsufficiency score, threshold: <10%) for deletions and probability of TS score (pTS threshold: >0.993) for duplications.

All eight CNVs were unique and thus none of these LP/P CNVs were overlapping and none had been reported previously on DECIPHER or ClinVar. Two CNVs were identified on chromosome 12; however, in different regions and thus no recurrent LP/P CNVs were identified in this cohort. One patient presented with a CNV including a deletion of the *ARID1B* gene associated with Coffin Siris syndrome which is one of the most common genes identified in DD patients via genotype-first approaches (Wright et al., 2015, van der Sluijs et al., 2019, Lu et al., 2021, Li et al., 2024). The presentation of this condition is non-specific, making it harder to diagnose with a phenotype-first approach. Clinical features most often associated with this disorder is developmental delay, microcephaly, abnormalities of the 5th fingers or toes and dysmorphic facial features (Santen and Clayton-Smith, 2014). Five of the eight CNVs are *de novo* and for the others the inheritance could not be confirmed as only one parent was tested.

Inspecting the classified variants present in all samples and not only within the affected individuals, highlighted additional LP/P CNVs. These CNVs were not further analysed as incidental findings in unaffected individuals were not included within the scope of this study. The proportion of pathogenic, likely pathogenic, VUS and benign variants classified by CNV-ClinViewer and identified with all three CNV calling tools is illustrated as a pie chart (Figure 3.13). A total of 20774 VUS CNVs were called and 7574 of these were called in affected participants. These VUSs represent ~76% of the total CNV calls reiterating the difficulty of dealing with these type of CNVs. The below numbers of CNVs identified represent the unfiltered data thus no exclusions based on the quality of the data, sample type or phenotypic match were made. The 148 LP/P CNVs were thus identified with CNV-ClinVIewer before manual filtering (step 1 of Figure 3.1B) and before shortlisting the final eight LP/P CNVs identified in affected individuals.

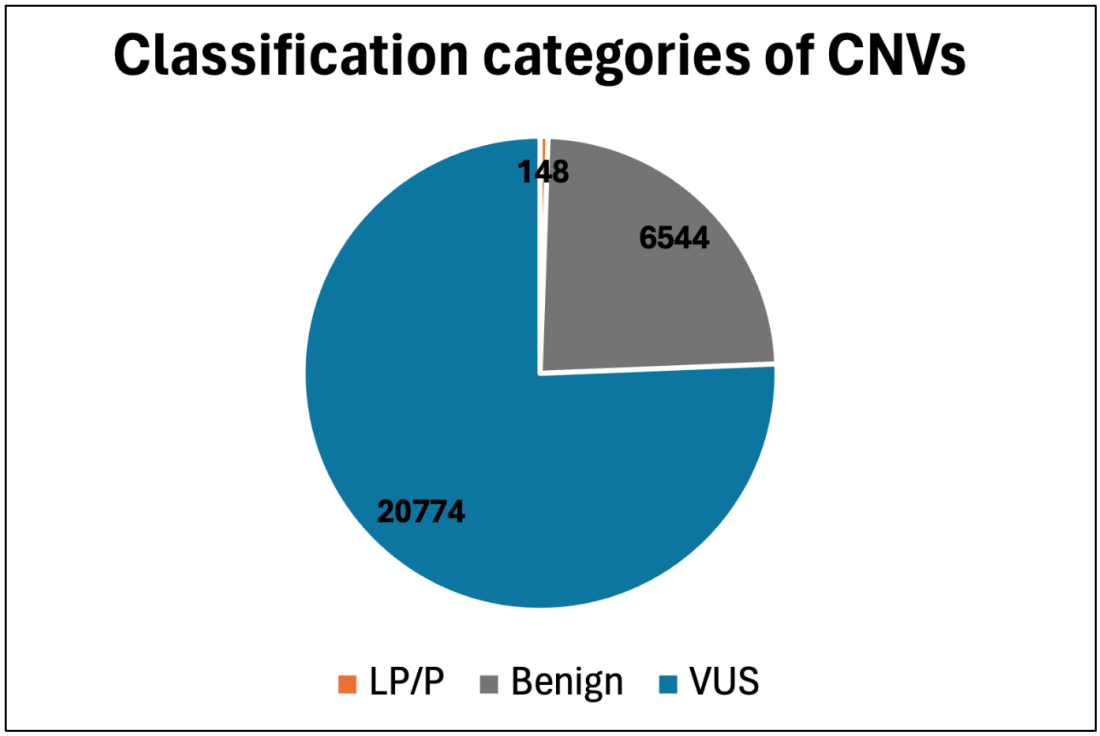


Figure 3.13: Number of CNVs classified as likely pathogenic/pathogenic, benign and VUS from all three CNV calling tools (CANOES, CLAMMS and XHMM) for all samples after classification with CNV-ClinViewer.

3.3.3 Additional analyses of shortlisted CNVs not meeting manual quality filtering parameters

The below CNVs are derived from the comparison between the results from CNV-ClinViewer and those from the manual classification process. An additional pathogenic CNV was highlighted during analysis with CNV-ClinViewer which was initially filtered out as it was small (<100kb) and only identified with one tool. Although the CNV from individual D3S_0044_01_1 [GRCh38/hg38] 16p13.3p13.3(3717298_3759144)x1] was identified using both CANOES and CLAMMS, it did not pass QC for CLAMMS. It did pass QC from CANOES but was not included in manual classification due to its size. It was decided to investigate the possibility of this CNV leading to development of the patient's phenotype. This CNV spans the first exon of the *TRAP1* (OMIM* 606219) gene encoding a mitochondrial chaperone protein that is member of the heat shock protein 90 family. This protein has ATPase activity and interacts with tumour necrosis factor type I (https://www.ncbi.nlm.nih.gov/gene/10131).

The second gene involved is he CREB binding protein (CREBBP) gene is an OMIM Morbid gene (OMIM* 600140) and linked to Rubeinstein-Taybi syndrome specifically due to microdeletions of chromosome 16pl3.3 (Petrij et al., 1995).This gene contains 31 exons of which 17-31 is involved in the above CNV. After clinicians evaluated the given information, it was excluded from feedback of findings as the phenotype of the patient (moderate developmental delay, delayed speech and language development, patent ductus arteriosus) did overlap sufficiently with that reported of patients with other similar CNVs in this region. This CNV was excluded from the final count of patient diagnoses and was flagged as a VUS.

One other CNV, also flagged for further investigation, was identified in patient D3S_0062_01_1 [GRCh38/hg38] Xq28(153723048_154653436)x1]. This CNV is located on the X chromosome and includes 54 genes of which six are HI, one being the well-known *MECP2* gene (OMIM* 300005). After a more thorough investigation, it was seen that while called as pathogenic, this patient is male and thus not expected to be viable with this large loss on the X chromosome. The IGV tool was also used to identify whether there were any reads in this region for this patient. If there were no reads in this entire region then a deletion would have been very probable; however, some reads were identified which cannot be the case if this region is deleted from the only X chromosome copy. This sample was also one which presented with more than 50 CNVs from the CLAMMS CNV output. This result will not be fed back to the family as it is a likely false positive result.

One VUS was highlighted during the manual classification step which contains genes involved in specific syndromes that could contribute to this patient's phenotype. This 345kb deletion identified in individual D3S_0001_01_1, [GRCh38/hg38]1q21.3q21.3(153930125_154275770)x1]**,** spans 21 genes including morbid genes *RPS27* (OMIM* 603702), *TPM3* (OMIM* 191030) and *HAX1* (OMIM* 605998). To investigate the CNV further, it was uploaded to ClinTAD to determine whether there is any additional evidence to link this variant to the patient phenotype. It was seen that the CNV does span a TAD boundary but does not seem to affect a gene matching the phenotype. No enhancer elements are included in this region and the weighted score also does not suggest that the phenotype is rare and specific to this region.

A conclusive result based on the ES result and available clinical data could unfortunately not be reached for this patient. This variant was also validated with CMA and the maximum size of the CNV on CMA was 586kb which is a total of 241kb larger than the original size as identified by ES CNV tools. Seven additional genes are present in this larger region, including *GATAD2B* (OMIM* 614998) which is also a morbid gene, causing GAND syndrome. This syndrome includes global developmental delay with moderate to severe intellectual disability, poor speech and language development as well as seizures which are overlapping with this patient's phenotype. After review by clinicians this CNV does partially contribute to the patient phenotype but the recommendation was made to send this sample for WGS to identify other possible variants contributing to the phenotype.

3.4 Discussion

3.4.1 Comparison of CNVs identified from the DDD-Africa dataset

After the three CNV tools were applied to the 287 samples, the tool with the most calls was CLAMMS followed by XHMM and lastly CANOES. It was also seen that there is an inverse relationship between the number of CNVs called and the average size. For the tool with a smaller average size of CNVs called (CLAMMS), more CNVs seem to have been called; however, the inverse is true for CANOES which has an overall larger average CNV size.

The majority of CNVs identified by CLAMMS were duplications rather than deletions in this dataset; however, for XHMM it was almost 50/50 which is also evident from previous studies (Tan et al., 2014, Hong et al., 2016, Pounraja et al., 2019). A majority of deletions were also identified by CANOES in this study. Most CNV calling algorithms have been shown to detect deletions over duplications, especially from read depth as the decreases relative to diploidy is easier to distinguish than increases (Gordeeva et al., 2021). The copy number change from N to N +1 is hard to distinguish when using dept of coverage especially in repeated regions where N is large (Teo et al., 2012).

The high variance in depth of coverage makes it even more difficult to detect duplications. It has been found that CLAMMS does have an increased rate of false positive duplications and thus could explain the outcome in this study (Packer et al., 2016). CANOES has been shown to provide superior calling compared to other tools when the average depth of coverage is lower and with fewer reference samples present. This is evident in this study where the average depth of coverage was 40X and thus lower than a usual exome with ~100X coverage. This could also have led to several CNVs not being detected by the other calling tools.

XHMM seems to be much more strict with filtering GC content variability as seen in the output, this tool excluded 102 samples before analysis. Accurate bioinformatics CNV calling can be hampered by factors such as sequencing difficulties in low complexity regions and GC content affecting the over- and underrepresentation of target regions which can then be interpreted as CNVs. Some CNVs might have been overlooked due to the fact that the tools require at least 20-30 mapping quality and some regions of sequence complexity or repetitive regions can influence the mapping quality and lead to false positive or – negative results. CNVs could also be missed due to the size as assessed in a previous study, callers mostly missed CNVs spanning less than 3 exons in size (Hong et al., 2016). In this study SVs and intragenic CNVs (< 1 exon) may have been missed even though the tools were still able to detect some smaller CNVs which would be overlooked by CMA.

Computational costs of these tools should be taken into account as evaluated in this study, CLAMMS was the least computationally exhaustive. It is recommended to use an ensemble approach and thus make use of different tools in order to maximise sensitivity and specificity of the output. However, this can be computationally intense and require more refined bioinformatics skills. It is thus not always feasible, especially in LMICs where resources are limited and thus choosing one tool with the best performance for the setting, would be sufficient. In settings where computational power and storage space is an issue or where large population studies are carried out, CLAMMS would be the best tool to implement. CANOES would be the recommended tool for rare disease studies and combined with XHMM this would lead to an even better yield. In cases where overdispersed data or datasets from different calling methods are available, XHMM would be best to implement.

Depending on sample size, resource availability and bioinformatics knowledge and experience, different tools might thus be recommended.

A total of eight LP/P CNVs were identified in eight different individuals resulting in a diagnostic yield of ~7%. This is additional to the yield observed after the SNV only pipeline was completed on the ES data. As mentioned in the methods section, most of the patients did have genetic testing prior to the ES. A total of 89% did indeed undergo MLPA testing and a further 3% CMA testing. It is possible that an increased diagnostic yield could have been achieved if patients with no prior genetic testing were included. These results were discussed internally with a multidisciplinary team who are experts in the field of genetics and variant interpretation before final classification decisions were made. The degree to which these patients are affected can perhaps also be expected when inspecting the CNV sizes and number of protein coding regions affected. All these CNVs were large (>1Mb) with an average of ~6.3Mb in size. Incidental findings were not explored in this study as it is not within the scope of this PhD. Data generated from this study could indeed be used for this purpose in future studies with appropriate ethics permissions.

As mentioned previously, most DD variants are *de novo* which is also evident from this study as the majority (5/8) were observed as such. Establishing whether the remaining CNVs are *de novo* was not possible as only one parent was available for recruitment. Sequencing more trios instead of single affected samples will not only assist with variant interpretation but also help solve more CNVs in the benign or pathogenic categories.

Evaluating the findings from this study, CANOES had the best output overall in detecting rare, true positive disease-causing CNVs. It would be more accurate to combine CANOES with XHMM which does identify CNVs from sex chromosomes and have been proven to complement each other (Backenroth et al., 2014). XHMM and CANOES have greater power to detect rare CNVs when compared to CLAMMS as seen from the fact that CLAMMS called more CNVs per sample overall and also fewer (6/8) overlapping with the shortlisted LP/P CNVs. The number of overlapping CNVs between all three tools are small in comparison to the total number of CNVs called; however, as with the truth set, most of the LP/P CNVs of interest were within this group.

It was seen that not all the shortlisted LP/P CNVs were within this group and thus some of the plausible disease-causing CNVs would have been missed if these were the only CNVs investigated. One tool did not identify all eight CNVs and thus it is important to not overlook individual results when using multiple CNV calling tools. Results from multiple CNV calling tools are merged in many cases and only overlapping CNVs identified; however, two LP/P CNVs would have been missed if this was done in this study.

The diagnostic yield of the current gold standard for CNV detection (CMA) is 15-20% (Miller et al., 2010), which is significantly lower than ES. In a recent scoping review, it was shown that ES for diagnosing neurodevelopmental disorders outperforms CMA by 10-28% (Srivastava et al., 2019), further supporting the combined SNV and CNV analysis approach from ES data. It is important to note that ES CNV tools have limitations due to their inability to detect specific types of variations for instance balanced structural variants (translocations and inversions), mosaicism and CNVs less than 50bp in size. Although analyses and technologies are improving to address these shortcomings, it should be considered when implementing these tools. While it would be ideal to validate exome CNVs with methods such as microarray, it is costly and often not feasible in LMICs. Accurate CNV calling incorporating thorough quality control can help limit false positive results. This is further evidenced by eradicating the need for Sanger sequencing validation of SNVs when proper quality control is carried out (Strom et al., 2014). Long-read sequencing has the ability to detect these structural variations as well as SNVs, making it ideal to implement as a single assay to replace all the above methods (Mantere et al., 2019, Oehler et al., 2023, Olivucci et al., 2024). At present, this technology is too costly to use as first-tier test; however, as costs decrease this will most likely become a possibility for future consideration.

3.4.2 CNV classification and interpretation

The importance of manual CNV interpretation was also highlighted in this study, not only using an automated classification tool, but also ensuring that CNV quality parameters and phenotypes of the patients match. Two separate smaller CNVs were called with one tool in this study, for one sample in particular, whereas other tools identified it as one large CNV. These two separate CNVs were classified as VUS by CNV-ClinViewer although this is one CNV which should be classified as LP.

This was picked up by manual CNV inspection and could have been missed if only the outcome of CNV-ClinViewer was used.

As with the truth set data, not all of the CNVs identified as LP/P initially were within the filtering parameters. A small number of CNVs identified as LP/P did not meet the filtering criteria; however, none of these were shortlisted as part of the CNVs contributing to the additional diagnostic yield. These variants will likely be reanalysed in future and should be validated with another method (i.e. CMA) to ensure they are not false positive results. These filtering parameters might be too strict, but due diligence was kept by classifying all the CNVs, not only those within the filtering parameters. In this way, the findings were not limited to only the strict filtering parameters, in case a possible disease-causing CNV was overlooked. Future studies exploring the best balance of filtering parameters versus increased yield making use of a larger sample set would be very useful for implementation.

None of the eight identified CNVs have been previously reported on DECIPHER; however, similar CNVs with large overlapping regions have been reported and in some instances these patients also have similar phenotypic features as the DDD-Africa patients. The results also reiterated that the genotype-first approach is more effective for many patients with DD. Known conditions, like Coffin Siris, can be missed by clinicians as these phenotypic features overlap, indicating the need for exome sequencing in DD patient cohorts.

Many identical CNVs were also identified in multiple individuals within this cohort which can be attributed to large population CNVs forming part of normal variation. Many unaffected individuals also carried CNVs present within affected individuals which led to the application of benign classification codes for these CNVs. These variants were not further investigated or highlighted in this study but should also be explored in future projects to contribute to public databases as part of normal population variation. This further highlights the importance of having databases like DECIPHER with previously reported variants from individuals of different origins and ethnic backgrounds. The accuracy of diagnosing disorders will increase as similar cases are reported and submitted to databases like DECIPHER, DGV, ClinVar and gnomAD.

The more CNVs reported on open-source databases, the better interpretation and classification of variants will be. Additional CNV publications and ClinVar submissions from understudied populations will expand the size, scope and improve the resolution of clinically relevant CNVs in the public domain. Although these public data repositories have contributed to diversify data, more effective production and sharing of genomic datasets is needed to advance genomic medicine globally. At present there is still a bias in the proportion of non-European population CNVs reported with an overrepresentation of the European population, making comparison with other populations more difficult.

Recent initiatives have been established to facilitate African data sharing and empower health experts by availing tools, training and coordination to strengthen laboratory and bioinformatics capacity (Mulder et al., 2016, Mulder et al., 2018, Makoni, 2020, Lumaka et al., 2022). There should thus be a focus on training in order to gain expertise for CNV interpretation and classification within Africa. More training opportunities and more skilled people in this specific niche will ensure additional diagnoses for patients. International collaborations and training could be crucial to resolve the true impact of CNVs and build strong core groups with expertise, experience and technical competence to accurately report on CNVs in diagnostic context within LMICs. CNV calling from existing ES datasets from non-European individuals may therefore be an important analysis to invest in. There is still limited data and genetic services in most of Africa, making it difficult to translate research into clinical healthcare services (Kamp et al., 2021). A current and thorough analysis of the cost-benefit for ES would be beneficial towards motivating the adoption of ES as a first-tier test in resource constrained environments.

Even with more variants being reported, there are still many difficulties in classifying CNVs. Multiple types of evidence need to be considered when classifying a CNV including protein coding regions, number of genes involved, inheritance pattern and population frequency. Additionally, interpreting CNVs from ES data is even more difficult as the breakpoints are not identified, especially if within intronic regions. Evidence of gene dosage is one of the main contributors to the pathogenicity of a CNV. Many genes have been established as HI or TS by a curation group like ClinGen which aids CNV analysis and classification if these genes are present within the given CNV.

Gene dosage studies and curation have unfortunately not been completed for all genes, contributing to the difficulty of CNV interpretation.

Development of a disease phenotype may also, in some cases, be caused by a combination of CNVs and other genetic variants rather than a single CNV. Crucial genes (OMIM Morbid) can be disrupted by a CNV causing development of a specific phenotype in patients; however, this is not always the case as there are many genes included which may not match the phenotype of the patient exactly. It is thus also of immense importance to have detailed clinical phenotyping for the interpretation of CNVs and determining the contribution to specific patient's phenotypic features or disease manifestation. The use of standard HPO terms should also be encouraged and should be entered together with the CNV on public databases. Currently, there are not enough HPO terms submitted on public online databases making it ineffective to diagnose Mendelian disorders (Fellner et al., 2021). An increased HPO submission will significantly improve the ability to evaluate clinical fit and identify new syndromes if publicly available CNVs have matching HPO data (Gargano et al., 2023, Maassen et al., 2023).

The use of multiple tools for accurate interpretation of the analysed CNVs is important as different information regarding the variants can be aggregated and a few such tools have been developed (Geoffroy et al., 2018, Gurbich and Ilinsky, 2020, Fan et al., 2021, Gažiová et al., 2022). ClassifyCNV was the first tool to automate the classification of CNVs using the updated ACMG/ClinGen guidelines (Gurbich and Ilinsky, 2020). These tools can reduce the time to diagnosis as it can be integrated into existing pipelines. This also assists scientists and clinicians by eliminating the need to search for gene content, population frequencies and gene dosage sensitivity on various genomic databases. A newer tool, MarCNV, combined with a machine learning-based pathogenicity predictor (ISV) has proven to reduce the number of VUS CNVs (Gažiová et al., 2022). CNV Extraction, Transformation, and Loading Artificial Intelligence (CNV-ETLAI), has been designed to automatically extract, transform and organise CNVs from literature into a database (Choi et al., 2022). As discussed in section 3.2.4 there are also TAD regions which can be disrupted and influence other nearby genes which may lead to the phenotype of the patient.

Tools like ClinTAD are helpful in compiling additional pathogenicity evidence for a variant which might not be identified as disease-causing at first. Even though this tool did not give any additional insight for the specific CNVs in this study, it will also become more important as additional information is gathered on genes and protein functions. Even though these tools are valuable, manual entry from users is still needed and thus this is an important factor for implementation within a clinical setting. These CNV interpretation tools should thus be used only for prediction as manual interpretation of the results is still needed especially for clinical decision-making.

The importance of reanalysis of data is also evident in this study as many CNVs classified as VUSs still remain concerning. Over 75% of the identified CNVs were classified as VUSs, leaving many patients without a definite diagnosis. A recent study reported that ~40% of the CNVs reported as VUS and reanalysed over an 8-year retrospective study have changed categories of which only 4.6% upgraded to LP/P. A total of 75% of the downgraded CNVs (to benign or likely benign) were below 500kb in size (Ji et al., 2021). Another, more recent study showed that comprehensive reanalysis of exome data from 5757 families with rare disease provided a molecular diagnosis for an additional ~1.5% of the families (Demidov et al., 2024). A total of 51 families received a diagnosis as well as 34 families with a partial diagnosis. Fifteen (0.26%) of these diagnoses can be attributed to CNVs.

A recent publication has shown that patients who still remain undiagnosed after ES, should first be considered for reanalysis of exome data rather than using WGS or other methods to expand the search (Schuermans et al., 2022). This is especially fitting in a LMIC setting where funds for additional testing are not available. Reanalysis has been carried out in recent years and in one study has led to an increased diagnostic yield of 12% (Ji et al., 2021). Manual reanalysis of exome data increased the diagnostic yield by 22% (CNVs=0.4%) in a previous study and a semiautomated reanalysis method increased the yield by 11.5% (CNVs=0.95%) showing much promise for implementation (Liu et al., 2019). Long term sustainability of exome reanalysis can be enhanced by better incorporation of automated workflows.

Newly discovered genes will result in the majority of new molecular diagnoses and databases like OMIM contributes greatly to these discoveries. An estimated 300 new phenotype entries per year are added to OMIM and due to continued sequencing efforts in large-scale this number continues to rise (Amberger et al., 2019, Ji et al., 2021). There are over 6500 genes with a phenotype description of which the molecular basis is known, available on the database since April 2024. Phenotypic expansion of previously identified disease genes is also a possibility which may lead to more clinical recognition of these phenotypes.

Incorporation of new clinically observed phenotypes into the reanalysis of exomes can expand the list of genes for analysis and potentially link these phenotypes to genes previously analysed. These VUS will remain a challenge to solve especially within populations where representation in public domain databases is inadequate. This further highlights the importance of collaboration and data sharing between research and diagnostic sites specifically from LMIC and the use of public domain databases like DECIPHER for submission of variants especially from understudied population groups. A wider adoption of CNV calling with ES data and use over time will allow for more opportunities to achieve this.

Incorporating the use of a newer reference genome, like the telomere-to-telomere reference genome can also help solve more CNVs and SVs and accurately determine the exact breakpoints (Rautiainen et al., 2023). More accurate read mapping can be achieved by using a pangenome which is representative of more populations (Wang et al., 2022). To this end a Pan-African genome has also been studied, showing that the use of population-specific genome graphs leads to more accurate variant calling and lower mapping errors (Tetikol et al., 2022).

Until the development of a comprehensive method for detection and classification of SNVs and SVs together, manual analysis will stay crucial for CNV classification. This being said, manual evaluation and interpretation of variants will most likely always stay relevant and serve as an important verification of classified variants. A recent advance called DRAGEN combines multi-genome mapping with pangenome references as well as machine learning-based variant detection, providing insights into individual genomes.

This tool can accurately detect all variant types (single-nucleotide variations, insertions or deletions, short tandem repeats, structural variations and copy number variations) and incorporates the analysis of medically relevant genes (Behera et al., 2024).

Ideally, we should be striving to develop a tool which could incorporate all aspects of CNV interpretation without having to search multiple different databases and online tools for additional evidence. Additional studies on diverse populations and the use of big data approaches will also aid in the development of more comprehensive variant classification tools. In the future, computational infrastructure and AI capabilities should allow for a single interpretation tool taking into account all CNV-relevant data available for assessment. The outcome would thus be more comprehensive and trustworthy and should ideally be developed to do systematic reanalysis as new information emerges.

Identifying a molecular diagnosis can inform decisions regarding future reproduction and recurrence risk for families and in previous studies there was an expectation to inform patients on recurrence risk and management of the condition (Diedericks et al., 2024). The fact that these families have an explanation for their child's DD and other difficulties not only ends the diagnostic odyssey, but also gives families peace of mind and end their anxiety and guilt they may have felt for years. Decisions regarding resources and support for specific conditions may also be guided and could provide families with a sense of empowerment due to the increased knowledge gained. Although these conditions cannot necessarily be cured, a diagnosis led to better management of the condition and appropriate clinical referrals. Understanding the prognosis will also in turn improve management of conditions which might develop later in life and identify areas where increased support is needed (Mollison et al., 2020, Savatt and Myers, 2021, Shingwenyana et al., 2023).

These results and increased molecular diagnostic yield reported from this study allows for better management and clinical care for DD families overall. It has been shown that there is need for community engagement to determine the need and preferences for the return of results. Appropriate ethical guidelines and practices are needed locally especially within LMICs in order to handle the return of genetic results (Mwaka et al., 2021).

As evidence and knowledge regarding population-specific variation and reclassification of variants increase over time, the abovementioned factors should be considered and incorporated during follow-up visits and return of results for patients and families (Fieggen et al., 2019).

The results presented in this study will all be returned to the patient and family by means of an individual feedback session conducted by a genetic counsellor. Three families had already received the individual patient reports and follow-up session. Two of these patients were referred to a cardiologist as part of management of the condition caused by the identified CNV. One of the two patients was also advised to consult an endocrine specialist when the patient is older as they may develop diabetes in early adulthood as part of the 17q deletion syndrome.

Returning results to patients is sensitive, but knowledge regarding the cause of the disease and a possible management plan and assessment of further health risks motivates for individual return of results (Shingwenyana et al., 2023). This also empowers the families by becoming more educated on the condition and thus provides clinical and personal utility for the families (Savatt and Myers, 2021, Diedericks et al., 2024).

Chapter 4: Study Conclusion

4 Conclusion

This study highlights the need for incorporating not only efficient, but appropriate exome pipelines in LMICs to further the implementation of genomic medicine and make it more attainable for all. Within LMICs where resources, funding and genetic services are scarce, it is imperative to identify an effective strategy for diagnosis of rare disease such as DD. Identifying a possible gold standard for CNV detection would be a much needed advancement to facilitate implementation. In this study, it was seen that exome CNV tools using read depth data show functional equivalence and could be used interchangeably without a major difference in diagnostic yield. This being said, using multiple tools together still showed superior results and maximises the yield. CANOES would be the recommendation from this study as a single exome CNV tool for implementation although only autosomal CNVs are identified. Combining this tool with another tool like XHMM thus ensures CNVs from sex chromosomes are also detected. Different tools should also be considered for implementation depending on available datasets and resources. All three tools implemented are based on read depth analysis as this is ES data which is limiting. If WGS data is available, tools using different methods of detection (read depth, split-read, paired-end and assembly based) could be applied to increase detection rate and subsequently diagnostic yield.

Although ES CNV analysis do not incur an additional sequencing cost, the cost of bioinformatics implementation should be considered. Tools like CLAMMS are much less computationally strenuous and would be ideal to implement in resource constrained environments. There thus seems to be the need for developing tools which are more user-friendly and less computational in order to allow more users within a range of settings.

There is also a need for smart, AI tools with exceptionally powerful computation, able to analyse larger cohorts of samples at once. These tools should ideally be able to simultaneously assess all the data dimensions needed to accurately classify CNVs. Many web-based tools are available for interpreting and classifying CNVs which has assisted in the CNV evaluation process. Manual evaluation of CNVs is still required in order to ensure correct classification and phenotypic match to the patient.

Improved diversity in population frequency databases will also provide access to key data needed for the clinical interpretation of CNVs. It is evident that a more diverse reference genome representing a larger range of population groups is required to improve CNV calling and classification. Multidisciplinary team discussions with experts in the field of genomics as well as collaboration is needed and will also allow better understanding and interpretation of CNVs.

Overall, simultaneous analysis of CNVs and SNVs through ES shows potential as a first-tier investigation for diagnosing rare monogenic disorders. This method does seem to be more sensitive in identifying a range of CNVs than CMA. In this study a total of eight LP/P CNVs were identified, increasing the diagnostic yield by ~7%. This additional yield was obtained without the need for additional sequencing or testing. Eight additional families thus received a molecular diagnosis through incorporating CNV analysis on existing ES data within a LMIC. These families were consulted at the genetics clinic for an average of seven years before the diagnoses were made. A molecular diagnosis can allow clinical and personal utility for patients and their families and end the diagnostic odyssey for many families affected with DD. A thorough explanation of the diagnosis by a genetic counsellor is important and also allow for better management and treatment of these patients.

5 Challenges and study limitations

As these exome bioinformatics CNV calling tools cannot identify the exact breakpoints, it is very challenging to establish the exact sizes of these variants. Small CNVs (<1 exon), inversions and translocations as well as mosaicism are not identified as part of this analysis. Using WGS instead of ES might alleviate some of these issues for future implementation in a setting where funds and resources are available. The truth set CRAM files downloaded from EGA database were aligned to the hg19 reference genome. This complicated analysis as the CRAM files had to be converted to FASTQ files and realigned to reference genome build GRCh38 before analysis could be done. Using BAM files aligned to the same reference genome for all samples is thus advisable to simplify analysis. To install and apply these tools to the data, a high level of bioinformatics skills was required.

In this study assistance from a bioinformatician was available; however, this would not be the case for many other resource limited settings. Using automated web-based analysis tools such as seqr (Pais et al., 2022) in future studies could diminish the need for high end bioinformatics input.

There are many limitations to the detection of CNVs from short read NGS data as discussed in chapter 3. Many CNVs may be missed or false positive calls made due to the lack of ES to perform optimally on low complexity/repeat regions. Many genes within the regions of the different CNVs are still not curated and dosage sensitivity not established. This made CNV interpretation and confirmation of pathogenicity very difficult.

African population CNVs are still in minority on public databases contributing to the difficulty of CNV interpretation in our specific cohort of patients. Limited clinical details and patient HPO terms further complicated CNV analysis and interpretation. This also highlights the importance of thorough clinical examination and detailed phenotyping of these patients.

The proband-only truth set was not ideal for the InDelible tool as the final and very important *de novo* analysis could not be completed. This could have led to some of these CNVs being missed. The sizes of the truth set CNVs were also not all suited to the specific tools. There were not enough CNVs within the range of InDelible and many were also out of the optimal size of detection for the other three CNV calling tools. The truth set CNVs were also not all disease causing and although a good way to measure the full scope of these tools, most were developed specifically for rare disease-causing CNVs. This could have led to fewer of the known CNVs being detected by these tools. A truth set with larger, more prominent disease-causing CNVs could have highlighted the true potential of these CNV calling tools to identify disease-causing CNVs with confidence.

Probes used during truth set sensitivity and specificity calculations were skewed toward true negative results due to all probes not within the known CNV thus being labeled true negative. This could thus have led to an overestimation of the true negative results for the three tools.

A number of the patients has been pre-screened by implementing Array CGH before this study as well. A cohort of patients without prior screening could have increased the diagnostic yield even further.

6 Future studies

Incorporation of the CNV tools into the DDD-Africa automatic variant calling pipeline is a next step which should be implemented to ease the process for future data analysis. This will allow SNV and CNV detection concurrently using all three tools together. A thorough cost analysis of ES should be carried out to motivate for routine implementation. This analysis should include the costs relating to additional sequencing tests (Sanger sequencing, CMA, targeted sequencing panels) and the fact that these costs would be eliminated when implementing ES.

Reanalysis and possible reclassification of VUS results should be completed and prioritised every ~2 years to ensure new data is considered and no diagnoses are missed (Deignan et al., 2019, Tan et al., 2020, Dai et al., 2022). All VUS CNVs matching patient phenotypic features could be uploaded to Matchmaker exchange to try and identify other patients with a similar variant (Philippakis et al., 2015). Recurrent CNVs within many participants have not been analysed in this study and this could be valuable to contribute benign CNVs to public domain databases like DECIPHER.

Newer tools such as GATK-gCNV used by seqr should be tested and compared to determine the value and detection rate. The purpose of this study was to determine whether an ensemble approach is more valuable than using one tool; however, if newer tools like GATK-gCNV performs well, it would be worth investigating. More advanced AI tools should also be applied to this dataset to ease the analysis and interpretation of CNVs from exome data. A tool such as VizCNV (Du et al., 2024) uses trio WGS data together with phased B-allele data to detect CNVs. This tool can analyse chromosomal abnormalities, exonic CNV and non-coding gene regulatory regions and prioritise *de novo* CNVs. Ideally, a tool should be developed to incorporate all aspects of CNV analysis and classification requiring minimal bioinformatics training for implementation. Long-read sequencing analysis can be applied to confirm the CNV sizes and compare results between ES data and long-read genomic data.

Samples without molecular diagnoses could also be subject to long-read analysis to ensure no SNVs or CNVs were missed and to be able to make any further diagnoses which was not within the scope of this study. Another future endeavor could also include a cost analysis of long-read WGS together with methylation analysis.

This study has allowed a closer inspection of the aetiology of CNVs in a cohort affected with DD in Africa. It has also identified an optimised approach to incorporate CNV analysis in ES studies within LMICs. Limitations within this field were explored and these results can be used for future endeavours aiming to improve CNV analysis from ES and NGS data in general. Future studies should focus on smaller intragenic CNVs or those including only a single exon to highlight the other key benefits of ES over microarray (Atik et al., 2024, Hahn et al., 2025). Many additional molecular diagnoses can be made by incorporating CNV analysis tools into existing NGS data (Pfundt et al., 2017, Testard et al., 2021, Hahn et al., 2025), ending the diagnostic odyssey for patients and giving their families a long-awaited answer.

7 References

Adeniyi, Y. C. & Adeniyi, A. F. 2020. Development of a community-based, one-stop service centre for children with developmental disorders: changing the narrative of developmental disorders in sub-Saharan Africa. *Pan Afr Med J,* 36**,** 164.

Alkan, C., Coe, B. P. & Eichler, E. E. 2011. Genome structural variation discovery and genotyping. *Nat Rev Genet,* 12**,** 363-76.

Amberger, J. S., Bocchini, C. A., Scott, A. F. & Hamosh, A. 2019. OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res,* 47**,** D1038-d1043.

Angelman, H. 1965. 'Puppet' Children A Report on Three Cases. *Developmental Medicine & Child Neurology,* 7**,** 681-688.

Aspden, J. L., Wallace, E. W. J. & Whiffin, N. 2023. Not all exons are protein coding: Addressing a common misconception. *Cell Genomics,* 3**,** 100296.

Atik, T., Avci Durmusalioglu, E., Isik, E., Kose, M., Kanmaz, S., Aykut, A., Durmaz, A., Ozkinay, F. & Cogulu, O. 2024. Diagnostic yield of exome sequencing-based copy number variation analysis in Mendelian disorders: a clinical application. *BMC Medical Genomics,* 17**,** 239.

Babadi, M., Fu, J. M., Lee, S. K., Smirnov, A. N., Gauthier, L. D., Walker, M., Benjamin, D. I., Karczewski, K. J., Wong, I., Collins, R. L., Sanchis-Juan, A., Brand, H., Banks, E. & Talkowski, M. E. 2022. GATK-gCNV: A Rare Copy Number Variant Discovery Algorithm and Its Application to Exome Sequencing in the UK Biobank. *bioRxiv***,** 2022.08.25.504851.

Backenroth, D., Homsy, J., Murillo, L. R., Glessner, J., Lin, E., Brueckner, M., Lifton, R., Goldmuntz, E., Chung, W. K. & Shen, Y. 2014. CANOES: detecting rare copy number variants from whole exome sequencing data. *Nucleic Acids Res,* 42**,** e97.

Baine-Savanhu, F., Macaulay, S., Louw, N., Bollweg, A., Flynn, K., Molatoli, M., Nevondwe, P., Seymour, H., Carstens, N., Krause, A. & Lombard, Z. 2023. Identifying the genetic causes of developmental disorders and intellectual disability in Africa: a systematic literature review. *Front Genet,* 14**,** 1137922.

Barbitoff, Y. A., Polev, D. E., Glotov, A. S., Serebryakova, E. A., Shcherbakova, I. V., Kiselev, A. M., Kostareva, A. A., Glotov, O. S. & Predeus, A. V. 2020. Systematic dissection of biases in whole-exome and whole-genome sequencing reveals major determinants of coding sequence coverage. *Scientific Reports,* 10**,** 2057.

Behera, S., Catreux, S., Rossi, M., Truong, S., Huang, Z., Ruehle, M., Visvanath, A., Parnaby, G., Roddey, C., Onuchic, V., Finocchio, A., Cameron, D. L., English, A., Mehtalia, S., Han, J., Mehio, R. & Sedlazeck, F. J. 2024. Comprehensive genome analysis and variant detection at scale using DRAGEN. *Nature Biotechnology*.

Biesecker, L. G. 2002. The end of the beginning of chromosome ends. *Am J Med Genet,* 107**,** 263-6.

Bigio, B., Seeleuthner, Y., Kerner, G., Migaud, M., Rosain, J., Boisson, B., Nasca, C., Puel, A., Bustamante, J., Casanova, J.-L., Abel, L. & Cobat, A. 2020. Detection of homozygous and hemizygous partial exon deletions by whole-exome sequencing. *bioRxiv***,** 2020.07.23.217976.

Bitta, M., Kariuki, S. M., Abubakar, A. & Newton, C. 2017. Burden of neurodevelopmental disorders in low and middle-income countries: A systematic review and meta-analysis. *Wellcome Open Res,* 2**,** 121.

Bragin, E., Chatzimichali, E. A., Wright, C. F., Hurles, M. E., Firth, H. V., Bevan, A. P. & Swaminathan, G. J. 2014. DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. *Nucleic Acids Res,* 42**,** D993-D1000.

Castellani, C., Melka, M., Wishart, A., Locke, M. E., Awamleh, Z., Reilly, L. & Singh, S. 2014. Biological relevance of CNV calling methods using familial relatedness including monozygotic twins. *BMC bioinformatics,* 15**,** 114.

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A. J., Kruglyak, S. & Saunders, C. T. 2016. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics,* 32**,** 1220-2.

Choi, J., Jeon, S., Kim, D., Chua, M. & Do, S. 2022. A scalable artificial intelligence platform that automatically finds copy number variations (CNVs) in journal articles and transforms them into a database: CNV extraction, transformation, and loading AI (CNV-ETLAI). *Comput Biol Med,* 144**,** 105332.

Collins, R. L., Brand, H., Karczewski, K. J., Zhao, X., Alföldi, J., Francioli, L. C., Khera, A. V., Lowther, C., Gauthier, L. D., Wang, H., Watts, N. A., Solomonson, M., O'donnell-Luria, A., Baumann, A., Munshi, R., Walker, M., Whelan, C., Huang, Y., Brookings, T., Sharpe, T., Stone, M. R., Valkanas, E., Fu, J., Tiao, G., Laricchia, K. M., Ruano-Rubio, V., Stevens, C., Gupta, N., Margolin, L., Team, G. a. D. P., Consortium, G. a. D., Taylor, K. D., Lin, H. J., Rich, S. S., Post, W., Chen, Y.-D. I., Rotter, J. I., Nusbaum, C., Philippakis, A., Lander, E., Gabriel, S., Neale, B. M., Kathiresan, S., Daly, M. J., Banks, E., Macarthur, D. G. & Talkowski, M. E. 2019. An open resource of structural variation for medical and population genetics. *bioRxiv***,** 578674.

Cooper, G. M., Coe, B. P., Girirajan, S., Rosenfeld, J. A., Vu, T. H., Baker, C., Williams, C., Stalker, H., Hamid, R., Hannig, V., Abdel-Hamid, H., Bader, P., Mccracken, E., Niyazov, D., Leppig, K., Thiese, H., Hummel, M., Alexander, N., Gorski, J., Kussmann, J., Shashi, V., Johnson, K., Rehder, C., Ballif, B. C., Shaffer, L. G. & Eichler, E. E. 2011. A copy number variation morbidity map of developmental delay. *Nat Genet,* 43**,** 838-46.

Cordoba, M., Rodriguez-Quiroga, S. A., Vega, P. A., Salinas, V., Perez-Maturo, J., Amartino, H., Vasquez-Dusefante, C., Medina, N., Gonzalez-Moron, D. & Kauffman, M. A. 2018. Whole exome sequencing in neurogenetic odysseys: An effective, cost- and time-saving diagnostic approach. *PLoS One,* 13**,** e0191228.

Coutelier, M., Holtgrewe, M., Jäger, M., Flöttman, R., Mensah, M. A., Spielmann, M., Krawitz, P., Horn, D., Beule, D. & Mundlos, S. 2022. Combining callers improves the detection of copy number variants from whole-genome sequencing. *European Journal of Human Genetics,* 30**,** 178-186.

D'arrigo, S., Gavazzi, F., Alfei, E., Zuffardi, O., Montomoli, C., Corso, B., Buzzi, E., Sciacca, F. L., Bulgheroni, S., Riva, D. & Pantaleoni, C. 2016. The Diagnostic Yield of Array Comparative Genomic Hybridization Is High Regardless of Severity of Intellectual Disability/Developmental Delay in Children. *J Child Neurol,* 31**,** 691-9.

Dai, P., Honda, A., Ewans, L., Mcgaughran, J., Burnett, L., Law, M. & Phan, T. G. 2022. Recommendations for next generation sequencing data reanalysis of

unsolved cases with suspected Mendelian disorders: A systematic review and meta-analysis. *Genet Med,* 24**,** 1618-1629.
Danecek, P., Gardner, E. J., Fitzgerald, T. W., Gallone, G., Kaplanis, J., Eberhardt, R. Y., Wright, C. F., Firth, H. V. & Hurles, M. E. 2024. Detection and characterization of copy-number variants from exome sequencing in the DDD study. *Genetics in Medicine Open,* 2**,** 101818.
De Goede, C., Yue, W. W., Yan, G., Ariyaratnam, S., Chandler, K. E., Downes, L., Khan, N., Mohan, M., Lowe, M. & Banka, S. 2016. Role of reverse phenotyping in interpretation of next generation sequencing data and a review of INPP5E related disorders. *Eur J Paediatr Neurol,* 20**,** 286-295.
Deciphering Developmental Disorders, S. 2015. Large-scale discovery of novel genetic causes of developmental disorders. *Nature,* 519**,** 223-8.
Deignan, J. L., Chung, W. K., Kearney, H. M., Monaghan, K. G., Rehder, C. W. & Chao, E. C. 2019. Points to consider in the reevaluation and reanalysis of genomic test results: a statement of the American College of Medical Genetics and Genomics (ACMG). *Genet Med,* 21**,** 1267-1270.
Demidov, G., Yaldiz, B., Garcia-Pelaez, J., De Boer, E., Schuermans, N., Van De Vondel, L., Paramonov, I., Johansson, L. F., Musacchia, F., Benetti, E., Bullich, G., Sablauskas, K., Beltran, S., Gilissen, C., Hoischen, A., Ossowski, S., De Voer, R., Lohmann, K., Oliveira, C., Topf, A., Vissers, L. & Laurie, S. 2024. Comprehensive reanalysis for CNVs in ES data from unsolved rare disease cases results in new diagnoses. *NPJ Genom Med,* 9**,** 49.
Dharmadhikari, A. V., Ghosh, R., Yuan, B., Liu, P., Dai, H., Al Masri, S., Scull, J., Posey, J. E., Jiang, A. H., He, W., Vetrini, F., Braxton, A. A., Ward, P., Chiang, T., Qu, C., Gu, S., Shaw, C. A., Smith, J. L., Lalani, S., Stankiewicz, P., Cheung, S.-W., Bacino, C. A., Patel, A., Breman, A. M., Wang, X., Meng, L., Xiao, R., Xia, F., Muzny, D., Gibbs, R. A., Beaudet, A. L., Eng, C. M., Lupski, J. R., Yang, Y. & Bi, W. 2019. Copy number variant and runs of homozygosity detection by microarrays enabled more precise molecular diagnoses in 11,020 clinical exome cases. *Genome Medicine,* 11**,** 30.
Diedericks, A., Bruwer, Z., Laing, N., Eastman, E., De Vries, J., Newton, C. R., Abubakar, A., Robinson, E. B., Donald, K. A. & On Behalf of the Neurodev, S. 2024. Parental perspectives regarding the return of genomic research results in neurodevelopmental disorders in South Africa: anticipated impact and preferences. *Journal of Community Genetics*.
Dong, X., Liu, B., Yang, L., Wang, H., Wu, B., Liu, R., Chen, H., Chen, X., Yu, S., Chen, B., Wang, S., Xu, X., Zhou, W. & Lu, Y. 2020. Clinical exome sequencing as the first-tier test for diagnosing developmental disorders covering both CNV and SNV: a Chinese cohort. *J Med Genet,* 57**,** 558-566.
Du, H., Lun, M. Y., Gagarina, L., Mehaffey, M. G., Hwang, J. P., Jhangiani, S. N., Bhamidipati, S. V., Muzny, D. M., Poli, M. C., Ochoa, S., Chinn, I. K., Linstrand, A., Posey, J. E., Gibbs, R. A., Lupski, J. R. & Carvalho, C. M. B. 2024. VizCNV: An integrated platform for concurrent phased BAF and CNV analysis with trio genome sequencing data. *bioRxiv*.
Ekure, E. N., Adeyemo, A., Liu, H., Sokunbi, O., Kalu, N., Martinez, A. F., Owosela, B., Tekendo-Ngongang, C., Addissie, Y. A., Olusegun-Joseph, A., Ikebudu, D., Berger, S. I., Muenke, M., Han, Z. & Kruszka, P. 2021. Exome Sequencing and Congenital Heart Disease in Sub-Saharan Africa. *Circulation: Genomic and Precision Medicine,* 14**,** e003108.

Fan, C., Wang, Z., Sun, Y., Sun, J., Liu, X., Kang, L., Xu, Y., Yang, M., Dai, W., Song, L., Wei, X., Xiang, J., Huang, H., Zhou, M., Zeng, F., Huang, L., Xu, Z. & Peng, Z. 2021. AutoCNV: a semiautomatic CNV interpretation system based on the 2019 ACMG/ClinGen Technical Standards for CNVs. *BMC Genomics,* 22**,** 721.

Fan, X., Abbott, T. E., Larson, D. & Chen, K. 2014. BreakDancer: Identification of Genomic Structural Variation from Paired-End Read Mapping. *Curr Protoc Bioinformatics,* 45**,** 15.6.1-11.

Fellner, A., Ruhrman-Shahar, N., Orenstein, N., Lidzbarsky, G., Shuldiner, A. R., Gonzaga-Jauregui, C., Brown-Shalev, H., Hagari-Bechar, O., Bazak, L. & Basel-Salmon, L. 2021. The role of phenotype-based search approaches using public online databases in diagnostics of Mendelian disorders. *Genetics in Medicine,* 23**,** 1095-1100.

Fieggen, K. J., Lambie, L. A. & Donald, K. A. 2019. *Investigating developmental delay in South Africa: A pragmatic approach*.

Filer, D. L., Kuo, F., Brandt, A. T., Tilley, C. R., Mieczkowski, P. A., Berg, J. S., Robasky, K., Li, Y., Bizon, C., Tilson, J. L., Powell, B. C., Bost, D. M., Jeffries, C. D. & Wilhelmsen, K. C. 2021. Pre-capture multiplexing provides additional power to detect copy number variation in exome sequencing. *BMC Bioinformatics,* 22**,** 374.

Firth, H. V., Richards, S. M., Bevan, A. P., Clayton, S., Corpas, M., Rajan, D., Van Vooren, S., Moreau, Y., Pettett, R. M. & Carter, N. P. 2009. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am J Hum Genet,* 84**,** 524-33.

Firth, H. V., Wright, C. F. & Study, D. D. D. 2011. The Deciphering Developmental Disorders (DDD) study. *Dev Med Child Neurol,* 53**,** 702-3.

Fritzen, D., Kuechler, A., Grimmel, M., Becker, J., Peters, S., Sturm, M., Hundertmark, H., Schmidt, A., Kreiß, M., Strom, T. M., Wieczorek, D., Haack, T. B., Beck-Wödl, S., Cremer, K. & Engels, H. 2018. De novo FBXO11 mutations are associated with intellectual disability and behavioural anomalies. *Human Genetics,* 137**,** 401-411.

Fromer, M., Moran, J. L., Chambert, K., Banks, E., Bergen, S. E., Ruderfer, D. M., Handsaker, R. E., Mccarroll, S. A., O'donovan, M. C., Owen, M. J., Kirov, G., Sullivan, P. F., Hultman, C. M., Sklar, P. & Purcell, S. M. 2012. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am J Hum Genet,* 91**,** 597-607.

Fromer, M. & Purcell, S. M. 2014. Using XHMM Software to Detect Copy Number Variation in Whole-Exome Sequencing Data. *Curr Protoc Hum Genet,* 81**,** 7.23.1-21.

Fung, J. L. F., Yu, M. H. C., Huang, S., Chung, C. C. Y., Chan, M. C. Y., Pajusalu, S., Mak, C. C. Y., Hui, V. C. C., Tsang, M. H. Y., Yeung, K. S., Lek, M. & Chung, B. H. Y. 2020. A three-year follow-up study evaluating clinical utility of exome sequencing and diagnostic potential of reanalysis. *npj Genomic Medicine,* 5**,** 37.

Gabrielaite, M., Torp, M. H., Rasmussen, M. S., Andreu-Sánchez, S., Vieira, F. G., Pedersen, C. B., Kinalis, S., Madsen, M. B., Kodama, M., Demircan, G. S., Simonyan, A., Yde, C. W., Olsen, L. R., Marvig, R. L., Østrup, O., Rossing, M., Nielsen, F. C., Winther, O. & Bagger, F. O. 2021. A Comparison of Tools for Copy-Number Variation Detection in Germline Whole Exome and Whole Genome Sequencing Data. *Cancers (Basel),* 13.

Gambin, T., Akdemir, Z. C., Yuan, B., Gu, S., Chiang, T., Carvalho, C. M. B., Shaw, C., Jhangiani, S., Boone, P. M., Eldomery, M. K., Karaca, E., Bayram, Y., Stray-Pedersen, A., Muzny, D., Charng, W. L., Bahrambeigi, V., Belmont, J. W., Boerwinkle, E., Beaudet, A. L., Gibbs, R. A. & Lupski, J. R. 2017. Homozygous and hemizygous CNV detection from exome sequencing data in a Mendelian disease cohort. *Nucleic Acids Res,* 45**,** 1633-1648.

Garcia, F. a. O., De Andrade, E. S. & Palmero, E. I. 2022. Insights on variant analysis in silico tools for pathogenicity prediction. *Front Genet,* 13**,** 1010327.

Gardner, E. J., Sifrim, A., Lindsay, S. J., Prigmore, E., Rajan, D., Danecek, P., Gallone, G., Eberhardt, R. Y., Martin, H. C., Wright, C. F., Fitzpatrick, D. R., Firth, H. V. & Hurles, M. E. 2021. Detecting cryptic clinically relevant structural variation in exome-sequencing data increases diagnostic yield for developmental disorders. *Am J Hum Genet,* 108**,** 2186-2194.

Gargano, M. A., Matentzoglu, N., Coleman, B., Addo-Lartey, E. B., Anagnostopoulos, Anna v., Anderton, J., Avillach, P., Bagley, A. M., Bakštein, E., Balhoff, J. P., Baynam, G., Bello, S. M., Berk, M., Bertram, H., Bishop, S., Blau, H., Bodenstein, D. F., Botas, P., Boztug, K., Čady, J., Callahan, T. J., Cameron, R., Carbon, Seth j., Castellanos, F., Caufield, J. H., Chan, L. E., Chute, Christopher g., Cruz-Rojo, J., Dahan-Oliel, N., Davids, J. R., De Dieuleveult, M., De Souza, V., De vries, B. B. A., De Vries, E., Depaulo, J. R., Derfalvi, B., Dhombres, F., Diaz-Byrd, C., Dingemans, A. J. M., Donadille, B., Duyzend, M., Elfeky, R., Essaid, S., Fabrizzi, C., Fico, G., Firth, H. V., Freudenberg-Hua, Y., Fullerton, J. M., Gabriel, D. L., Gilmour, K., Giordano, J., Goes, F. S., Moses, R. G., Green, I., Griese, M., Groza, T., Gu, W., Guthrie, J., Gyori, B., Hamosh, A., Hanauer, M., Hanušová, K., He, Y., Hegde, H., Helbig, I., Holasová, K., Hoyt, C. T., Huang, S., Hurwitz, E., Jacobsen, J. O. B., Jiang, X., Joseph, L., Keramatian, K., King, B., Knoflach, K., Koolen, D. A., Kraus, Megan l., Kroll, C., Kusters, M., Ladewig, M. S., Lagorce, D., Lai, M.-C., Lapunzina, P., Laraway, B., Lewis-Smith, D., Li, X., Lucano, C., Majd, M., Marazita, M. L., Martinez-Glez, V., Mchenry, T. H., Mcinnis, M. G., Mcmurry, J. A., Mihulová, M., Millett, C. E., Mitchell, P. B., Moslerová, V., Narutomi, K., Nematollahi, S., Nevado, J., et al. 2023. The Human Phenotype Ontology in 2024: phenotypes around the world. *Nucleic Acids Research,* 52**,** D1333-D1346.

Gažiová, M., Sládeček, T., Pös, O., Števko, M., Krampl, W., Pös, Z., Hekel, R., Hlavačka, M., Kucharík, M., Radvánszky, J., Budiš, J. & Szemes, T. 2022. Automated prediction of the clinical impact of structural copy number variations. *Scientific Reports,* 12**,** 555.

Geoffroy, V., Herenger, Y., Kress, A., Stoetzel, C., Piton, A., Dollfus, H. & Muller, J. 2018. AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics,* 34**,** 3572-3574.

Gilissen, C., Hehir-Kwa, J. Y., Thung, D. T., Van De Vorst, M., Van Bon, B. W., Willemsen, M. H., Kwint, M., Janssen, I. M., Hoischen, A., Schenck, A., Leach, R., Klein, R., Tearle, R., Bo, T., Pfundt, R., Yntema, H. G., De Vries, B. B., Kleefstra, T., Brunner, H. G., Vissers, L. E. & Veltman, J. A. 2014. Genome sequencing identifies major causes of severe intellectual disability. *Nature,* 511**,** 344-7.

Gillet-Markowska, A., Richard, H., Fischer, G. & Lafontaine, I. 2015. Ulysses: accurate detection of low-frequency structural variations in large insert-size sequencing libraries. *Bioinformatics,* 31**,** 801-8.

Gordeeva, V., Sharova, E., Babalyan, K., Sultanov, R., Govorun, V. M. & Arapidi, G. 2021. Benchmarking germline CNV calling tools from exome sequencing data. *Sci Rep,* 11**,** 14416.

Gregor, A., Sadleir, L. G., Asadollahi, R., Azzarello-Burri, S., Battaglia, A., Ousager, L. B., Boonsawat, P., Bruel, A.-L., Buchert, R., Calpena, E., Cogné, B., Dallapiccola, B., Distelmaier, F., Elmslie, F., Faivre, L., Haack, T. B., Harrison, V., Henderson, A., Hunt, D., Isidor, B., Joset, P., Kumada, S., Lachmeijer, A. M. A., Lees, M., Lynch, S. A., Martinez, F., Matsumoto, N., Mcdougall, C., Mefford, H. C., Miyake, N., Myers, C. T., Moutton, S., Nesbitt, A., Novelli, A., Orellana, C., Rauch, A., Rosello, M., Saida, K., Santani, A. B., Sarkar, A., Scheffer, I. E., Shinawi, M., Steindl, K., Symonds, J. D., Zackai, E. H., Reis, A., Sticht, H. & Zweier, C. 2018. De Novo Variants in the F-Box Protein FBXO11 in 20 Individuals with a Variable Neurodevelopmental Disorder. *The American Journal of Human Genetics,* 103**,** 305-316.

Guo, Y., Long, J., He, J., Li, C. I., Cai, Q., Shu, X. O., Zheng, W. & Li, C. 2012. Exome sequencing generates high quality data in non-target regions. *BMC Genomics,* 13**,** 194.

Guo, Y., Sheng, Q., Samuels, D. C., Lehmann, B., Bauer, J. A., Pietenpol, J. & Shyr, Y. 2013. Comparative study of exome copy number variation estimation tools using array comparative genomic hybridization as control. *Biomed Res Int,* 2013**,** 915636.

Gurbich, T. A. & Ilinsky, V. V. 2020. ClassifyCNV: a tool for clinical annotation of copy-number variants. *Scientific Reports,* 10**,** 20375.

Hagberg, B., Aicardi, J., Dias, K. & Ramos, O. 1983. A progressive syndrome of autism, dementia, ataxia, and loss of purposeful hand use in girls: Rett's syndrome: report of 35 cases. *Ann Neurol,* 14**,** 471-9.

Hahn, E., Dharmadhikari, A. V., Markowitz, A. L., Estrine, D., Quindipan, C., Maggo, S. D. S., Sharma, A., Lee, B., Maglinte, D. T., Shams, S., Deardorff, M. A., Biegel, J. A., Gai, X., Sun, M., Schmidt, R. J., Raca, G. & Ji, J. 2025. Copy number variant analysis improves diagnostic yield in a diverse pediatric exome sequencing cohort. *npj Genomic Medicine,* 10**,** 16.

Hall, B. D. 1979. Choanal atresia and associated multiple anomalies. *J Pediatr,* 95**,** 395-8.

Hamosh, A., Scott, A. F., Amberger, J., Bocchini, C., Valle, D. & Mckusick, V. A. 2002. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res,* 30**,** 52-5.

Handsaker, R. E., Korn, J. M., Nemesh, J. & Mccarroll, S. A. 2011. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nature Genetics,* 43**,** 269-276.

Harrison, P. W., Amode, M. R., Austine-Orimoloye, O., Azov, Andrey g., Barba, M., Barnes, I., Becker, A., Bennett, R., Berry, A., Bhai, J., Bhurji, S. K., Boddu, S., Branco lins, P. R., Brooks, L., Ramaraju, Shashank b., Campbell, Lahcen i., Martinez, M. C., Charkhchi, M., Chougule, K., Cockburn, A., Davidson, C., De silva, Nishadi h., Dodiya, K., Donaldson, S., El Houdaigui, B., Naboulsi, Tamara e., Fatima, R., Giron, C. G., Genez, T., Grigoriadis, D., Ghattaoraya, Gurpreet s., Martinez, J. G., Gurbich, Tatiana a., Hardy, M., Hollis, Z., Hourlier, T., Hunt, T., Kay, M., Kaykala, V., Le, T., Lemos, D., Lodha, D., Marques-Coelho, D., Maslen, G., Merino, Gabriela a., Mirabueno, Louisse p., Mushtaq, A., Hossain, Syed n., Ogeh, Denye n., Sakthivel, M. P., Parker, A., Perry, M., Piližota, I., Poppleton, D., Prosovetskaia, I., Raj, S., Pérez-Silva, José g.,

Salam, Ahamed imran a., Saraf, S., Saraiva-Agostinho, N., Sheppard, D., Sinha, S., Sipos, B., Sitnik, V., Stark, W., Steed, E., Suner, M.-M., Surapaneni, L., Sutinen, K., Tricomi, F. F., Urbina-Gómez, D., Veidenberg, A., Walsh, T. A., Ware, D., Wass, E., Willhoft, Natalie l., Allen, J., Alvarez-Jarreta, J., Chakiachvili, M., Flint, B., Giorgetti, S., Haggerty, L., Ilsley, Garth r., Keatley, J., Loveland, Jane e., Moore, B., Mudge, Jonathan m., Naamati, G., Tate, J., Trevanion, Stephen j., Winterbottom, A., Frankish, A., Hunt, S. E., Cunningham, F., Dyer, S., Finn, Robert d., Martin, Fergal j. & Yates, Andrew d. 2023. Ensembl 2024. *Nucleic Acids Research,* 52**,** D891-D899.

Hiz Kurul, S., Oktay, Y., Töpf, A., Szabó, N. Z., Güngör, S., Yaramis, A., Sonmezler, E., Matalonga, L., Yis, U., Schon, K., Paramonov, I., Kalafatcilar I, P., Gao, F., Rieger, A., Arslan, N., Yilmaz, E., Ekinci, B., Edem, P. P., Aslan, M., Özgör, B., Lochmüller, A., Nair, A., O'heir, E., Lovgren, A. K., Maroofian, R., Houlden, H., Polavarapu, K., Roos, A., Müller, J. S., Hathazi, D., Chinnery, P. F., Laurie, S., Beltran, S., Lochmüller, H. & Horvath, R. 2022. High diagnostic rate of trio exome sequencing in consanguineous families with neurogenetic diseases. *Brain,* 145**,** 1507-1518.

Holliday, R. & Pugh, J. E. 1975. DNA modification mechanisms and gene activity during development. *Science,* 187**,** 226-32.

Hong, C. S., Singh, L. N., Mullikin, J. C. & Biesecker, L. G. 2016. Assessing the reproducibility of exome copy number variations predictions. *Genome Med,* 8**,** 82.

Hu, X., Guo, R., Guo, J., Qi, Z., Li, W. & Hao, C. 2020. Parallel Tests of Whole Exome Sequencing and Copy Number Variant Sequencing Increase the Diagnosis Yields of Rare Pediatric Disorders. *Front Genet,* 11**,** 473.

Hu, X., Li, N., Xu, Y., Li, G., Yu, T., Yao, R. E., Fu, L., Wang, J., Yin, L., Yin, Y., Wang, Y., Jin, X., Wang, X., Wang, J. & Shen, Y. 2018. Proband-only medical exome sequencing as a cost-effective first-tier genetic diagnostic test for patients without prior molecular tests and clinical diagnosis in a developing country: the China experience. *Genet Med,* 20**,** 1045-1053.

Ibn-Salem, J., Köhler, S., Love, M. I., Chung, H. R., Huang, N., Hurles, M. E., Haendel, M., Washington, N. L., Smedley, D., Mungall, C. J., Lewis, S. E., Ott, C. E., Bauer, S., Schofield, P. N., Mundlos, S., Spielmann, M. & Robinson, P. N. 2014. Deletions of chromosomal regulatory boundaries are associated with congenital disease. *Genome Biol,* 15**,** 423.

Jang, W., Kim, Y., Han, E., Park, J., Chae, H., Kwon, A., Choi, H., Kim, J., Son, J. O., Lee, S. J., Hong, B. Y., Jang, D. H., Han, J. Y., Lee, J. H., Kim, S. Y., Lee, I. G., Sung, I. K., Moon, Y., Kim, M. & Park, J. H. 2019. Chromosomal Microarray Analysis as a First-Tier Clinical Diagnostic Test in Patients With Developmental Delay/Intellectual Disability, Autism Spectrum Disorders, and Multiple Congenital Anomalies: A Prospective Multicenter Study in Korea. *Ann Lab Med,* 39**,** 299-310.

Jansen, S., Van Der Werf, I. M., Innes, A. M., Afenjar, A., Agrawal, P. B., Anderson, I. J., Atwal, P. S., Van Binsbergen, E., Van Den Boogaard, M.-J., Castiglia, L., Coban-Akdemir, Z. H., Van Dijck, A., Doummar, D., Van Eerde, A. M., Van Essen, A. J., Van Gassen, K. L., Guillen Sacoto, M. J., Van Haelst, M. M., Iossifov, I., Jackson, J. L., Judd, E., Kaiwar, C., Keren, B., Klee, E. W., Klein Wassink-Ruiter, J. S., Meuwissen, M. E., Monaghan, K. G., De Munnik, S. A., Nava, C., Ockeloen, C. W., Pettinato, R., Racher, H., Rinne, T., Romano, C., Sanders, V. R., Schnur, R. E., Smeets, E. J., Stegmann, A. P. A., Stray-

Pedersen, A., Sweetser, D. A., Terhal, P. A., Tveten, K., Vannoy, G. E., De Vries, P. F., Waxler, J. L., Willing, M., Pfundt, R., Veltman, J. A., Kooy, R. F., Vissers, L. E. L. M. & De Vries, B. B. A. 2019. De novo variants in FBXO11 cause a syndromic form of intellectual disability with behavioral problems and dysmorphisms. *European Journal of Human Genetics,* 27**,** 738-746.

Ji, J., Leung, M. L., Baker, S., Deignan, J. L. & Santani, A. 2021. Clinical Exome Reanalysis: Current Practice and Beyond. *Mol Diagn Ther,* 25**,** 529-536.

Jiang, Y., Wang, Y. & Brudno, M. 2012. PRISM: pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants. *Bioinformatics,* 28**,** 2576-83.

Kallioniemi, A., Kallioniemi, O. P., Sudar, D., Rutovitz, D., Gray, J. W., Waldman, F. & Pinkel, D. 1992. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science,* 258**,** 818-21.

Kamp, M., Krause, A. & Ramsay, M. 2021. Has translational genomics come of age in Africa? *Human Molecular Genetics,* 30**,** R164-R173.

Kaplanis, J., Samocha, K. E., Wiel, L., Zhang, Z., Arvai, K. J., Eberhardt, R. Y., Gallone, G., Lelieveld, S. H., Martin, H. C., Mcrae, J. F., Short, P. J., Torene, R. I., De Boer, E., Danecek, P., Gardner, E. J., Huang, N., Lord, J., Martincorena, I., Pfundt, R., Reijnders, M. R. F., Yeung, A., Yntema, H. G., Study, D., Vissers, L. E. L. M., Juusola, J., Wright, C. F., Brunner, H. G., Firth, H. V., Fitzpatrick, D. R., Barrett, J. C., Hurles, M. E., Gilissen, C. & Retterer, K. 2020. Integrating healthcare and research genetic data empowers the discovery of 28 novel developmental disorders. *bioRxiv***,** 797787.

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., Walters, R. K., Tashman, K., Farjoun, Y., Banks, E., Poterba, T., Wang, A., Seed, C., Whiffin, N., Chong, J. X., Samocha, K. E., Pierce-Hoffman, E., Zappala, Z., O'donnell-Luria, A. H., Minikel, E. V., Weisburd, B., Lek, M., Ware, J. S., Vittal, C., Armean, I. M., Bergelson, L., Cibulskis, K., Connolly, K. M., Covarrubias, M., Donnelly, S., Ferriera, S., Gabriel, S., Gentry, J., Gupta, N., Jeandet, T., Kaplan, D., Llanwarne, C., Munshi, R., Novod, S., Petrillo, N., Roazen, D., Ruano-Rubio, V., Saltzman, A., Schleicher, M., Soto, J., Tibbetts, K., Tolonen, C., Wade, G., Talkowski, M. E., Aguilar Salinas, C. A., Ahmad, T., Albert, C. M., Ardissino, D., Atzmon, G., Barnard, J., Beaugerie, L., Benjamin, E. J., Boehnke, M., Bonnycastle, L. L., Bottinger, E. P., Bowden, D. W., Bown, M. J., Chambers, J. C., Chan, J. C., Chasman, D., Cho, J., Chung, M. K., Cohen, B., Correa, A., Dabelea, D., Daly, M. J., Darbar, D., Duggirala, R., Dupuis, J., Ellinor, P. T., Elosua, R., Erdmann, J., Esko, T., Färkkilä, M., Florez, J., Franke, A., Getz, G., Glaser, B., Glatt, S. J., Goldstein, D., Gonzalez, C., Groop, L., et al. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature,* 581**,** 434-443.

Kipkemoi, P., Kim, H. A., Christ, B., O'heir, E., Allen, J., Austin-Tse, C., Baxter, S., Brand, H., Bryant, S., Buser, N., De Menil, V., Eastman, E., Murugasen, S., Galvin, A., Kombe, M., Ngombo, A., Mkubwa, B., Mwangi, P., Kipkoech, C., Lovgren, A., Macarthur, D. G., Melly, B., Mwangasha, K., Martin, A., Nkambule, L. L., Sanchis-Juan, A., Singer-Berk, M., Talkowski, M. E., Vannoy, G., Van Der Merwe, C., Newton, C., O'donnell-Luria, A., Abubakar, A., Donald, K. A. &

Robinson, E. B. 2023. Phenotype and genetic analysis of data collected within the first year of NeuroDev. *Neuron,* 111**,** 2800-2810.e5.
Klambauer, G., Schwarzbauer, K., Mayr, A., Clevert, D. A., Mitterecker, A., Bodenhofer, U. & Hochreiter, S. 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**,** e69.
Korbel, J. O., Abyzov, A., Mu, X. J., Carriero, N., Cayting, P., Zhang, Z., Snyder, M. & Gerstein, M. B. 2009. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biology,* 10**,** R23.
Krumm, N., Sudmant, P. H., Ko, A., O'roak, B. J., Malig, M., Coe, B. P., Project, N. E. S., Quinlan, A. R., Nickerson, D. A. & Eichler, E. E. 2012. Copy number variation detection and genotyping from exome sequence data. *Genome Res,* 22**,** 1525-32.
Laduca, H., Farwell, K. D., Vuong, H., Lu, H. M., Mu, W., Shahmirzadi, L., Tang, S., Chen, J., Bhide, S. & Chao, E. C. 2017. Exome sequencing covers >98% of mutations identified on targeted next generation sequencing panels. *PLoS One,* 12**,** e0170843.
Laframboise, T. 2009. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Res,* 37**,** 4181-93.
Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., Mcdaniel, K., Ovetsky, M., Riley, G., Zhou, G., Holmes, J b., Kattman, B. L. & Maglott, D. R. 2017. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research,* 46**,** D1062-D1067.
Lappalainen, I., Almeida-King, J., Kumanduri, V., Senf, A., Spalding, J. D., Ur-Rehman, S., Saunders, G., Kandasamy, J., Caccamo, M., Leinonen, R., Vaughan, B., Laurent, T., Rowland, F., Marin-Garcia, P., Barker, J., Jokinen, P., Torres, A. C., De Argila, J. R., Llobet, O. M., Medina, I., Puy, M. S., Alberich, M., De La Torre, S., Navarro, A., Paschall, J. & Flicek, P. 2015. The European Genome-phenome Archive of human data consented for biomedical research. *Nat Genet,* 47**,** 692-5.
Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. 2014. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biology,* 15**,** R84.
Lee, H., Deignan, J. L., Dorrani, N., Strom, S. P., Kantarci, S., Quintero-Rivera, F., Das, K., Toy, T., Harry, B., Yourshaw, M., Fox, M., Fogel, B. L., Martinez-Agosto, J. A., Wong, D. A., Chang, V. Y., Shieh, P. B., Palmer, C. G., Dipple, K. M., Grody, W. W., Vilain, E. & Nelson, S. F. 2014. Clinical exome sequencing for genetic identification of rare Mendelian disorders. *JAMA,* 312**,** 1880-7.
Lemire, G., Sanchis-Juan, A., Russell, K., Baxter, S., Chao, K. R., Singer-Berk, M., Groopman, E., Wong, I., England, E., Goodrich, J., Pais, L., Austin-Tse, C., Ditroia, S., O'heir, E., Ganesh, V. S., Wojcik, M. H., Evangelista, E., Snow, H., Osei-Owusu, I., Fu, J., Singh, M., Mostovoy, Y., Huang, S., Garimella, K., Kirkham, S. L., Neil, J. E., Shao, D. D., Walsh, C. A., Argilli, E., Le, C., Sherr, E. H., Gleeson, J. G., Shril, S., Schneider, R., Hildebrandt, F., Sankaran, V. G., Madden, J. A., Genetti, C. A., Beggs, A. H., Agrawal, P. B., Bujakowska, K. M., Place, E., Pierce, E. A., Donkervoort, S., Bönnemann, C. G., Gallacher, L., Stark, Z., Tan, T. Y., White, S. M., Töpf, A., Straub, V., Fleming, M. D., Pollak, M. R., Õunap, K., Pajusalu, S., Donald, K. A., Bruwer, Z., Ravenscroft, G.,

Laing, N. G., Macarthur, D. G., Rehm, H. L., Talkowski, M. E., Brand, H. & O'donnell-Luria, A. 2024. Exome copy number variant detection, analysis, and classification in a large cohort of families with undiagnosed rare genetic disease. *The American Journal of Human Genetics,* 111**,** 863-876.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & Subgroup, G. P. D. P. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25**,** 2078-2079.

Li, Z., Liu, F., Wan, R., Wu, Y. & Liu, J. 2024. [Genetic analysis of two children with Coffin-Siris syndrome due to variants of ARID1B gene]. *Zhonghua Yi Xue Yi Chuan Xue Za Zhi,* 41**,** 67-74.

Lim, M., Carollo, A., Neoh, M. J. Y., Sacchiero, M., Azhari, A., Balboni, G., Marschik, P., Nordahl-Hansen, A., Dimitriou, D. & Esposito, G. 2023. Developmental disabilities in Africa: A scientometric review. *Research in Developmental Disabilities,* 133**,** 104395.

Lincoln, S. E., Hambuch, T., Zook, J. M., Bristow, S. L., Hatchell, K., Truty, R., Kennemer, M., Shirts, B. H., Fellowes, A., Chowdhury, S., Klee, E. W., Mahamdallie, S., Cleveland, M. H., Vallone, P. M., Ding, Y., Seal, S., Desilva, W., Tomson, F. L., Huang, C., Garlick, R. K., Rahman, N., Salit, M., Kingsmore, S. F., Ferber, M. J., Aradhya, S. & Nussbaum, R. L. 2021. One in seven pathogenic variants can be challenging to detect by NGS: an analysis of 450,000 patients with implications for clinical sensitivity and genetic test implementation. *Genet Med,* 23**,** 1673-1680.

Liu, P., Meng, L., Normand, E. A., Xia, F., Song, X., Ghazi, A., Rosenfeld, J., Magoulas, P. L., Braxton, A., Ward, P., Dai, H., Yuan, B., Bi, W., Xiao, R., Wang, X., Chiang, T., Vetrini, F., He, W., Cheng, H., Dong, J., Gijavanekar, C., Benke, P. J., Bernstein, J. A., Eble, T., Eroglu, Y., Erwin, D., Escobar, L., Gibson, J. B., Gripp, K., Kleppe, S., Koenig, M. K., Lewis, A. M., Natowicz, M., Mancias, P., Minor, L., Scaglia, F., Schaaf, C. P., Streff, H., Vernon, H., Uhles, C. L., Zackai, E. H., Wu, N., Sutton, V. R., Beaudet, A. L., Muzny, D., Gibbs, R. A., Posey, J. E., Lalani, S., Shaw, C., Eng, C. M., Lupski, J. R. & Yang, Y. 2019. Reanalysis of Clinical Exome Sequencing Data. *New England Journal of Medicine,* 380**,** 2478-2480.

Liu, X., Li, C., Mou, C., Dong, Y. & Tu, Y. 2020. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Medicine,* 12**,** 103.

Lohmann, K. & Klein, C. 2014. Next Generation Sequencing and the Future of Genetic Diagnosis. *Neurotherapeutics,* 11**,** 699-707.

Lu, C., Black, M. M. & Richter, L. M. 2016. Risk of poor development in young children in low-income and middle-income countries: an estimation and analysis at the global, regional, and country level. *Lancet Glob Health,* 4**,** e916-e922.

Lu, G., Peng, Q., Wu, L., Zhang, J. & Ma, L. 2021. Identification of de novo mutations for ARID1B haploinsufficiency associated with Coffin-Siris syndrome 1 in three Chinese families via array-CGH and whole exome sequencing. *BMC Med Genomics,* 14**,** 270.

Lumaka, A., Carstens, N., Devriendt, K., Krause, A., Kulohoma, B., Kumuthini, J., Mubungu, G., Mukisa, J., Nel, M., Olanrewaju, T. O., Lombard, Z. & Landouré, G. 2022. Increasing African genomic data generation and sharing to resolve rare and undiagnosed diseases in Africa: a call-to-action by the H3Africa rare diseases working group. *Orphanet J Rare Dis,* 17**,** 230.

Maassen, W., Legger, G., Kul Cinar, O., Van Daele, P., Gattorno, M., Bader-Meunier, B., Wouters, C., Briggs, T., Johansson, L., Van Der Velde, J., Swertz, M., Omoyinmi, E., Hoppenreijs, E., Belot, A., Eleftheriou, D., Caorsi, R., Aeschlimann, F., Boursier, G., Brogan, P., Haimel, M. & Van Gijn, M. 2023. Curation and expansion of the Human Phenotype Ontology for systemic autoinflammatory diseases improves phenotype-driven disease-matching. *Front Immunol,* 14**,** 1215869.

Macdonald, J. R., Ziman, R., Yuen, R. K., Feuk, L. & Scherer, S. W. 2014. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res,* 42**,** D986-92.

Macnee, M., Pérez-Palma, E., Brünger, T., Klöckner, C., Platzer, K., Stefanski, A., Montanucci, L., Bayat, A., Radtke, M., Collins, R. L., Talkowski, M., Blankenberg, D., Møller, R. S., Lemke, J. R., Nothnagel, M., May, P. & Lal, D. 2022. CNV-ClinViewer: Enhancing the clinical interpretation of large copy-number variants online. *medRxiv***,** 2022.03.23.22272818.

Makoni, M. 2020. Africa's $100-million Pathogen Genomics Initiative. *The Lancet Microbe,* 1**,** e318.

Manickam, K., Mcclain, M. R., Demmer, L. A., Biswas, S., Kearney, H. M., Malinowski, J., Massingham, L. J., Miller, D., Yu, T. W. & Hisama, F. M. 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). *Genetics in Medicine,* 23**,** 2029-2037.

Mantere, T., Kersten, S. & Hoischen, A. 2019. Long-Read Sequencing Emerging in Medical Genetics. *Front Genet,* 10**,** 426.

Marchuk, D. S., Crooks, K., Strande, N., Kaiser-Rogers, K., Milko, L. V., Brandt, A., Arreola, A., Tilley, C. R., Bizon, C., Vora, N. L., Wilhelmsen, K. C., Evans, J. P. & Berg, J. S. 2018. Increasing the diagnostic yield of exome sequencing by copy number variant analysis. *PLoS One,* 13**,** e0209185.

Masri, A. & Hamamy, H. 2021. Cost Effectiveness of Whole Exome Sequencing for Children with Developmental Delay in a Developing Country: A Study from Jordan. *Journal of Pediatric Neurology,* 20**,** 020-023.

Medvedev, P., Fiume, M., Dzamba, M., Smith, T. & Brudno, M. 2010. Detecting copy number variation with mated short reads. *Genome Research,* 20**,** 1613-1622.

Mellone, S., Puricelli, C., Vurchio, D., Ronzani, S., Favini, S., Maruzzi, A., Peruzzi, C., Papa, A., Spano, A., Sirchia, F., Mandrile, G., Pelle, A., Rasmini, P., Vercellino, F., Zonta, A., Rabbone, I., Dianzani, U., Viri, M. & Giordano, M. 2022. The Usefulness of a Targeted Next Generation Sequencing Gene Panel in Providing Molecular Diagnosis to Patients With a Broad Spectrum of Neurodevelopmental Disorders. *Front Genet,* 13**,** 875182.

Merino Elia, A. & Coghill, J. 2021. A practical guide to assessing and investigating developmental delay. *Paediatrics and Child Health,* 31**,** 335-339.

Merkel, D. 2014. Docker: lightweight Linux containers for consistent development and deployment. *Linux J.,* 2014**,** Article 2.

Miclea, D., Szucs, A., Mirea, A., Stefan, D. M., Nazarie, F., Bucerzan, S., Lazea, C., Grama, A., Pop, T. L., Farcas, M., Zaharie, G., Matyas, M., Mager, M., Vintan, M., Popp, R. & Alkhzouz, C. 2021. Diagnostic Usefulness of MLPA Techniques for Recurrent Copy Number Variants Detection in Global Developmental Delay/Intellectual Disability. *Int J Gen Med,* 14**,** 4511-4515.

Miller, D. T., Adam, M. P., Aradhya, S., Biesecker, L. G., Brothman, A. R., Carter, N. P., Church, D. M., Crolla, J. A., Eichler, E. E., Epstein, C. J., Faucett, W. A., Feuk, L., Friedman, J. M., Hamosh, A., Jackson, L., Kaminsky, E. B., Kok, K., Krantz, I. D., Kuhn, R. M., Lee, C., Ostell, J. M., Rosenberg, C., Scherer, S. W., Spinner, N. B., Stavropoulos, D. J., Tepperberg, J. H., Thorland, E. C., Vermeesch, J. R., Waggoner, D. J., Watson, M. S., Martin, C. L. & Ledbetter, D. H. 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**,** 749-64.

Mithyantha, R., Kneen, R., Mccann, E. & Gladstone, M. 2017. Current evidence-based recommendations on investigating children with global developmental delay. *Archives of Disease in Childhood,* 102**,** 1071-1076.

Moeschler, J. B. & Shevell, M. 2014. Comprehensive evaluation of the child with intellectual disability or global developmental delays. *Pediatrics,* 134**,** e903-18.

Mollison, L., O'daniel, J. M., Henderson, G. E., Berg, J. S. & Skinner, D. 2020. Parents' perceptions of personal utility of exome sequencing results. *Genet Med,* 22**,** 752-757.

Monroe, G. R., Frederix, G. W., Savelberg, S. M., De Vries, T. I., Duran, K. J., Van Der Smagt, J. J., Terhal, P. A., Van Hasselt, P. M., Kroes, H. Y., Verhoeven-Duif, N. M., Nijman, I. J., Carbo, E. C., Van Gassen, K. L., Knoers, N. V., Hovels, A. M., Van Haelst, M. M., Visser, G. & Van Haaften, G. 2016. Effectiveness of whole-exome sequencing and costs of the traditional diagnostic trajectory in children with intellectual disability. *Genet Med,* 18**,** 949-56.

Moosa, S., Coetzer, K. C., Lee, E. & Seo, G. H. 2022. Undiagnosed disease program in South Africa: Results from first 100 exomes. *Am J Med Genet A,* 188**,** 2684-2692.

Moosmann, J., Uebe, S., Dittrich, S., Rüffer, A., Ekici, A. & Toka, O. 2015. Novel Loci for Non-Syndromic Coarctation of the Aorta in Sporadic and Familial Cases. *PLoS ONE,* 10.

Mpangase, P., Frost, J., Tikly, M., Ramsay, M. & Hazelhurst, S. 2021. nf-rnaSeqCount: A Nextflow pipeline for obtaining raw read counts from RNA-seq data. *South African Computer Journal,* 33.

Mulder, N., Abimiku, A., Adebamowo, S. N., De Vries, J., Matimba, A., Olowoyo, P., Ramsay, M., Skelton, M. & Stein, D. J. 2018. H3Africa: current perspectives. *Pharmgenomics Pers Med,* 11**,** 59-66.

Mulder, N. J., Adebiyi, E., Alami, R., Benkahla, A., Brandful, J., Doumbia, S., Everett, D., Fadlelmola, F. M., Gaboun, F., Gaseitsiwe, S., Ghazal, H., Hazelhurst, S., Hide, W., Ibrahimi, A., Jaufeerally Fakim, Y., Jongeneel, C. V., Joubert, F., Kassim, S., Kayondo, J., Kumuthini, J., Lyantagaye, S., Makani, J., Mansour Alzohairy, A., Masiga, D., Moussa, A., Nash, O., Ouwe Missi Oukem-Boyer, O., Owusu-Dabo, E., Panji, S., Patterton, H., Radouani, F., Sadki, K., Seghrouchni, F., Tastan Bishop, Ö., Tiffin, N. & Ulenga, N. 2016. H3ABioNet, a sustainable pan-African bioinformatics network for human heredity and health in Africa. *Genome Res,* 26**,** 271-7.

Mwaka, E. S., Sebatta, D. E., Ochieng, J., Munabi, I. G., Bagenda, G., Ainembabazi, D. & Kaawa-Mafigiri, D. 2021. Researchers' perspectives on return of individual genetics results to research participants: a qualitative study. *Glob Bioeth,* 32**,** 15-33.

Nijkamp, J. F., Van Den Broek, M. A., Geertman, J. M., Reinders, M. J., Daran, J. M. & De Ridder, D. 2012. De novo detection of copy number variation by co-assembly. *Bioinformatics,* 28**,** 3195-202.

Noonan, J. A. 1968. Hypertelorism with Turner phenotype. A new syndrome with associated congenital heart disease. *Am J Dis Child,* 116**,** 373-80.

Nyangiri, O. A., Noyes, H., Mulindwa, J., Ilboudo, H., Kabore, J. W., Ahouty, B., Koffi, M., Asina, O. F., Mumba, D., Ofon, E., Simo, G., Kimuda, M. P., Enyaru, J., Alibu, V. P., Kamoto, K., Chisi, J., Simuunza, M., Camara, M., Sidibe, I., Macleod, A., Bucheton, B., Hall, N., Hertz-Fowler, C., Matovu, E. & Trypanogen Research Group, A. M. O. T. H. a. C. 2020. Copy number variation in human genomes from three major ethno-linguistic groups in Africa. *BMC Genomics,* 21**,** 289.

O'fallon, B., Durtschi, J., Kellogg, A., Lewis, T., Close, D. & Best, H. 2022. Algorithmic improvements for discovery of germline copy number variants in next-generation sequencing data. *BMC Bioinformatics,* 23**,** 285.

Oehler, J. B., Wright, H., Stark, Z., Mallett, A. J. & Schmitz, U. 2023. The application of long-read sequencing in clinical settings. *Human Genomics,* 17**,** 73.

Olivucci, G., Iovino, E., Innella, G., Turchetti, D., Pippucci, T. & Magini, P. 2024. Long read sequencing on its way to the routine diagnostics of genetic diseases. *Front Genet,* 15**,** 1374860.

Olusanya, B. O., Davis, A. C., Wertlieb, D., Boo, N.-Y., Nair, M. K. C., Halpern, R., Kuper, H., Breinbauer, C., De Vries, P. J., Gladstone, M., Halfon, N., Kancherla, V., Mulaudzi, M. C., Kakooza-Mwesige, A., Ogbo, F. A., Olusanya, J. O., Williams, A. N., Wright, S. M., Manguerra, H., Smith, A., Echko, M., Ikeda, C., Liu, A., Millear, A., Ballesteros, K., Nichols, E., Erskine, H. E., Santomauro, D., Rankin, Z., Smith, M., Whiteford, H. A., Olsen, H. E. & Kassebaum, N. J. 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**,** e1100-e1121.

Packer, J. S., Maxwell, E. K., O'dushlaine, C., Lopez, A. E., Dewey, F. E., Chernomorsky, R., Baras, A., Overton, J. D., Habegger, L. & Reid, J. G. 2016. CLAMMS: a scalable algorithm for calling common and rare copy number variants from exome sequencing data. *Bioinformatics,* 32**,** 133-5.

Pais, L. S., Snow, H., Weisburd, B., Zhang, S., Baxter, S. M., Ditroia, S., O'heir, E., England, E., Chao, K. R., Lemire, G., Osei-Owusu, I., Vannoy, G. E., Wilson, M., Nguyen, K., Arachchi, H., Phu, W., Solomonson, M., Mano, S., O'leary, M., Lovgren, A., Babb, L., Austin-Tse, C. A., Rehm, H. L., Macarthur, D. G. & O'donnell-Luria, A. 2022. seqr: A web-based analysis and collaboration tool for rare disease genomics. *Hum Mutat,* 43**,** 698-707.

Pang, A. W., Macdonald, J. R., Pinto, D., Wei, J., Rafiq, M. A., Conrad, D. F., Park, H., Hurles, M. E., Lee, C., Venter, J. C., Kirkness, E. F., Levy, S., Feuk, L. & Scherer, S. W. 2010. Towards a comprehensive structural variation map of an individual human genome. *Genome Biology,* 11**,** R52.

Park, K. B., Nam, K. E., Cho, A. R., Jang, W., Kim, M. & Park, J. H. 2019. Effects of Copy Number Variations on Developmental Aspects of Children With Delayed Development. *Ann Rehabil Med,* 43**,** 215-223.

Patel, R. Y., Shah, N., Jackson, A. R., Ghosh, R., Pawliczek, P., Paithankar, S., Baker, A., Riehle, K., Chen, H., Milosavljevic, S., Bizon, C., Rynearson, S., Nelson, T., Jarvik, G. P., Rehm, H. L., Harrison, S. M., Azzariti, D., Powell, B., Babb, L.,

Plon, S. E., Milosavljevic, A. & On Behalf of the Clingen, R. 2017. ClinGen Pathogenicity Calculator: a configurable system for assessing pathogenicity of genetic variants. *Genome Medicine,* 9**,** 3.

Petrij, F., Giles, R. H., Dauwerse, H. G., Saris, J. J., Hennekam, R. C., Masuno, M., Tommerup, N., Van Ommen, G. J., Goodman, R. H., Peters, D. J. & Et Al. 1995. Rubinstein-Taybi syndrome caused by mutations in the transcriptional co-activator CBP. *Nature,* 376**,** 348-51.

Pfundt, R., Del Rosario, M., Vissers, L., Kwint, M. P., Janssen, I. M., De Leeuw, N., Yntema, H. G., Nelen, M. R., Lugtenberg, D., Kamsteeg, E. J., Wieskamp, N., Stegmann, A. P. A., Stevens, S. J. C., Rodenburg, R. J. T., Simons, A., Mensenkamp, A. R., Rinne, T., Gilissen, C., Scheffer, H., Veltman, J. a. P. D. & Hehir-Kwa, J. Y. 2017. Detection of clinically relevant copy-number variants by exome sequencing in a large cohort of genetic disorders. *Genet Med,* 19**,** 667-675.

Philippakis, A. A., Azzariti, D. R., Beltran, S., Brookes, A. J., Brownstein, C. A., Brudno, M., Brunner, H. G., Buske, O. J., Carey, K., Doll, C., Dumitriu, S., Dyke, S. O., Den Dunnen, J. T., Firth, H. V., Gibbs, R. A., Girdea, M., Gonzalez, M., Haendel, M. A., Hamosh, A., Holm, I. A., Huang, L., Hurles, M. E., Hutton, B., Krier, J. B., Misyura, A., Mungall, C. J., Paschall, J., Paten, B., Robinson, P. N., Schiettecatte, F., Sobreira, N. L., Swaminathan, G. J., Taschner, P. E., Terry, S. F., Washington, N. L., Züchner, S., Boycott, K. M. & Rehm, H. L. 2015. The Matchmaker Exchange: a platform for rare disease gene discovery. *Hum Mutat,* 36**,** 915-21.

Plagnol, V., Curtis, J., Epstein, M., Mok, K. Y., Stebbings, E., Grigoriadou, S., Wood, N. W., Hambleton, S., Burns, S. O., Thrasher, A. J., Kumararatne, D., Doffinger, R. & Nejentsev, S. 2012. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics,* 28**,** 2747-54.

Pös, O., Radvanszky, J., Styk, J., Pös, Z., Buglyó, G., Kajsik, M., Budis, J., Nagy, B. & Szemes, T. 2021. Copy Number Variation: Methods and Clinical Applications. *Applied Sciences,* 11**,** 819.

Potti, T., Petty, E. & Lesperance, M. 2011. A Comprehensive Review of Reported Heritable Noggin-Associated Syndromes and Proposed Clinical Utility of One Broadly Inclusive Diagnostic Term: NOG-Related-Symphalangism Spectrum Disorder (NOG-SSD). *Human mutation,* 32**,** 877-86.

Pounraja, V. K., Jayakar, G., Jensen, M., Kelkar, N. & Girirajan, S. 2019. A machine-learning approach for accurate detection of copy-number variants from exome sequencing. *bioRxiv***,** 460931.

Pranav Chand, R., Vinit, W., Vaidya, V., Iyer, A. S., Shelke, M., Aggarwal, S., Magar, S., Danda, S., Moirangthem, A., Phadke, S. R., Goyal, M., Ranganath, P., Mistri, M., Shah, P., Shah, N. & Kotecha, U. H. 2023. Proband only exome sequencing in 403 Indian children with neurodevelopmental disorders: Diagnostic yield, utility and challenges in a resource-limited setting. *European Journal of Medical Genetics,* 66**,** 104730.

Quenez, O., Cassinari, K., Coutant, S., Lecoquierre, F., Le Guennec, K., Rousseau, S., Richard, A.-C., Vasseur, S., Bouvignies, E., Bou, J., Lienard, G., Manase, S., Fourneaux, S., Drouot, N., Nguyen-Viet, V., Vezain, M., Chambon, P., Joly-Helas, G., Le Meur, N., Castelain, M., Boland, A., Deleuze, J.-F., Génin, E., Campion, D., Dartigues, J.-F., Deleuze, J.-F., Lambert, J.-C., Redon, R., Ludwig, T., Grenier-Boley, B., Letort, S., Lindenbaum, P., Meyer, V., Quenez,

O., Dina, C., Bellenguez, C., Charbonnier, C., Giemza, J., Chatel, S., Férec, C., Le marec, H., Letenneur, L., Nicolas, G., Rouault, K., Bacq, D., Boland, A., Lechner, D., Tournier, I., Charbonnier, F., Kasper, E., Bougeard, G., Frebourg, T., Saugier-Veber, P., Baert-Desurmont, S., Campion, D., Rovelet-Lecrux, A., Nicolas, G. & Consortium, F. 2021. Detection of copy-number variations from NGS data using read depth information: a diagnostic performance evaluation. *European Journal of Human Genetics,* 29**,** 99-109.

Quinlan, A. R., Clark, R. A., Sokolova, S., Leibowitz, M. L., Zhang, Y., Hurles, M. E., Mell, J. C. & Hall, I. M. 2010. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res,* 20**,** 623-35.

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

Rajagopalan, R., Murrell, J. R., Luo, M. & Conlin, L. K. 2020. A highly sensitive and specific workflow for detecting rare copy-number variants from exome sequencing data. *Genome Med,* 12**,** 14.

Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V. & Korbel, J. O. 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics,* 28**,** i333-i339.

Rautiainen, M., Nurk, S., Walenz, B. P., Logsdon, G. A., Porubsky, D., Rhie, A., Eichler, E. E., Phillippy, A. M. & Koren, S. 2023. Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nature Biotechnology,* 41**,** 1474-1482.

Ravnan, J. B., Tepperberg, J. H., Papenhausen, P., Lamb, A. N., Hedrick, J., Eash, D., Ledbetter, D. H. & Martin, C. L. 2006. Subtelomere FISH analysis of 11 688 cases: an evaluation of the frequency and pattern of subtelomere rearrangements in individuals with developmental disabilities. *J Med Genet,* 43**,** 478-89.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., González, J. R., Gratacòs, M., Huang, J., Kalaitzopoulos, D., Komura, D., Macdonald, J. R., Marshall, C. R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M. J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D. F., Estivill, X., Tyler-Smith, C., Carter, N. P., Aburatani, H., Lee, C., Jones, K. W., Scherer, S. W. & Hurles, M. E. 2006. Global variation in copy number in the human genome. *Nature,* 444**,** 444-54.

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K. & Rehm, H. L. 2015. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med,* 17**,** 405-24.

Riggs, E. R., Andersen, E. F., Cherry, A. M., Kantarci, S., Kearney, H., Patel, A., Raca, G., Ritter, D. I., South, S. T., Thorland, E. C., Pineda-Alvarez, D., Aradhya, S. & Martin, C. L. 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**,** 245-257.

Riggs, E. R., Nelson, T., Merz, A., Ackley, T., Bunke, B., Collins, C. D., Collinson, M. N., Fan, Y. S., Goodenberger, M. L., Golden, D. M., Haglund-Hazy, L., Krgovic,

D., Lamb, A. N., Lewis, Z., Li, G., Liu, Y., Meck, J., Neufeld-Kaiser, W., Runke, C. K., Sanmann, J. N., Stavropoulos, D. J., Strong, E., Su, M., Tayeh, M. K., Kokalj Vokac, N., Thorland, E. C., Andersen, E. & Martin, C. L. 2018. Copy number variant discrepancy resolution using the ClinGen dosage sensitivity map results in updated clinical interpretations in ClinVar. *Hum Mutat,* 39**,** 1650-1659.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G. & Mesirov, J. P. 2011. Integrative genomics viewer. *Nat Biotechnol,* 29**,** 24-6.

Romdhane, L., Mezzi, N., Dallali, H., Messaoud, O., Shan, J., Fakhro, K. A., Kefi, R., Chouchane, L. & Abdelhak, S. 2021. A map of copy number variations in the Tunisian population: a valuable tool for medical genomics in North Africa. *npj Genomic Medicine,* 6**,** 3.

Rossum, G. V. & Drake, F. L. 2009. *Python 3 Reference Manual*, CreateSpace.

Royer-Bertrand, B., Cisarova, K., Niel-Butschi, F., Mittaz-Crettol, L., Fodstad, H. & Superti-Furga, A. 2021. CNV Detection from Exome Sequencing Data in Routine Diagnostics of Rare Genetic Disorders: Opportunities and Limitations. *Genes (Basel),* 12.

Santen, G. W. & Clayton-Smith, J. 2014. The ARID1B phenotype: what we have learned so far. *Am J Med Genet C Semin Med Genet,* 166c**,** 276-89.

Satam, H., Joshi, K., Mangrolia, U., Waghoo, S., Zaidi, G., Rawool, S., Thakare, R. P., Banday, S., Mishra, A. K., Das, G. & Malonia, S. K. 2023. Next-Generation Sequencing Technology: Current Trends and Advancements. *Biology (Basel),* 12.

Savatt, J. M. & Myers, S. M. 2021. Genetic Testing in Neurodevelopmental Disorders. *Frontiers in Pediatrics,* 9.

Schuermans, N., Hemelsoet, D., Terryn, W., Steyaert, S., Van Coster, R., Coucke, P. J., Steyaert, W., Callewaert, B., Bogaert, E., Verloo, P., Vanlander, A. V., Debackere, E., Ghijsels, J., Leblanc, P., Verdin, H., Naesens, L., Haerynck, F., Callens, S., Dermaut, B., Poppe, B., De Bleecker, J., Santens, P., Boon, P., Laureys, G., Kerre, T. & For, U. D. P. 2022. Shortcutting the diagnostic odyssey: the multidisciplinary Program for Undiagnosed Rare Diseases in adults (UD-PrOZA). *Orphanet Journal of Rare Diseases,* 17**,** 210.

Schwarze, K., Buchanan, J., Taylor, J. C. & Wordsworth, S. 2018. Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *Genet Med,* 20**,** 1122-1130.

Shaffer, J. G., Mather, F. J., Wele, M., Li, J., Tangara, C. O., Kassogue, Y., Srivastav, S. K., Thiero, O., Diakite, M., Sangare, M., Dabitao, D., Toure, M., Djimde, A. A., Traore, S., Diakite, B., Coulibaly, M. B., Liu, Y., Lacey, M., Lefante, J. J., Koita, O., Schieffelin, J. S., Krogstad, D. J. & Doumbia, S. O. 2019. Expanding Research Capacity in Sub-Saharan Africa Through Informatics, Bioinformatics, and Data Science Training Programs in Mali. *Frontiers in Genetics,* Volume 10 - 2019.

Sheth, F., Shah, J., Jain, D., Shah, S., Patel, H., Patel, K., Solanki, D., Iyer, A., Menghani, B., Mhatre, P., Mehta, S., Bajaj, S., Patel, V., Pandya, M., Dhami, M., Patel, D., Sheth, J. & Sheth, H. 2023. Comparative yield of molecular diagnostic algorithms for autism spectrum disorder diagnosis in India: Evidence supporting whole exome sequencing as first tier test PREPRINT (Version 1) available at: Research Square 10.21203/rs.3.rs-2888202/v1 (Accessed June 13, 2023).

Shingwenyana, B., Rossouw, B., Thom, J., Louw, N., Krause, A. & Lombard, Z. 2023. Research participants' perspectives regarding the feedback of secondary findings-A cohort from the DDD-Africa study, South Africa. *J Genet Couns*.

Spector, J. D. & Wiita, A. P. 2019. ClinTAD: a tool for copy number variant interpretation in the context of topologically associated domains. *J Hum Genet,* 64**,** 437-443.

Srivastava, S., Love-Nichols, J. A., Dies, K. A., Ledbetter, D. H., Martin, C. L., Chung, W. K., Firth, H. V., Frazier, T., Hansen, R. L., Prock, L., Brunner, H., Hoang, N., Scherer, S. W., Sahin, M. & Miller, D. T. 2019. Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders. *Genet Med,* 21**,** 2413-2421.

Stark, Z., Tan, T. Y., Chong, B., Brett, G. R., Yap, P., Walsh, M., Yeung, A., Peters, H., Mordaunt, D., Cowie, S., Amor, D. J., Savarirayan, R., Mcgillivray, G., Downie, L., Ekert, P. G., Theda, C., James, P. A., Yaplito-Lee, J., Ryan, M. M., Leventer, R. J., Creed, E., Macciocca, I., Bell, K. M., Oshlack, A., Sadedin, S., Georgeson, P., Anderson, C., Thorne, N., Melbourne Genomics Health, A., Gaff, C. & White, S. M. 2016. A prospective evaluation of whole-exome sequencing as a first-tier molecular test in infants with suspected monogenic disorders. *Genet Med,* 18**,** 1090-1096.

Strom, S. P., Lee, H., Das, K., Vilain, E., Nelson, S. F., Grody, W. W. & Deignan, J. L. 2014. Assessing the necessity of confirmatory testing for exome-sequencing results in a clinical molecular diagnostic laboratory. *Genet Med,* 16**,** 510-5.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T. & Collins, R. 2015. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med,* 12**,** e1001779.

Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. 2016. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLOS Computational Biology,* 12**,** e1004873.

Tan, N. B., Stapleton, R., Stark, Z., Delatycki, M. B., Yeung, A., Hunter, M. F., Amor, D. J., Brown, N. J., Stutterd, C. A., Mcgillivray, G., Yap, P., Regan, M., Chong, B., Fanjul Fernandez, M., Marum, J., Phelan, D., Pais, L. S., White, S. M., Lunke, S. & Tan, T. Y. 2020. Evaluating systematic reanalysis of clinical genomic data in rare disease from single center experience and literature review. *Mol Genet Genomic Med,* 8**,** e1508.

Tan, R., Wang, Y., Kleinstein, S. E., Liu, Y., Zhu, X., Guo, H., Jiang, Q., Allen, A. S. & Zhu, M. 2014. An evaluation of copy number variation detection tools from whole-exome sequencing data. *Hum Mutat,* 35**,** 899-907.

Teo, S. M., Pawitan, Y., Ku, C. S., Chia, K. S. & Salim, A. 2012. Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics,* 28**,** 2711-8.

Testard, Q., Vanhoye, X., Yauy, K., Naud, M.-E., Vieville, G., Rousseau, F., Dauriat, B., Marquet, V., Bourthoumieu, S., Genevieve, D., Gatinois, V., Wells, C., Willems, M., Coubes, C., Pinson, L., Dard, R., Tessier, A., Hervé, B., Vialard, F., Harzallah, I., Touraine, R., Cogné, B., Deb, W., Besnard, T., Pichon, O., Laudier, B., Mesnard, L., Doreille, A., Busa, T., Missirian, C., Satre, V., Coutton, C., Celse, T., Harbuz, R., Raymond, L., Taly, J.-F. & Thevenon, J. 2021. Exome sequencing as a first-tier test for copy number variant detection : retrospective

evaluation and prospective screening in 2418 cases. *medRxiv*, 2021.10.14.21264732.

Testard, Q., Vanhoye, X., Yauy, K., Naud, M. E., Vieville, G., Rousseau, F., Dauriat, B., Marquet, V., Bourthoumieu, S., Geneviève, D., Gatinois, V., Wells, C., Willems, M., Coubes, C., Pinson, L., Dard, R., Tessier, A., Hervé, B., Vialard, F., Harzallah, I., Touraine, R., Cogné, B., Deb, W., Besnard, T., Pichon, O., Laudier, B., Mesnard, L., Doreille, A., Busa, T., Missirian, C., Satre, V., Coutton, C., Celse, T., Harbuz, R., Raymond, L., Taly, J. F. & Thevenon, J. 2022. Exome sequencing as a first-tier test for copy number variant detection: retrospective evaluation and prospective screening in 2418 cases. *J Med Genet,* 59**,** 1234-1240.

Tetikol, H. S., Turgut, D., Narci, K., Budak, G., Kalay, O., Arslan, E., Demirkaya-Budak, S., Dolgoborodov, A., Kabakci-Zorlu, D., Semenyuk, V., Jain, A. & Davis-Dusenbery, B. N. 2022. Pan-African genome demonstrates how population-specific genome graphs improve high-throughput sequencing data analysis. *Nature Communications,* 13**,** 4384.

Thevenon, J., Callier, P., Andrieux, J., Delobel, B., David, A., Sukno, S., Minot, D., Mosca Anne, L., Marle, N., Sanlaville, D., Bonnet, M., Masurel-Paulet, A., Levy, F., Gaunt, L., Farrell, S., Le Caignec, C., Toutain, A., Carmignac, V., Mugneret, F., Clayton-Smith, J., Thauvin-Robinet, C. & Faivre, L. 2013. 12p13.33 microdeletion including ELKS/ERC1, a new locus associated with childhood apraxia of speech. *Eur J Hum Genet,* 21**,** 82-8.

Thormann, A., Halachev, M., Mclaren, W., Moore, D. J., Svinti, V., Campbell, A., Kerr, S. M., Tischkowitz, M., Hunt, S. E., Dunlop, M. G., Hurles, M. E., Wright, C. F., Firth, H. V., Cunningham, F. & Fitzpatrick, D. R. 2019. Flexible and scalable diagnostic filtering of genomic variants using G2P with Ensembl VEP. *Nature Communications,* 10**,** 2373.

Tibiri, E. B., Boua, P. R., Soulama, I., Dubreuil-Tranchant, C., Tando, N., Tollenaere, C., Brugidou, C., Nanema, R. K. & Tiendrebeogo, F. 2025. Challenges and opportunities of developing bioinformatics platforms in Africa: the case of BurkinaBioinfo at Joseph Ki-Zerbo University, Burkina Faso. *Briefings in Bioinformatics,* 26.

Trappe, K., Emde, A.-K., Ehrlich, H.-C. & Reinert, K. 2014. Gustaf: Detecting and correctly classifying SVs in the NGS twilight zone. *Bioinformatics,* 30**,** 3484-3490.

Truty, R., Paul, J., Kennemer, M., Lincoln, S. E., Olivares, E., Nussbaum, R. L. & Aradhya, S. 2019. Prevalence and properties of intragenic copy-number variation in Mendelian disease genes. *Genet Med,* 21**,** 114-123.

Tuzun, E., Sharp, A. J., Bailey, J. A., Kaul, R., Morrison, V. A., Pertz, L. M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D., Olson, M. V. & Eichler, E. E. 2005. Fine-scale structural variation of the human genome. *Nat Genet,* 37**,** 727-32.

Välipakka, S., Savarese, M., Sagath, L., Arumilli, M., Giugliano, T., Udd, B. & Hackman, P. 2020. Improving Copy Number Variant Detection from Sequencing Data with a Combination of Programs and a Predictive Model. *The Journal of Molecular Diagnostics,* 22**,** 40-49.

Van Der Sanden, B. P. G. H., Schobers, G., Corominas Galbany, J., Koolen, D. A., Sinnema, M., Van Reeuwijk, J., Stumpel, C. T. R. M., Kleefstra, T., De Vries, B. B. A., Ruiterkamp-Versteeg, M., Leijsten, N., Kwint, M., Derks, R., Swinkels, H., Den Ouden, A., Pfundt, R., Rinne, T., De Leeuw, N., Stegmann, A. P., Stevens, S. J., Van Den Wijngaard, A., Brunner, H. G., Yntema, H. G., Gilissen,

C., Nelen, M. R. & Vissers, L. E. L. M. 2023. The performance of genome sequencing as a first-tier test for neurodevelopmental disorders. *European Journal of Human Genetics,* 31**,** 81-88.

Van Der Sluijs, P. J., Jansen, S., Vergano, S. A., Adachi-Fukuda, M., Alanay, Y., Alkindy, A., Baban, A., Bayat, A., Beck-Wödl, S., Berry, K., Bijlsma, E. K., Bok, L. A., Brouwer, A. F. J., Van Der Burgt, I., Campeau, P. M., Canham, N., Chrzanowska, K., Chu, Y. W. Y., Chung, B. H. Y., Dahan, K., De Rademaeker, M., Destree, A., Dudding-Byth, T., Earl, R., Elcioglu, N., Elias, E. R., Fagerberg, C., Gardham, A., Gener, B., Gerkes, E. H., Grasshoff, U., Van Haeringen, A., Heitink, K. R., Herkert, J. C., Den Hollander, N. S., Horn, D., Hunt, D., Kant, S. G., Kato, M., Kayserili, H., Kersseboom, R., Kilic, E., Krajewska-Walasek, M., Lammers, K., Laulund, L. W., Lederer, D., Lees, M., López-González, V., Maas, S., Mancini, G. M. S., Marcelis, C., Martinez, F., Maystadt, I., Mcguire, M., Mckee, S., Mehta, S., Metcalfe, K., Milunsky, J., Mizuno, S., Moeschler, J. B., Netzer, C., Ockeloen, C. W., Oehl-Jaschkowitz, B., Okamoto, N., Olminkhof, S. N. M., Orellana, C., Pasquier, L., Pottinger, C., Riehmer, V., Robertson, S. P., Roifman, M., Rooryck, C., Ropers, F. G., Rosello, M., Ruivenkamp, C. a. L., Sagiroglu, M. S., Sallevelt, S. C. E. H., Sanchis Calvo, A., Simsek-Kiper, P. O., Soares, G., Solaeche, L., Sonmez, F. M., Splitt, M., Steenbeek, D., Stegmann, A. P. A., Stumpel, C. T. R. M., Tanabe, S., Uctepe, E., Utine, G. E., Veenstra-Knol, H. E., Venkateswaran, S., Vilain, C., Vincent-Delorme, C., Vulto-Van Silfhout, A. T., Wheeler, P., Wilson, G. N., Wilson, L. C., Wollnik, B., Kosho, T., Wieczorek, D., et al. 2019. The ARID1B spectrum in 143 patients: from nonsyndromic intellectual disability to Coffin–Siris syndrome. *Genetics in Medicine,* 21**,** 1295-1307.

Vaseghi, H., Akrami, S. M. & Rashidi-Nezhad, A. 2023. The challenges in the interpretation of genetic variants detected by genomics techniques in patients with congenital anomalies. *Journal of Clinical Laboratory Analysis,* 37**,** e24967.

Vermeesch, J. R., Fiegler, H., De Leeuw, N., Szuhai, K., Schoumans, J., Ciccone, R., Speleman, F., Rauch, A., Clayton-Smith, J., Van Ravenswaaij, C., Sanlaville, D., Patsalis, P. C., Firth, H., Devriendt, K. & Zuffardi, O. 2007. Guidelines for molecular karyotyping in constitutional genetic diagnosis. *Eur J Hum Genet,* 15**,** 1105-14.

Wang, T., Antonacci-Fulton, L., Howe, K., Lawson, H. A., Lucas, J. K., Phillippy, A. M., Popejoy, A. B., Asri, M., Carson, C., Chaisson, M. J. P., Chang, X., Cook-Deegan, R., Felsenfeld, A. L., Fulton, R. S., Garrison, E. P., Garrison, N. A., Graves-Lindsay, T. A., Ji, H., Kenny, E. E., Koenig, B. A., Li, D., Marschall, T., Mcmichael, J. F., Novak, A. M., Purushotham, D., Schneider, V. A., Schultz, B. I., Smith, M. W., Sofia, H. J., Weissman, T., Flicek, P., Li, H., Miga, K. H., Paten, B., Jarvis, E. D., Hall, I. M., Eichler, E. E., Haussler, D. & The Human Pangenome Reference, C. 2022. The Human Pangenome Project: a global resource to map genomic diversity. *Nature,* 604**,** 437-446.

Wiener, E. K., Buchanan, J., Krause, A. & Lombard, Z. 2023. Retrospective file review shows limited genetic services fails most patients - an argument for the implementation of exome sequencing as a first-tier test in resource-constraint settings. *Orphanet J Rare Dis,* 18**,** 81.

Wilczewski, C. M., Obasohan, J., Paschall, J. E., Zhang, S., Singh, S., Maxwell, G. L., Similuk, M., Wolfsberg, T. G., Turner, C., Biesecker, L. G. & Katz, A. E. 2023. Genotype first: Clinical genomics research through a reverse phenotyping approach. *The American Journal of Human Genetics,* 110**,** 3-12.

Wright, C. F., Campbell, P., Eberhardt, R. Y., Aitken, S., Perrett, D., Brent, S., Danecek, P., Gardner, E. J., Chundru, V. K., Lindsay, S. J., Andrews, K., Hampstead, J., Kaplanis, J., Samocha, K. E., Middleton, A., Foreman, J., Hobson, R. J., Parker, M. J., Martin, H. C., Fitzpatrick, D. R., Hurles, M. E. & Firth, H. V. 2023. Genomic Diagnosis of Rare Pediatric Disease in the United Kingdom and Ireland. *New England Journal of Medicine,* 388**,** 1559-1571.

Wright, C. F., Fitzgerald, T. W., Jones, W. D., Clayton, S., Mcrae, J. F., Van Kogelenberg, M., King, D. A., Ambridge, K., Barrett, D. M., Bayzetinova, T., Bevan, A. P., Bragin, E., Chatzimichali, E. A., Gribble, S., Jones, P., Krishnappa, N., Mason, L. E., Miller, R., Morley, K. I., Parthiban, V., Prigmore, E., Rajan, D., Sifrim, A., Swaminathan, G. J., Tivey, A. R., Middleton, A., Parker, M., Carter, N. P., Barrett, J. C., Hurles, M. E., Fitzpatrick, D. R. & Firth, H. V. 2015. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *The Lancet,* 385**,** 1305-1314.

Wright, C. F., Fitzpatrick, D. R. & Firth, H. V. 2018. Paediatric genomics: diagnosing rare disease in children. *Nature Reviews Genetics,* 19**,** 253-268.

Xiang, J., Ding, Y., Yang, F., Gao, A., Zhang, W., Tang, H., Mao, J., He, Q., Zhang, Q. & Wang, T. 2021. Genetic Analysis of Children With Unexplained Developmental Delay and/or Intellectual Disability by Whole-Exome Sequencing. *Frontiers in Genetics,* 12.

Yaldiz, B., Kucuk, E., Hampstead, J., Hofste, T., Pfundt, R., Corominas Galbany, J., Rinne, T., Yntema, H. G., Hoischen, A., Nelen, M., Gilissen, C., Riess, O., Haack, T. B., Graessner, H., Zurek, B., Ellwanger, K., Ossowski, S., Demidov, G., Sturm, M., Schulze-Hentrich, J. M., Schüle, R., Xu, J., Kessler, C., Wayand, M., Synofzik, M., Wilke, C., Traschütz, A., Schöls, L., Hengel, H., Lerche, H., Kegele, J., Heutink, P., Brunner, H., Scheffer, H., Hoogerbrugge, N., Hoischen, A., Hoen, P. a. C. T., Vissers, L. E. L. M., Gilissen, C., Steyaert, W., Sablauskas, K., De Voer, R. M., Kamsteeg, E.-J., Van De Warrenburg, B., Van Os, N., Te Paske, I., Janssen, E., De Boer, E., Steehouwer, M., Yaldiz, B., Kleefstra, T., Brookes, A. J., Veal, C., Gibson, S., Maddi, V., Mehtarizadeh, M., Riaz, U., Warren, G., Dizjikan, F. Y., Shorter, T., Töpf, A., Straub, V., Bettolo, C. M., Manera, J. D., Hambleton, S., Engelhardt, K., Clayton-Smith, J., Banka, S., Alexander, E., Jackson, A., Faivre, L., Thauvin, C., Vitobello, A., Denommé-Pichon, A.-S., Duffourd, Y., Bruel, A.-L., Peyron, C., Pélissier, A., Beltran, S., Gut, I. G., Laurie, S., Piscia, D., Matalonga, L., Papakonstantinou, A., Bullich, G., Corvo, A., Fernandez-Callejo, M., Hernández, C., Picó, D., Paramonov, I., Lochmüller, H., Gumus, G., Bros-Facer, V., Rath, A., Hanauer, M., Lagorce, D., Hongnat, O., Chahdil, M., Lebreton, E., Stevanin, G., et al. 2023. Twist exome capture allows for lower average sequence coverage in clinical exome sequencing. *Human Genomics,* 17**,** 39.

Yang, X.-A., Hao, H. & Liao, C. 2023. Editorial: Next generation sequencing (NGS) for rare diseases diagnosis - Volume II. *Frontiers in Genetics,* Volume 14 - 2023.

Yang, Y., Muzny, D. M., Reid, J. G., Bainbridge, M. N., Willis, A., Ward, P. A., Braxton, A., Beuten, J., Xia, F., Niu, Z., Hardison, M., Person, R., Bekheirnia, M. R., Leduc, M. S., Kirby, A., Pham, P., Scull, J., Wang, M., Ding, Y., Plon, S. E., Lupski, J. R., Beaudet, A. L., Gibbs, R. A. & Eng, C. M. 2013. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N Engl J Med,* 369**,** 1502-11.

Ye, K., Guo, L., Yang, X., Lamijer, E. W., Raine, K. & Ning, Z. 2018. Split-Read Indel and Structural Variant Calling Using PINDEL. *Methods Mol Biol,* 1833**,** 95-105.

Yilmaz, F., Null, M., Astling, D., Yu, H.-C., Cole, J., Santorico, S. A., Hallgrimsson, B., Manyama, M., Spritz, R. A., Hendricks, A. E. & Shaikh, T. H. 2021. Genome-wide copy number variations in a large cohort of bantu African children. *BMC Medical Genomics,* 14**,** 129.

Zare, F., Dow, M., Monteleone, N., Hosny, A. & Nabavi, S. 2017. An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. *BMC Bioinformatics,* 18**,** 286.

Zarrei, M., Burton, C. L., Engchuan, W., Young, E. J., Higginbotham, E. J., Macdonald, J. R., Trost, B., Chan, A. J. S., Walker, S., Lamoureux, S., Heung, T., Mojarad, B. A., Kellam, B., Paton, T., Faheem, M., Miron, K., Lu, C., Wang, T., Samler, K., Wang, X., Costain, G., Hoang, N., Pellecchia, G., Wei, J., Patel, R. V., Thiruvahindrapuram, B., Roifman, M., Merico, D., Goodale, T., Drmic, I., Speevak, M., Howe, J. L., Yuen, R. K. C., Buchanan, J. A., Vorstman, J. a. S., Marshall, C. R., Wintle, R. F., Rosenberg, D. R., Hanna, G. L., Woodbury-Smith, M., Cytrynbaum, C., Zwaigenbaum, L., Elsabbagh, M., Flanagan, J., Fernandez, B. A., Carter, M. T., Szatmari, P., Roberts, W., Lerch, J., Liu, X., Nicolson, R., Georgiades, S., Weksberg, R., Arnold, P. D., Bassett, A. S., Crosbie, J., Schachar, R., Stavropoulos, D. J., Anagnostou, E. & Scherer, S. W. 2019. A large data resource of genomic copy number variation across neurodevelopmental disorders. *NPJ Genom Med,* 4**,** 26.

Zarrei, M., Macdonald, J. R., Merico, D. & Scherer, S. W. 2015. A copy number variation map of the human genome. *Nature Reviews Genetics,* 16**,** 172-183.

Zhai, Y., Zhang, Z., Shi, P., Martin, D. M. & Kong, X. 2021. Incorporation of exome-based CNV analysis makes trio-WES a more powerful tool for clinical diagnosis in neurodevelopmental disorders: A retrospective study. *Human Mutation,* 42**,** 990-1004.

Zhang, J., Wang, J. & Wu, Y. 2012. An improved approach for accurate and efficient calling of structural variations with low-coverage sequence data. *BMC Bioinformatics,* 13**,** S6.

Zhang, W., Li, D., Pang, N., Jiang, L., Li, B., Ye, F., He, F., Chen, S., Liu, F., Peng, J., Yin, J. & Yin, F. 2022. The second-tier status of fragile X syndrome testing for unexplained intellectual disability/global developmental delay in the era of next-generation sequencing. *Front Pediatr,* 10**,** 911805.

Zhao, L., Liu, H., Yuan, X., Gao, K. & Duan, J. 2020. Comparative study of whole exome sequencing-based copy number variation detection tools. *BMC Bioinformatics,* 21**,** 97.

Zhao, M., Wang, Q., Wang, Q., Jia, P. & Zhao, Z. 2013. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics,* 14**,** S1.

8 Appendices

Appendix I

EGA data access agreement

WELLCOME SANGER INSTITUTE

DDD DATA ACCESS AGREEMENT

These terms and conditions govern access to the managed access datasets (details of which are set out in Appendix I) to which the User Institution has requested access. The User Institution agrees to be bound by these terms and conditions. DDD data should only be used for the analysis of developmental disorders and parental data should only be used to interpret variants in the child.

Applicant

Name of applicant (User), including affiliation and contact details

Name with Title: Prof Zané Lombard

Position: Principal Medical Scientist

Affiliation: University of the Witwatersrand

Institution's Legal Name (if differing from affiliation): Wits Health Consortium (Pty) Ltd

Institutional postal address: University of the Witwatersrand, 147 de Korte street, Johannesburg, South Africa

Institutional E-mail Address: zane.lombard@wits.ac.za

Authorised Representative

Name of authorised representative of the User Institution, including affiliation and contact details:

Name with Title: Mr Alfred Farrell

Position: Chief Executive Officer

Affiliation: Wits Health Consortium

Institutional postal address: Wits Health Consortium, 31 Princess of Wales Terrace, Johannesburg, 2193, Gauteng, South Africa

Institutional E-mail Address: afarrell@witshealth.co.za

Specific limitations on areas of research

Genome Research Limited. Registered Office: 215 Euston Road London NW1 2BE. A company registered in England (No. 2742969) and a charity registered in England (No. 1021457).

DDD data should only be used for the analysis of developmental disorders and parental data should only be used to interpret variants in the child.

APPENDIX I - DATASET DETAILS

Dataset reference

All DDD datasets

Name of project that created the dataset

Deciphering Developmental Disorders

Security Level

We ask that you provide a level of security similar to that described below when storing and using the data you have applied for.

File access: Data should only be accessible to named users. Files should either have only user Unix read/write access, not group or world access, or project specific Unix groups should be used for group access that contain only those names authorised to access the data. User IDs within groups should be reviewed at 6 monthly intervals by the applicant. Data kept on laptops should be encrypted when not in active use, either in individual encrypted files or in encrypted directories/partitions. Data should not be held on USB keys or other portable hard drives. Users may be asked to sign an agreement addressing their responsibilities with respect to access to such data.

Registered Users

Individuals who the User Institution wishes to have access to the Data

If any other individual (e.g. members of your research team) will also require access to the data, please provide their details below. Please note that any individual who is not listed as a Registered User will not be permitted access to the datasets.

Registered User

Name: Nadja Louw

Institutional Email Address: nadja.louw@wits.ac.za

Position: PhD Student

Affiliation: University of the Witwatersrand

APPENDIX II - LAY DESCRIPTION OF YOUR PROJECT

Genome Research Limited. Registered Office: 215 Euston Road London NW1 2BE. A company registered in England (No. 2742969) and a charity registered in England (No. 1021457).

Title of the project

The role of copy number variants in the aetiology of developmental disorders in South Africa - a whole exome sequencing study

Description of the project

The reason we are applying for data access is due to the EGA dataset used in the InDelible paper mainly (EGAS00001000775-https://doi.org/10.1101/2020.10.02.20194241). We aim to identify the most appropriate bioinformatics approach to detect CNVs from exome data. Subsequently, this approach will be implemented in a developmental disorder variant analysis pipeline for WES data generated by the DDD-Africa study, to establish the role that CNVs play in developmental disorders in this African cohort. The DDD-UK data will be used as control samples in order to test the efficacy of the tools before incorporating into our DDD- Africa dataset. Incorporation of computational tools will be carried out using the BAM and/or CRAM files created from the DDD-UK WES data. This dataset will be our truth set as InDelible is one of the tools we will test and consists of a set samples with high confidence CNVs which have been wet-lab validated. The specific objectives will be: 1. To evaluate and compare selected bioinformatics tools for functional equivalence in order to select the best tool or combination of tools for bioinformatics pipelines 2. To implement the chosen bioinformatics approach to evaluate developmental disorder patient data in the DDD-Africa WES dataset 3. To annotate and interpret identified putative disease-causing CNVs. The dataset will also be utilized as truth-set, to evaluate and determine appropriate quality control metric cut-offs for causative variants.

APPENDIX III - PUBLICATION POLICY

WSI are committed to the principles of rapid data release. WSI intend to publish the results of our analysis of this data set and do not consider its deposition into public databases to be the equivalent of such publications. WSI anticipate that the data set could be useful to other qualified researchers for a variety of purposes. However, some areas of work are therefore subject to a publication moratorium.

The publication moratorium covers any publications (including oral communications) that describe the use of the dataset. For research papers, submission for publication should not occur until 12 months after these data were first made available on the relevant hosting database, unless WTSI has provided written consent to earlier submission.

In any publications based on this data, please describe how the data can be accessed, including the name of the hosting database (e.g., The European Genome-phenome Archive at the European Bioinformatics Institute) and its accession numbers (e.g., EGAS00000000029), and acknowledge its use as follows:

"The DDD study presents independent research commissioned by the Health Innovation Challenge Fund [grant number HICF-1009-003], a parallel funding partnership between Wellcome and the Department of Health, and the Wellcome Sanger Institute [grant numberWT098051]. The views expressed in this publication are those of the author(s) and not necessarily those of Wellcome or the Department of Health. The study has UK Research Ethics Committee approval (10/H0305/83, granted by the Cambridge South REC, and GEN/284/12 granted by the Republic of Ireland REC). The research team acknowledges the support of the National Institute for Health Research, through the Comprehensive Clinical Research Network"

EGA Account

Once your application has been approved, an EGA account will be generated for named EGA Account Holder(s) to allow the download of the requested dataset(s) from the EGA. The account which is generated is a private account and login details must not be shared. Sharing Login details will be treated as a breach of security and could result in the account being blocked.

Please provide the name and email address of up to 2 people named in the application (including yourself, if required) who will require an EGA account for the purpose of downloading the data.

Nadja Louw: nadja.louw@wits.ac.za Zane Lombard: zane.lombard@wits.ac.za

Definitions

Genome Research Limited. Registered Office: 215 Euston Road London NW1 2BE. A company registered in England (No. 2742969) and a charity registered in England (No. 1021457).

Authorised Personnel: The individuals at the User Institution to whom WSI grants access to the Data. This includes the User, the individuals listed on the User Institution's initial request for access to the Data and any other individuals for whom the User Institution subsequently requests access to the Data. Details of the initial Authorised Personnel are set out in Appendix I.

Data: The managed access datasets to which the User Institution has requested access.

Data Producers: WSI and the collaborators listed in Appendix I responsible for the development, organisation, and oversight of the Data.

External Collaborator: A collaborator of the User, working for an institution other than the User Institution.

Project: The project for which the User Institution has requested access to the Data. A description of the Project is set out in Appendix II.

Publications: Includes, without limitation, articles published in print journals, electronic journals, reviews, books, posters and other written and verbal presentations of research.

Research Participant: An individual whose data form part of the Data.

Research Purposes: shall mean research that is seeking to advance the understanding of genetics and genomics, including the treatment of disorders, and work on statistical methods that may be applied to such research.

User: The principal investigator for the Project.

User Institution(s): The Institution that has requested access to the Data.

WSI: Genome Research Limited, operating as the Wellcome Sanger Institute

Terms of the Agreement

1. The User Institution agrees to only use the Data for the purpose of the Project (described in Appendix II) and only for Research Purposes. The User Institution further agrees that it will only use the Data for Research Purposes which are within the limitations (if any) set out in Appendix I.

2. The User Institution agrees to preserve, at all times, the confidentiality of the Data. In particular, it undertakes not to use, or attempt to use the Data to compromise or otherwise infringe the confidentiality of information on Research Participants. Without prejudice to the generality of the foregoing, the User Institution agrees to use at least the measures set out in Appendix I to protect the Data.

3. The User Institution agrees to protect the confidentiality of Research Participants in any research papers or publications that they prepare by taking all reasonable care to limit the possibility of identification.

4. The User Institution agrees not to link or combine the Data to other information or archived data available in a way that could re-identify the

Research Participants, even if access to that data has been formally granted to the User Institution or is freely available without restriction.

5. The User Institution agrees only to transfer or disclose the Data, in whole or part, or any material derived from the Data, to the Authorised Personnel. Should the User Institution wish to share the Data with an External Collaborator, the External Collaborator must complete a separate application for access to the Data.

6. The User Institution agrees that the Data Producers, and all other parties involved in the creation, funding or protection of the Data: a) make no warranty or representation, express or implied as to the accuracy, quality or comprehensiveness of the Data; b) exclude to the fullest extent permitted by law all liability for actions, claims, proceedings, demands, losses (including but not limited to loss of profit), costs, awards damages and payments made by the Recipient that may arise (whether directly or indirectly) in any way whatsoever from the Recipient's use of the Data or from the unavailability of, or break in access to, the Data for whatever reason and; c) bear no responsibility for the further analysis or interpretation of these Data.

7. The User Institution agrees to follow the Fort Lauderdale Guidelines (https://www.wtccc.org.uk/wtccc/assets/wtd003207.pdf) and the Toronto Statement (http://www.nature.com/nature/journal/v461/n7261/full/461168a.html). This includes but is not limited to recognising the contribution of the Data Producers and including a proper acknowledgement in all reports or publications resulting from the use of the Data.

8. The User Institution agrees to follow the Publication Policy in Appendix III. This includes respecting the moratorium period for the Data Producers to publish the first peer-reviewed report describing and analysing the Data.

9. The User Institution agrees not to make intellectual property claims on the Data and not to use intellectual property protection in ways that would prevent or block access to, or use of, any element of the Data, or conclusion drawn directly from the Data.

10. The User Institution can elect to perform further research that would add intellectual and resource capital to the data and decide to obtain intellectual property rights on these downstream discoveries. In this case, the User Institution agrees to implement licensing policies that will not obstruct further research and to follow the U.S. National Institutes of Health Best Practices for the Licensing of Genomic Inventions (2005) (https://www.icgc.org/files/daco/NIH_Best PracticesLicensingGenomicInventions_2005_en.pdf) in conformity with the Organisation for Economic Co-operation and Development Guidelines for the Licensing of the Genetic Inventions (2006) (http://www.oecd.org/science/biotech/36198812.pdf).

11. WSI is funded by the Wellcome Trust whose charitable objective is to improve health. If results arising from the User Institution's use of the Data could provide health solutions for the benefit of people in the developing world, the User Institution agrees to offer non-exclusive licenses to such results on a reasonable basis for use in low income and low-middle income countries (as defined by the World Bank) to any party that requests such a license solely for uses within these territories.

12. The User Institution agrees to destroy/discard the Data held, once it is no

longer used for the Project, unless obliged to retain the data for archival purposes in conformity with audit or legal requirements.

13. The User Institution will notify WSI within 30 days of any changes or departures of Authorised Personnel.

14. The User Institution will notify WSI prior to any significant changes to the protocol for the Project.

15. The User Institution will notify WSI as soon as it becomes aware of a breach of the terms or conditions of this agreement.

16. WSI may terminate this agreement by written notice to the User Institution. If this agreement terminates for any reason, the User Institution will be required to destroy any Data held, including copies and backup copies. This clause does not prevent the User Institution from retaining the data for archival purpose in conformity with audit or legal requirements.

17. The User Institution accepts that it may be necessary for the Data Producers to alter the terms of this agreement from time to time. As an example, this may include specific provisions relating to the Data required by Data Producers other than WTSI. In the event that changes are required, the Data Producers or their appointed agent will contact the User Institution to inform it of the changes and the User Institution may elect to accept the changes or terminate the agreement.

18. If requested, the User Institution will allow data security and management documentation to be inspected to verify that it is complying with the terms of this agreement.

19. The User Institution agrees to distribute a copy of these terms to the Authorised Personnel. The User Institution will procure that the Authorised Personnel comply with the terms of this agreement.

20. This agreement (and any dispute, controversy, proceedings or claim of whatever nature arising out of this agreement or its formation) shall be construed, interpreted and governed by the laws of England and Wales and shall be subject to the exclusive jurisdiction of the English courts.

Agreement

I have read and agree to abide by the terms and conditions outlined in the Data Access Agreement.

Yes

Date: 15 March 2022

Genome Research Limited. Registered Office: 215 Euston Road London NW1 2BE. A company registered in England (No. 2742969) and a charity registered in England (No. 1021457).

Appendix II

DDD-Africa Ethics certificates

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

R49 Professors A Krause & Z Lombard; Dr N Carstens; Ms N Louw

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
CLEARANCE CERTIFICATE NO. M230567

NAME: (Principal Investigator)	Professors A Krause & Z Lombard; Dr N Carstens; Ms N Louw
DEPARTMENT:	School of Pathology Division of Human Genetics National Health Laboratory Service
PROJECT TITLE:	*Decyphering Development Disorders in Africa (DDD-Africa) - evaluating clinical exome sequencing in an African setting*
DATE CONSIDERED:	Ad hoc
DECISION:	Approved unconditionally
CONDITIONS:	Renewal of M18/06/78 - expired on 2023/07/05 NIH Study No. UO1MH115483
NOTE:	If contact information regarding student study participants is required, please contact the Registrar's office - <Nicoleen.Potgieter@wits.ac.za>
SUPERVISOR:	Not applicable
APPROVED BY:	Dr M Vorster, Co-Chairperson, HREC (Medical)
DATE OF APPROVAL:	2023/06/20 **EXPIRY DATE:** 2028/06/19

This Clearance Certificate is valid for 5 years from the date of approval. An extension may be applied for.

DECLARATION OF INVESTIGATORS

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

I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated from the research protocol as approved, I/we undertake to submit details to the Committee. **I agree to submit a yearly progress report**. When a funder requires annual re-certification, the application date will be one year after the date when the study was initially reviewed. In this case, the study was initially reviewed in **May** and therefore reports and re-certification will be due in the month of **May** each year. Unreported changes to the study may invalidate the clearance given by the HREC (Medical).

Signature of Principal Investigator — Date: Aug 17, 2023

UNIVERSITY OF THE WITWATERSRAND JOHANNESBURG

R49 Prof. A Krause; Drs N Carstens & Z Lombard; Ms N Louw

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
CLEARANCE CERTIFICATE NO. M180678

NAME: (Principal Investigator)	Prof. A Krause; Drs N Carstens & Z Lombard; Ms N Louw
DEPARTMENT:	School of Pathology Department of Human Genetics National Health Laboratory Service
PROJECT TITLE:	*Decyphering Development Disorders in Africa (DDD-Africa) - evaluating clinical exome sequencing in an African setting*
DATE CONSIDERED:	Ad hoc
DECISION:	Approved unconditionally
CONDITIONS:	Sub-study under M160830 New investigator added 2021/09/06 NIH Study No. UO1MH115483
NOTE:	If contact information regarding student study participants is required, please contact the Registrar's office - <Nicoleen.Potgieter@wits.ac.za>
SUPERVISOR:	Not applicable
APPROVED BY:	Dr CB Penny, Chairperson, HREC (Medical)
DATE OF APPROVAL:	06/07/2018

This Clearance Certificate is valid for 5 years from the date of approval. An extension may be applied for.

DECLARATION OF INVESTIGATORS

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

I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated from the research protocol as approved, I/we undertake to submit details to the Committee. **I agree to submit a yearly progress report**. When a funder requires annual re-certification, the application date will be one year after the date when the study was initially reviewed. In this case, the study was initially reviewed in **«Missing mail merge field»** and therefore reports and re-certification will be due in the month of **«Missing mail merge field»** each year. Unreported changes to the study may invalidate the clearance given by the HREC (Medical).

Signature of Principal Investigator Date

Appendix III

Truth set CNVs summary

Table III.1: Truth set CNV information

SIZE (KB)	TYPE	SEX	CLASSIFICATION
126.26	DEL	Female	Likely benign
542.42	DUP	Female	Likely benign
12.89	DEL	Male	Likely pathogenic
436.99	DEL	Male	Benign
423.19	DUP	Female	Likely benign
504.12	DEL	Female	Likely benign
6808.10	DEL	Female	Likely pathogenic
50.09	DUP	Male	VUS
1516.62	DEL	Female	Likely pathogenic
167.84	DUP	Female	Likely benign
29794.73	DEL	Female	Likely pathogenic
9.82	DUP	Female	Likely pathogenic
2555.27	DEL	Male	Likely pathogenic
51.61	DUP	Male	Likely benign
222.16	DEL	Male	Likely pathogenic
31.67	DEL	Male	Likely pathogenic
1472.70	DUP	Female	Likely pathogenic
1691.65	DEL	Male	Likely benign
76.75	DEL	Female	Likely benign
1.69	DEL	Male	Likely benign
0.44	DUP	Male	Benign
2812.92	DEL	Male	Likely pathogenic
333.80	DUP	Male	Benign
0.32	DEL	Male	Likely pathogenic
51.74	DEL	Male	Likely benign
194.97	DEL	Male	Benign
1349.90	DEL	Female	Likely benign
557.12	DEL	Male	Likely pathogenic
7994.90	DEL	Female	Likely pathogenic
550.30	DEL	Female	Benign
201.51	DEL	Female	Likely benign
194.10	DEL	Male	Benign
1418.36	DEL	Female	Likely pathogenic
9.11	DEL	Female	Likely benign
22.11	DUP	Male	Likely benign
346.71	DEL	Male	Likely benign
4.30	DEL	Female	Likely benign
8.70	DUP	Male	VUS
5376.07	DEL	Female	Likely pathogenic
0.43	DUP	Male	VUS
59.33	DEL	Female	Likely benign
0.51	DEL	Female	Likely benign
186.74	DUP	Female	Likely benign
12.79	DEL	Female	Likely benign
8.11	DEL	Male	Likely pathogenic
171.99	DEL	Male	Likely pathogenic
7802.04	DEL	Female	Likely pathogenic
27.02	DEL	Female	Benign
82.27	DEL	Female	Likely pathogenic
1530.13	DEL	Female	Likely pathogenic
76.85	DUP	Male	Likely benign
14.69	DEL	Female	Likely benign
1561.94	DEL	Male	Likely benign
0.44	DEL	Male	Likely pathogenic
85.80	DUP	Female	VUS
0.87	DUP	Female	Likely benign
1095.51	DEL	Male	Benign
11.55	DEL	Male	Likely pathogenic
142.35	DEL	Female	Likely pathogenic
4964.03	DEL	Female	Likely pathogenic
141.71	DEL	Male	Benign
38.01	DUP	Male	Likely benign
310.66	DUP	Female	Likely benign
1731.46	DEL	Male	Benign
8712.88	DUP	Male	Likely pathogenic
250.71	DUP	Female	Likely benign
286.66	DEL	Male	Benign
186.03	DUP	Female	Likely benign
57.96	DEL	Male	Benign
236.75	DEL	Female	Likely pathogenic
0.54	DEL	Female	VUS
11.82	DEL	Female	Likely pathogenic
9744.38	DEL	Female	Likely pathogenic
0.47	DEL	Female	Likely benign
4.02	DEL	Male	Likely benign
56.60	DEL	Male	Likely benign
2537.68	DEL	Male	Likely pathogenic
49.25	DUP	Male	Likely benign
0.00	DEL	Female	Likely pathogenic
0.44	DUP	Female	Benign
57.48	DEL	Male	VUS
194.10	DEL	Female	Benign
125.87	DEL	Male	VUS
40.91	DEL	Male	Likely benign
601.82	DEL	Female	Likely pathogenic
470.39	DUP	Female	Likely benign
0.44	DUP	Male	Benign
46.63	DUP	Male	Likely pathogenic
4.46	DEL	Male	Likely pathogenic
12230.12	DUP	Female	Likely pathogenic

Appendix IV

InDelible SV output file

Table IV.1: TSV file table of contents for InDelible

Column Name	Column #	Description
chrom	1	Chromosome of breakpoint
position	2	Position of breakpoint
coverage	3	total number of reads covering breakpoint
insertion_context	4	total number of insertions (cigar "I") in reads overlapping this breakpoint
deletion_context	5	total number of deletions (cigar "D") in reads overlapping this breakpoint
sr_total	6	total number of split reads (cigar "S") in reads overlapping this breakpoint
sr_total_long	7	Number of reads with SR length ≥ MINIMUM_LENGTH_SPLIT_READ
sr_total_short	8	Number of reads with SR length < MINIMUM_LENGTH_SPLIT_READ
sr_long_5	9	sr_total_long for 5' end of reads
sr_short_5	10	sr_total_short for 5' end of reads
sr_long_3	11	sr_total_long for 3' end of reads
sr_short_3	12	sr_total_short for 3' end of reads
sr_entropy	13	Sequence entropy of the longest SR sequence given by the formula from Schmitt and Herzel (1997)
context_entropy	14	Sequence entropy of the ±20bp from the breakpoint position
entropy_upstream	15	Sequence entropy of the +20bp from the breakpoint position
entropy_downstream	16	Sequence entropy of the -20bp from the breakpoint position
sr_sw_similarity	17	Smith-Waterman based similarity of split reads from the breakpoint

Column Name	Column #	Description
avg_avg_sr_qual	18	Average sequence quality of split bases
avg_mapq	19	Average mapping quality of reads supporting the breakpoint
seq_longest	20	longest split sequence
pct_double_split	21	Number of reads with both 5' and 3' split reads
prob_N	22	Probability of the breakpoint being a false positive based on the adaptive learning model (1 - prob_Y)
prob_Y	23	Probability of the breakpoint being a true positive based on the adaptive learning model
predicted	24	Is prob_Y > prob_N?
ddg2p	25	Does this breakpoint intersect any genes given by ddg2p_bed file in config.yml
hgnc	26	Does this breakpoint intersect any genes given by hgnc_file in config.yml
hgnc_constrained	27	Does this breakpoint intersect any genes given by hgnc_constrained in config.yml
exonic	28	Does this breakpoint intersect any exons given by ensembl_exons in config.yml
transcripts	29	What transcripts does this breakpoint intersect? If > 10 transcripts, will return 'multiple_transcripts'
maf	30	"Allele Frequency" based on the InDelible database provided with --d
mode	31	How did bwa alignment perform? One of: BLAST_REPEAT (Aligned to a repeat/ME sequence), REALN (Aligned to unique sequence), REALN_CHR (Aligned to unique sequence on another chromosome), REALN_XL (Aligned to unique sequence on the same chromosome, but was flagged as an "improper pair by bwa"), FAIL_ALIGNMENT (split sequence did not align at all), FAIL_LOWMAPQ (split sequence aligned with MAPQ = 0), FAIL_MULTISPLIT (InDelible could not decide whether the sequence was in the 5' or 3' direction), FAIL_REFERENCE (anchoring reference read aligned in the wrong place).
otherside	32	Putative coordinate for alternate breakpoint
svtype	33	Putative SV class. Possible values are DUP (duplication), DEL (deletion), INS (followed by either

Column Name	Column #	Description
		the assembled sequence OR the type of repeat insertion [i.e. Alu, L1, etc.]), CMPLX (followed by DEL/DUP and assembled additional insertion sequence), or TRANSSEGDUP (segmental duplication or translocation). SEGDUP_TRANS represents either a segmenetal duplication or translocation. As we cannot discern with short read data between a SEGDUP or translocation, we list both here.
size	34	Distance to otherside from 'position'
variant_coord	35	Simply 'chrom' : 'position' - 'otherside', where possible
otherside_found	36	Was InDelible sucessful in identifying the other breakpoint?
is_primary	37	If otherside found or blast_hit = "repeats_hit" potential SV type
aln_length	38	Length of the aligned sequence
mum_sr	39	Number of SRs in the bam/cram provided to --m with the same 'position'
dad_sr	40	Number of SRs in the bam/cram provided to --d with the same 'position'
mum_indel_context	41	Number of reads in the bam/cram provided to --m with cigar 'I/D' values
dad_indel_context	42	Number of reads in the bam/cram provided to --d with cigar 'I/D' values
mum_cov	43	Coverage in in the bam/cram provided to --m
dad_cov	44	Coverage in in the bam/cram provided to --d

Appendix V

XHMM CNV output file

Table V.1: Table of contents for XHMM output file

SAMPLE	sample name in which CNV was called
CNV	type of copy number variation (DEL or DUP)
INTERVAL	genomic range of the called CNV
KB	length in kilobases of called CNV
CHR	chromosome name on which CNV falls
MID_BP	the midpoint of the CNV (to have one genomic number for plotting a single point, if desired)
TARGETS	the range of the target indices over which the CNV is called (NOTE: considering only the **FINAL** set of post-filtering targets)
NUM_TARG	# of exome targets of the CNV
Q_EXACT	Phred-scaled quality of the exact CNV event along the entire interval - Identical to *EQ* in .vcf output from genotyping
Q_SOME	Phred-scaled quality of some CNV event in the interval - Identical to *SQ* in .vcf output from genotyping
Q_NON_DIPLOID	Phred-scaled quality of not being diploid, i.e., DEL or DUP event in the interval - Identical to *NDQ* in .vcf output from genotyping
Q_START	Phred-scaled quality of "left" breakpoint of CNV - Identical to *LQ* in .vcf output from genotyping
Q_STOP	Phred-scaled quality of 'right" breakpoint of CNV - Identical to *RQ* in .vcf output from genotyping
MEAN_RD	Mean normalized read depth (z-score) over interval - Identical to *RD* in .vcf output from genotyping
MEAN_ORIG_RD	Mean read depth (# of reads) over interval - Identical to *ORD* in .vcf output from genotyping

Appendix VI

CANOES CNV output file

Table VI.1: Contents of the CANOES output file

Data frame with the following columns:
SAMPLE: name of sample
CNV: DEL of DUP
INTERVAL: CNV coordinates in the form chr:start-stop
KB: length of CNV in kilobases
CHR: chromosome
MID_BP: middle base pair of CNV
TARGETS: target numbers of CNV in the form start..stop
NUM_TARG: how many targets are in the CNV
Q_SOME: a Phred-scaled quality score for the CNV

Appendix VII

CLAMMS CNV output file

List of contents of the CLAMMS output file:

1. chromosome
2. window start coordinate
3. window end coordinate
4. interval (chr:start-end)
5. sample name/id
6. DEL or DUP
7. most likely integer copy number
8. number of windows in the call
9. Q_SOME: Phred-scaled quality of any CNV being in this interval.
10. Q_EXACT: a non-Phred-scaled quality score that measures how closely the coverage profile matches the exact called CNV state and breakpoints. Will document in greater detail later. Any call with Q_EXACT < 0 is of questionable quality.
11. Q_LEFT_EXTEND: Phred-scaled quality of the left breakpoint (based on the likelihood ratio of the stated breakpoint compared to extending the call by 1 window on the left)
12. LEFT_EXTEND_COORD: add this to the CNV start coordinate to get the start coordinate of the first window to the left of the called CNV
13. Q_RIGHT_EXTEND: phred-scaled quality of the right breakpoint (based on the likelihood ratio of the stated breakpoint compared to extending the call by 1 window on the right)
14. RIGHT_EXTEND_COORD: add this to the CNV end coordinate to get the end coordinate of the first window to the right of the called CNV
15. Q_LEFT_CONTRACT: phred-scaled quality of the left breakpoint (based on the likelihood ratio of the stated breakpoint compared to shrinking the call by 1 window on the left)
16. LEFT_CONTRACT_COORD: add this to the CNV start coordinate to get the start coordinate of the second window of the called CNV
17. Q_RIGHT_CONTRACT: phred-scaled quality of the right breakpoint (based on the likelihood ratio of the stated breakpoint compared to shrinking the call by 1 window on the right)
18. RIGHT_CONTRACT_COORD: add this to the CNV end coordinate to get the end coordinate of the second-to-last window of the called CNV

Appendix VIII

Script created to identify overlapping CNVs between CNV tools

```
#!/usr/bin/env python3

import os
import sys

canoes_file = sys.argv[1] #importing canoes cnv file
clamms_file = sys.argv[2]
#xhmm_file = sys.argv[3]

f = open(canoes_file, "r")
g = open(clamms_file, "r")

canoes_list = [] #empty list in order to keep adding to list

for line in f:
    if line.startswith("D3S"):
        line = line.strip().split() #strip removes new line character, split convert line to list
        canoes_list.append(line)

clamms_list = []

for line in g:
    if line.startswith("D3S"):
        line = line.strip().split()
        clamms_list.append(line)

#out_canoes_clamms = []

outlist = []

print('sample_ID_CANOES' + '\t' + 'Interval_CANOES' + '\t' + 'sample_ID_clamms' +
'\t' + 'CNV' + '\t' + 'Interval_CLAMMS')

for i in canoes_list:
    interval_canoes = i[2]
    interval = interval_canoes.split(':')
    chr = interval[0]
    coordinates = interval[1]
    coordinates = coordinates.split('-')
    coordstart = int(coordinates[0])
    coordend = int(coordinates[1])

    out_canoes_clamms = [i[0], interval_canoes]
    #print(out_canoes_clamms)

    for j in clamms_list:
        #out_canoes_clamms = []
        interval_clamms = j[2]
        interval_clamms = interval_clamms.split(':')
        chr_clamms = interval_clamms[0]
```

```
        coord_clamms = interval_clamms[1]
        coord_clamms = coord_clamms.split('-')
        coordstart_clamms = int(coord_clamms[0])
        coordend_clamms = int(coord_clamms[1])

        if chr_clamms != chr:
            pass
        elif coordend_clamms - coordstart_clamms <= 500:
            if  coordstart_clamms in range(coordstart - 50, coordstart + 50) and
coordend_clamms in range(coordend - 50, coordend + 50):
                out_canoes_clamms.append(j[:3])
                # out_canoes_clamms.append(i[0])
                # out_canoes_clamms.append(interval_canoes)
        elif coordend_clamms - coordstart_clamms <= 1000:
            if  coordstart_clamms in range(coordstart - 200, coordstart + 200) and
coordend_clamms in range(coordend -200, coordend + 200):
                out_canoes_clamms.append(j[:3])
                # out_canoes_clamms.append(i[0])
                # out_canoes_clamms.append(interval_canoes)
        elif coordend_clamms - coordstart_clamms <= 10000:
            if  coordstart_clamms in range(coordstart - 800, coordstart + 800) and
coordend_clamms in range(coordend - 800, coordend + 800):
                out_canoes_clamms.append(j[:3])
                # out_canoes_clamms.append(i[0])
                # out_canoes_clamms.append(interval_canoes)
        elif coordend_clamms - coordstart_clamms <= 50000:
            if  coordstart_clamms in range(coordstart - 3000, coordstart + 3000) and
coordend_clamms in range(coordend - 3000, coordend + 3000):
                out_canoes_clamms.append(j[:3])
                # out_canoes_clamms.append(i[0])
                # out_canoes_clamms.append(interval_canoes)
        elif coordend_clamms - coordstart_clamms <= 250000:
            if  coordstart_clamms in range(coordstart - 10000, coordstart + 10000) and
coordend_clamms in range(coordend - 10000, coordend + 10000):
                out_canoes_clamms.append(j[:3])
                # out_canoes_clamms.append(i[0])
                # out_canoes_clamms.append(interval_canoes)
        elif coordend_clamms - coordstart_clamms <= 500000:
            if  coordstart_clamms in range(coordstart - 20000, coordstart + 20000) and
coordend_clamms in range(coordend - 20000, coordend + 20000):
                out_canoes_clamms.append(j[:3])
                # out_canoes_clamms.append(i[0])
                # out_canoes_clamms.append(interval_canoes)
        elif coordend_clamms - coordstart_clamms <=1000000:
            if  coordstart_clamms in range(coordstart - 40000, coordstart + 40000) and
coordend_clamms in range(coordend - 40000, coordend + 40000):
                out_canoes_clamms.append(j[:3])
                # out_canoes_clamms.append(i[0])
                # out_canoes_clamms.append(interval_canoes)
        elif coordend_clamms - coordstart_clamms <= 5000000:
```

```
            if  coordstart_clamms in range(coordstart - 50000, coordstart + 50000) and
coordend_clamms in range(coordend - 50000, coordend + 50000):
                out_canoes_clamms.append(j[:3])
                # out_canoes_clamms.append(i[0])
                # out_canoes_clamms.append(interval_canoes)
        elif coordend_clamms - coordstart_clamms > 5000000:
            if  coordstart_clamms in range(coordstart - 100000, coordstart + 100000) and
coordend_clamms in range(coordend -100000, coordend + 100000):
                out_canoes_clamms.append(j[:3])
                # out_canoes_clamms.append(i[0])
                # out_canoes_clamms.append(interval_canoes)
        else :
            pass
    #print(out_canoes_clamms)
    if len (out_canoes_clamms) == 3:
        final = []
        final.append (out_canoes_clamms [0])
        final.append (out_canoes_clamms [1])
        final.append (out_canoes_clamms [2] [0])
        final.append (out_canoes_clamms [2] [1])
        final.append (out_canoes_clamms [2] [2])
        print ("\t".join (final))
```

Appendix IX

Additional information for patients with LP/P CNVs identified

Patient D3S_0009_01_1

This patient is a 5 year old female (at time of recruitment) and presented with moderate ID/DD, dysmorphic features, a patent ductus arterium defect, pectus excavatum and strabismus. She was included within category A and C due to the severity of ID/DD and major as well as minor malformations. This deletion on chromosome 16 [GRCh38/hg38] 16q22.1q22.3(69245485_72960252)x1 spans 3,71Mb and covers region 16q22.1-22.3. Both parents were also subjected to ES and this deletion is not present thus this is a *de novo* variant. There are 77 genes involved including *AP1G1* gene associated with Usmani-Riazzudin syndrome SNVs. This CNV was validated by CMA which also concluded that this is a pathogenic duplication.

Patient D3S_0040_01_1

This 4 year, 5 month old female at recruitment presented with global developmental delay, dysmorphic, coarse facial features, minor skin pigmentary changes and an umbilical hernia, macroglossia, broad hallux, protruding ear. She was Included for severe ID/DD within category A. This large deletion [GRCh38/hg38] 6q25.2q25.3(153282150_157897057)x1 spans 4.61MB and 34 genes which included the *ARID1B* gene associated with Coffin-Siris. Coffin-Siris is inherited in an autosomal dominant manner and cases usually present with mild to severe DD/ID, coarse facial features, microcephaly, feeding difficulties and hypoplasia. This is a *de novo* CNV as both the parents did not present with the deletion and was also confirmed with CMA.

Patient D3S_0055_01_1

This patient is a 3 year 8 month old female who presented with Brain atrophy, microcephaly, dysmorphic features, epicanthus, sparse lateral eyebrow, depressed nasal bridge, micrognathia, long philtrum, midface retrusion, short nose, narrow forehead, periorbital fullness. She was included within category A, for moderate ID/DD. A duo (mother and child) was recruited and this CNV is not present in the mother. It might thus be a *de novo* CNV although we will be unable to confirm this until the father is also tested. This 4,82Mb deletion on chromosome 8, [GRCh38/hg38] 8p21.2p12(26354309_31173210)x1 includes 80 genes.

Patient D3S_0070_01_1

This 8 year old male was recruited within inclusion category A and presented with microcephaly, downslanted palpebral fissures, cryptorchidism, short toe, proximal placement of thumb, cataract, aniridia, corneal opacity. This is a *de novo* 8,65Mb deletion [GRCh38/hg38] 11p14.1p13(30010610_33142168)x1 (8645kb) which spans 67 genes including *PAX6* and *WT1*. The *PAX6* gene is associated with WAGR 11p13 deletion syndrome which seems to fit the patient's phenotype. The main phenotypic features of this syndrome is aniridia, Wilms tumour, genitourinary abnormalities and developmental delay. The *WT1* gene is associated with Wilms tumour which is also inherited in an autosomal dominant manner. This CNV was confirmed with CMA.

Patient D3S_0085_01_1

This 5 year old male was included for moderate to severe ID/DD within category A. This patient presented with microcephaly, bilateral hearing loss, hypotonia and dysmorphic features. A total of 61 genes are included in this 4,39Mb *de novo* deletion [GRCh38/hg38] 12p13.33p13.31(1262908_5647947)x1 which also includes twelve dosage sensitive genes. This CNV includes the 12p13.33 microdeletion region, including *ELKS/ERC1* genes, associated with childhood apraxia of speech (Thevenon et al., 2013) and was also confirmed with CMA.

Patient D3S_0101_01_1

This 11 year old female was included in category A and presented with failure to thrive, dysmorphic features. A *de novo* deletion spanning 4,93Mb, [GRCh38/hg38] 17q22q23.2(53822906_58734380)x1, has been identified and classified as pathogenic (2.05). Sixty four genes are involved in this deletion, including forty protein coding genes as well as the *NOG* gene which is confirmed as haploinsufficient. Many different phenotypes are involved and has been termed NOG-Related-Symphalangism Spectrum Disorder (NOG-SSD) (Potti et al., 2011). A confirmatory CMA was completed.

Patient D3S_0109_01_1

This female patient was recruited within category A at 5 years of age presenting with seizures, cortical blindness, hypotonia, coarse facial features and dysmorphic features. A duplication of 9,08Mb has been identified on chromosome 2 [GRCh38/hg38] 2p21p16.1(41947349_51028389)x3, involving 104 genes. This CNV was not identified in the mother, but as the father has not been recruited, it cannot be confirmed as *de novo*. Similar duplications with overlapping regions have been uploaded to DECIPHER and classified as LP/P. One TS gene, *FBXO11* is included in this duplication which is involved in intellectual developmental disorder with dysmorphic facies and behavioural abnormalities. Most of the variants within the *FBXO11* gene are *de novo* deletions and there is no clear evidence or functional studies completed to show what effect a duplication will have on this gene (Fritzen et al., 2018, Gregor et al., 2018, Jansen et al., 2019). However, another smaller duplication was also identified, [GRCh38/hg38]2p16.2p16.1(53670417_55322795)x3, thus combining these duplications could suggest that a larger region is duplicated (~13.38Mb). This does not affect the CNV classification as it still stays likely pathogenic even though 37 additional genes are included in the entire region to a total of 141 duplicated genes.

Patient D3S_0114_01_1

This 6 year old female was included within inclusion category D and presented with microcephaly (postnatal onset), failure to thrive, strabismus, brisk reflexes and dysmorphic features. The CNV identified, spans 6,25Mb and includes 56 genes, [GRCh38/hg38] 12p12.1p11.22(23534153_29783848)x1. It was classified as pathogenic (1.35) and includes the *SOX5* gene associated with Lamb-Shaffer syndrome. This CNV is not present in the mother and cannot be confirmed as *de novo* since the father was not recruited. A confirmatory CMA was completed.

Appendix X

Website Links

DDD-Africa project page (https://h3africa.org/index.php/ddd-africa/)

European Genome-phenome Archive (EGA) (https://ega-archive.org/)

Nextflow workflow (https://github.com/phelelani/nf-exomecnv)

DDG2P gene list (https://panelapp.genomicsengland.co.uk/panels/484/)

InDelible Github (https://github.com/HurlesGroupSanger/indelible)

XHMM Github (https://github.com/RRafiee/XHMM)

CANOES Github (https://github.com/ShenLab/CANOES)

CLAMMS Github (https://github.com/rgcgithub/clamms)

ClinGen CNV pathogenicity calculator (https://cnvcalc.clinicalgenome.org/cnvcalc/)

Clinical Genome (ClinGen) Resource (https://www.genome.gov/Funded-Programs-Projects/ClinGen-Clinical-Genome-Resource)

Database of Genomic Variants (DGV) (https://dgv.tcag.ca/dgv/app/home)

ClinVar Database (https://www.ncbi.nlm.nih.gov/clinvar/)

Genome Aggregation Database (gnomAD) (https://gnomad.broadinstitute.org/)

CNV-ClinViewer (https://cnv-clinviewer.broadinstitute.org/)

Appendix XI

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 Nadja Louw (Student number: 2153705) am a student registered for the degree of PhD Human Genetics in the academic year 2024.

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: 12 May 2025

Appendix XII

Turnitin report

Turnitin Originality Report

Processed on: 11-Dec-2024 10:08 AM SAST
ID: 2548893370
Word Count: 38617
Submitted: 1

N.Louw (2153705) Final Thesis .docx By Nadja Louw

Although a 28% match is seen, a total of 17% can be explained by matches to my published mini review (13% in red below) and a further 4% (orange) relates to the output file format as given in the Appendix for tools InDelible and CLAMMS.

Similarity Index	Similarity by Source	
28%	Internet Sources:	25%
	Publications:	20%
	Student Papers:	2%

Matches relating to my own publication

6% match (Internet from 20-Mar-2024)
https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2023.1277784/pdf?isPublishedV2=false

5% match ()
Nadja Louw, Nadia Carstens, Zané Lombard, . "Incorporating CNV analysis improves the yield of exome sequencing for rare monogenic disorders—an important consideration for resource-constrained settings", Frontiers in Genetics

Matches relating to Appendix III and VI showing the output file format

3% match (Internet from 26-Dec-2022)
https://github.com/HurlesGroupSanger/indelible

1% match (Internet from 29-Dec-2023)
https://www.frontiersin.org/articles/10.3389/fgene.2023.1277784/full

1% match (Internet from 12-Jan-2023)
https://github.com/rgcgithub/clamms

1% match (Internet from 06-Apr-2024)
https://www.wits.ac.za/media/wits-university/faculties-and-schools/health-sciences/documents/ABSTRACTS.%20MOLECULAR%20AND%20COMPARATIVE%20BIOSCIENCES.pdf

< 1% match (Internet from 15-Mar-2024)
https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2021.738561/pdf?isPublishedV2=false

< 1% match (Internet from 20-Mar-2024)
https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2023.1277784/epub?isPublishedV2=false

< 1% match (Internet from 20-Mar-2024)
https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2023.1137922/pdf?isPublishedV2=false

< 1% match (Internet from 13-Dec-2023)
https://www.frontiersin.org/articles/10.3389/fgene.2023.1277948/full

< 1% match (Internet from 20-Mar-2024)
https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2022.940292/pdf?isPublishedV2=false

< 1% match (Internet from 01-Oct-2022)
https://www.frontiersin.org/articles/10.3389/fgene.2022.933381/full

< 1% match (Internet from 12-Oct-2023)
https://www.frontiersin.org/articles/10.3389/fgene.2023.1137922/full

< 1% match (Internet from 13-Dec-2022)
https://www.frontiersin.org/articles/10.3389/fped.2021.526779/full

< 1% match (Internet from 07-Jul-2024)
https://www.frontiersin.org/journals/bioinformatics/articles/10.3389/fbinf.2022.969247/full