Evaluating whole exome sequencing on the Ion Torrent S5™ as a potential diagnostic tool for developmental disorders

Kaitlyn Ashley Flynn

Student Number: 1035820

Supervisor: Prof Zané Lombard (PhD)

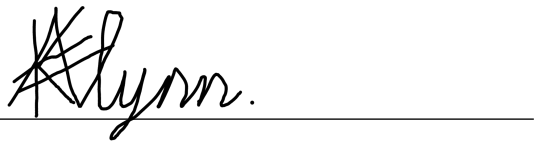
Co-Supervisor: Dr Lindie Lamola (PhD)

A Dissertation submitted to the Faculty of Health Science, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Master of Science in Medicine

Johannesburg, 2020

Declaration

I, Kaitlyn Ashley Flynn, declare that this Dissertation is my own, unaided work. It is being submitted for the Degree of Master of Science in Medicine at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.



(Signature of candidate)

28th day of October 20 20 in Bedfordview

Abstract

Developmental disorders (DD) are a group of heterogeneous disorders characterised by delayed physical and mental development. It is thought that 25-50% of DD have a genetic aetiology, yet a majority of patients remain undiagnosed through current testing procedures. In South Africa, chromosomal karyotyping (diagnostic rate of 3%) and MLPA (diagnostic rate of 5-10%) are the routine diagnostic testing approach for patients presenting with DD. Globally, experts are recommending WES as a first-line diagnostic test for DD, mainly due to its superior diagnostic rate (±40%), and the fact that it outperforms CMA (current gold-standard) for evaluation of unexplained DD. However, the utility of this test has not been investigated in a limited resource setting. This study aimed to implement a WES pipeline and investigate the diagnostic rate, clinical utility, and associated financial and computational costs, using the Thermo Fisher Ion Torrent S5™ NGS instrument. This was done by performing WES on seven DD patients and one sequencing control DNA sample. Data analysis was performed across three investigative platforms namely Moon (Diploid); an in-house, bespoke bioinformatics pipeline, and VarAFT; and all were complemented with manual analysis to identify possible putative disease-causing mutations. Identified variants were validated with Sanger sequencing. Of the seven patients tested we identified disease-causing mutations in six individuals and validated said variants in five individuals. We determined that WES of a single patient cost approximately ZAR13,000.00 (laboratory cost and data analysis). The diagnostic rate was greater than 70% for this project and significant clinical interventions were identified for most patients. We refined a bioinformatic workflow for the effective analysis of WES data to identify putative disease-causing mutations in South African DD patients and concluded that WES would be a highly beneficial tool in the investigation of DD within the local context.

Acknowledgments

I would like to thank my supervisors Prof Zané Lombard and Dr Lindie Lamola for their unwavering support and guidance throughout this project and for providing me with the opportunity to conduct this project for my Master's degree. I would also like to thank Dr Candice Feben and Dr Amanda Krause for their invaluable clinical insight during this project. A special thanks to Mr Shaun Aron for his help with our bioinformatics pipeline set-up and additional training. Thank you to the Thermo Fisher technicians Mmatshepho Pasha-Muchemenye and Burger van Jaarsveld for the NGS training. Thank you to Miss Odirile Tabane and Miss Jessica Levesley for their assistance with Sanger sequencing. I would also like to thank the funders of this project, without whom this project would not have been possible. Finally, I would like to thank my mother, Samantha Flynn, and the rest of my family, for their continuing love, support, and utmost confidence in my abilities.

Funders:

H3Africa NIMH/NIH: U01MH115483

Medical Research Council (MRC)

National Research Foundation (NRF)

National Health Laboratory Service Research Trust (NHLSRT)

Faculty of Health Science Research Council (FRC)

Table of Contents

List of Abbreviations and Symbols

A, C, G, T: Adenine, Cytosine, Guanine, Thymine

ACMG: American College of Medical Genetics

AGVP: African Genome Variation Project

BAM: Binary Alignment Map

cDNA: complementary Deoxyribonucleic Acid

CMA: Chromosomal Microarray

CNV(s): Copy Number Variant(s)

CT: Computerised tomography

dbSNP: Single Nucleotide Polymorphism Database

DD: Developmental Disorders

DDD: Deciphering Developmental Delay

DNA: Deoxyribonucleic Acid

dNTP: deoxyribonucleotide triphosphate

ExAC: Exome Aggregation Consortium

FISH: Fluorescence in situ Hybridization

FRAX: Fragile X

gDNA: genomic Deoxyribonucleic Acid

GnomAD: The Genome Aggregation Database

GRCh37: Genome Reference Consortium Human Build 37

HGVS: Human Genome Variation Society

HIV: Human Immunodeficiency Virus

HPO: Human Phenotype Ontology

HREC: Human Research Ethics Committee

IRGV: Ion Reporter Genomic Viewer

MAF: Minor Allele Frequency

MLPA: Multiplex Ligation-dependant Probe Amplification

NCBI: National Center for Biotechnology Information

NEB: New England Biolabs

NGS: Next Generation Sequencing

NHLS: National Health Laboratory Service

PCR: Polymerase Chain Reaction

PolyPhen-2: Polymorphism Phenotyping version 2

QC: Quality Control

SIFT: Sorting Intolerant from Tolerant

SNP: Single Nucleotide Polymorphism

SNV: Single Nucleotide Variant

TCO: Total Cost of Ownership

UCSC: University of California Santa Cruz

UK: United Kingdom

USA: United States of America

US$: United States Dollar

VarAFT: Variant Annotation and Filter Tool

VCF: Variant Call Format

VEP: Variant Effect Predictor

VUS: Variant of Unknown Significance

WES: Whole Exome Sequencing

WGS: Whole Genome Sequencing

ZA: South Africa

ZAR: South African Rand

List of Figures

List of Tables

Preface

This project is a sub-study of the larger Deciphering Development Disorders in Africa (DDD-Africa) project currently ongoing in the Division of Human Genetics, University of the Witwatersrand and the National Health Laboratory Service (NHLS) Braamfontein. This Division performs both research and diagnostic genetic services in South Africa. Individuals from both research (senior researchers, lecturers, and postgraduate students) and diagnostics (medical geneticists, clinicians, and genetic counsellors) are involved in this project. DDD-Africa is a research project that aims to better understand the genetic contributions to developmental disorders in Africa and effectively implement whole exome sequencing in a diagnostic setting to allow for better precision public health in Africa. The partners of this project include collaborators from the UK (the Wellcome Trust Sanger Institute and DDD-UK) and the Democratic Republic of Congo (University of Kinshasa); as well as other contributing institutions namely H3ABioNet, Wits Health Consortium, and the National Institute of Biomedical Research (INRB).

1. Introduction

Quick, accurate, and affordable molecular testing is necessary for the better diagnosis and management of inherited diseases. Next generation sequencing (NGS), also known as massively parallel sequencing, stands out as a comprehensive and effective tool with significant diagnostic potential. The rapid evolution of NGS has stimulated conversation on how impactful this technology may be at increasing the diagnostic yield and clinical utility of the testing options for South African patients. Whole exome sequencing (WES, the unbiased analysis of all protein-coding sequences (Fogel et al., 2016)) has been increasingly reported over the past few years as a possible candidate for testing of Mendelian disorders (Need et al., 2012) and has been suggested as the first-tier testing method for neurodevelopmental disorders (Srivastava et al., 2019). The use of this technology may shine a light on genetic variants that may be responsible for the symptoms found in previously undiagnosed patients (Wright et al., 2015). In addition, it has been shown that WES has an increased diagnostic rate (25-40%) across multiple Mendelian disorders (Córdoba et al., 2018; Need et al., 2012; Wright et al., 2015) when compared to traditional testing methods such as karyotyping and Sanger sequencing, that may end the diagnostic odyssey some patients have found themselves trapped in. In knowing how useful this technology may prove to be, it is necessary to have an estimate of the costing, timing, and workflow necessary to make this an effective diagnostic option in a local, often resource-poor setting.

A significant number of patients who do not currently receive a diagnosis through locally available testing are those affected with developmental disorders (DD). Developmental disorders encompass a wide range of genetic diseases (Miclea et al., 2015), however, the specific phenotypes expressed in these patients are usually typical of genetic heterogeneity (Soden et al., 2014), making the diagnosis of these patients highly taxing. The use of WES may allow us the possible identification of the causal variant that went undetected in previous, targeted diagnostic tests performed on these patients (Ewans et al., 2018).

1.1 Developmental disorders

Developmental disorders are characterised by abnormal and/or delayed mental or physical development (Boyle et al., 2011). These developmental impairments typically result in severe and chronic disabilities that impact a patient's daily functioning. A

common symptom of many patients with DD is intellectual disability which is suggested to have a prevalence of roughly 3% worldwide (Shevell et al., 2003). The prevalence in South Africa is not known; however, limited data suggest it matched the global statistics (Fieggen et al., 2019). Intellectual disability is used when referring to children over the age of five and is not to be confused with global developmental delay, used when referring to children under the age of five. Global developmental delay is often identified as a notable delay in two or more developmental domains (Srour and Shevell, 2015).

Developmental disorders arise from a wide range of both genetic and environmental elements but it is estimated that underlying genetic causes are a major contributor causing 25-50% of cases (Gilissen et al., 2014). Environmental causes contributing to the remainder of cases include infections and exposure to environmental and chemical toxins; malnutrition and nutritional deficiencies; perinatal and neonatal factors; and poverty and trauma. Although it is understood that the majority of cases have an underlying genetic cause, many of these cases are overlooked in low-to-middle income countries based on the prevalence of environmental impact. Due to low income and limited resources, much of Africa is still plagued by infectious diseases and malnutrition. For example, in 2019 it was reported the 20.7 million individuals infected with Human Immunodeficiency Virus (HIV) resided in east and southern Africa. This accounts for 54% of the world's HIV infected population (UNAIDS 2020). HIV is a known cause of DD and is particularly noted as an issue in areas where it affects a large proportion of childbearing women who have limited access to antiretroviral therapies, caesarean deliveries and alternatives to breast-feeding (Institute of Medicine (US) Committee on Nervous System Disorders in Developing Countries 2001).

Apart from the large number of environmental impacts leading to DD in the African population, the genetic causes of DD are overlooked based on their difficulty to be diagnosed, and as such a majority of cases go undiagnosed. The main reason for this difficulty can be attributed to the heterogeneity of these disorders both clinically and genetically (Soden et al., 2014). Genetic heterogeneity occurs when mutations at two or more distinct genetic loci result in the same or similar phenotypes. This increases the complexity of genetic testing because medical scientists are required to test multiple genes or loci in order to identify possible disease-causing mutations. This

direct/targeted testing is only possible when clinicians can narrow down a patient's phenotype in order to identify a likely syndrome.

The majority of DD show overlapping features and phenotypes along with clinical heterogeneity. Only a few syndromes have well-defined features and/or unique characteristics to drive molecular investigations (Larizza and Finelli 2018). Clinical heterogeneity can be understood as diverse phenotypic representations of the same disorder/syndrome (Ball et al., 2016). Due to this heterogeneity, it is always beneficial to rule out the most prevalent genetic aetiologies of DD (where relevant) to avoid unnecessary treatments, testing, and intervention and narrow down possible diagnoses.

1.2 Prevalent genetic aetiologies of Developmental Disorders

Chromosomal abnormalities are responsible for the majority (25%) of DD cases with a genetic aetiology, with trisomy 21 being the most common (Miclea et al., 2015). Trisomy 21, also known as Down Syndrome, is caused by the presence of an additional chromosome 21 (Hayes and Batshaw 1993). Additional chromosomal abnormalities such as trisomy 13 and 18, microdeletions, and partial duplications are also known to be associated with DD (Miclea et al., 2015). Chromosomal anomalies are tested for using karyotyping and multiplex ligation probe-dependent amplification (MLPA).

Monogenic causes of DD are responsible for 10% of cases, where Fragile X syndrome (the most common associated disorder) contributes 5% of DD cases (Miclea et al., 2015). While these statistics and prevalence rates are said to be a 'global' representation, it is important to note that the figures are based on European data and there is no clear representation of these figures in Africans. Fragile X syndrome (FRAX) has a prevalence of 1 in 5000 males affected and 1 in 8000 females affected (Rajaratnam et al., 2017). FRAX is caused by mutations in the *FMR1* gene which generates a protein responsible for synapse development (Essop and Krause, 2013). Due to the high prevalence of this syndrome and the availability of a robust testing method for FRAX, it is necessary to implement this as a first-line test when a patient presents with the associated phenotype. Other monogenic causes of DD which have well-defined phenotypes can be tested for using a range of single-gene tests which

are typically requested by the referring clinician. In some cases of known genetic heterogeneity, multi-gene panels may be available for testing.

1.3 Diagnostic testing for Developmental Disorders

Until recently, the recommended first-tier genetic test, globally, for patients presenting with unexplained DD was a chromosomal microarray (CMA) (Kharbanda et al., 2015). This testing method identifies changes in copy number variants (CNVs) and its high sensitivity allows the detection of aberrations at a submicroscopic level. This method reportedly allows the detection of changes spanning only 5Kb in size indicating a resolution that is 1000 times higher than traditional karyotyping (Batzir et al., 2015). The use of microarray to replace the original routine testing method of karyotyping resulted in an increase in the diagnostic rate by approximately 10% (Miller et al., 2010).

In South Africa, microarray is not currently routinely offered as a diagnostic test to state patients. The Division of Human Genetics (University of the Witwatersrand and National Health Laboratory Service), is one of the major service providers of clinical and laboratory-based genetic tests in South Africa and has been conducting research into DD for the past 35 years (Essop and Krause, 2013). Here, the current standard of testing for patients with DD involves karyotyping followed by MLPA analysis. Approximately 20% of patients receive a positive diagnosis from the above testing methods (personal communication from this laboratory). Many patients unfortunately remain undiagnosed after the first round of testing, and due to limited resources, the possibility of additional tests is out-of-reach for most, leaving them with no information or options for treatment and management of their disease.

1.4 Evaluating Whole Exome Sequencing as a diagnostic test for Developmental Disorders

The Deciphering Developmental Disorders study (DDD-UK) is a multidisciplinary research project with the aim of advancing clinical genetic practice for children with DD in the United Kingdom. This study began in 2011 and comprehensively investigated the clinical, ethical, and genomic components of DD. It aimed to minimise incidental findings and effectively showed the benefits of incorporating new genetic technologies into the research of DD. The study evaluated the effectiveness of WES as a diagnostic tool and achieved a 27% diagnostic yield for the 1,133 patients tested

(Wright et al., 2015). This group's more recent publication showed that the continual generation of new genetic and disease information allows for the reanalysis of previously analysed data. Through the reanalysis of the original 1,133 patients tested, the overall diagnostic yield of the study rose to 40%, with an additional 182 individuals receiving a positive diagnosis (Wright et al., 2018).

Other groups have emerged to investigate this technology as a means of effectively diagnosing individuals with DD; one of these is the NeuroDev study. The aim of their study involves the genetic characterisation of neurodevelopmental disorders in sub-Saharan Africa. To achieve this, they plan to perform WES on 1800 individuals with neurodevelopmental disorders from South Africa and Kenya as well as 1800 matched controls and 1900 parental samples. This project is a collaboration between the Broad Institute in Cambridge Massachusetts; the University of Cape Town and the KEMRI-Wellcome Trust Research Programme (de Menil et al., 2019).

In support of the research being seen to date showing the use of WES as a potentially powerful tool in the diagnosis of DD, the consensus statement released by Srivastava et al., suggested WES be the preferred first-tier testing method in the diagnosis of DD as the method was shown to outperform CMA (2019).

1.5 Next Generation Sequencing Platforms

WES is a genomic sequencing technique used to sequence the protein-coding regions of a gene which is known as the exome (Arts et al., 2019). This method falls under the umbrella term of NGS alongside whole genome sequencing. The ability to sequence multiple samples in parallel and generate mass amounts of data in a relatively quick time-frame makes this technology highly desirable in a diagnostic and research setting (Goodwin et al., 2016). With the demand for the technology steadily rising over the past ten years there has been a dramatic expansion in the NGS market, with multiple suppliers offering a wide range of instruments and software that is constantly being upgraded and improved to match their competitors (Pereira et al., 2020).

The most commonly referred to sequencing platforms are supplied by Illumina (USA) and Ion Torrent (Thermo Fisher, USA). These two companies have designed unique platforms using different chemistries. However, both work on the same second-generation sequencing workflow of library preparation, amplification and clonal

expansion, sequencing, and analysis (Pereira et al., 2020). The chemistry of each company differs in that Illumina systems makes use of sequencing-by-synthesis (SBS) and Ion Torrent systems uses semiconductor sequencing (Bentley et al., 2008; Lahens et al., 2017). Many studies have compared the two systems through various applications and reported on the advantages and disadvantages. Due to rapid system updates and technological improvements, both are constantly upgrading and reporting better/newer features. This MSc project made use of Ion Torrent™ systems and technologies, due to the availability of the platform within the department. Therefore, the use of this platform will be discussed throughout and compared to other known systems later on.

1.6 Ion Torrent S5™ system

The Ion Torrent system sequences DNA using a metal-oxide-semiconductor chip (Merriman et al., 2012). Initially, DNA is fragmented, and each fragment is bound to a microbead. This fragment then replicates until the entire microbead is covered in the DNA fragment. Each bead then fills a microwell situated on the semiconductor chip. These microwells are flooded with a single deoxynucleoside triphosphate (dNTP) solution at a time (known as a flow) and incorporation of a nucleotide is noted by the release of a hydrogen ion (H^+). A built-in pH meter (ion sensor) is used to detect pH changes and convert these chemical signals into digital voltage signals. This direct sequencing approach allows for ‘real-time’ base calling resulting in rapid sequencing technology (Ambardar et al., 2016).

Limitations of the technology have been reported and include false positives; specifically, the calling of indels (insertion or deletion of bases in the genome) in homopolymer regions (Damiati, et al., 2016). A homopolymer occurs when one type of chemical unit or monomer (in this case a nitrogenous base A, C, G, or T) is repeated multiple times in succession. When sequencing these repeat regions, the same base is incorporated multiple times which releases a high hydrogen ion concentration during a single flow. A higher voltage will be recorded indicating the incorporation of more than one nucleotide. When long homopolymers are encountered the quantification of the voltage change is typically inaccurate due to the inefficiency of the system leading to the high error rate of indels being called within these regions (Quail et al., 2012).

These limitations need to be kept in mind when evaluating the system as well as other considerations associated with implementing new technologies.

1.7 Considerations in evaluating WES as a diagnostic test

1.7.1 The cost implications of WES

Whilst the initial investment in sequencing infrastructure and reagents is quite high, the price per sample sequencing has decreased dramatically over the years. Commercial service providers have offered promotional sequencing at costs of US$250 in the past ("BGI- Exome Seq Fall Pricing Promotion," 2015., Steenhuysen 2015). A systematic review conducted by Schwarze et al. looked at the costs of WES listed in current literature. The review investigated 18 papers that had WES costs listed. These costs ranged from US$555 to US$5169 for a single WES test (Schwarze et al., 2018). A more recent review by Suwinski et al., reports the average cost of exome generation to be US$400 but the addition of data analysis and clinical interpretation typically results in costs greater than US$1000 (Suwinski et al., 2019).

The implementation of WES would involve costs outside of the test itself as well. A study conducted by Monroe et al. suggested that the use of WES, as opposed to the traditional course of diagnostic tests, results in significantly lower costs. The difference was substantial enough, that WES was suggested as a cheaper first-line genetic test overall. In addition, the implementation of WES would not only reduce the cost of genetic testing but also the number of unnecessary non-genetic tests patients undergo; thus, reducing the overall health-care costs of some patients by 50% (Monroe et al., 2016).

Although the price has decreased and appears more feasible, it is necessary to explore the cost-effectiveness of such a test in a resource-limited. Understanding the costs associated with doing these tests 'in-house' would enable laboratories to determine whether or not it was feasible to set up the infrastructure locally or more cost-effective to outsource the testing. Local infrastructure has added benefits of knowledge generation, skill development, student training, job creation, and increasing local resources. However, due to the country's low economic status, this can only be implemented if proved to be cost-effective.

1.7.2 Data storage, processing and analysis challenges

The ability to sequence multiple samples in parallel and generate mass amounts of data is highly useful in a diagnostic and research setting. However, the consequence of large data generation is the necessity of computational hardware; software; and individuals who are trained computationally and bioinformatically to manage, analyse, and interpret this data.

Next generation sequencing data generation and management are typically divided into three phases known as primary, secondary, and tertiary analysis. Primary analysis involves the analysis of raw data, base calling, and then scoring the called base quality. On the Ion Torrent system, this is outputted as an unmapped binary alignment file (uBAM). Overall, this process is relatively complicated and involves nucleotide incorporations being identified by the ion sensor and converted into general data (DAT) files. These files are inputted into the server and converted into an alternative file format. The Ion Torrent Basecaller module is able to read the files and generates the uBAM file. The uBAM file is then mapped and produces a binary alignment file (BAM) in which the sequencing data is stored (Ion Torrent 2014).

This is followed by secondary analysis in which the sequence is aligned to a selected reference and then variants are called. For this project the hg19 reference, also known as GRCh37 (Genome Reference Consortium Human Build 37) was selected. This is the human reference genome to which the Ion Torrent system is currently compatible and is still used by the majority of downstream analysis software. The Ion Torrent system uses the Torrent Mapping Alignment Program (TMAP) for this step (Homer et al., 2009). The TMAP uses the Smith-Waterman algorithm for initial mapping (Smith and Waterman 1981), after which the alignment is refined and a post-alignment processing step is performed (Pereira et al., 2020). The last step of secondary analysis is variant calling; for this Ion Torrent incorporates the Torrent Variant Caller (TVC) which uses a Bayesian statistical method and flow space data to generate a Variant Call Format (VCF) file.

Tertiary analysis is the final major step in the analysis pipeline and mainly involves data interpretation. This step can be broken down into variant annotation (determining the biological context of variants), variant filtering (narrowing down a list of possible

putative disease-causing mutations based on the annotations), and prioritisation and visualisation (identifying a few variants suggestive to be causative and validating them further with molecular techniques) (Pereira et al., 2020).

These three main areas of analysis are large and complex in and of themselves and yet still provide additional tasks, challenges, and limitations towards the researchers that need to be considered. Beginning with primary analysis, the amount of data generated from WES is much larger than any other genetic test that this Division currently offers, thus data management may be a challenge initially.

This includes the long-term storage of the generated data as well. This data includes primary sequencing data such as BAM files; and secondary data such as annotation files, and reports (Shabani, et al., 2018). While the secondary data is relatively simple to store using online cloud services, local computers, and hard drive backups; primary data files are much larger and therefore require more intensive storage solutions. This may not present as a challenge initially, however, the more exomes sequenced, the more data generated, and the more storage space required. On average a single exome of 40X coverage (average number of reads that align to a target/reference), with 110,000,000 reads at a read length of 75 typically generated a BAM file of 5.7-10GB (Strand Life Sciences 2016). Therefore, the data generated from 200 exomes would be roughly 1.6TB. If one where to generate and store backups of this data for safety and security measures the data would likely double in storage requirements (Strand Life Sciences 2016). With fluctuating read lengths and exome coverage, BAM files may reach up to 30GB, which poses a greater challenge to data storage.

The term of how long data is stored is typically determined by the service provider of the test, and can vary based on whether the data will be subjected to reanalysis at a later date and if it will be used for means of validation and/or research (Shabani, et al., 2018). It is the recommendation of the American College of Medical Genetics (ACMG) clinical laboratory standards for next-generation sequencing (Rehm, et al., 2013), that laboratories enforce a minimum term of two years for the storage of primary data (e.g. BAM files).

A final consideration for when looking at storage solutions and analysis of such large datasets, is the security associated with sharing and access to the data. Anonymised

patient and sequencing data need to be stored on secure servers/systems where access is granted and monitored by local administrators to ensure the privacy and confidentiality of the data as well as the security of the system (Kulkarni and Frommolt 2017; Moorthie et al., 2013).

Secondary analysis is a largely automated process, this leaves only a few challenges that a bioinformatician or researcher would need to consider. The main challenges that arise here are errors during sequencing and quality control procedures. The errors that arise during sequencing cannot typically be prevented however they can be monitored with the use of base quality scores known as Phred scores. The scores are based on a logarithmic error probability so, for example, a Phred score of 20(Q20) would have a 1 in 100 probability of being inaccurate indicating accuracy of 99% (Ewing et al., 1998). An additional consideration would be the use of a sequencing control to assess sample quality. A sequencing control is designed or selected to contain DNA sequences from a known species (specifically Homo Sapien for this project) that is processed jointly with selected samples during sequencing. The use of a sequencing control will allow one to detect or estimate sequencing errors through bioinformatic analysis following sequencing (Cao et al., 2017).

Tertiary analysis is the final step and involves active participation on the part of bioinformaticians and researchers. Data analysis and interpretation is a procedure containing multiple steps and makes use of various software to generate a manageable and interpretable set of results. As one of the main aspects of this project this topic will be discussed in-depth later on. Some of the broad categories used to narrow down or prioritise possible causative variants include: ethnically specific allele frequencies of variants; the possible impact of a variant on protein function; evolutionary sequence conservation and comparison with existing disease-specific databases containing known variant-disease/phenotype associations (Lee et al., 2014). One of the biggest challenges of this step, as suggested by researchers, is the determination of causal versus non-casual variants and understanding the clinical/medical significance of the information (Swaminathan et al., 2012; Daber et al., 2013).

1.7.3 Reanalysis of inconclusive results

The new developments and hope that arrive in the presence of WES being offered to families with patients suffering from undiagnosed genetic disorders, although astounding is not always successful or promising (Rosell et al., 2016). This technology provides a unique look at the human exome which may shed light on the disease of undiagnosed patients, however, it does not guarantee a diagnosis. In some cases, it may shed no light on the disease of the patient in question, because much of the impact of variation in the human genome is still unknown. However, new information on multiple Mendelian disorders (including DD), and their causative variants is emerging monthly within the field of genetics (Xiao et al., 2018). This allows scientists to reanalyse previously generated data to search for a possible diagnosis. As mentioned above, the work by Wright and colleagues showed that three years after initial testing, 182 patients received a diagnosis through reanalysis (Wright et al., 2018). With this in mind, the challenge becomes setting up a system whereby the data, of individuals who did not receive a diagnosis, is stored, tracked, managed, and reanalysed at certain intervals in the hopes of identifying new information associated to variants found within that individual.

1.8 Diagnostic Yield and Clinical utility

Diagnostic yield is defined as the likelihood of a test/procedure providing the information necessary and adequate to indicate a diagnosis and is a statistic commonly associated with the effectiveness of diagnostic technology. However, a diagnosis is not always the outcome for each patient, thus the clinical utility of the test may prove more beneficial in the overall treatment and management of patients. Clinical utility is the personal benefit that a patient receives due to an offered test and/or medical intervention (Lesko et al., 2010). While the concepts are somewhat connected the figures have shown to be discordant in multiple studies with the clinical utility of the tests appearing much lower than the diagnostic yields (Iglesias et al., 2014; Thevenon et al., 2016; Srivastava et al., 2014) which may impact the perceived utility of the test. The utility of WES in a low-resource setting may additionally be impacted by a lack of practical experience; transferable skills; bioinformatics knowledge and experience; monetary resources; data management and storage solutions; and the inaccurate clinical phenotyping of African DD patients who are vastly understudied.

1.9 Rationale

Deciphering Developmental Disorders-Africa is a research project based in the Division of Human Genetics, in partnership with DDD-UK. The overarching aim of this multifaceted project is to generate a pipeline for research and diagnostic purposes that will enable a better understanding of the genetic contribution of DD among African patients. This pipeline will be structured around the use of WES as a method of pathogenic variant detection.

Whole exome sequencing has emerged as a possible first-line genetic test for the diagnosis of individuals affected with DD. This technology has been shown to have a higher diagnostic rate than regular testing procedures, allowing for better detection of the causal variants of these heterogeneous disorders. The utility of this test has not been investigated in a limited resource setting. The availability of an Ion S5™ Semiconductor sequencer within the Division makes it possible to analyse the feasibility, diagnostic rate and clinical utility of WES in identifying pathogenic mutations in South African DD patients. Therefore, the research question for this study is: Is WES a viable and cost-effective approach to diagnose DD in a South African setting?

1.10 Aims and Objective

1.10.1 Aim

To implement a WES workflow and evaluate the financial and computational costs involved in this technique.

1.10.2 Objectives

1. Plan an effective WES laboratory and data analysis workflow using the Ion AmpliSeq™ and appropriate bioinformatics tools.
2. Perform WES on 8 DD patient samples to evaluate and refine the workflow.
3. Analyse WES data generated to identify any putative disease-causing mutations and optimize the bioinformatics pipeline.
4. Test parental samples for identified putative disease-causing mutations to establish variant inheritance.
5. Perform a cost and diagnostic yield analysis for WES to enable an indication of the cost-to-benefit ratio for this strategy and compare the Ion Torrent S5™ system to other available systems.

2. Methods

This chapter will describe the methods used to achieve the aim and objectives set out in Chapter 1 (Section 1.11). The chapter discusses the ethics clearance received for this project; the planning of a WES workflow necessary to conduct this project; the WES procedure on the Ion Torrent S5™ system; data analysis; and the validation procedures using Sanger sequencing. Lastly, this chapter addresses the methodology of performing a cost and diagnostic yield analysis.

2.1 Ethics

Ethics clearance for this project was granted by the Human Research Ethics Committee (Medical) of the University of the Witwatersrand, under certificate number M190578, see Appendix 1. This study is operating under the larger DDD-Africa study with ethics certificate number M180678, see Appendix 2.

2.2 Objective 1: Planning a whole exome sequencing workflow/pipeline

A basic outline of the laboratory and analytical procedure expected to be performed in the project can be seen below in Figure 1. This workflow was adapted from the work published by the DDD-UK project (Wright et al., 2015), however, it contains information more directed to the objectives of this project and the use of Ion Torrent platforms.

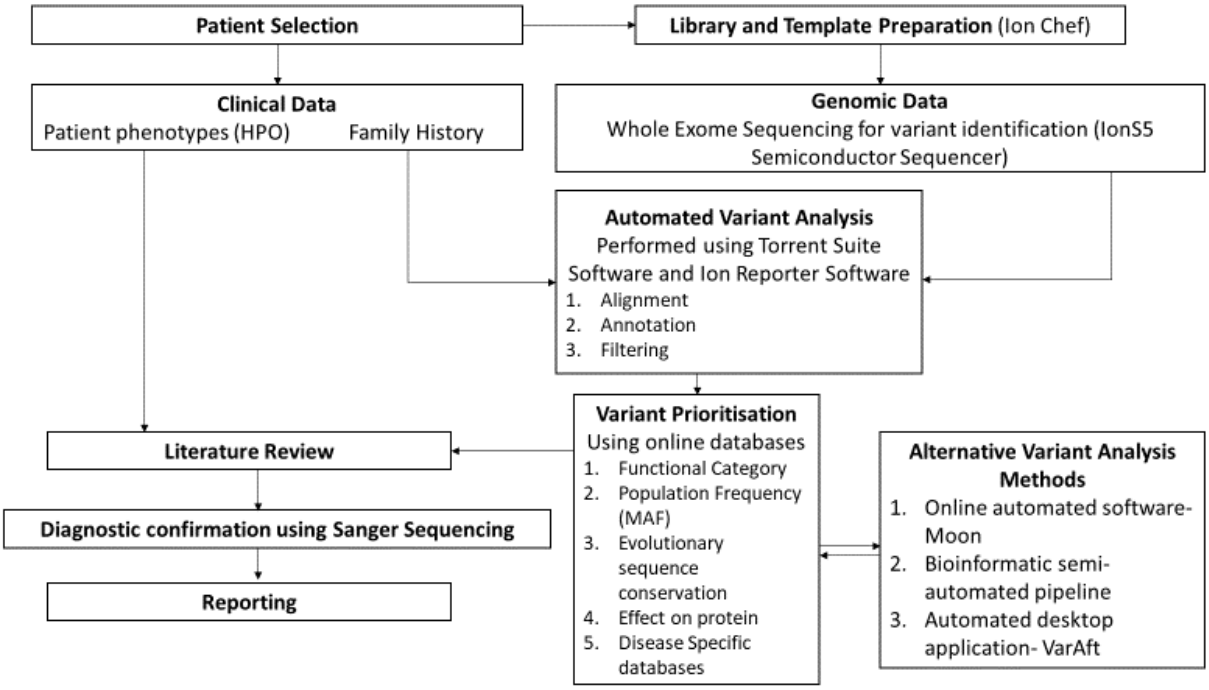


Figure 1. Planned workflow of whole exome sequencing. The above flow diagram illustrates the procedure of whole exome sequencing from the first step of patient/sample selection until the final step of reporting results. Sequencing is performed using the Ion Torrent S5™ platform and sequencing data will undergo primary and secondary analysis using the Torrent Suite™ and Ion Reporter™ software. Tertiary analysis will be performed using multiple variant analysis methods in conjunction with online databases for manual investigation. Significant variants are confirmed using Sanger sequencing before reporting. Adapted from (Wright et al., 2015).

2.3 Objective 2: Performing whole exome sequencing on 8 DD patient samples to evaluate and refine the workflow

2.3.1 Sample Selection

The patient samples used for this study were selected from a pre-approved list (provided by the clinicians and senior researchers in the Division) based on deoxyribonucleic acid (DNA) quality and concentration. These patients had already been recruited for the DDD-Africa study. Seven samples were selected for WES with the addition of one control sample to account for sequencing quality (control DNA from CEPH individual 1347-02, Thermo Fisher, USA). DNA extraction was previously performed using the salting-out method (Miller et al., 1988) and stored at 4°C in Tris-EDTA Buffer (TE Buffer). Quality and quantity were ascertained using a UV spectrophotometric approach using the Thermo Fisher™ NanoDrop 2000 (Thermo Fisher, USA). DNA quality was further visualised using gel electrophoresis, where 1µl

of DNA was visualised on a 0.8% (w/v) agarose gel that was electrophoresed at 5V/cm. The samples for this study were selected based on this information. DNA quantification of the selected samples was then performed using a Qubit™ 3.0 fluorometer; this was one of the suggested methods of quantification as recommended by the Ion AmpliSeq™ Exome RDY Library Preparation User Guide (Thermo Fisher, USA). The samples were initially quantified using the Qubit™ dsDNA BR (broad range) Assay Kit (Thermo Fisher, USA) which has a range of 2-1000ng. The second quantification used the Qubit™ dsDNA HS (high sensitivity) Assay Kit (Thermo Fisher, USA) which has a range of 0.2-100ng. Both quantification kits were used to allow for an additional dilution step for more accurate dilutions. DNA was diluted to ~50ng genomic DNA (gDNA), using nuclease-free water as recommended by the reference guide.

2.3.2 Whole Exome Sequencing

WES was performed using the Ion Torrent S5™ system housed at the Division of Human Genetics, the procedure is summarised in Figure 2. The procedure itself involves three main steps namely library preparation, library templating, and finally sequencing.

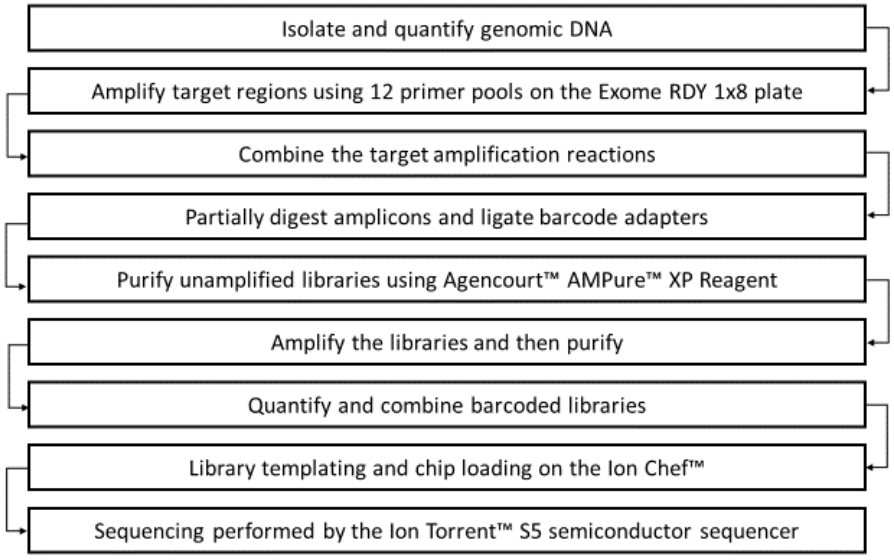


Figure 2. Flow diagram illustrating the procedure of whole exome sequencing on the Ion Torrent S5™ platform. Adapted from the Ion AmpliSeq™ Exome RDY Library Preparation User Guide (Thermo Fisher, USA).

i. Library Preparation

Library preparation is the first step of NGS and produces a collection of targeted DNA fragments for sequencing. This was performed following the procedure outlined in the Ion AmpliSeq™ Exome RDY Library Preparation User Guide (Thermo Fisher, USA); using an Ion AmpliSeq™ Exome RDY Kit 1x8 (Thermo Fisher, USA), and Ion Xpress™ Barcode Adaptors 1-16 (Thermo Fisher, USA). Target regions were initially amplified using 12 primer pools in 1x8 Exome RDY Plate. The target amplification reactions were then combined and partially digested. Barcoded adaptors were ligated to the partially digested amplicons and this was followed by two purification steps. The exome libraries require amplification before quantification for enrichment and to ensure there is sufficient material available for quantification. The barcoded libraries were thus amplified and underwent an additional two rounds of purification. The exome libraries were then stored at -20°C until used for sequencing.

The amplified libraries were quantified using a Qubit™ 3.0 fluorometer and a Qubit™ dsDNA HS Assay Kit (Thermo Fisher, USA). Exome libraries typically have yields of 300-1500ng/ml, therefore the Qubit™ dsDNA HS Assay Kit (Thermo Fisher, USA) was used. Libraries were diluted (based on the calculated concentrations) to a final concentration of ~22ng/ml (100pM) using the Low TE buffer provided in the Ion AmpliSeq™ Library Kit Plus (Thermo Fisher, USA).

The reference guide used suggested that two barcoded libraries could be pooled onto a single 540™ Chip (Thermo Fisher, USA) for a targeted average coverage depth of 100X. This recommendation was followed and exome libraries 1 and 3; 2 and 4; 5 and 6; 7 and 8 (control) were paired off and combined respectively. Libraries were paired based on quantified concentrations that were closely matched. The individual libraries of a designated pair were separately diluted to the same concentration and the combined. Each combined library was then diluted to a final concentration of 75pM (this was mid-range of the suggested final concentration range 50-100pM).

Library dilution and combination were performed directly before template preparation was initialised and therefore not all libraries were diluted and combined at the same time.

ii. Library Templating

Library templating is the second step of NGS and this step aims to generate multiple copies of each amplicon to increase the strength of nucleotide base detection during sequencing. A planned run was created using the Torrent Suite™ Software to provide instructions directing the templating, performed using the Ion Chef™, and the sequencing, performed using the Ion S5™. Template preparation was performed on the Ion Chef™ using Ion 540™ Chef Reagents (Thermo Fisher, USA), Ion S5™ Chef supplies (Thermo Fisher, USA), and Ion S5™ Chef Solutions (Thermo Fisher, USA). This was performed using the Ion 540™ Kit-Chef Quick Reference as a guide. The reagents, combined barcoded libraries, and chips were loaded onto the Ion Chef™. This process involves an emulsion PCR step that fragments and amplifies exome targets and binding these fragments to microbeads. Following this, the templated libraries are loaded onto the 540™ chips. Upon completion of the Ion Chef™ run, the chips and reagents were unloaded. One chip was moved to the sequencer to begin sequencing (only one chip can be sequenced at a time), whilst the other was placed in a storage container and refrigerated at 4°C until it was ready to be sequenced.

iii. Sequencing

Whole exome sequencing was performed using the Ion S5™ System. After an initialisation step, the first templated chip removed from the Ion Chef™ was loaded into the Ion S5™ for sequencing. This was performed following the guidelines set out in the Ion 540™ Kit-Chef Quick Reference and used Ion S5™ sequencing solutions (Thermo Fisher, USA) and Ion S5™ Sequencing Reagents (Thermo Fisher, USA). The sequencing of the prepared whole exome libraries is based on 200bp chemistry and uses a *Samba* flow order. This flow order is a step sequence of 32 nucleotides that is repeated 15.63 times to reach a total of 500 flows (Sarkozy et al., 2014). Upon completion of this sequencing (2.5 hours), the first chip was removed and the second was loaded to begin sequencing. The sequencing of 4 chips took two days to complete, upon which Binary Alignment Map (BAM) files were generated containing the sequence output.

2.4 Objective 3: Analysing whole exome sequencing data generated to identify any putative disease-causing mutations and optimise the bioinformatics pipeline

2.4.1 Data Processing

A flow diagram illustrating the data processing of the BAM files and generation of the Variant Call Files (VCFs) can be seen below in Figure 3. The Ion Torrent associated software(s): Torrent Suite™ and Ion Reporter™ (version 5.12.0 Thermo Fisher, USA) were used for these processes. VCF files were analysed using multiple software and pipelines and the results compared in order to identify the most effective and time-efficient programs to identify possible putative disease-causing variants in South African DD patients.

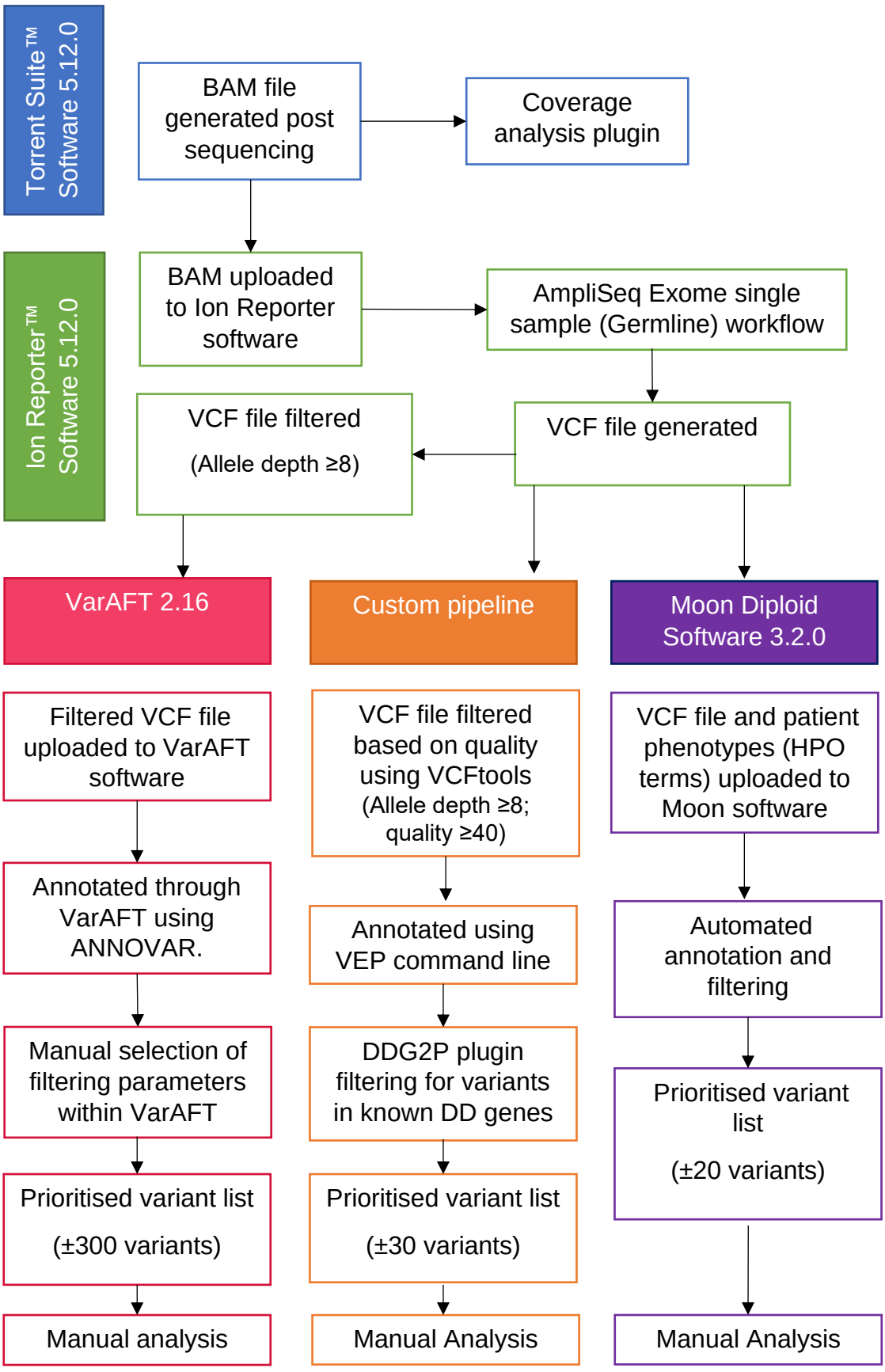


Figure 3. Summary of data analysis pipelines created to investigates exomes of DD patient samples.

i. Torrent Suite™ Software

The Torrent Suite™ Software (version 5.12.0) allows a user to view current sequencing activity, plan and monitor sequencing runs, and review sequencing data post-run. The 'planned run' was created prior to sequencing whereby the sequencing run information, sample identifiers, and analysis preferences were selected. Sequencing can be monitored in real time using key metrics provided by this software; and by including the Coverage Analysis Plugin, run statistics can be viewed and data quality analysis performed. The BAM file generated during sequencing is accessible through this software and allows one to directly upload said file to the Ion Reporter™ software (version 5.12.0, Thermo Fisher, USA) and request the analysis pipeline of interest. The AmpliSeq Exome single sample (germline) workflow was selected, this was the recommended pipeline for proband exome analysis as suggested by the Thermo Fisher technicians. The software produced an annotated VCF for each patient that was then downloaded. Filtering parameters are available within the Ion Reporter™ software; however, they were not capable of performing all of the filtering steps required and different software and platforms were investigated.

ii. Ion Reporter™ Software

The Ion Reporter™ software (version 5.12.0) is a collection of bioinformatics tools that allow for data analysis, annotation, filtering, and the export of sequencing data. The BAM file generated by the Torrent Suite™ was uploaded to the Ion Reporter™. The data was run through the AmpliSeq Exome single sample (Germline) workflow. This workflow uses the 'variantcaller' plugin to generate a VCF file, after which it runs this file through multiple annotation databases. The VCF files for one of the downstream software applications required pre-filtered files based on quality control; the built-in filter system of the Ion Reporter™ Software was used for this. The parameters included a CNV confidence range filter ($10.0 \leq$ CNV confidence range$\leq 1.x10^{7}$) to return CNVs with confidence ranging between 10 and 10,000,000. As no CNV baseline is currently set up on the specific system used for this project, CNVs could not confidently be called and were therefore filtered out. The second parameter involved filtering for quality, thus an allele read count was set to filter out variants with an allele depth of ≤8. Unfortunately, no filter was available to remove variants of a low-quality score (PHRED score ≤40) therefore, this parameter was left out.

iii. Moon Software

Moon (Diploid) software (version 3.2.0) is a web-based software that uses artificial intelligence to diagnoses rare diseases from NGS data (Clark et al., 2019). WES VCF files were uploaded to Moon and the patient's disease age of onset, gender, and symptoms (Human Phenotype Ontology (HPO) terms, see Table 1) were entered. The Moon software analyses this input and suggests a prioritised possible causal variant. Additional variants were also suggested for review. The variants listed contain annotation data from multiple sources which can be seen in Table 1 (associated URLs can be found in Appendix 3). Filter parameters were set to remove all variants with an allele depth of ≤8; a quality score of ≤40 and/or a frequency of ≥1% seen in GnomAD (see Table 1). These were the default filter parameters suggested by Moon and were kept constant across the different analysis platforms used. Specific filter values are useful as they narrow down the potential list of causative variants. Filtering for allele depth allows the removal of variants that were poorly covered during sequencing and similarly, filtering with a quality score removes variants of poor quality leaving a list of variants with sufficient coverage and good quality for investigation (Harel and Lupski 2017). Further, filtering according to allele frequencies is greatly beneficial in reducing the list of potentially causative variants as it is known that variants likely to be causative of rare genetic diseases are typically rare within the population and are expected to have an allele frequency of less than 1% (Gibson 2011). Filters were also set to only include non-synonymous coding variants and all zygosity options (homozygous, heterozygous, hemizygous, compound heterozygous, homoplasmic and heteroplasmic).

Table 1. Variant annotation tools used by Moon (v3.2.0)

Tool	Description	Reference
ClinVar	Online database of genomic variation and associations to human health	Landrum and Kattman 2018
dbNSFP (v4.0)	Database for the functional prediction and annotation of non-synonymous SNVs in the human genome	Liu et al., 2016
dbSNP (v151)	Public database of known genetic variation across different species	Sherry et al., 2001
Apollo	Web-based genome annotation editor	Dunn et al., 2019
RefSeq (v37)	Database of curated nucleotide sequences	Haft et al., 2018
GnomAD (v2.1.1)	Aggregation of whole exome and genome data providing allele frequencies	Karczewski et al., 2020
HPO	A vocabulary of phenotypic abnormalities associated with human disease	Kohler et al., 2018
DGV	Database of human genomic structural variation	MacDonald et al., 2014
dbVar	Database of human genomic structural variation	Lappalailen et al., 2013
Mitomap	A human mitochondrial genome database	Lott et al., 2013
Mitimpact (v2.9.1)	A collection of pathogenicity prediction for variants of the human genome and mitochondrial genome	Castellana et al., 2015
Mastermind	Genomic search engine identifying publications associated to known variants	Genomenon 2020

iv. Customised Pipeline

The Wits Core Cluster is a computer cluster shared between different academic units at the University of the Witwatersrand. A computer cluster is an assemble of connected computers that work together to essentially be viewed as a single unit. Computer clusters can have multiple uses but are highly beneficial in a genomic setting, as the processing power, analysis speed, and availability of data storage are drastically increased in comparison to that of a single computer. Due to the mass of data generated by NGS, the availability of a computer cluster would be a valuable resource for researchers and bioinformaticians. The Wits Core Cluster contains two log-in

nodes, 9 storage servers, and 45 worker nodes with approximately 1PB of globally addressable disk space. These features allowed for the rapid analysis and processing of the WES data generated by this project.

The use of this computer cluster allowed for automated batch sample annotation and filtering of WES data. VCFs were uploaded onto the cluster where they underwent analysis. VCFs were initially run through a 'VCFtools' script that filtered out variants based on allele depth (≤8) and quality (PHRED scores ≤40). This script is shown below:

```
#!/bin/bash
#PBS -N vcftools
#PBS -l walltime=2:00:00,mem=4GB
#PBS -q batch
#PBS -o /home/kait/VCFs
#PBS -e /home/kait/VCFs

DATA=/home/kait/VCFs
cd $DATA
module load vcftools

for K in {1..8};do vcftools --vcf Sample_${K}_vcf.vcf --min-
meanDP 8 --minQ 40 --recode --recode-INFO-all --out
vcftools_${K}.vcf;
done
```

The quality-filtered VCFs generated from the above script were then run through a second script with variant annotation using the command line version of VEP (McLaren et al., 2016) with an additional plugin to assist filtering. A gene2phenotype (G2P) VEP plugin was added to the script to filter for variants in likely disease-causing genes found in the G2P database (see Appendix 3 for URL). G2P is a dataset constructed from published literature that combines phenotype, gene, and variant data (Thormann et al., 2019). The database contains 2500 genes known to be associated with DD and the allelic requirements necessary for disease expression, variants that overlap the genes on this list will be prioritised and variants within genes not found on the list will be filtered out. This was developed by EMBL-EBI in association with the Wellcome Sanger Institute and TGMI and operates as a part of VEP (Wright et al., 2015). These filtering parameters were ideal for our investigation into identifying possible disease-causing mutations in South African DD patients. This script showing our use of the plugin (written by Mr Shaun Aron) is shown below:

```
#!/bin/bash
#PBS -N vep_g2p
#PBS -l walltime=200:00:00,mem=100GB
#PBS -q batch
#PBS -o /home/kait/VCFs
#PBS -e /home/kait/VCFs

DATA=/home/kait/VCFs
cd $DATA
module load perl526

for K in {1..8};do vep -i Sample_${K}_vcf.vcf --
output_file="variant_effect_${K}.txt" --
dir_plugins="/opt/exp_soft/bioinf/ensemble-
vep/plugins/VEP_plugins_Jan_2019' --plugin
G2P,file="DDG2P.csv",log_dir="g2p_log_dir_${K}",txt_report="
repoort_${K}.txt",html_report="report_${K}.html" --
dir_cache="/dataB/aux/37/cadd/vep_caches" --cache --port
3337;
done
```

v. VarAFT

VarAFT is a downloadable software that was created to improve the annotation and filtrations steps associated with the data analysis of NGS data (see Appendix 3 for URL). The software is freely available to non-commercial users and was developed by the Genetics and Bioinformatics team at INSERM UMR 1251, Medical School, la Timone Aix-Marseille University in Marseille, France (Desvignes, et al., 2018). VCF files were uploaded and annotated through multiple annotation tools including ANNOVAR (16/04/2018; Wang et al., 2010) and UMD-Predictor (v2; Salgado et al., 2016). Filtering parameters were manually selected and included variant type (only exonic variants were selected); variant effect (nonsynonymous, stop loss, stop gain, frameshift insertion, frameshift deletion, and frameshift substitution were selected for); frequency (variants with frequencies ≤0.01 found in GnomAD and 1000 Genomes were included); prediction (all indicators of polymorphisms, neutral and benign variants were filtered out); and gene information (this filter was inapplicable in this study as no disease, pathway or gene associations were known). The list of prioritised variants generated from this software was too large for manual review and these results were therefore not taken further.

2.4.2 Variant Prioritisation Through Manual Analysis

Manual analysis was performed on all variants prioritised by the Moon (Diploid) software and custom pipeline for each patient. Once variants were prioritised to roughly 30 variants per sample the automatic filtering chains and programs were no longer able to prioritise the lists further. At this stage, each variant on the list was investigated to determine its possible validity in being the most likely disease-causing variant. Each variant was analysed following the guidelines for variant interpretation set out by the ACMG (Richards et al., 2015). Online software and web-browsers were used to analyse all criteria for each variant, these criteria can be seen in Appendix 4. Examples of online software and databases that were used can be seen below in Table 2 (the associated URLs can be found in Appendix 3). These software have multiple purposes such as variant visualisation with tools such as the Ion Reporter Genomic Viewer (version 5.12.3.0); variant annotation using the Variant Effect Predictor; pathogenicity prediction using PolyPhen-2 and SIFT; and additional tools for literature review, variant evolutionary conservation analysis, frequency analysis, and genotype-phenotype correlation.

Variants classified as benign or likely benign were ruled out to be causative and thus removed. Variants identified with an unknown significance were removed in most cases unless the implicated gene was strongly indicative of being the causative gene, in this case, the variant went through a more thorough investigation to determine whether or not it could be upgraded to likely pathogenic. Pathogenic and likely pathogenic variants were prioritised further in cases where the variant and affected gene were highly likely to be causative of the disease. If the phenotype of the patient was a strong match to that of the suggested disorder, they were presented to the clinical and senior research team for discussion. The outcome of this discussion determined which variants were validated using Sanger sequencing.

Table 2: Variant annotation and analysis tools used during manual analysis

Tool or Website	Functionality of tool	Version	Reference
GnomAD	Aggregation of whole exome and genome data providing allele frequencies	v2.1.1	Karczewski et al., 2020
VEP	Variant annotation tool	Release 99	McLaren et al., 2016
ENSEMBL	Genome browser with annotation, alignment and prediction functions	Release 99	Cunningham et al., 2019
IRGV	Genomic viewer used to visualise sequencing data within BAM files	5.12	Thermo Fisher Scientific Inc. 2019
Varsome	Variant search engine containing annotated and current variant classifications	-	Kopanos et al., 2018
Mutalyzer	Checking of sequence variant nomenclature	2.0.32	Wildeman et al., 2008
DECIPHER	Online collection of tools to aid genomic variant interpretation	v9.31	Firth et al., 2009
OMIM	Database of human genes and associated genetic diseases	-	McKusick 1998
ClinVar	Online database of genomic variation and associations to human health	-	Landrum and Kattman 2018
PubMed	Online literature search engine	-	NCBI Research Coordinators 2018
Mastermind	Genomic search engine identifying publications associated to known variants	-	Genomenon 2020
PolyPhen-2	Prediction tool of variant impact on a protein	v2	Adzhubei et al., 2010
SIFT	Prediction tool of variant impact on a protein	-	Sim et al., 2012
Pfam	Database of protein families	32.0	El-Gebali et al., 2019
CADD	Tool for scoring the deleteriousness of variants	GRCh37-v1.6	Rentzsch et al., 2018
Moon	Software able to identify rare diseases from NGS data	3.2.0	Clark et al., 2019

2.5 Objective 4: Variant validation and testing parental samples to determine the mode of inheritance

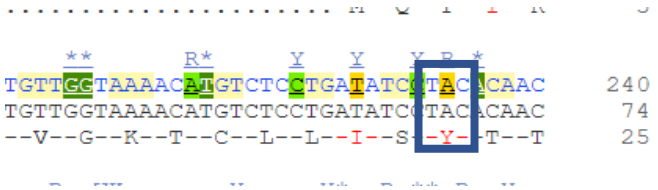
2.5.1 Variant Validation

Prioritised variants underwent Sanger sequencing as a means of validation. When available, the parental samples were also sequenced in order to determine the mode of inheritance.

i. Primer Design

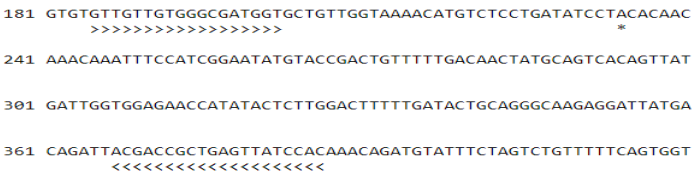
A summarised flow diagram of the primer design procedures can be seen in Figure 4. Variants were identified on the Ensembl (GRCh37-Release 99) transcript from which they were called by the Moon analysis software. Ensembl is a genome browser that supports a collection of vertebrate genomes including Homo Sapiens (Cunningham et al., 2018). The RefSeq gene reference sequence was obtained from NCBI RefSeq (accessed 20 February 2020) in order to design primers and downloaded as a FASTA file. The National Center for Biotechnology Information (NCBI) is an online database containing genomic and biomedical information. Included in this resource are the reference sequences of multiple genes within the human genome (NCBI Research Coordinators 2018). The sequence surrounding the variant (taken from Ensembl) was then identified in the NCBI RefSeq sequence, this area and surrounds were used for primer design. A hundred base pairs were selected before and after the variant (where possible) to ensure enough sequence space for disrupted sequencing that occurs at the beginning of Sanger sequencing. The selected sequence was then entered into Primer3 (v4.1.0). Primer3 is an online automatic primer design tool used to design primers for PCR (Untergasser et al., 2012). Selected primers generated by the tool were then run through primer specificity tools such as Primer BLAST and UCSC in

Variant was found and viewed on the corresponding ENSEMBL transcript

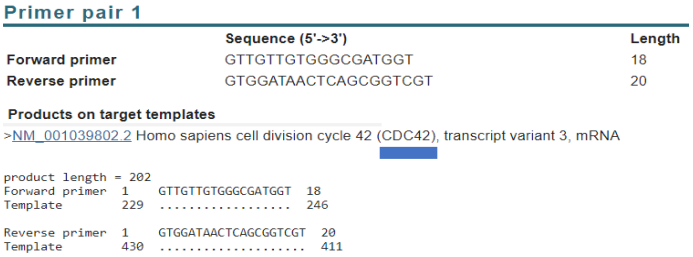


Variant was found in FASTA file retrieved from NCBI Reference Sequence

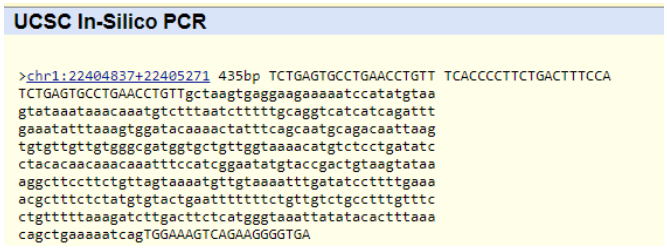
FASTA file copied to Primer3 to generate primers



Primer BLAST used to ensure selected primer were unique to the gene of interest



Primers were checked using UCSC *in silico* PCR to determine target compatibility



Primers were checked using SNPCheck to ensure no SNPs were present within the primers

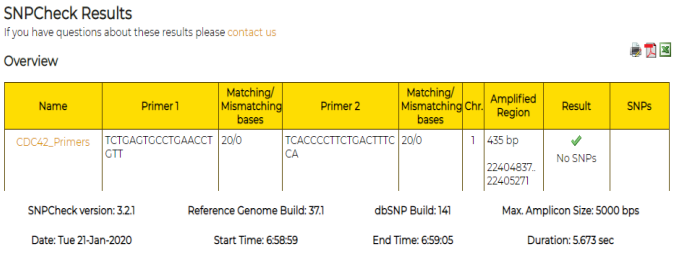


Figure 4. Flow diagram illustrating the steps taken in PCR primer design.

silico PCR. Primer BLAST is an online tool used to design PCR primers and check them for specificity. In our case, the tool was only used to determine specificity. UCSC *in silico* PCR is an online tool that determines primer specificity, much like Primer BLAST. These two tools were used to ensure all primers isolated the specific target region on the current chromosomal location. Lastly, primers were entered into SNPCheck, which is an online web tool used to identify SNPs within oligonucleotide primers designed for PCR that may prevent the binding of the primer to the target sequence (Lai and Love 2012). The tool checks primers against data from dbSNP, 1000 genomes, and the Exome Sequencing Project, these databases contain European, African, Asian, and American genome data. The presence of African data comparisons makes the tool applicable to our data. The URLs for all of the above-mentioned tools can be found in Appendix 3.

ii. PCR Optimisation and DNA Amplification

Primers were supplied by Inqaba Biotec™ (Pretoria, South Africa) and are shown in Appendix 5. The primers were resuspended in the recommended volumes of TE Buffer for a stock concentration to 100ng/µl; and then diluted to 10ng/µl for use in PCR. The annealing temperature for each primer pair was determined using a temperature gradient experiment ranging ±5°C from the primer's melting point, and a control DNA sample. Specific annealing temperatures were selected for each primer pair based on the results of the temperature gradient. PCR was performed using a Bio Rad T100™ Thermal Cycler (Thermo Fisher, USA) and an Applied Biosystems SimpliAmp Thermal Cycler (Thermo Fisher, USA). PCR reagents and thermal cycling conditions can be found in Tables 3 and 4 respectively.

Table 3. Polymerase Chain Reaction (PCR) protocol for DNA amplification

PCR Component	**Volume (µl)**
DNA [100ng/µl]	1
Forward Primer [10ng/µl]	1
Reverse Primer [10ng/µl]	1
dNTP Mix [10mM per dNTP]	2.5
AmpliTaq Gold Buffer II [10X]	2.5
AmpliTaq Gold MgCl2 [25nM]	2.5
ddH2O	14.3
AmpliTaq Gold DNA Polymerase [5U/ µl]	0.2
TOTAL	25

Table 4. PCR cycling conditions

Cycling Conditions			
Step(s)	**Temperature (°C)**	**Duration**	**Cycles**
Initial denaturation	95	3min	1
Denaturation	95	30sec	35X
Annealing	51-55*	30sec	
Elongation	72	30sec	
Final Elongation	72	1min	1
Hold	5	∞	1

*Annealing temperatures were unique to each primer pair but were all between the range of 51-55°C.

PCR products were run on a 3% (w/v) agarose gel to ensure amplification of the desired DNA product, SeaKem LE Agarose (Lonza, Basil, Switzerland) was used to make the gel. Two microlitres of PCR product was added to 3µl of agarose gel loading buffer (6X, 15%Ficoll Dye) and run on the agarose gel at 5V/cm.

Post-PCR and cycle sequencing clean-up were performed using double distilled water (ddH_2O) and Millipore Injection Solution respectively in PCR Cleanup Filter Plates on the Millipore Millivac Maxi SD1P014M04 (MilliporeSigma, USA). Cycle sequencing was set up using the BigDye™ Terminator v3.0 Cycle Sequencing Kit and run on a Bio Rad T100™ Thermal Cycler (Thermo Fisher, USA). All patient samples were sequenced in both forward and the reverse directions for the specific variants being

validated. Cycle sequencing reagents and thermal cycler conditions can be seen in Tables 5 and 6 respectively.

Table 5. Cycle Sequencing Protocol

Reagent	**Volume**
Cleaned PCR Product	5µL
BigDye™ Terminator 3.1 Ready Reaction Mix	2µL
BigDye™ Terminator 5X Sequencing Buffer	1µL
Primer (forward or reverse) [10ng/µl]	1µL
ddH2O	1µL
Total	10µL

Table 6. Thermal Cycler Conditions for Cycle Sequencing

Temperature	**Time**	**Cycles**
96°C	1 min	1
96°C	30 sec	25
50°C	15 sec	
60°C	4 min	
5°C	∞	1

Sanger sequencing was performed on cycle sequencing product suspended in Millipore Injection Solution using an Applied Biosystems 3500XL Genetic Analyzer (Thermo Fisher, USA). Sequencing electropherograms were then analysed using the CLC Main Workbench (version 20.0.3, Qiagen, the URL can be found in Appendix 3). Sequences were uploaded onto the CLC Main Workbench where they were visualised and then trimmed. Trimming is performed to remove the poor sequencing data at the ends of the sequence, these trimmed areas are then excluded from further analysis. After trimming, sequences were selected for secondary peak calling. Secondary peak calling identified locations where initially only one peak was called but there are two peaks present, this function allows for the detection of heterozygous mutations (Qiagen 2020). Once complete, reference sequences (obtained from NCBI RefSeq 20 March 2020) were uploaded and used to align the generated sequencing data. Once aligned, conflicts could be identified (areas where the generated sequence data differed from that of the reference), should a conflict exist at the specific variant location then the variant was confirmed (Qiagen 2020). If no conflict was present then the variant was not valid and cannot be attributed to the cause of the disease.

2.6 Objective 5: Cost and Diagnostic Yield Analysis

Costing was determined with the NHLS costing tool and included the costs of WES reagents, consumables and labour. The pricing of WES was also obtained from other commercial companies providing the service in South Africa and internationally for comparison.

The diagnostic yield was determined by dividing the number of positive results by the total number of patients tested (de Haan et al., 2012). A positive diagnosis was confirmed by a group of clinicians and researchers. The confirmation was based on the identification of a pathogenic or likely pathogenic variant that was validated through Sanger sequencing. The variant needed to suggest a phenotype the was a clinical match or close clinical match to that of the patient.

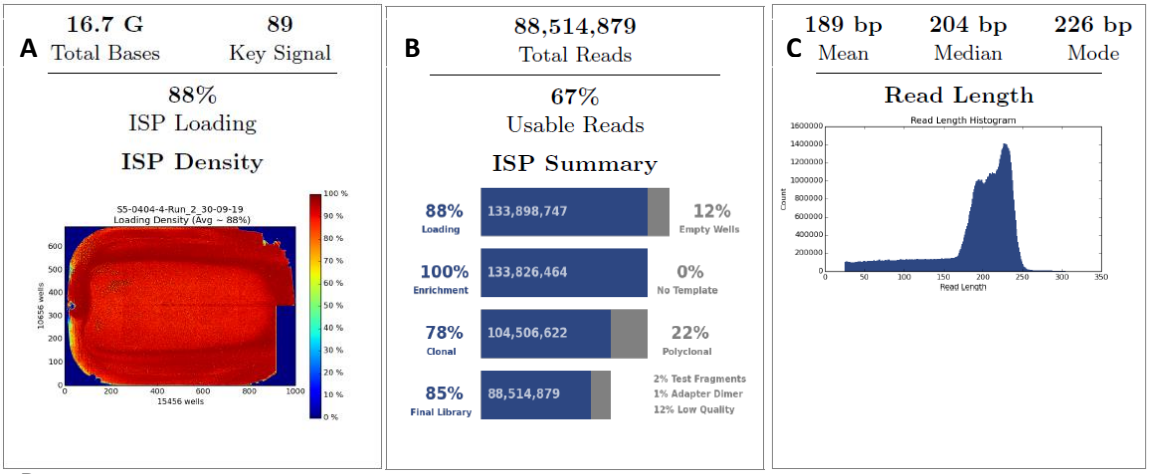
Finally, the specifications of the Ion Torrent S5™ were compared to that of the Illumina MiSeq (an alternative, locally available NGS platform) to understand how the output of each platform may differ to its competitor.

3. Results

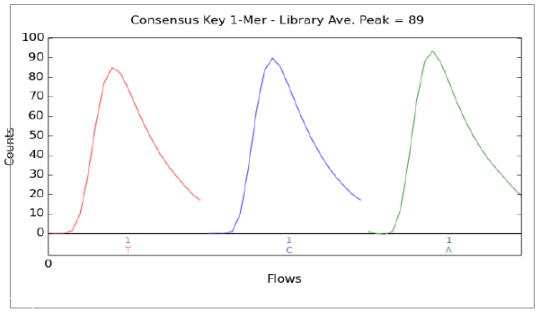
The results obtained from the methodologies implemented in Chapter 2 are presented in this chapter. NGS data quality and run results, followed by the laboratory work timelines are shown. The data analysis platforms used and their performance leading to the patient-specific results. Lastly, the chapter shows the results of the cost and diagnostic analysis and the comparison of the Ion Torrent S5™ system to the Illumina MiSeq system.

3.1 NGS Quality and Run Results

Upon completion of sequencing, the Torrent Suite™ Software produces a 'Run Report' containing important quality control and sequencing run statistics. The results in this report are separate from those reported from the Coverage Analysis Plugin which will be discussed later. This 'Run Report' can be separated into two sections namely the quality control and run statistics before alignment (Figure 5) and the quality control and run statistics post-alignment (Figure 6). The statistics provided before alignment include the chip loading statistics (Figure 5A), showing the Ion Sphere Particle (ISP) loading percentage. This is the percentage of wells in the chip that contain a live ISP. A heat map shows the loading across the physical surface of the chip. The chip loading step is performed by the Ion Chef™ and was, in this case, of good quality with 88% of the chip showing wells containing live ISPs (Figure 5A).



D



E

Addressable Wells	**151,543,787**	
With ISPs	133,898,747	88.4%
Live	133,826,464	99.9%
Test Fragment	1,855,793	01.4%
Library	131,970,671	98.6%
Library ISPs	**131,970,671**	
Filtered: Polyclonal	29,319,842	22.2%
Filtered: Low Quality	12,975,126	09.8%
Filtered: Adapter Dimer	1,160,824	00.9%
Final Library ISPs	**88,514,879**	**67.1%**

F

Barcode Name	Sample	Bases	≥ Q20	Reads	Mean Read Length	Read Length Histogram
No barcode	none	161,289,862	131,451,934	970,747	166 bp	
IonXpress_007	Sample 7	8,057,415,064	7,024,445,656	41,981,176	191 bp	
IonXpress_008	Control	8,492,855,554	7,387,798,817	45,556,813	186 bp	

G

Test Fragment	Reads	Percent 50AQ17	Read Length Histogram
TF_1	**775,990**	**88**	

Figure 5. Quality control and run statistics before alignment. This figure is a collection of the quality control and run statistics of Chip 3 (containing Sample 7 and the Control sample) before alignment to a reference. **A**: Graphical representation of chip loading performed by the Ion Chef™ and the percentage of Ion Sphere Particles (ISP) loaded on the chip. **B**: Bar graph showing an ISP summary and the percentage of usable reads detected. **C**: A histogram showing the read length distribution and average. **D**: Average signal readings for the flows of nitrogenous bases T (red), C (blue), and A (green). **E**: Numerical representation of the available wells for ISP loading on the given chip and filtering steps leading to the final library ISPs with usable reads. **F**: Run statistics of barcoded samples and reads identified without a barcode. **G**: Run statistics and quality control of the Test Fragment (positive control).

The second statistic is an ISP summary (Figure 5B). This summary includes a percentage of usable reads which is the percentage of library ISPs that pass the polyclonal, low quality, and primer dimer filters. These filters are illustrated as a bar graph, however, the factors taken into account to determine the percentage of usable reads is better depicted by the information provided in Figure 5E. This figure shows that of the initial 151,543,787 available wells initially present on the chip, 88.4% were loaded with ISPs, and of those, 99.9% were live. The ISPs containing Test Fragments (1.4%) were then subtracted leaving a library containing 131,970,671 wells containing live ISPs. This number was further filtered to remove polyclonal ISPs (22.2%). An ISP is reported as clonal if all of its DNA fragments are cloned from a single original template, thus all the fragments on the bead are identical. Polyclonal ISPs arise when DNA fragments are cloned from two or more templates. This disrupts sequencing and downstream analysis as the fragments won't respond in unison to each nucleotide flowed across the chip. Therefore, they are removed from the final library. Additional filters were implemented to remove low-quality reads (9.8%) and adapter dimers (0.9%). This left the final library of ISPs at 88,514,879 (67.1% of Library ISPs).

Read length statistics (Figure 5C) show an average read length of 189bp aligning to the 200bp chemistry that was used for exome sequencing. The nucleotide flow statistics can be seen in Figure 5D. The average signal readings for the flows of nucleotides T, C, and A are shown; where the count is dependent on the number of times the nucleotide appears in the *Samba* flow and how many flows were sent across the chip.

The sample-specific statistics for each chip are available due to the demultiplexing of the barcoded samples, this can be seen in Figure 5F. Here, the total number of bases called per sample can be found; and the number of bases with a Phred quality score greater than 20 (≥Q20). This is suggestive of an error rate of less than 1%. The mean read length and a read length histogram are provided for each sample. The read length histograms show a distribution skewed to the right; this is indicative of longer read lengths. Lastly, in this section, run statistics of the Test Fragment are provided, see Figure 5G. A Test Fragment is a positive control incorporated into the Ion Torrent™ chemistry. The 'Run Summary' provided the number of filtered reads identified for the Test Fragment (775,990) and the percentage of reads for the Test Fragment with a

minimum of 50 base pairs in length and an error rate of 1 in 50 or a Phred score of greater than 17, this is depicted as 50AQ17 (88). A read length histogram is provided, and shows the distribution skewed to the right in the sample-specific histograms.

Post-alignment of sequencing data to the hg19 (GRCh37) human reference sequence resulted in an alignment summary found within the 'Run Report'. This summary provided alignment quality data (Figure 6A). The data covers three different quality criteria namely:

- AQ17: alignment data with a Phred quality score of greater than 17, suggestive of an error rate less than 2%
- AQ20- alignment data with a Phred score of greater than 20, suggestive of an error rate less than 1%
- 'Perfect' alignment data which provided information on the longest perfectly aligned segment.

Post-alignment read length histograms were also provided showing how many reads of a certain read length were obtained under different quality filters (Figures 6 B, C, D, and E). Across these different quality cut-offs, the distributions are consistently skewed towards the right indicating reads of greater length. However, as the quality cut-offs increase (AQ47) we can see that the amount of reads at the desired 200bp length decrease by half, and the amount of low read length (50bp) has doubled.

A

Alignment Quality

	AQ17	AQ20	Perfect
Total Number of Bases [Mbp]	15.5 G	14.1 G	10.5 G
Mean Length [bp]	185	175	139
Longest Alignment [bp]	355	351	332
Mean Coverage Depth	5.0	4.6	3.4

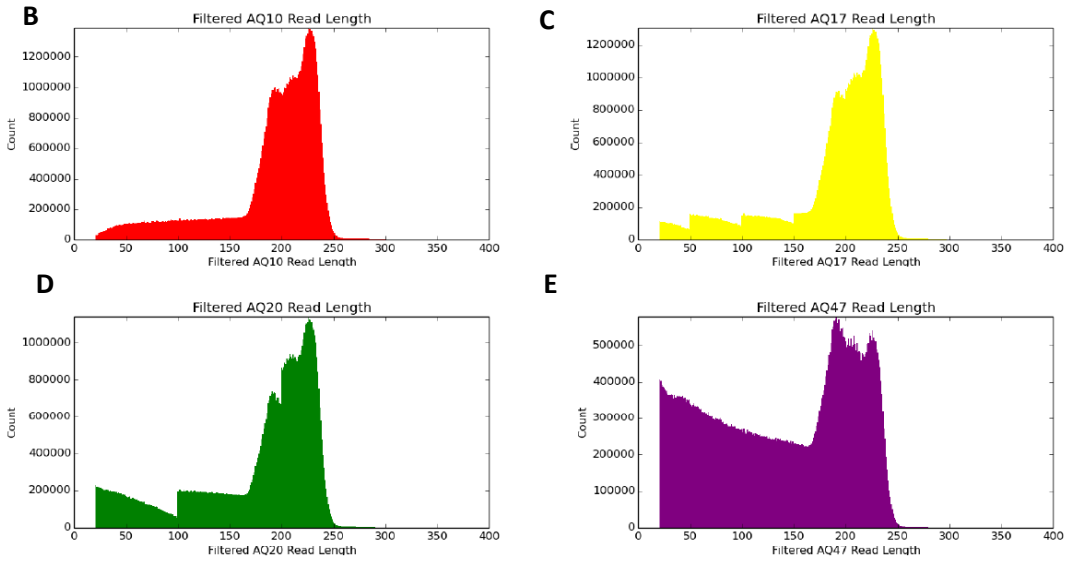


Figure 6. Quality control and run statistics post-alignment. This figure is a collection of the quality control and run statistics of Chip 3 (containing Sample 7 and the Control sample) post-alignment to a reference. **A**: Numerical representation of the alignment quality. Statistics provided at three different quality intervals: AQ17*, AQ20* and Perfect. **B**: Filtered read length histogram after alignment with a quality of 10 (AQ10*). **C**: Filtered read length histogram after alignment with a quality of 17 (AQ17*). **D**: Filtered read length histogram after alignment with a quality of 20 (AQ20*). **E**: Filtered read length histogram after alignment with a quality score of 47 (AQ47*).
*Quality scores provided are based on Phred scores.

The overall NGS run statistics for all four chips used can be seen below in Table 7. Two samples were loaded per chip by the Ion Chef™ and a high loading percentage can be seen across all four chips. The clonality was consistent across all four chips as well as the percentage of library consisting of Test Fragments. Low quality, however, is shown to have doubled in chips 3 and 4 in comparison to chips 1 and 2, including the percentage of adaptor dimers identified.

Table 7: NGS sequencing run statistics

Chip/Run	Samples loaded on chip	Chip Loading %	Enrichment %	Clonal %	Polyclonal %	Final Library %	Test Fragment %	Adapter Dimer %	Low-Quality %
1	5 and 6	90	100	78	22	84	2	1	13
2	7 and control	88	100	78	22	85	2	1	12
3	1 and 3	84	100	80	20	68	2	2	28
4	2 and 4	84	100	79	21	73	2	2	23

The NGS run statistics of each sample are further broken down in Table 8. The mapped reads per sample vary between 15 and 60 million reads. With samples of lower mapped reads, such as Sample 2 and 3, showing a decrease in mean read depth and a sampler BAM file size (GB). The mean read length is consistent aligning to the 200bp chemistry used by the Ion Torrent™ exome platform, as well as the percentage of reads on target and the uniformity of the data.

Table 8: Individual NGS sample sequencing statistics

Sample	Mapped Reads	Mean Read Length	On Target	Mean Read Depth	Uniformity	BAM File Size (GB)
1	50,309,845	187bp	92,85%	151,2	94,97%	32,04
2	15,055,080	193bp	95,09%	47,18	94,18%	9,59
3	16,938,249	189bp	92,77%	51,42	94,87%	10,67
4	56,585,087	189bp	93,97%	173,6	95,19%	36,18
5	26,673,505	192bp	94,72%	83,41	94,44%	17.37
6	60,666,515	190bp	93,75%	186,3	95,02%	38,96
7	41,685,579	191bp	94,86%	129,8	92,67%	27,2
Control	45,097,195	186bp	93,09%	135,3	94,99%	28,92

3.1.1 Coverage Analysis Plugin Results

The report generated by the Coverage Analysis Plugin contains some of the results generated by the ‘Run Report’ discussed above, however, it is more specifically focussed on each sample sequenced. Such information includes the base read depth coverage across each chromosome (Figure 7A); the average amplicon reads per primer pool (Figure 7B); and the number of amplicons according to increasing amplicon C/G base content (Figure 7C). From these graphs, it is seen that the base

read depth is relatively constant with chromosome 19 showing a higher than average read depth, and chromosomes 13 and 18 show less than average coverage. The mean amplicon reads were consistent across primer pools. In the third graph we can see a typical distribution of the number of amplicons according to an increasing amplicon C/G base content, with only minimal amplicon failure and amplicon drop out only identifiable at a high C/G base content (≥75%).

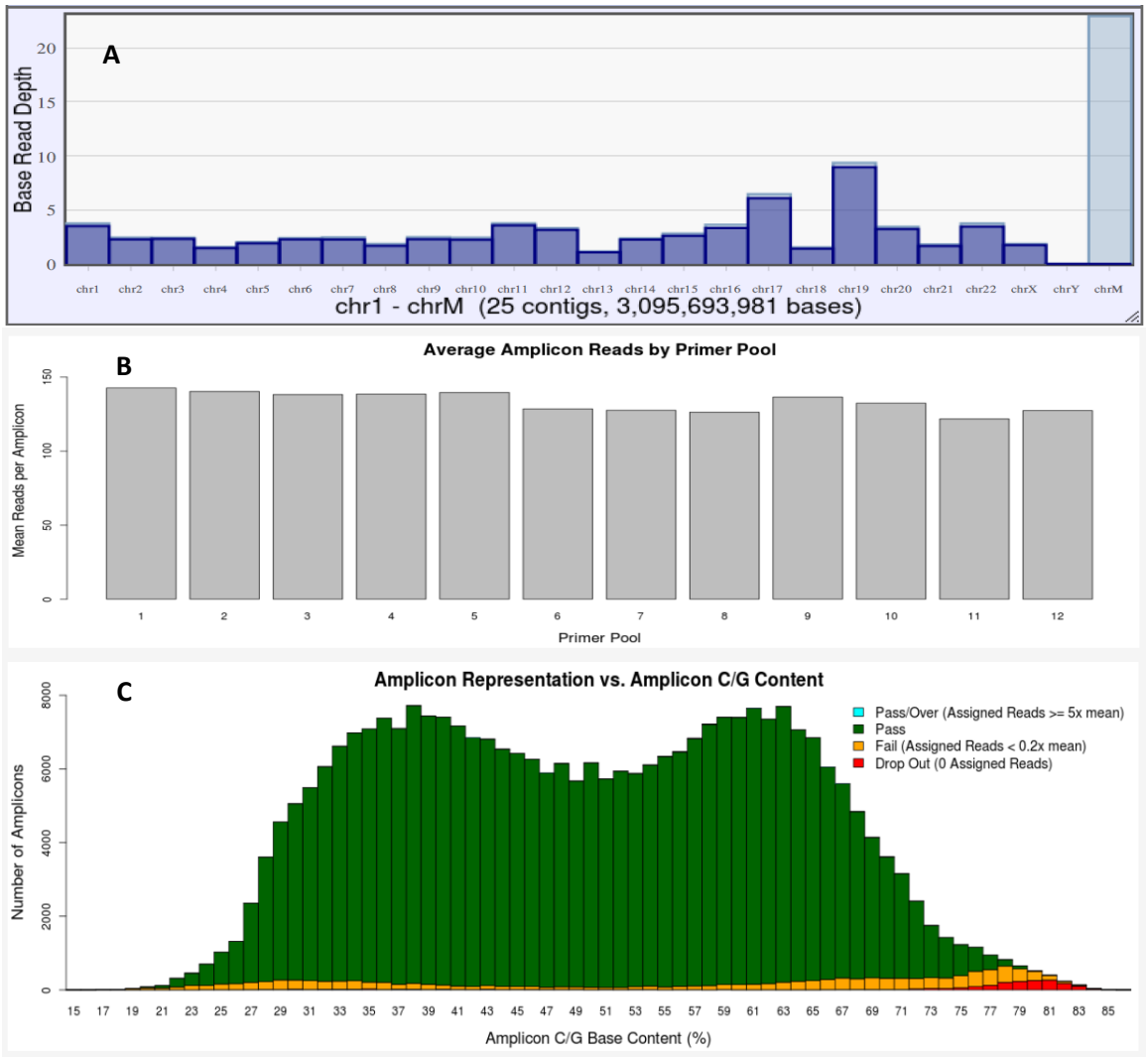


Figure 7. Graphical results of the Coverage Analysis Plugin performed on the control sample. **A**: Bar graph of the base read depth across human chromosomes 1-22 and chromosomes X and Y. Dark blue bars show actual base read depth, light blue bars show the expected base read depth according to the reference. ChrM represents the mitochondrial genome (was not sequenced). **B**: Bar graph illustrating the average amplicon reads across 12 primer pools used in the Ion Torrent™ Exome RDY Kit. **C**: Bar graph illustrating the number of amplicons according to increasing amplicon C/G Base content (%), passed amplicons are shown in green, failed amplicons are shown in yellow and drop-out amplicons are shown in red.

3.2 Whole Exome Sequencing Laboratory Work Timeline

The process of library preparation through to sequencing completion took approximately 50 hours. A breakdown of the time taken to complete each step can be seen in Table 9 below. The laboratory work was conducted across four 12-hour workdays but could be extended over more days when working less hours.

Table 9: Approximate duration of whole exome sequencing procedures from library preparation to sequencing of a 1X8 reaction plate.

	Procedure	**Duration**	**Repeated**	**Total Duration**
Hands-on	Library Preparation	18hrs	-	18hrs
	Quantification; dilution; combining libraries	1hr	X2	2hrs
	Creating a planned run	20min		40min
Automated	Library templating and chip loading	8-12hrs		16-24hrs
	Sequencer initialisation	1hr		2hrs
	Sequencing	2.5-3hrs	X4	10-12hrs
Total duration of whole exome sequencing in hours				48.6-58.6hrs
Total duration of whole exome sequencing in 12-hour work days				±4 days

3.3 Data Analysis Platform Performance

Three alternative data analysis methods were used in this project and the inputs, outputs, and performance of each method are compared in Table 10 below. The first-hand interaction with all three analysis platforms allowed the design of a recommended WES pipeline for the investigation of possible putative-disease causing mutations in South African DD patients, which is outlined in Figure 8.

Table 10: Comparison of data analysis platforms.

	Moon (Diploid) Software	**Custom Pipeline**	**VarAFT**
Hardware and software requirements	Online web-based software	Program made from BASH scripting run on a computer cluster/server	Downloadable desktop application with supporting software downloads
Data Security	Password protected (software account)	Password protected with secure server access control	Password protected (computer-user account)
Data accessibility	Remote access	Remote access	On-site access
Clinical information used to inform prioritisation	Combination of multiple HPO terms illustrating patient's phenotype	G2P list of 2500 known associated DD genes	Gene and disease phenotype information
Software user requires training prior to use	No	Yes	No
Linux coding skills required	No	Yes	No
Capable of identifying compound heterozygotes	Yes	Yes	Yes
Quality-based filter parameters available	Yes	Yes	No
Number of prioritised variants per sample post-analysis	±20	±30	±300
Time taken for analysis per sample*	2 minutes	1 day	2 hours
Cost of analysis per sample	US$49	N/A	US$0
Number of patients in which a prioritised/causative variant was identified	86% (6/7)	71% (5/7)	29% (2/7)

* The time estimates in the table for analysis did not include the manual analysis of prioritised variants.

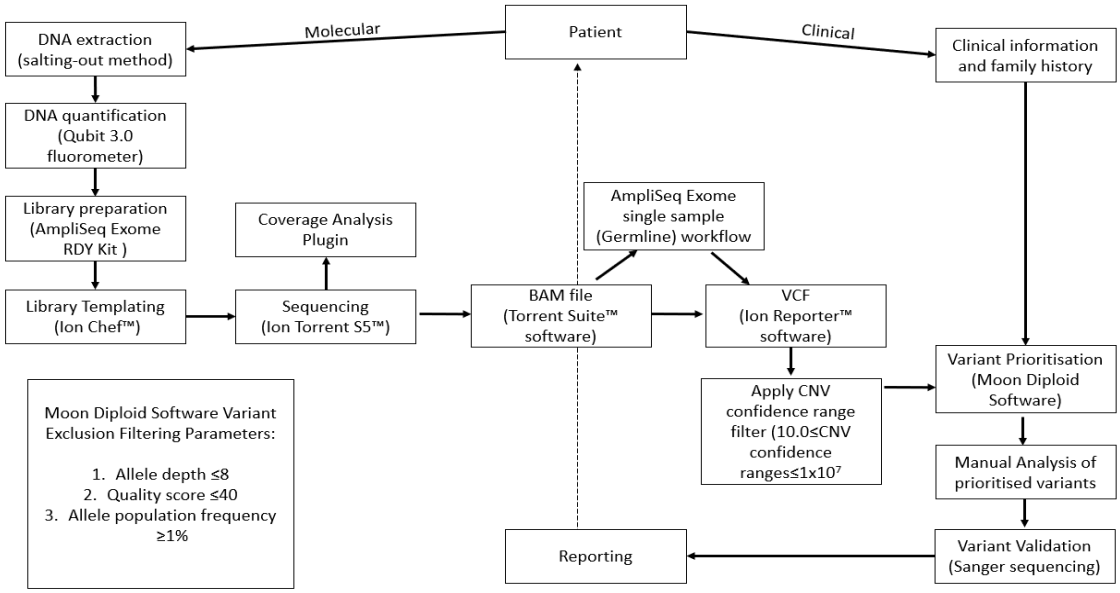


Figure 8. Whole exome sequencing pipeline for the genetic testing of patients with developmental disorders.

3.4 Patient Results

3.4.1 Patient 1

i. Clinical Description

This is a 9-year-old, female patient who was born to a phenotypically normal, non-consanguineous couple. Ultrasounds performed during the pregnancy found no abnormalities, but a complication of threatened miscarriage was noted. The child was born at full gestation through caesarean section and was postnatally admitted for two weeks due to congenital cardiac defects. Prior to enrolment in the DDD-Africa study, karyotyping, MLPA, FISH (22q11), and microarray diagnostic tests were performed, yielding normal results. The reported system anomalies identified in this patient and their associated HPO codes are shown in Table 11. Neurodevelopment assessment identified inappropriate fine motor development, abnormal behaviour (occasionally aggressive and hyperactive), and moderate intellectual disability. A CT-scan was performed but found to be normal. No clinical diagnosis was provided, but the suggested working differential were ‘Turner like’ and Noonan syndrome noting that the facial features were not typical of either of these syndromes.

Table 11. System anomalies recorded for Patient 1

HPO Term	HPO ID
Thickened nuchal skin	HP:0000286
Short nose	HP:0003196
Anteverted nares	HP:0000463
Low-set posteriorly rotated ears	HP:000368
Short neck	HP:0000470
Epicanthus	HP:0000286
Generalised Hypotonia	HP:0001290
Transposition of the great arteries	HP:0001669
Atrial septal defect	HP:0001631
Ventricular septal defect	HP:0001629
Pulmonic stenosis	HP:0001642
Wide intermammillary distance	HP:0006610
Blue nevus	HP:0100814
Sparse scalp hair	HP:0002209

ii. Variant Prioritisation

The results of the Moon (Diploid) software for Patient 1 prioritized a missense mutation in the Cell Division Control Protein 42 Homolog (*CDC42)* gene. The classification of this variant can be seen in Table 12 and the variant information is summarised in Appendix 6. The variant was classified as likely pathogenic due to three or more moderate criteria supporting pathogenicity. The corresponding possible diagnosis (as suggested by Moon) was Takenouchi-Kosaki syndrome (OMIM:616737), which has overlapping clinical features with those seen in this patient.

Table 12: ACMG classification of *CDC42* c.68A>G variant.

ACMG Code	**Reason**
PS3	Pathogenic Supporting. A well-conducted functional study shows a deleterious effect (Martinelli et al., 2018). Downgraded to supporting evidence based on the fact that the study has not been replicated (Brnich et al., 2020).
PM2	Pathogenic Moderate. Variant not found in GnomAD exomes 15/11/2019.
PM1	Pathogenic Moderate. Variant located in the N-terminal α Helix (Martinelli et al., 2018) of the Ras domain (Pfam 11/02/2020).
PM6	Pathogenic Moderate. *De novo*, without maternity and paternity, confirmed. Determined with Sanger sequencing.
PP2	Pathogenic Supporting. 11 non-VUS missense variants in this gene are pathogenic. (Varsome- 15/11/2019).
PP3	Pathogenic Supporting. 10/14 pathogenicity predictors determined the variant to be pathogenic/deleterious. (Moon- 15/11/2019).
PP5	Pathogenic Supporting. ClinVar classifies the variant as pathogenic 15/11/2019.

This variant was covered 120 times during WES and can be seen visually through the Ion Reporter Genomic Viewer (IRGV) screenshot below in Figure 9.

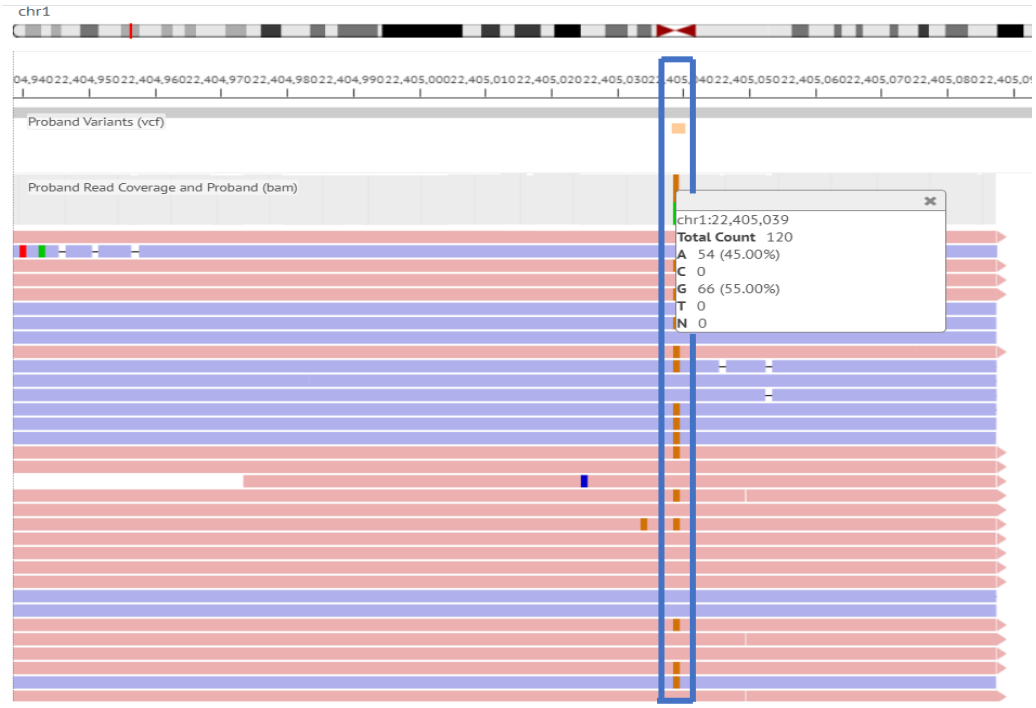


Figure 9. IRGV image of *CDC42* variant found in Patient 1. This image shows a screenshot of the IRGV output illustrating the *CDC42* c.68A>G variant on chromosome 1 that was identified in Patient 1. Pink tracks represent forward strands and blue tracks represent reverse strands. The location was covered 120 times; an A allele was recorded 54 times and a G allele was recorded 66 times. This shows Patient 1 to be an A/G heterozygote at this position.

The custom pipeline results for this patient identified variants in 23 genes, see Appendix 7 (Table A7A). After manual analysis, only one variant was prioritised with a likely pathogenic status in the Lysine Methyltransferase 2D (*KMT2D)* gene, where the suggested disease phenotype was somewhat similar to that of the patient. Further investigation of this variant did not provide sufficient support in pathogenicity; thus the variant was excluded as being causative. The variant identified by Moon (*CDC42* c.68A>G) for this patient was filtered-out in the custom pipeline results as the specific gene was not listed on the DD-G2P list of 2500 associated DD genes used for filtering. None of the variants identified by the custom pipeline were taken forward for validation in this patient.

iii. Variant Validation

The *CDC42* variant identified in this patient was selected for variant validation as the suggested phenotype was most closely matched to the phenotype of the patient. The results of the validation can be seen below in Figure 10. The parents of the patient were also sequenced for this variant and were found to be homozygous for the wild-type allele (A) indicating a *de novo* mutation in the child.

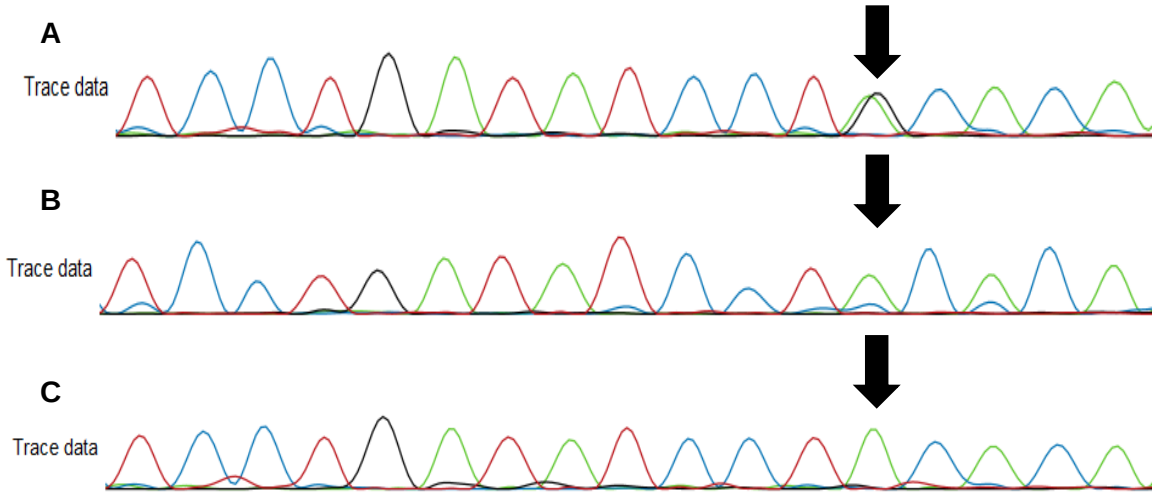


Figure 10. *CDC42* c.68A>G variant validation through Sanger sequencing. **A**: Electropherogram identifying Patient 1 as an A/G heterozygote (variant is present). **B**: Electropherogram identifying Patient 1's mother as an A/A homozygote (variant is absent). **C**: Electropherogram identifying Patient 1's father as an A/A homozygote (variant is absent). Variant location is indicated by the black arrows.

3.4.2 Patient 2

i. Clinical Description

A 6-year-old, male patient born to non-consanguineous parents. The father was unavailable for recruitment, thus only the child and his mother were involved in the study. The mother reported being concerned about decreased foetal movements during the pregnancy, however, all prenatal scans performed returned normal. The child was born through normal vaginal delivery, without complications, at 36 weeks' gestation, and no congenital abnormalities were identified. Neurodevelopment assessment identified delayed gross motor development; delayed speech and language; and moderate intellectual disability in the child. System anomalies that were

later identified are shown in Table 13. A CT-scan was performed and found to be normal. Prior to recruitment in the DDD-Africa study, karyotyping, microarray, FISH and MLPA were performed but returned normal results. Clinicians suspected possible Williams syndrome; however, the test results were negative.

Table 13. System anomalies recorded for Patient 2

HPO Term	HPO ID
Prominent metopic ridge	HP:0005487
Down slanted palpebral fissures	HP:000494
Epicanthus	HP:0000286
Depressed nasal bridge	HP:0005280
Anteverted nares	HP:0000463
Micrognathia	HP:0000347
Abnormality of the helix	HP:0011039
Postnatal microcephaly	HP:0005484
Brisk reflexes	HP:0001348
Generalised hypotonia	HP:0001290
Umbilical hernia	HP:0001537
Cryptorchidism	HP:0000028
Decreased palmar creases	HP:0006184
Café-au-lait spot	HP:0000957

ii. Variant Prioritisation

The Moon (Diploid) software identified a missense mutation in the Casitas B-lineage Lymphoma (*CBL*) gene in Patient 2. The classification criteria applied to this variant can be seen in Table 14 and detailed variant information is summarised in Appendix 6. The variant was classified as a VUS due to contradictory supporting criteria and thus was not investigated further.

Table 14: ACMG classification of *CBL* c.2520T>G variant.

ACMG Code	**Reason**
PP2	Pathogenic Supporting. 39/50 non VUS variants in this gene are pathogenic (Varsome 18/11/2019).
PP1	Pathogenic Supporting. Co-segregation with disease reported by Tekendo-Ngogang, et al., 2019.
BS2	Benign Strong. Observed in healthy individuals, GnomAD allele count=8. (1 African individual reported 02/12/2019)
BP4	Benign Supporting. 9/14 pathogenicity predictors reported the variant as benign/neutral (Moon 02/12/2019).

The custom pipeline results for this patient identified variants in 18 genes, see Appendix 7 (Table A7B). After manual analysis no variants were prioritised further as

all were identified as VUS (without enough evidence to support pathogenicity or further investigation), likely benign or benign.

iii. Variant Validation

No variants were selected for validation for this patient.

3.4.3 Patient 3

i. Clinical Description

A 14-year-old, male patient born to a phenotypically normal, non-consanguineous couple. Reduced foetal movement was noted during prenatal scans, and the child was born through normal vaginal delivery, without complications, at 36 weeks' gestation presenting with a cleft palate. The patient reportedly had a twin brother who was stillborn from an unknown cause, however, the mother did not report any abnormalities in the stillborn. Neurodevelopmental assessment identified profoundly delayed gross motor development, fine motor development, speech and language, and severe intellectual disability in the child. Social behaviour was described as inappropriate along with abnormal behaviour, specifically sleep disturbance. System anomalies are reported in Table 15. Prior to recruitment in the DDD-Africa study, MLPA and karyotyping were conducted and yielded normal results.

Table 15. System anomalies recorded for Patient 3

HPO Term	HPO ID
Microcephaly	HP:000252
Microtia	HP:0008551
Abnormality of the pinna	HP:0000377
Cleft palate	HP:0000175
Microdontia	HP:0000691
Wide nasal bridge	HP:0000431
Underdeveloped nasal alae	HP:0000430
Coarse facial features	HP:0000280
Seizures	HP:0001250
Brain atrophy	HP:0012444
Cryptorchidism	HP:0000028
Unilateral conductive hearing impairment	HP:0040119

ii. Variant Prioritisation

The Moon (Diploid) software identified a stop gain mutation in the Elongation Factor Tu GTP Binding Domain Containing 2 (*EFTUD2*) gene in Patient 3. The classification of this variant can be seen in Table 16 and the variant information is summarised in

Appendix 6. The variant was classified as pathogenic due to one very strong and two moderate pathogenicity criteria. The possible diagnosis (as suggested by Moon) was Mandibulofacial dysostosis with microcephaly (OMIM:610536).

Table 16: AMCG classification of *EFTUD2* c.1732C>T variant.

ACMG Code	Reason
PVS1	Pathogenic Very Strong. Null variant in gene, which is a known mechanism of disease associated with Mandibulofacial dysostosis. (Varome-15/11/2019).
PM2	Pathogenic Moderate. Variant not found in GnomAD-15/11/2019.
PM6	Pathogenic Moderate. *De novo*, without maternity and paternity confirmed. Determined with Sanger Sequencing.
PP3	Pathogenic Supporting. 11/13 pathogenicity predictors identify the variant as pathogenic/deleterious. (Moon-15/11/2019).
PP5	Pathogenic Supporting. ClinVar classifies the variant as pathogenic 15/11/2019.

As a stop gain the variant would result in a truncated protein, this variant was viewed in Mutalyzer to examine the effect on the protein. The results can be seen in Figure 11 below.

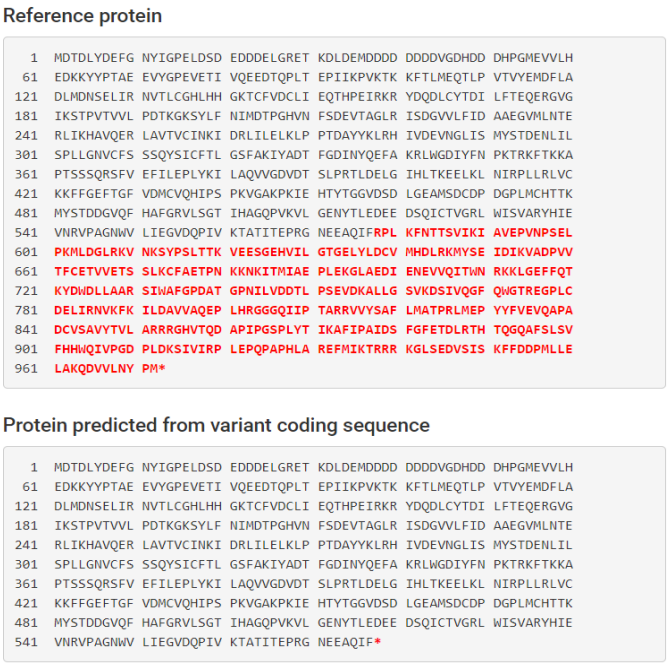


Figure 11. Screenshot of Mutalyzer results for *EFTUD2* c.1732C>T variant.

The reference protein shows a protein of 972 amino acids in length. The letters in red represent the amino acids affected by the variant. The protein predicted from the variant coding sequence shows a protein of only 577 amino acids in length.

This variant was covered 52 times during WES and can be seen visually in Figure 12 below.

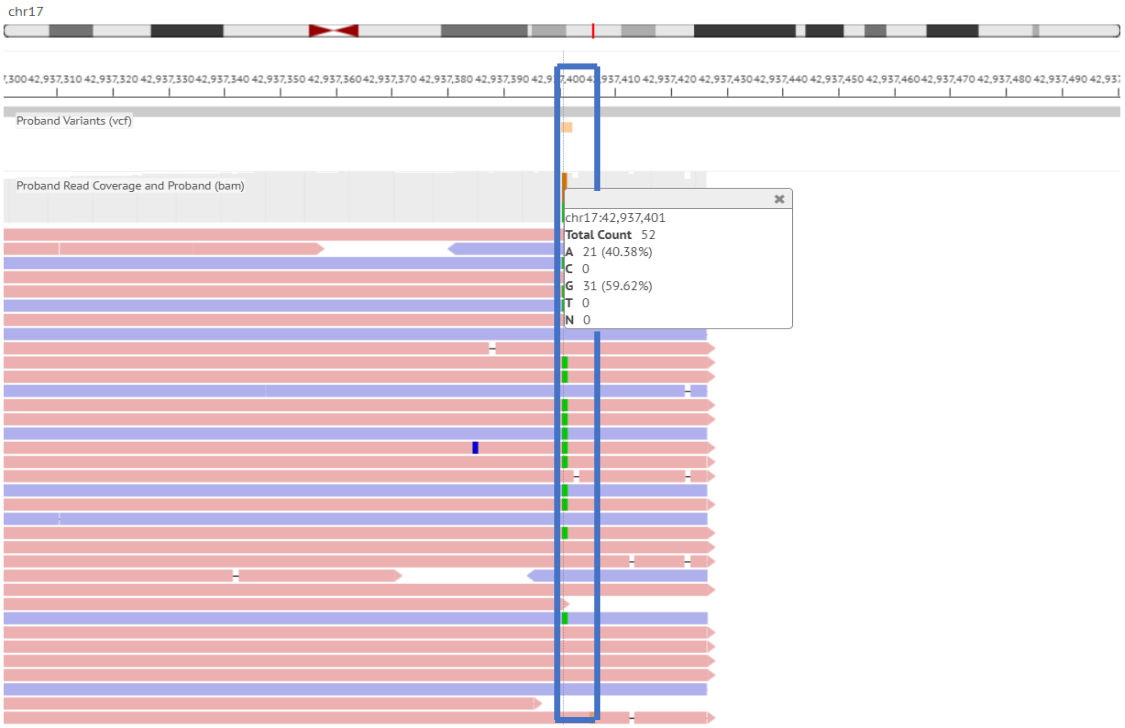


Figure 12. IRGV image of *EFTUD2* variant found in Patient 3. This image shows a screenshot of the IRGV output illustrating the *EFTUD2* c.1732C>T variant on chromosome 17 that was identified in Patient 3. Pink tracks represent forward strands and blue tracks represent reverse strands. The location was covered 52 times; an A allele was recorded 21 times and a G allele was recorded 31 times. This shows Patient 3 to be an A/G heterozygote at this position.

The custom pipeline results for this patient showed variants identified in 37 genes, see Appendix 7 (Table A7C). The same variant was identified in the *EFTUD2* gene as found by Moon. An additional variant identified in the Treacle Ribosome Biogenesis Factor 1 (*TCOF1*) gene was initially prioritised, but after a thorough investigation it was determined to be an indel in a homopolymer and was excluded. All other variants were excluded after being identified as VUS, likely benign or benign.

iii. Variant Validation

The *EFTUD2* variant identified in this patient was selected for variant validation using Sanger sequencing. The results can be seen below in Figure 13 below. The parents of this patient were also sequencing and found to be homozygous for the wild-type allele (C), indicating a *de novo* mutation within the child.

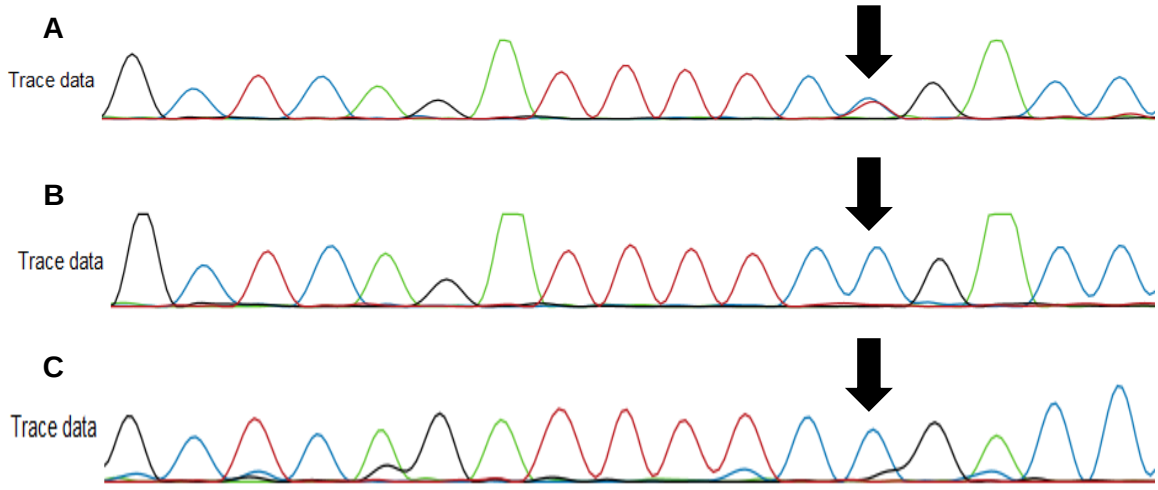


Figure 13. *EFTUD2* c.1732C>T variant validation through Sanger sequencing. **A**: Electropherogram identifying Patient 3 as a C/T heterozygote (variant is present). **B**: Electropherogram identifying Patient 3's mother as a C/C homozygote (variant is absent). **C**: Electropherogram identifying Patient 3's father as a C/C homozygote (variant is absent). Variant location is indicated by the black arrows.

3.4.4 Patients 4 and 5

Patients 4 and 5 are presented together as they are fraternal twins and the same possible putative disease-causing variants presented here were found in both patients.

i. Patients 4 and 5 Clinical Descriptions

Two 4-year-old males, born to a phenotypically normal non-consanguineous couple. The father was unavailable for recruitment, thus, only the mother was included. No abnormalities were detected in ultrasounds during the pregnancy, and a caesarean section was performed at full gestation, without complications. Neurodevelopmental assessment of both patients identified delayed motor development; speech and language; and inappropriate social behaviour. The overall assessment determined severe intellectual disability in both patients. The system anomalies of Patient 4 can be seen in Table 17 and the anomalies of Patient 5 can be seen in Table 18. Prior to recruitment for the DDD-Africa study, MLPA was performed but identified no anomalies. The referring clinicians raised the concern of possible congenital myopathy and suggested a differential diagnosis of congenital muscular dystrophy. A creatinine kinase test was requested and showed elevated levels in both patients' indicative of muscle damage. A CT scan (performed in 2016) showed global atrophy of the cerebral cortex.

Table 17. System anomalies recorded for Patient 4

HPO Term	HPO ID
Microcephaly	HP:000252
Sparse scalp hair	HP:0002209
Synophrys	HP:0000664
Facial hirsutism	HP:0009937
Bitemporal hollowing	HP:0025386
Curly eyelashes	HP:0007665
Upslanted palpebral fissure	HP:0000582
Muscular hypotonia	HP:0001258
Lower limb hypertonia	HP:0006895
Skeletal muscle atrophy	HP:0003202
Knee flexion contracture	HP:0006380
Hip contracture	HP:0003273
Ankle contracture	HP:0006466
Slender finger	HP:0001238
Blue nevus	HP:0100814
Café-au-lait spot	HP:0000957

Table 18. System anomalies recorded for Patient 5

HPO Term	HPO ID
Microcephaly	HP:000252
Upslanted palpebral fissure	HP:0000582
Narrow forehead	HP:0000341
Frontal hirsutism	HP:0011335
Exaggerated cupid’s brow	HP:0002263
Reduced tendon reflexes	HP:0001315
Muscle flaccidity	HP:0010547

As fraternal twins, patients four and five are expected to share 50% of their DNA the same as ‘normal’ siblings. The number of variants shared between these Patients and the number of unique variants can be seen in Figure 14.

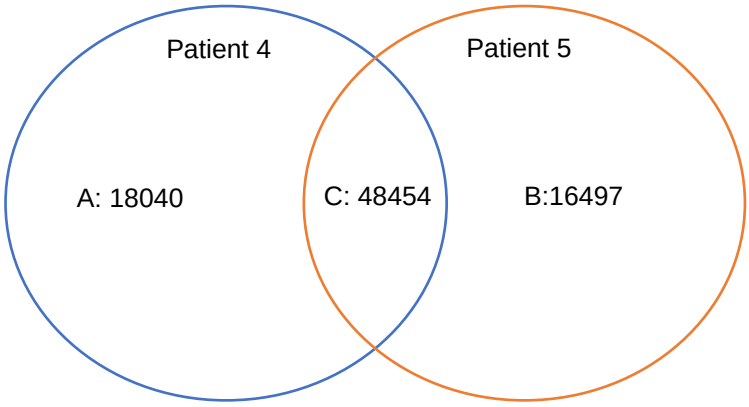


Figure 14. Venn diagram illustrating the shared and unique variants of patients 4 and 5 (fraternal twins). **A**: Variants unique to Patient 4 (18040). **B**: Variants unique to Patient 5 (16497). **C**: Variants shared by both patients (48454).

ii. Variant Prioritisation

The Moon (Diploid) Software identified two missense mutations in the Protein O-mannosyl-transferase 2 (*POMT2*) gene of Patients 4 and 5 presenting as a compound heterozygote. The classification of these variants can be seen in Tables 19 and 20 and the specific variant information is shown in Appendix 6. Both variants are classified as likely pathogenic due to two moderate and two supporting pathogenicity criteria. The possible diagnosis (as suggested by Moon) is Muscular Dystrophy-Dystroglycanopathy with Mental Retardation, type B2 (OMIM:613156).

Table 19: ACMG classification of *POMT2* c.2232G>T variant.

ACMG Code	**Reason**
PM1	Pathogenic Moderate. Hot-spot, length of 61 bases with 6 known pathogenic variants and 0 benign (Varsome 18/11/2019). Variant lies within the PMT 4TMC domain (Pfam 04/12/2019).
PM2	Pathogenic Moderate. Variant not found in GnomAD 18/11/2019.
PP3	Pathogenic Supporting. 14/14 pathogenicity predictors support pathogenic/deleterious effect (Moon 18/11/2019).
PP4	Pathogenic Supporting. Patient's phenotype or family history is highly specific for the gene (applied at the discretion of the referring clinician).

Table 20: ACMG classification of *POMT2* c.1424G>C variant.

ACMG Code	Reason
PM1	Pathogenic Moderate. Variant lies within the MIR3 domain (Pfam 04/12/2019).
PM2	Pathogenic Moderate. Variant not found in GnomAD 18/11/2019.
PP3	Pathogenic Supporting. 14/14 pathogenicity predictors support pathogenic/deleterious effect (Moon 18/11/2019).
PP4	Pathogenic Supporting. Patient's phenotype or family history is highly specific for the gene (applied at the discretion of the referring clinician).

The *POMT2* c.1424G>C variant was covered 156 times in Patient 4 and 67 times in Patient 5. The *POMT2* c.2232G>T variant was covered 161 times in Patient 4 and 49 times in Patient 5 through WES. These variants were visualised and IRGV screenshots can be seen in Figures 15 (Patient 4) and 16 (Patient 5).

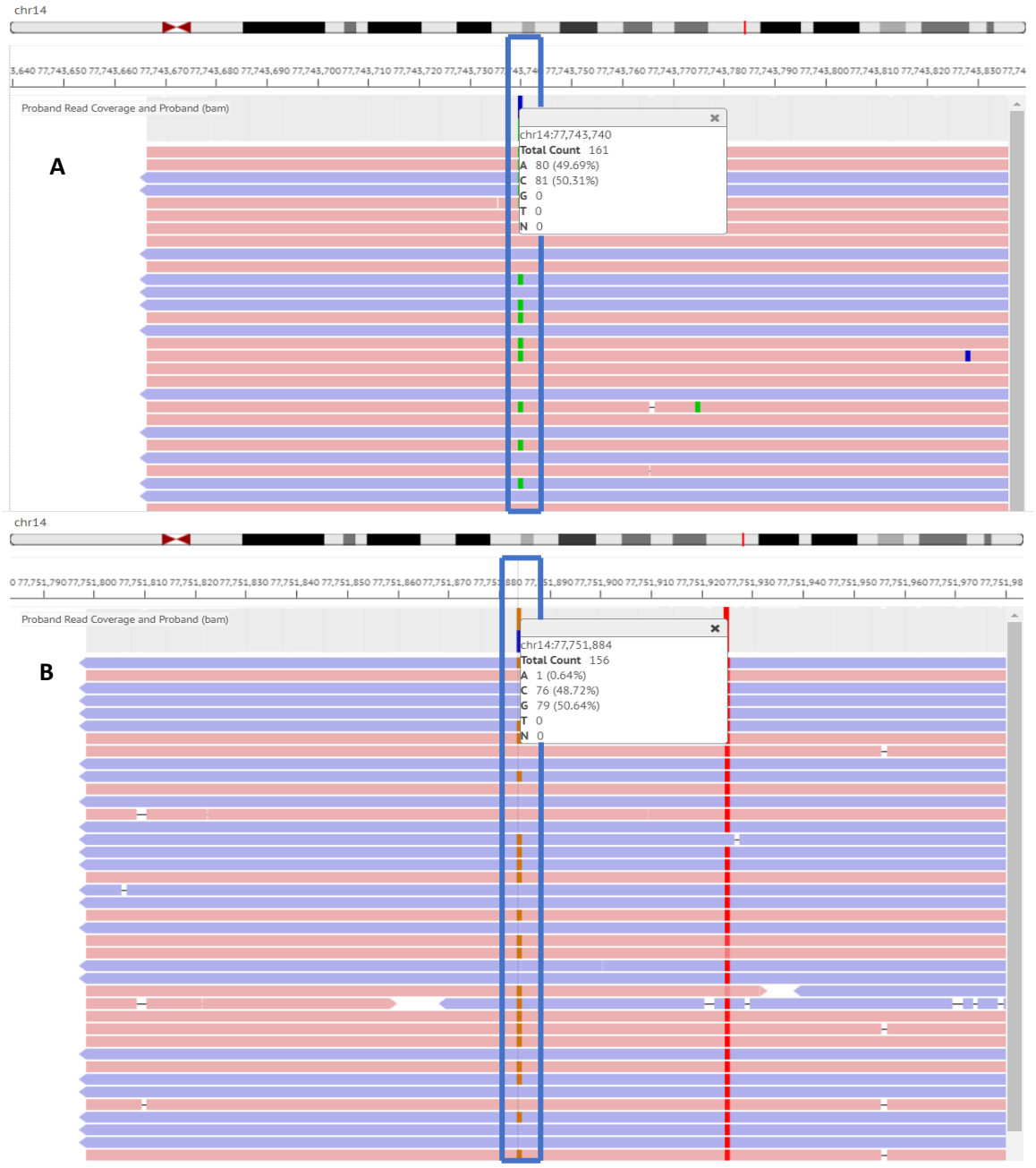


Figure 15. IRGV image of *POMT2* variants found in Patient 4. This image shows two screenshots of the IRGV outputs illustrating the *POMT2* **A**: c.2232G>T and **B**: c.1424G>C variants on chromosome 14 that were identified in Patient 4. Pink tracks represent forward strands and blue tracks represent reverse strands. The first location **A** was covered 161 times; an A allele was recorded 80 times and a C allele was recorded 81 times. This shows Patient 4 to be an A/C heterozygote at this position. The second location **B** was covered 156 times; a C allele was recorded 76 times and a G allele was recorded 79 times. An A allele was erroneously recorded once, thus Patient 4 is a C/G heterozygote at this position.

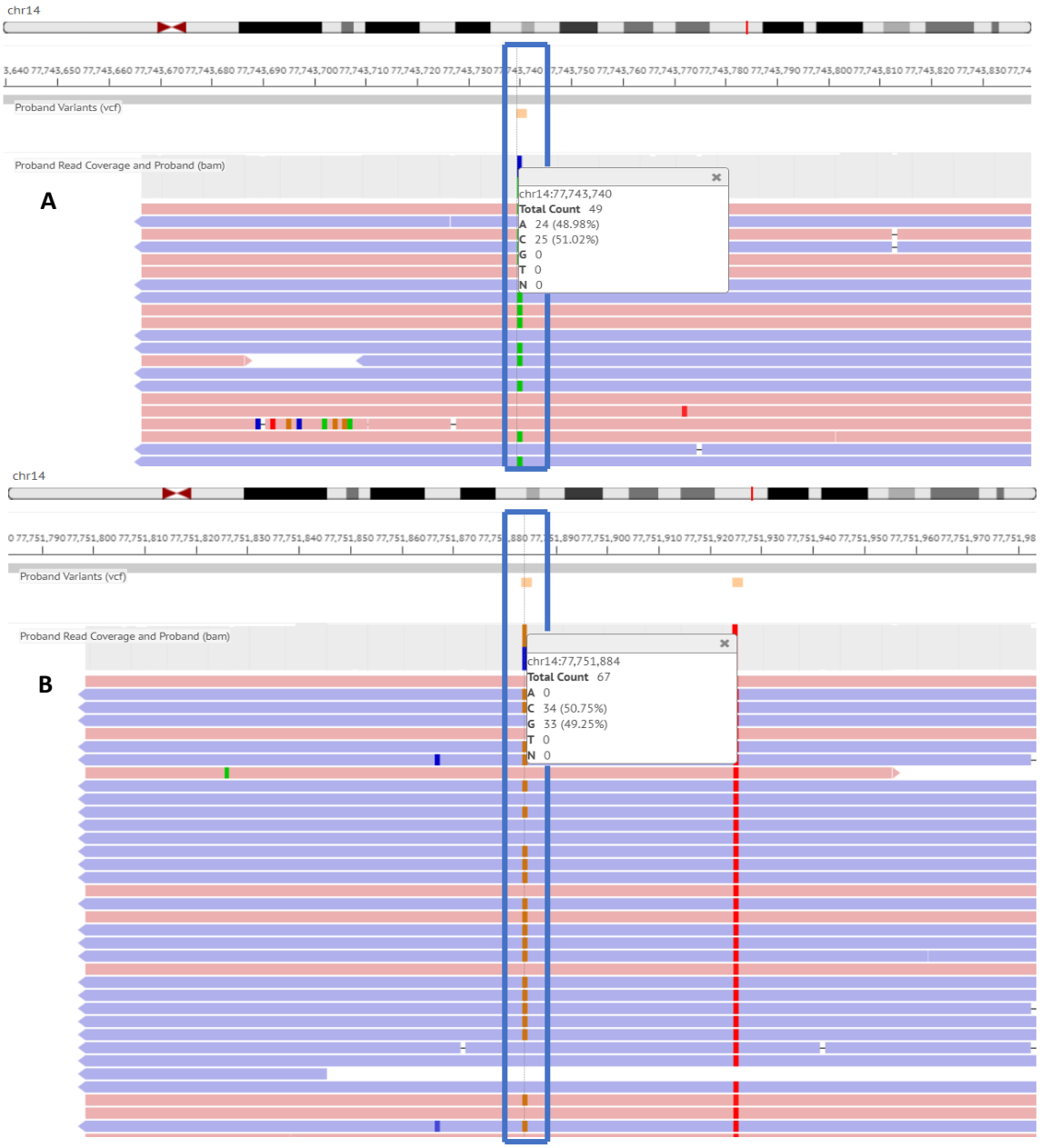


Figure 16. IRGV image of *POMT2* variants found in Patient 5. This image shows two screenshots of the IRGV outputs illustrating the *POMT2* **A**: c.2232G>T and **B**: c.1424G>C variants on chromosome 14 that were identified in Patient 5. Pink tracks represent forward strands and blue tracks represent reverse strands. The first location **A** was covered 49 times; an A allele was recorded 24 times and a C allele was recorded 25 times. This shows Patient 5 to be an A/C heterozygote at this position. The second location **B** was covered 67 times; a C allele was recorded 34 times and a G allele was recorded 33 times. This shows Patient 5 to be a C/G heterozygote at this position.

The custom pipeline results showed variants in 18 genes in Patient 4 see Appendix 7 (Table A7D) and in 19 genes in Patient 5 (Table A7E). The *POMT2* variants were found in both patients in the custom pipeline results and all other variants identified were excluded as VUS, benign, or likely benign.

iii. Variant Validation

Both *POMT2* variants were selected for validation through Sanger sequencing. The results of this sequencing can be seen below in Figure 17. Only the maternal sample of these patients was available and sequenced for both variants. The maternal sample was heterozygote for the one variant (c.1424G>C) but a wild-type homozygote for the second variant (c.2232G>T). Patients 4 and 5 were heterozygote for both variants. From this, we can confirm that the first variant (c.1424G>C) was inherited by the children from the mother.

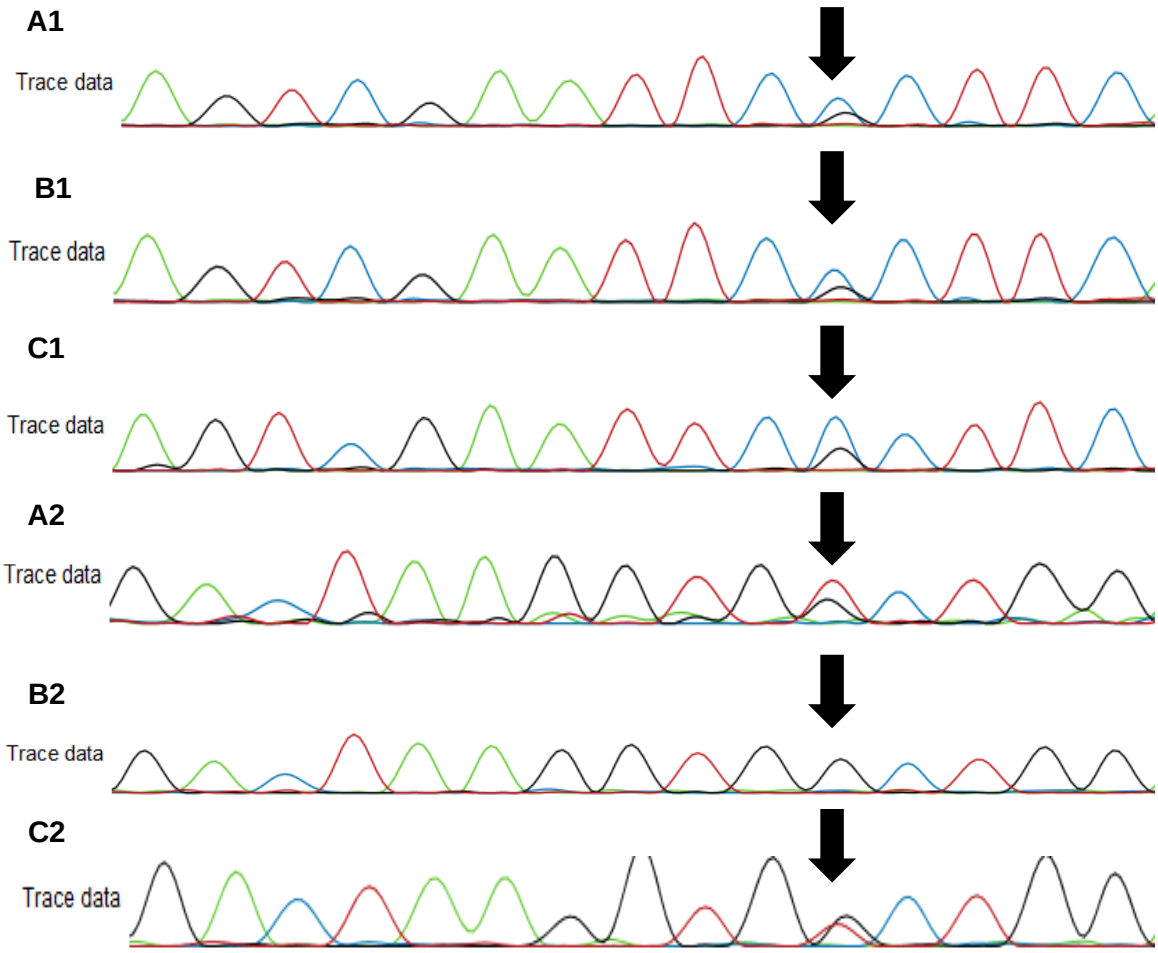


Figure 17. *POMT2* c.1424G>C and c.2232G>T variant validation through Sanger sequencing. **A1**: Electropherogram identifying Patient 4 as a G/C heterozygote at position c.1424 (variant present). **B1**: Electropherogram identified Patient 4 and 5's mother as a G/C heterozygote at position c.1424 (variant present). **C1**: Electropherogram identifying Patient 5 as a G/C heterozygote at position c.1424 (variant present). **A2**: Electropherogram identifying Patient 4 as a G/T heterozygote at position c.2232 (variant present). **B2**: Electropherogram identifying Patient 4 and 5's mother as a G/G homozygote at position c.2232 (variant not present). **C2**: Electropherogram identifying Patient 5 as a G/T heterozygote at position c.2232 (variant present). Variant locations are indicated by the black arrows.

3.4.5 Patient 6

i. Clinical Description

A 13-year-old male born to a phenotypically normal non-consanguineous couple by caesarean section, performed at 34 weeks gestation due to oligohydramnios. The pregnancy was only diagnosed during labour, as the mother did not experience any pregnancy symptoms. The mother was 40-years-old at the time of conception of this pregnancy, and she had been previously diagnosed with hyperthyroidism, for which she was taking neomercazole during pregnancy. Neurodevelopmental assessment

identified inappropriate social behaviour and abnormal behaviour, with an overall conclusion of moderate intellectual disability. System anomalies reported in this patient can be seen in Table 21. Prior to recruitment in the DDD-Africa study, karyotyping and MLPA were performed but returned normal results.

Table 21. System anomalies recorded for Patient 6

HPO Term	HPO ID
Scaphocephaly	HP:0030799
Narrow forehead	HP:0000341
Oval face	HP:0000300
Maxillozygomatic hypoplasia	HP:0005439
High anterior hairline	HP:0009890
Upslanted palpebral fissure	HP:0000582
Epicanthus palpebralis	HP:0031770
Bilateral ptosis	HP:0001488
Proptosis	HP:0000520
Long nose	HP:0003189
Lower limb hyperreflexia	HP:0002395
Hallux valgus	HP:0001822
Tapered finger	HP:0001182

ii. Variant Prioritisation

The Moon (Diploid) software results identified a frameshift deletion in the proto-oncogene B-Raf (*BRAF*) of Patient 6 and a frameshift insertion in the *KMT2D* gene. The classification of these variants can be found in Tables 22 and 23 respectively and the variant information can be found in Appendix 6. Both variants are classified as likely pathogenic due to one very strong and one moderate supporting criteria for each. The associated disorders (as suggested by Moon) are Cardiofaciocutaneous syndrome (OMIM:115150) for the *BRAF* variant and Kabuki syndrome (OMIM:147920) for the *KMT2D* variant.

Table 22: ACMG classification of *BRAF* c. 87_88delCG variant.

ACMG Code	**Reason**
PVS1	Pathogenic Very Strong. Null variant caused by frameshift affecting gene, which is a known mechanism of disease associated to Noonan syndrome, Cardiofaciocutaneous syndrome, and LEOPARD syndrome (Varsome 15/11/19).
PM2	Pathogenic Moderate. Variant not found in GnomAD 15/11/2019.

This variant is a frameshift deletion leading to the insertion of a stop codon which would result in a truncated protein. The variant was examined using Mutalyzer to examine the effect of the variant on the protein and results can be seen in Figure 18 below.

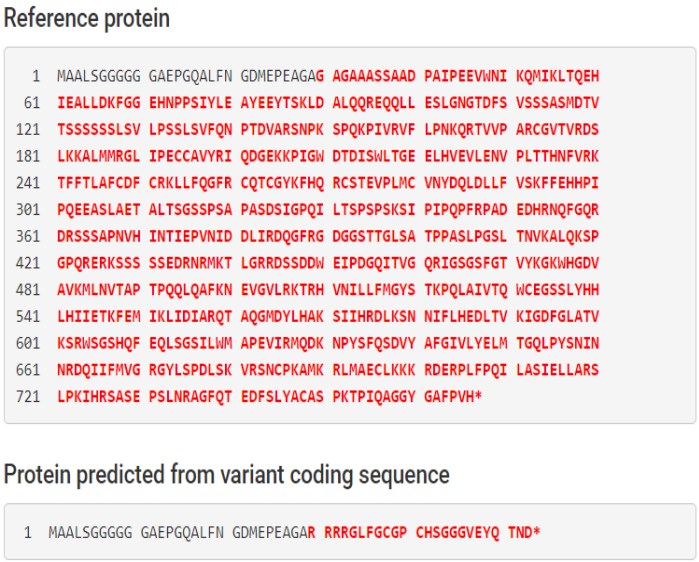


Figure 18. Screenshot of Mutalyzer results showing *BRAF* c.87_88delCG variant.

The reference protein shows a protein of 766 amino acids in length. The letters in red represent the amino acids affected by the variant. The protein predicted from the variant coding sequence shows a protein of only 53 amino acids in length.

Table 23: ACMG classification of *KMT2D* c.2243_2244insC variant.

ACMG Code	Reason
PVS1	Pathogenic Very Strong. Null variant caused by frameshift affecting gene, which is a known mechanism of disease associated to Kabuki Syndrome (Varsome 15/11/2019).
PM2	Pathogenic Moderate. Variant not found in GnomAD (15/11/2019).

This variant is a frameshift insertion which introduces a premature stop codon resulting in a truncated protein. The variant was viewed in Mutalyzer and the results can be seen in Figure 19 below.

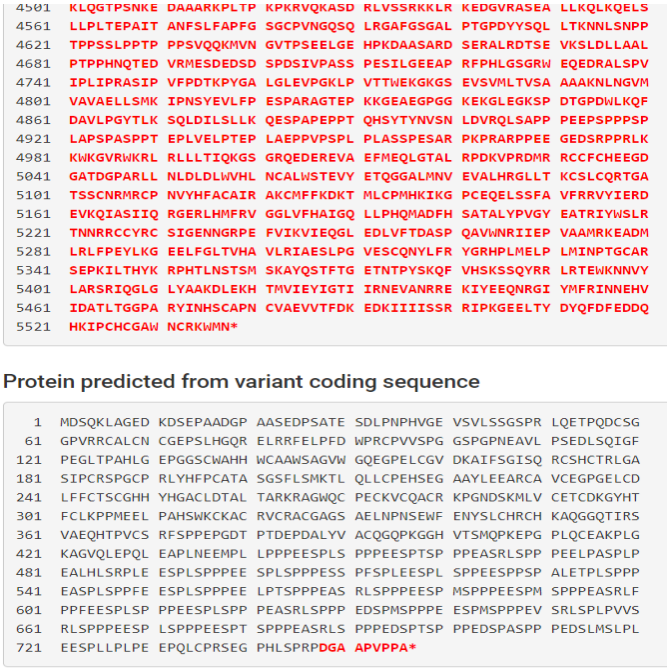


Figure 19. Screenshot of Mutalyzer results of *KMT2D* c.2243_2244insC variant.

The reference protein (top) shows a protein of 5537 amino acids in length. The letters in red represent the amino acids affected by the variant. The protein predicted from the variant coding sequence shows a protein of only 756 amino acids in length. Image begins at amino acid 4501. Due to the size of the protein the complete image could not be included.

The *BRAF* variant was covered 10 times during WES and the *KMT2D* variant was covered 185 times, they can be seen visually below in Figure 20.

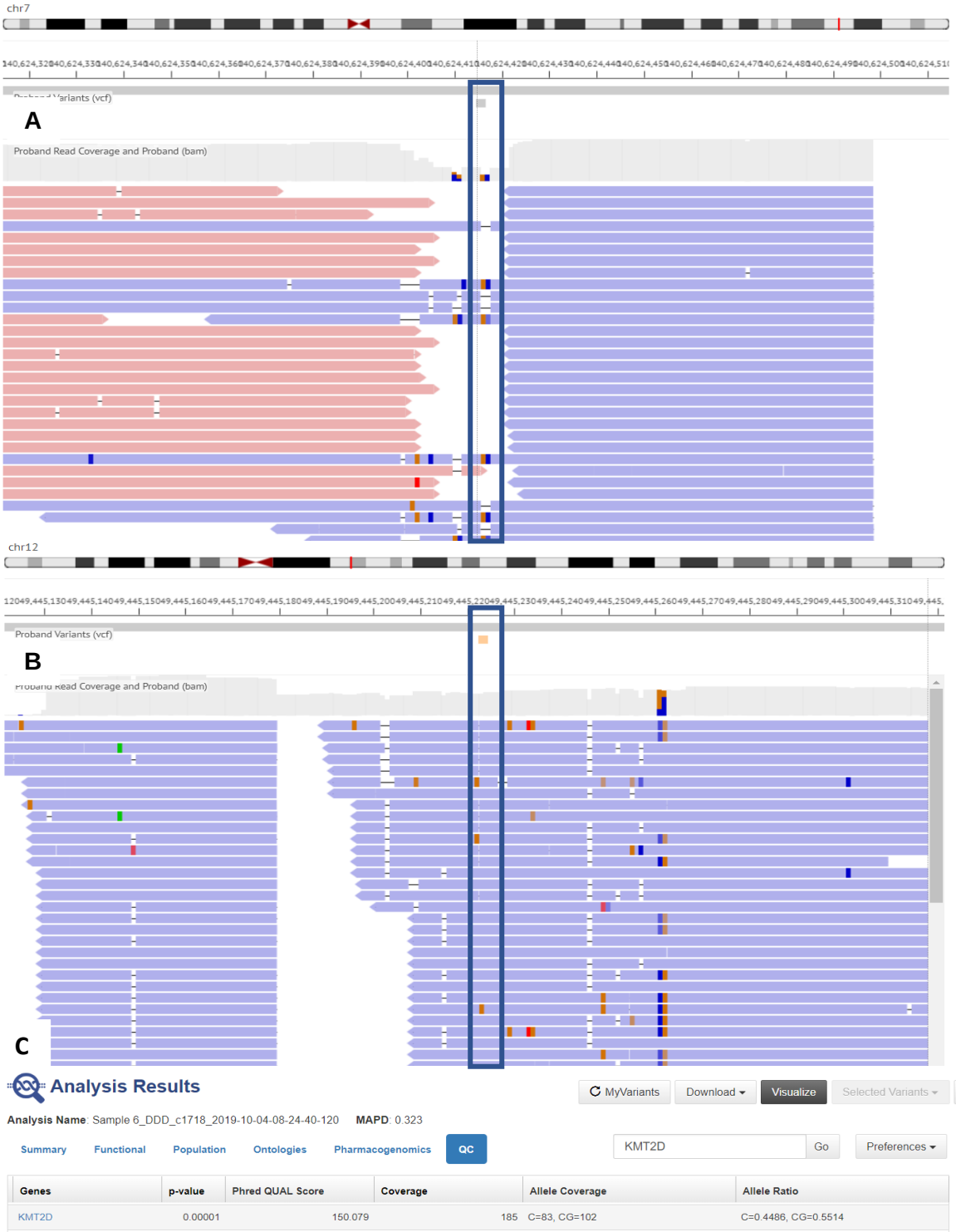


Figure 20. IRGV image of *BRAF* and *KMT2D* variants found in Patient 6. This image shows two screenshots of the IRGV outputs illustrating the **A**: *BRAF* c.87_88delCG and **B**: *KMT2D* c.2243_2244insG variants on chromosomes 7 and 12 respectively that were identified in Patient 6. Pink tracks represent forward strands and blue tracks represent reverse strands. The first location **A** was covered 10 times; the deletion was identified 5 times (represented as a dash). This shows Patient 6 to be a heterozygote for the deletion at this position. The second location **B** was covered 187 times; the insertion was reported 102 times as shown in **C**. Thus Patient 6 is a heterozygote for the insertion at this position.

The custom pipeline results showed variants in 26 genes in Patient 6, see Appendix 7 (Table A7F). The same *BRAF* and *KMT2D* variants were identified as found by Moon. All other variants were excluded as being VUS, benign, or likely benign.

iii. Variant Validation

Both the *BRAF* and *KMT2D* variants were selected for validation through Sanger sequencing. The Sanger sequencing could not confirm the presence of the *BRAF* variant. The sequencing results for this variant (*BRAF* c.87_88delCG) can be seen below in Figure 21. The assay to validate the *KMT2D* variant could not be optimised after multiple optimisation attempts and thus the variant could not be validated. The optimisation strategies undertaken can be seen in Appendix 8.

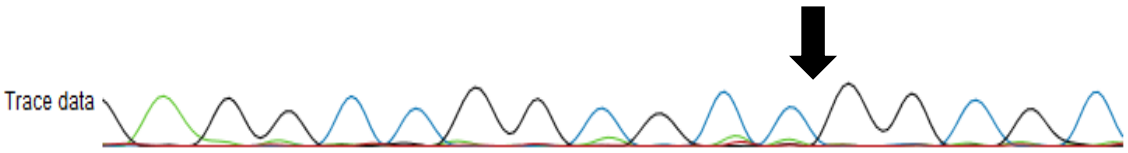


Figure 21. *BRAF* c.87_88delCG variant validation through Sanger sequencing. Electropherogram showing the absence of the suspected 2bp *BRAF* deletion in patient 6, location indicated by the black arrow.

3.4.6 Patient 7

i. Clinical Description

A 33-year-old female, born to a phenotypically normal non-consanguineous couple. The father was unavailable for recruitment; only the mother was included. Ultrasound scans were reported as normal, and the child was born through normal vaginal delivery at full gestation, without complications. Congenital anomalies were present in the form of a soft palate cleft and bilateral ear anomalies. Neurodevelopmental assessment identified delayed motor development; speech and language; and overall mild intellectual disability. The reported system anomalies can be seen in Table 24. Prior to recruitment in the DDD-Africa study, karyotyping and MLPA were performed and returned normal results. A CT scan was performed in 2015 and was normal.

Table 24. System anomalies recorded for Patient 7

HPO Term	HPO ID
Cleft Palate	HP:0000175
Widow's peak	HP:0000349
Slender nose	HP:0000417
Microretrognathia	HP:0000308
Abnormality of the ear	HP:0000598
Hearing impairment	HP:0000365
Microcephaly	HP:000252
Short stature	HP:0004322
Intellectual disability, mild	HP:0001256

ii. Variant prioritisation

The Moon (Diploid) software results identified a frameshift deletion in the *EFTUD2* gene of Patient 7. The ACMG classification of this variant can be seen in Table 25 below and the variant information can be found in Appendix 6. The variant was classified as likely pathogenic due to one very strong and one moderate criterion supporting pathogenicity. The possible diagnosis associated to the variant (as suggested by Moon) was Mandibulofacial dysostosis with microcephaly (OMIM:610536).

Table 25: ACMG classification of *EFTUD2* c.263_264delCT variant.

ACMG Code	**Reason**
PVS1	Pathogenic Very Strong. Null variant caused by frameshift affecting gene, which is a known mechanism of disease associated to Mandibulofacial dysostosis, Guion-Almeida type and Esophageal atresia, syndromic (Varsome 16/11/2019).
PM2	Pathogenic Moderate. Variant not found in GnomAD 16/11/2019.

This variant is a frameshift deletion that results in a premature stop-codon. The variant was viewed in Mutalyzer and the results can be seen in Figure 22 below.

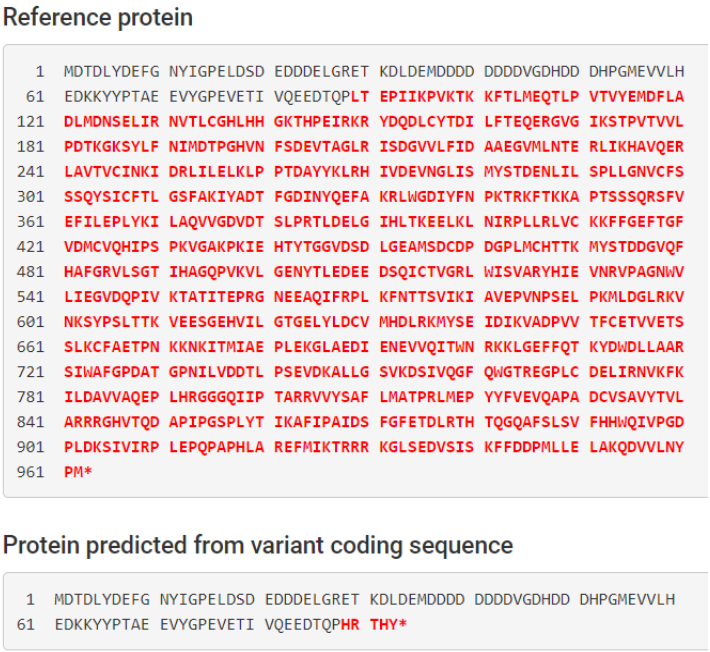


Figure 22. Screenshot of Mutalyzer of *EFTUD2* c.263_264delCT variant.

The reference protein shows a protein of 962 amino acids in length. The letters in red represent the amino acids affected by the variant. The protein predicted from the variant coding sequence shows a protein of only 93 amino acids in length.

The variant was covered 159 times by WES and can be seen visually in Figure 23 below.

Figure 23. IRGV image of *EFTUD2* variant found in Patient 7. This image shows a screenshot of the IRGV output illustrating the EFTUD2 c.263_264delCT variant on chromosome 17 that was identified in Patient 7. Pink tracks represent forward strands and blue tracks represent reverse strands. The location was covered 159 times; the deletion was identified 77 times. Thus Patient 7 is a heterozygote for the deletion at this position.

The custom pipeline results showed variants in 15 genes in patient 7, see Appendix 7 (Table A7G). The same *EFTUD2* variant was identified as found by Moon and the additional variants were all excluded as VUS, benign, or likely benign.

iii. Variant Validation

The *EFTUD2* variant was selected for validation through Sanger sequencing. The results of the sequencing can be seen below in Figure 24. Only the maternal sample of this patient was available for validation and was sequenced for the variant. Both the patient and her mother indicated the presence of the deletion in the sequencing results, suggesting the child inherited the deletion from her mother.

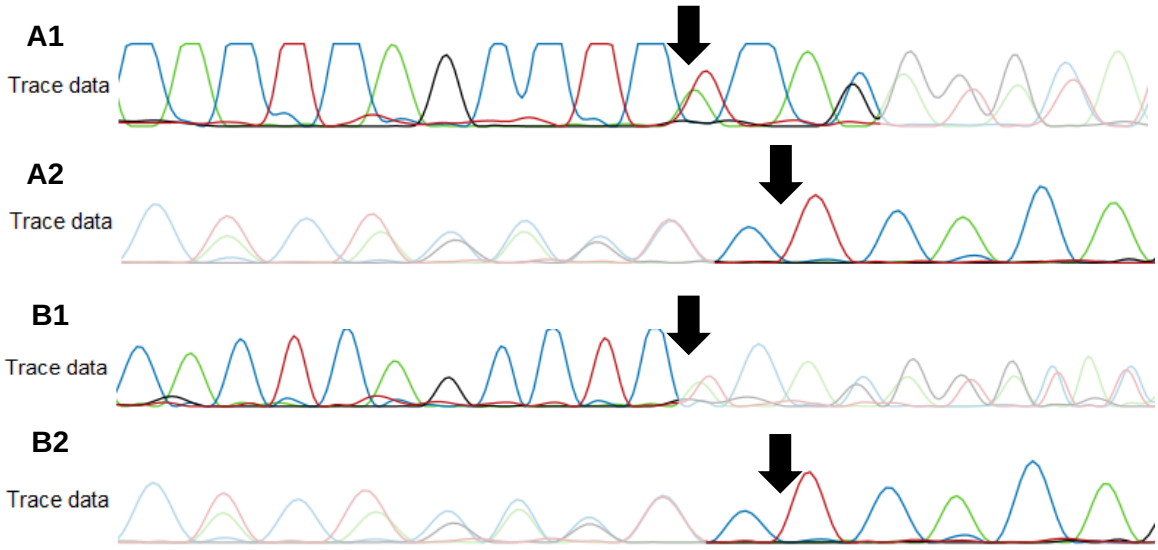


Figure 24. *EFTUD2* c.263_264delCT variant validation through Sanger sequencing. **A1**: Electropherogram identifying CT deletion in Patient 7 through forward strand sequencing. **A2**: Electropherogram identifying CT deletion in Patient 7 through reverse stand sequencing. **B1**: Electropherogram identifying CT deletion in Patient 7's mother through forward strand sequencing. **B2**: Electropherogram identifying CT deletion in Patient 7's mother through reverse strand sequencing. Variant location is indicated by the black arrows.

3.4.7 Same Variants identified in Multiple Patients

Through manual analysis, it was found that 15 variants in different genes appeared in multiple patients (Table 26). Of the 15 variants, 3 were identified as likely pathogenic whilst the others were all VUS, benign, or likely benign. The three variants classified as likely pathogenic were investigated further and found to be indels lying within homopolymers. As the Ion Torrent is known to have a high error rate for calling indels within homopolymers, one of these variants in the IQ Motif and Sec7 Domain ArfGEF 2 (*IQSEC2*) gene was selected for investigation. Classification of this variant is shown in Table 27 and additional information is in Appendix 6.

Table 26. Variants that were identified in multiple patients through manual analysis.

Gene	Variant location and nucleotide change	ENSEMBL Transcript	Patients in which variant was found	ACMG classification
COL18A1	chr21:46924425:46924434:CGGCCCCCCA:-	ENST00000342220	1, 2, 3, 4, 5, 7	VUS
RTTN	chr18:67863852:67863851:-:TCC	ENST00000437017	1, 2, 3, 4, 5, 6, 7	Benign
ATN1	chr12:7045892:7045903:CAGCAGCAGCAG:CAG	ENST00000356654	1, 4, 6, 7	VUS
TH	chr11:2189728:2189728:C:-	ENST00000333684	1, 6, 7	Likely Pathogenic
PKD1L1	chr7:47913579:47913580:TG:CG	ENST00000289672	1, 2, 5, 6, 7	VUS
ARMC9	chr2:232087474:232087475:AT:GA	ENST00000349938	1, 2, 3, 4, 5, 6, 7, 8	VUS
CLCNKB	chr1:16375063:16375064:CA:GC	ENST00000375679	1, 4, 5, 6, 8	Benign
IQSEC2	chrX:53285132:53285133:GC:-	ENST00000375365	1, 3, 4, 5, 7	Likely Pathogenic
SIK1	chr21:44838333:44838333:G:-	ENST00000270162	1, 3, 4, 5	Likely Pathogenic
SMARCA2	chr9:2039815:2039817:GCA:-	ENST00000357248	1, 3, 6, 8	Benign
CUL7	chr6:43014298:43014299:TT:CC	ENST00000265348	1, 2, 3, 4, 5, 6, 7, 8	VUS
HYDIN	chr16:70975565:70975565:T:C	ENST00000393567	1, 2, 3, 4, 5, 6, 7	Likely Benign
KIF1A	chr2:241696871:241696873:TCC:-	ENST00000404283	1, 2, 3, 4, 5, 6, 8	Benign
CRYBB1	chr22:27012207:27012207:G:C	ENST00000215939	2, 3, 4, 5, 7	VUS
WRAP53	chr17:7606722:7606722:C:-	ENST00000457584	4, 5, 6	VUS

Table 27: ACMG classification of *IQSEC2* c.848_849delGC variant.

ACMG Code	Reason
PVS1	Pathogenic Very Strong. Null variant in gene, which is a known mechanism of disease. Associated with X-linked mental retardation (Mutalyzer 18/11/2019).
PM2	Pathogenic Moderate. Variant not found in GnomAD 18/11/2019.

The variant was covered 140 times in Patient 1; 32 times in Patient 3; 63 times in Patient 4; 50 times in Patient 5 and 160 times in Patient 7 by WES. A visual of the variant in Patient 5 can be seen below in Figure 25.

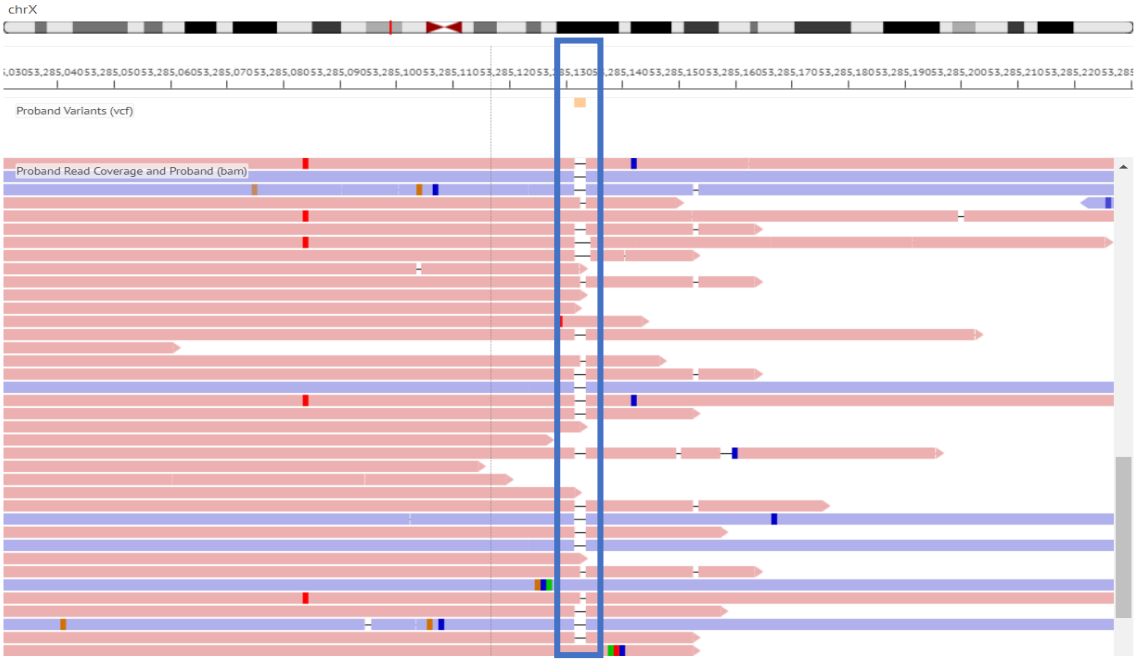


Figure 25. IRGV image of *IQSEC2* variant found in Patient 5. This diagram shows a screenshot of the IRGV output illustrating the *IQSEC2* c.848_849delGC variant on chromosome X that was identified in Patient 5. Pink tracks represent forward strands and blue tracks represent reverse strands. The location was covered 50 with a GC deletion identified 48 times.

The *IQSEC2* variant was selected for validation through Sanger sequencing. The variant was sequenced in Patients 1, 3, 4, 5, and 7. The results can be seen below in Figure 26 and show that no deletion was identified in any of the samples tested.

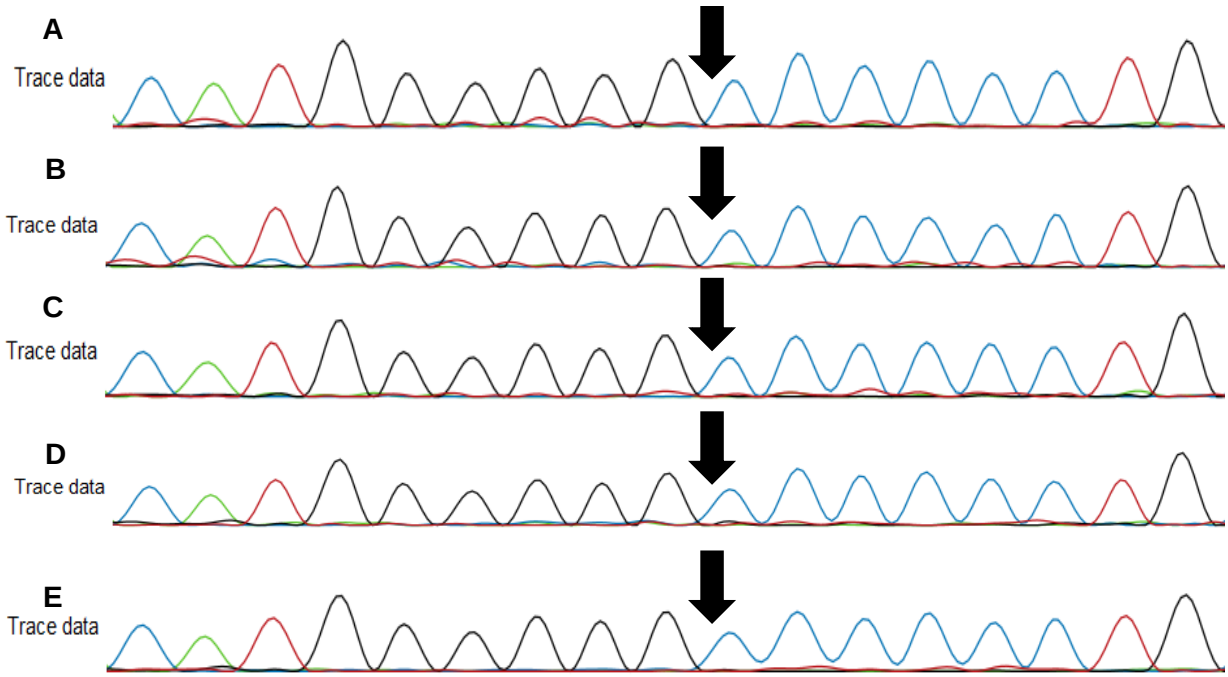


Figure 26. *IQSEC2* c.848_849delGC variant validation through Sanger sequencing. **A**: Electropherogram showing no deletion at position c.848_849 in Patient 1. **B**: Electropherogram showing no deletion at position c.848_849 in Patient 3. **C**: Electropherogram showing no deletion at position c.848_849 in Patient 4. **D**: Electropherogram showing no deletion at position c.848_849 in Patient 5. **E**: Electropherogram showing no deletion at position c.848_849 in Patient 7. The variant location is indicated by the black arrows.

3.5 Whole exome sequencing cost analysis

The costing for whole exome sequencing was determined using the NHLS costing model. This model incorporates the cost of reagents and consumables and includes a consumable wastage fee. The labour charges attributed to the test are determined by the number of Medical Scientists required to perform the test, their corresponding skill level, and the percentage of time required to run the test and analyse the data. Fixed equipment costs and their associated depreciation rates are also included. The overall cost per annum is calculated from which the costing of a single test is then determined. VAT is added to the total price at 15%, resulting in the final testing cost.

Through the use of the NHLS Costing Model, the cost of performing WES on the Ion Torrent™ S5 within this Division would amount to R8989.67 per sample sequenced.

However, this is not the total cost inferred by each patient tested as the pricing for DNA extraction, Moon software sample analysis, and Sanger sequencing for variant validation need to be accounted for. The Moon software cost was included as a cost estimation for data analysis. The addition of these subsequent tests/factors totals the price at R13,128.80 per patient tested (Table 28).

Table 28. Costing per sample for WES according to the NHLS Costing Model

Test or Procedure	**Cost per sample (South African Rands)**
DNA Extraction	700
Whole Exome Sequencing	8989.67
Moon Sample Analysis	806
Sanger Sequencing	2633.13
TOTAL	**13,128.8**

3.6 Diagnostic Yield Analysis

The diagnostic yield analysis of this test was determined by dividing the number of positive results by the total number of cases tested. The calculation can be seen below:

$$\frac{\#\ of\ cases\ with\ a\ positive\ result}{Total\ \#\ of\ cases\ tested} X100 = Diagnostic\ Yield\ (\%)$$

$$\frac{5}{7}\ X\ 100 = 71.43\%$$

3.7 Comparing Ion Torrent S5™ to the Illumina MiSeq™

With the information gained about performing WES on human samples using the Ion Torrent™ S5 system, it was decided to compare the specification of this procedure on another available system. The only other system to which we would have direct access is an Illumina MiSeq™ (Illumina, USA) housed at the University of the Witwatersrand Medical School. The specifications of this system were obtained from the Illumina website (Illumina Inc. 2020) and compared to the specifications of the Ion Torrent™ S5 system used for this project. These specifications can be seen in Table 29 below.

Table 29. Specification comparison of WES on the Ion Torrent S5™ ad Illumina MiSeq™ systems.

	Ion Torrent S5™	Illumina MiSeq™
Number of exomes sequenced per run	2-3	1
Read length sequencing chemistry	200bp	2x300bp
Number of mapped reads per sequence	15 million-35 million	±25 million
BAM file size (GB) per sample	10-30	±15
Library preparation time	18 hours (8 samples)	48 hours (8 samples)
Library templating, sequencing and base calling time	±34 hours per chip (2-3 samples)	±112 hours (2 samples)

4. Discussion

This project aimed to implement a WES workflow on the Ion Torrent S5™ platform and evaluate the financial and computational costs involved in this technique. This was carried out by following the methodology outlined in Chapter 2 that subsequently provided the results shown in Chapter 3. This chapter will go on to discuss the results and overall outcomes of this project, by firstly discussing the quality control and WES performed on the Ion Torrent S5™ system. This will be followed by discussing the bioinformatic pipelines used and their associated performance, leading to the Patient results and diagnosis. Lastly, the cost analysis; diagnostic yield analysis; and the comparison of the system to an alternative available system. Concluding the chapter by discussing the project limitations, future directions, and overall conclusion.

4.1 Quality Control and Whole Exome Sequencing on the Ion Torrent S5™

According to the information provided by the coverage analysis report and the WES run report, the Ion Torrent™ S5 efficiently sequenced the exomes of seven DD patient samples and one sequencing control. Each individually sequenced sample produced libraries of ≥78% clonality and final libraries (usable reads) of ≥68%. It was noted that the percentage of low-quality reads doubled in the sequencing of the third and fourth chips as opposed to the first two, however, we are unaware as to the cause of this increase. All eight sample libraries were prepared together on the same 1x8 exome RDY plate (Thermo Fisher, USA) and all samples were quantified and diluted using the same method. The only difference in procedure was that chips one and two were sequenced first and the libraries of the remaining chips were frozen for an additional two-three days before sequencing. However, as recommended by the manufacturer, libraries can safely be stored at -20°C for two to three months prior to sequencing we do not believe this to be the cause. An attempt was made to reach out to the technicians at Thermo Fisher to provide some insight into the cause of this.

The average read length across all samples was 185bp which aligns with the 200bp chemistry used by the Ion Torrent™ platform for exome sequencing. As this technology makes use of short-length reads, this average is expected. The mean read depth of each sample was also well above that of the recommended mean read depth for diagnostic exome sequencing of 35X (Ajay et al., 2011). In our samples, the mean read depth ranged from 47X to 186X. While all samples were above the recommended

value, there was a noticeable difference in the mean read depth across samples. This was most likely caused by dilution errors introduced when combining the final libraries. Library quantification was performed using the Qubit™ 3.0 fluorometer, one of the recommended quantification methods as suggested by the Ion AmpliSeq™ Exome RDY Library Preparation User Guide (Thermo Fisher, USA). Dilutions were performed based on these results. This quantification method has many benefits including low cost and rapid quantification in comparison to other methods such as qPCR. It is shown to have quantitative results with comparable sensitivities to these other methods (Robin et al., 2016). However, errors were introduced as a result of this quantification method most likely due to diluting libraries of a higher concentration to a much lower concentration in a single dilution step. Alternatively, accurate Qubit quantification of frozen DNA samples has been reported to require sufficient salt (NaCl) concentrations in DNA solutions (Nakayama et al., 2016). Our libraries were frozen prior to quantification; thus, this may have influenced the effectiveness of the quantification method. To avoid this during future sequencing endeavours, different quantification methods could be investigated. Quantitative PCR, the use of the Agilent™ 2100 Bioanalyzer™, or additional dilutions (serial dilution) could be used with the Qubit™ system to ensure a more accurate final concentration of each library.

The mean read depth indicates the possibility of including an additional sample to each chip, thus enabling the sequencing of three exomes per run. However, in order to implement this, the quantification and dilution method would need to be as accurate as possible to ensure equal coverage of each sample above the recommended 35X. Sequencing three samples instead of two would also decrease the size of the BAM file generated per sample as this is dependent on the number of reads generated through sequencing. The BAM files generated through this project's sequencing ranged between 10GB-40GB, this large range was dependent on the extent to which each sample was covered during sequencing. The equal loading of two samples (seen in chip 4) showed an average BAM file size of 30GB. If three samples were loaded this file size would drop to roughly 20GB per sample which may help with storage capacity.

Additional considerations of data quality involve the presence of sequencing errors. Sequencing errors are found across all next generation sequencing platforms and thus quality scores (typically Phred scores) have been implemented during base calling to

determine the accuracy of called bases. To address this, we filtered out all variants with a quality score of less than Q40, such that all remaining variants had a predicted accuracy of 99.99%. Sanger sequencing was also conducted on prioritised variants to ensure the validity of the call. We did not initially filter out indels identified in homopolymer regions. We had not come across any recommendations to do so, even though we were aware that the Ion Torrent system is notorious for the false error rate of calling indels in long homopolymer stretches of DNA (Yeo et al., 2012; Bragg et al., 2013; Song et al., 2017). One indel in a homopolymer region was identified through variant prioritisation and selected for validation. Through this validation, it was identified as an erroneous call. Based on this evidence and the newly published suggestion by Pereira et al., we would recommend filtering out indels called in homopolymer stretches of longer than five base pairs in future sequencing endeavours using this platform (Pereira et al., 2020).

4.2 Bioinformatic Workflow

The data analysis of this project consisted of multiple steps performed across a wide range of software and platforms (Chapter 2, Figure 3). We made use of the Torrent Suite™ and Ion Reporter™ software (Thermo Fisher, USA) for primary and secondary analysis as this software was designed for the Ion Torrent platform with the ability to adequately analyse Ion Torrent sequencing data. These software performed well at these stages of analysis, however, several limitations were seen in the software's tertiary analysis capabilities, more specifically the variant filtering capabilities. Filtering based on quality control (Phred scores) was not available at the time of investigation and the filtering according to other parameters was not always efficient or easily manipulated to the requirements of the researcher. A CNV confidence range filter was applied to all VCFs prior to download. The CNV calling baseline has not been set up on the Ion Torrent system used for this project, therefore we did not have the capabilities to call and investigate CNVs. We used alternative tertiary analysis software based on our interaction with the Torrent Suite and Ion Reporter software. It would thus be our recommendation to use alternative data analysis methods for tertiary analysis of human exome data.

In conducting tertiary analysis, we compared three annotation and filtering platforms, namely the Moon Diploid software (online automated web-service), a bioinformatic

pipeline (referred to as the Custom Pipeline), and lastly VarAft (a downloaded desktop application). Each filtering platform generated a prioritised variant list for each patient that then underwent manual analysis using a multitude of online tools and software.

The automated online web-service tool known as Moon (Diploid) was the first investigated platform in this project. As previously mentioned, Moon is a relatively recent software package that makes use of artificial intelligence to rapidly analyse and interpret variants and has been seen in literature as an aid in the diagnosis of rare disease and cases of intellectual disability and DD (Clark et al., 2019; Lumaka et al., 2018). This software prioritised 10-20 variants per patient from the ±50000 variants in the each uploaded VCF file. A list of 20 prioritised variants is more manageable to manually investigate, and significantly reduces the labour costs of data analysis. However, perhaps the most beneficial aspect of the software is its ability to generate a prioritised variant list from WES data in mere minutes. The machine-learning-based functionality of the software along with the built-in filtering parameters allows for rapid variant prioritisation. This reduces the typical manual investigation time per sample, from between 20 and 40 hours (Wenger et al., 2017) to three to five hours.

In addition to providing rapid analysis, the Moon (Diploid) output was highly accurate with the possible putative disease-causing mutation found in Moon's top three listed variants in six of the seven patients. Similarly, the software developers tested Moon on 30 WES cases, and in all cases, the causal variant was identified in the top three variants listed. A study by Clark et al. used the Moon (Diploid) software to analyse 100 WGS cases and showed a 99% accuracy rate (Clark et al., 2019). As the software was only launched in 2017, very few publications involving the use of Moon (Diploid) software are available. However, many diagnostic laboratories and facilities have posted promising reviews of the software and incorporated the software into their own NGS pipelines. These facilities include Leiden University Medical Center (Netherlands); Antwerp University Hospital (Belgium); and the Center for Cardiovascular Genetics and Gene Diagnostics (Switzerland).

This software is not freely available and works on a credit-based system. At the time of this investigation (November 2019), a single analysis run per sample cost US$49 which amounted to approximately ZAR806. The software was developed by a company called Diploid (Belgium) which has recently been purchased by the American

genomics company Invitae and it is unknown how the analysis costing may be affected by the change of ownership. Nonetheless, the ability of the software to rapidly reduce analysis time ensures the cost-effectiveness of the program, as the process becomes much less labour intensive on the part of the medical scientists analysing the data. Using the NHLS costing model, the labour costs of a Senior Medical Scientist performing 20 hours of manual analysis would amount to ZAR5,000.00. This makes the software highly desirable in a diagnostic setting where there are multiple cases awaiting analysis. Medical scientists are inundated with cases for sequencing and other genetic test which prevents the undivided attention of exome data analysis. This was the main contributing factor to our decision that the Moon (Diploid) software would be the most beneficial to the implementation of this test locally.

The second tertiary analysis method investigated was the use of a bioinformatics pipeline. The bioinformatics pipeline designed in this project applies specifically to identifying possible putative disease-causing variants in patients presenting with DD. The pipeline consists of two separate BASH scripts that are run one after the other. The application of these scripts prioritised approximately 30 variants per patient that then underwent manual analysis. The possible putative disease-causing mutations identified in each patient from this application aligned to the variants prioritised by the Moon software with a few exceptions. One major limitation of this plugin application is that the final list of prioritised variants is determined only by the 2500 genes within the database, so any causative variants found in genes, not on this list would be filtered out. This was seen in Patient 1 were the causative variant identified by Moon was in a gene not found on the DDG2P list (namely the *CDC42* gene) and was thus filtered out and missed in the initial custom pipeline analysis. As the DDG2P list expands over time to include more DD associated genes, this limitation may be overcome. However, until such time, the list could serve as a starting point in variant prioritisation and if no causal variants are identified then analysts would need to broaden the scope of variant investigation to variants in genes not found on the list.

Apart from this limitation, the pipeline performed well overall and accomplished the aim of filtering variants into a manageable data-set for manual investigation, and would most likely serve as a beneficial tertiary analysis method in the investigation of WES data in DD patients. The DDG2P list is currently contributing to a wide range of DD research on-going globally. Many pathogenic variants associated to DD patients are

found in genes on this list (Gazdagh et al., 2015; Choi et al., 2017; Wright et al., 2018) and other lists/databases that are being compiled, combined and curated (Erger et al., 2019). An additional benefit of this platform that should be mentioned is that the online web-tools and software used during the application are free. Apart from the initial hardware requirements and labour costs there is no service fee per sample as there is with the Moon (Diploid) software. While the costing is a major advantage of this custom pipeline, the lack of the applicable Linux coding skills necessary to operate this program locally is a major deterrent to its implementation. Thus, system administrators and trained bioinformaticians with the necessary skills would be needed to effectively establish and implement this pipeline.

The VarAft software was the third analysis platform investigated, and while the software is free and relatively user-friendly it did not meet the needs of our data analysis requirements. One of the major limitations encountered with this software was the ‘gene information’ section within the filtering pipeline. HPO terms could be incorporated as well as suspected disorders, affected organ function, and suspected molecular pathways. We made use of the HPO term input as we only had phenotypic data for each patient under the umbrella term of DD. Due to the heterogeneous nature of these disorders, we did not have any suspected genes or molecular pathway implications upon which we could act. The limitation arose as only a single HPO term could be selected at a time, where variants were prioritised based on a small list of genes associated with the specific term. In cases where there are no major malformations to lead this search, the filtering is then ineffective as was the case with the patients in this study.

For Patients 4 and 5, we had a working differential of congenital muscular dystrophy; when this term was selected, the causative variant was identified (which had been confirmed with the previously mentioned platforms). Thus, with specific disease suspicions’ or gene information this software may be beneficial in identifying putative disease-causing mutations. As such, it could be implemented into the investigation of WES data of samples with suspected clinical diagnoses.

Overall, we had a prioritised list of variants totalling at approximately 300 variants per patient using this platform. It was decided that the time taken for manual analysis of

300 variants per patient would not be feasible in either a research or diagnostic setting when better-performing platforms had already been identified.

4.3 Patient Results and Diagnoses

We identified possible putative disease-causing mutations in six of the seven DD patients sequenced in this project; five of these were subsequently validated.

4.3.1 Patient 1

In Patient 1 a possible pathogenic variant was identified in the *CDC42* gene. The gene encodes a small GTPase protein of the Rho-subfamily, which is known to regulate signalling pathways, controlling multiple cellular functions such as migration, morphology, cell cycle progression, and endocytosis (Johnson and Pringle 1990). Mutations in this gene have been associated with two different autosomal dominant disease phenotypes commonly known as Takenouchi-Kosaki syndrome (Takenouchi et al., 2015) and Wiskott-Aldrich syndrome (Ochs 2009). We identified a missense mutation (c.68A>G) in the *CDC42* gene that affects the N-terminal α Helix of the Ras domain within the associated protein (Pfam 11/02/2020). This mutation has previously been described in a functional study, that performed *in silico*, *in vitro,* and *in vivo* analysis, and it was found to be pathogenic illustrating a phenotype closely resembling Takenouchi-Kosaki syndrome (OMIM:616737), also known as Macrothrombocytopenia and Mental Retardation syndrome (Martinelli et al., 2018).

Following the guidelines set out by the ACMG for the interpretation of sequence variants, we applied three moderate criteria supporting pathogenicity and three supporting criteria supporting pathogenicity (Chapter 3, Table 12). The PS3 code was also applied, which is a strong pathogenicity indication based on a well-conducted functional study (*in vivo*/*in vitro*) showing a damaging effect on a gene or gene product. However, based on the recommendation published earlier this year (Brnich et al., 2020), this code was downgraded to supporting for this variant based on the fact that the functional study in question (Martinelli et al., 2018) has not yet been replicated. A second functional study (*in vivo*) of this gene was identified (Uehara et al., 2019) however, the study only looked at the effect of a single variant (p.Tyr64Cys) and while the same hypomorphic effect was identified for this variant as seen in the variant in Patient 1, it was decided to not upgrade the code, until further functional studies of the same variant have been performed. Thus, the variant currently presents as likely

pathogenic according to our interpretation. The variant was validated using Sanger sequencing and the parental samples were also sequenced showing the variant to be *de novo* in the child as the parents were both homozygous for the wild-type allele.

Upon consultation with the referring clinicians and senior research team, it was determined that this variant could be the causative variant of this patient's phenotype. As this disease has only recently been identified and described and not yet seen in African patients (to our knowledge), further clinical investigation will be performed before providing a definite diagnosis. A blood test testing for enlarged platelets (macrothrombocytopenia) will be requested as this appears to be a common phenotype and a more comprehensive clinical exam to identify any phenotypes that may have been overlooked initially. The clinicians will contact the authors of the functional study (Martinelli et al., 2018) to discuss the phenotype in our patient in order to gain a more experienced opinion. There are additional complications associated with this disorder, such as renal and urogenital anomalies which would need to be investigated and monitored in this patient due to the findings of this study.

4.3.2 Patient 2

In Patient 2 no possible putative disease-causing variants were identified during the initial analysis. Of the prioritised variants that were identified, none of the associated diseases were a match to our patient's phenotype or else the variants were classified as benign, likely benign, or VUS and thus disregarded. The top prioritised variant identified in this patient was a missense mutation in the *CBL* gene (c.2520T>G). The protein encoded by this gene is a negative regulator of many tyrosine kinase signalling pathways as it functions as an E3 ubiquitin ligase (Thien et al., 2001). Mutations in this gene have been associated with Noonan syndrome-like disorders (OMIM:607721) which explains why this variant may have been prioritised by Moon as our patient has many features typical of Noonan syndrome. Additionally, this variant was erroneously or perhaps rather ambiguously described as pathogenic in an abstract of a recently published paper that Moon would have taken into account (Tekendo-Ngongang et al., 2019). The authors of this paper were contacted and confirmed the classification of the variant as a VUS due to contradictory pathogenicity criteria applied using the ACMG guidelines (Chapter 3, Table 14).

The VCF was recently reanalysed which identified additional variants in four genes namely: *TOP2B, PPP1R12A, TTN*, and *EXOC8*. All variants were determined to have unknown significance; however, it does support that reanalysis at a later date may provide more insight into possible pathogenic variants within patients who do not receive a diagnosis through WES (Xiao et al., 2017; Ewans et al., 2018; Baker et al., 2019).

4.3.3 Patient 3

In Patient 3 a known pathogenic mutation in the *EFTUD2* gene was identified. This gene encodes a subunit of two complexes known as the minor and major spliceosome (Beauchamp et al., 2019). Mutations in this gene are known to cause an autosomal dominant disorder called Mandibulofacial Dysostosis with Microcephaly (OMIM:610536), also referred to as the Guion-Almeida type (Sarkar et al., 2014). We identified a stop-gain mutation (c.1732C>T) that was previously reported in literature as pathogenic (Sarkar et al., 2014). Using the ACMG guidelines we applied one very strong criteria of pathogenicity, two moderate criteria, and two supporting criteria resulting in an overall classification as pathogenic (Chapter 3, Table16). Validation was performed using Sanger sequencing and the parents of the child were also sequenced for the variant, but were both homozygous for the wild-type, thus identifying the mutation to be *de novo* within the child. Based on this evidence and consultation with the referring clinicians and senior research team it was determined that this was the disease-causing variant of this patient.

4.3.4 Patients 4 and 5

In Patients 4 and 5 (fraternal twins) the same compound heterozygous variants were identified in both patients. These variants were found in the *POMT2* gene and have not previously been reported in literature. This gene encodes a portion of the protein O-mannosyltransferase enzyme complex which is responsible for α-dystroglycan protein modification and is found in a multitude of body tissues (Akasaka-Manya et al., 2006). Mutations in this gene are known to cause disorders of congenital muscular dystrophy that can range in severity (Yanagisawa et al., 2009). More specifically the phenotype expressed within these patients is suggestive of congenital muscular dystrophy with mental retardation, type B2 (OMIM:613156). As this disorder follows an autosomal recessive mode of inheritance, two defective gene copies/alleles would be necessary to invoke the disease phenotype. As mentioned, we identified a

compound heterozygous mutation which is the presence of two different mutations at a particular gene locus with one on each chromosome pair (Zhong et al., 2017). Both variants identified were likely pathogenic, and are found in two adjacent functional domains namely the MIR3 domain (variant c.1424G>C) and the PMT 4TMC domain (variant c.2232G>T).

Following the ACMG guidelines, we applied two moderate criteria and two supporting criteria supporting pathogenicity to each variant thus, providing an overall classification as likely pathogenic (Chapter 3, Tables 19 and 20). One of the supporting codes, namely PP4, was applied at the discretion of the clinicians as this code makes use of phenotypic data to support the variant classification (Richards et al., 2015). Prior to WES, the working differential for these patients was congenital muscular dystrophy. A high creatine kinase value indicated muscle damage; and together with other evidence supported application of PP4.

Sanger sequencing was used for variant validation and to determine the mode of inheritance. The father was unavailable for testing, thus, only the patients' and maternal samples were tested. The first variant (c.1424G>C) was identified in both twins and the mother, all three samples were heterozygous for the mutation and from this we can assume that they inherited the mutation from their mother. The second variant (c.2232G>C) was identified in Patients 4 and 5 (both heterozygous) and absent in the mother (homozygous wild-type). As the mother was negative for the presence of the second mutation, we can assume that this mutation would have been inherited from the father; and with each parent only carrying one mutation, neither were affected. As the two mutations were found in both patients, we can apply the ACMG code PP1, which is a supporting criterion of pathogenicity and indicates the co-segregation of a mutation within multiple affected family members.

4.3.5 Patient 6

In Patient 6, two possible putative disease-causing variants in different genes were identified; a frameshift deletion in the *BRAF* gene (c.87_88delCG) and a frameshift insertion in the *KMT2D* gene (c.2243_2244insC). The variants were classified as likely pathogenic following the application of ACMG codes (Chapter 3, Tables 22 and 23).

The *BRAF* gene variant lies within a GC-rich region which made it challenging to select suitable primers for amplification and sequencing (Flores-Juárez et al., 2016). In the

WES the coverage around this variant was considerably lower than neighbouring gene regions, and the variant itself was only covered ten times during WES, with the deletion seen in five of the ten reads. We determined this variant to be a false positive call as the deletion was not confirmed by Sanger sequencing, and as such was not investigated further.

The variant identified in the *KMT2D* gene was better covered during WES, with a total read depth of 185 and the insertion was seen in 102 of the 185 reads which increases the likelihood of the variant being a 'true' call. Mutations in this gene have been associated with Kabuki syndrome (Cheon and Ko 2015). Kabuki syndrome is an autosomal dominant disorder characterised by intellectual disability and multiple congenital abnormalities; however, the phenotypic features of the disorder have not been well characterised or defined which increases the difficulty of providing a clinical diagnosis (Cheon and Ko 2015; Wang et al., 2019). The assay for the validation of this variant could not be optimised after multiple attempts (see Appendix 7) due to the variant's location in a GC-rich region. Therefore, we cannot confirm whether or not this is the possible putative disease-causing mutation in this patient at this time.

4.3.6 Patient 7

In Patient 7 a possible putative disease-causing mutation was identified in the *EFTUD2* gene. This is the same gene in which a variant was identified for Patient 3, and is associated with Mandibulofacial Dysostosis with Microcephaly (OMIM:610536). A frameshift deletion (c.263_264delCT) was identified in this patient that was classified as likely pathogenic following the guidelines set out by the ACMG. We applied one very strong criterion and one moderate criterion supporting pathogenicity (Chapter 3, Table 25).

We validated this variant using Sanger sequencing and sequenced the maternal sample for the variant. The mutation was identified in the child through both forward and reverse sequencing and was also identified in the maternal sample. The maternal sample was sequenced a second time to ensure that the mother did possess the mutation and there was not any sample cross-contamination during the initial sequencing. Through this, it was confirmed that the mother does have the variant and passed it on to the child.

Mandibulofacial Dysostosis with Microcephaly is an autosomal dominant disorder, only one mutated allele is necessary for the development of the disorder. While a majority of the cases reported have shown a *de novo* occurrence of the disorder (Luquetti et al., 2013; Matsuo et al., 2017) some cases of autosomal dominant inheritance have been reported (Lines et al., 2014). The disease has been reported in literature as highly penetrant (meaning the trait expressed by a variant allele will almost always be present in individuals with that allele) however, it presents with a large variable expressivity (meaning the degree to which the trait is expressed is largely variable among affected individuals) (Huang et al., 2016). One such case was illustrated by Voight and colleagues, where they presented the case of an intellectually-normal mother, with two affected children, that did carry the same mutation as her offspring and was later described as having sub-clinical features (Voight et al., 2013), and we hypothesise a similar scenario in Patient 7 and her mother.

4.4 Cost Analysis

The price of a single sample WES comprehensive diagnostic test was determined to total at approximately ZAR13,128.00 if performed locally in this laboratory. The implementation of this test in a limited resource setting would provide unique benefits. These benefits include job creation for local scientists and technologists; and the implementation of a platform from which students and researchers can gain practical, hands-on experience in this important genetic testing niche. However, at the same time with resources being limited it is necessary to provide an adequate test at an affordable price (for state laboratories servicing state patients). The service should allow for the testing of the majority of the population as opposed to a minority due to costing restraints. Thus, if the use of an outsourced testing provider would be more cost-effective then it would be necessary for us to investigate this option.

Comparative costs were requested from both local and international exome service providers whose services encompass both research and diagnostic needs. At present there are no local diagnostic exome sequencing service providers within South Africa (ZA), all diagnostic WES tests are outsourced to international companies. For this reason, quotes were requested from two international service providers, one based in the USA and one based in the Netherlands. Quotes were requested for the exome

sequencing of a single sample (proband only) to compare with the sequencing costs determined within this laboratory. The USA company quoted US$1250 for a 'Boosted Exome, Proband-Only test'. This package includes DNA extraction, sequencing, data analysis, and report generation. With the exchange rate at the time of the in-house estimation (1US$/ZAR15.20- 09/10/2019) this amounts to ZAR19,000.00 per sample and additional courier chargers would need to be added. The Netherlands' company offers two possible exome options that could apply to DD, namely a WES Intellectual Disability test and a WES Mendelian inherited disorders test. Both of these options are priced at €990, with the exchange rate at the time of the in-house estimation (1€/ZAR16.71- 09/10/2019) this would cost ZAR16,542.90 per patient. This service would include DNA extraction, WES, and data analysis with additional courier fees required.

Additional quotes were requested from local research facilities in order to have a local comparison on testing costs within South Africa. The first quote for whole exome sequencing for a single sample was ZAR22,776.70 (28/04/2020). This price covered only the sequencing itself. DNA extraction, data analysis, and courier charges were not included. A second company quoted exome sequencing at ZAR6840 (based on 2019 prices) per patient and covered only the sequencing itself. Based on these prices, WES on the Ion Torrent S5™ is a reasonably priced option and may prove to be more feasible than outsourcing sequencing to international service providers.

As mentioned above the Moon software data analysis cost was included in the total WES cost calculation, and currently stands at US$49 per sample analysed. However, the cost of using the bioinformatics pipeline generated in this project was also investigated to provide a comprehensive understanding of the total costs necessary if this workflow were to be implemented. This project was conducted in a research environment and as such the Wits Core Cluster was used to meet the processing power requirements of the pipeline. A similar set-up would need to be established in-house with the necessary security and storage measures to align with a diagnostic environment. We would need to invest in an on-site server that met the processing power requirements of the program. A standard quad-core server with 2.10GHz 2MB processing power was determined to be sufficient for the power requirements of the programs within this workflow and the base cost for such a server is ZAR12,906.00. Additional requirements would be necessary such as the purchase of additional RAM,

as the VEP script used a relatively large amount of RAM during analysis. It was suggested that the minimum amount of RAM needed would be 30GB. We requested additional RAM and hard drive storage, and other server kits necessary for installation and set-up, and the price rose to ZAR40,829.00. While this price appears large, it would be a beneficial start-up investment as this is only a once-off cost.

Subsequent costs would include yearly maintenance and additional storage if needed. Staffing requirements for effective implementation and upkeep would also be necessary, such as system administrators, server managers, and bioinformaticians. Some of these services, such as server management, can be outsourced to professional companies. This would be preferable as the service involves 24/7 monitoring, auditing, running updates, ensuring backups are working, and additional administrative tasks (SolarWinds Worldwide 2019). Whether the service is outsourced or monitored by an in-house staff member, it is of high importance as the service maintains server security and functionality. Data breaches and server malfunction would be of great risk when dealing with sensitive diagnostic testing. Bioinformaticians and data analysts would also be required to manage and analyse the large data sets generated through this type of diagnostic testing. The actual programs and web-sources used when running the programs hold no additional charges.

The final cost considerations involved the long-term storage of raw data (BAM files). It is typically the decision of the laboratory conducting the tests, as to how long the raw data should be stored and the NHLS in South Africa requires the data is stored indefinitely. As this data would contain sensitive and confidential information, on-site storage would provide the greatest security measures, and as such two on-site storage solutions were investigated namely a physical storage server and tape drives. A storage server was quoted at ZAR221,986.95 with 192TB storage capacity (24 x 8Tb) and this would allow the indefinite storage of between 30-50 years' worth of data, depending on the fluctuation of file sizes. Alternatively, an LTO-8 tape drive kit was quoted at ZAR66,706.50, capable of storing 30Tb of compressed data per LTO-8 cartridge. A single LTO-8 cartridge was quoted at ZAR2383.42, capable of storing five to seven years' worth of data. The total cost of ownership (TCO) following the tape drive solution is much less than that of the storage server and has been shown to have and 86% lower TCO than 'all-disk' storage solutions and 66% lower TCO than cloud storage solutions (IBM 2020). The use of tape drives also offers added data security

as it a stable storage solution designed for archiving data offline. As the raw data (BAM files) won't be accessed often a tape drive storage solution is most suitable to meet the storage requirements of this data.

4.5 Diagnostic yield analysis

The diagnostic yield for this small sub-study was calculated to be 71.43%. This was determined by dividing the number of positive cases by the total number of cases tested. A positive case was defined as one in which a possible putative disease-causing mutation was identified and classified as either likely pathogenic or pathogenic according to the ACMG guidelines; validated using Sanger sequencing and then confirmed through consultation with a referring clinician. The diagnostic yield of this project is much higher than the global diagnostic rate of approximately 40% (Córdoba et al., 2018; Need et al., 2012; Wright et al., 2015). Two to three patients were expected to be diagnosed in this study as per this statistic however, causal variants were identified and validated in five of the patients.

This was likely due to the pre-selection of patients with well-defined phenotypes capable of leading the variant investigation and prioritisation. HPO terms allow the linking of phenotypes to genes and disease (Köhler et al., 2017), thus the well-structured phenotyping and incorporation of HPO terms into variant investigation is an effective approach in narrowing down possible putative disease-causing variants and improving the diagnostic rate (Bone et al., 2016; Frésard and Montgomery 2018).

4.6 Comparing the Ion Torrent S5™ to the Illumina MiSeq Platform

As a final consideration in the investigation of WES on the Ion Torrent S5™ as a potential tool for diagnosing DD patients in South Africa, we compared the specifications of our system to the Illumina MiSeq platform (Illumina, USA). The comparison made use of the specifications identified in our project for the Ion Torrent S5™ system and the specifications described on the Illumina website for the Illumina MiSeq system (Illumina Inc. 2020); additional comparisons were sought after in literature, however, there is very limited published data available. A cost comparison was not performed however, the price of high-throughput sequencing on both platforms has been reported as 'cost-competitive' (Marine et al., 2020). The major differences seen between the two systems were the number of samples that could be

sequenced per run and the time taken for library preparation, templating, and sequencing. The Ion Torrent system outperformed the Illumina system on both counts making it the ideal system for implementation in this laboratory.

4.7. Project Limitations

One major limitation identified in our project was the lack of African databases in which we could compare the allelic frequency of the variants identified in our patients. We did make use of GnomAD and the 1000 Genomes data which has limited African data. However, we are well aware of the large amount of genetic variation among African patients (Campbell and Tishkoff 2010; Gomez et al., 2014; Rotimi et al., 2017) therefore, access to African specific data would be greatly beneficial in understanding the frequencies of these variants within the general population (Shearer et al., 2014; Deepti et al., 2019). We are currently in the process of requesting access to the African Genome Variation Project (AGVP) data which consists of genotyping data collected from Sub-Saharan African patients (Gurdasani et al., 2015), in order to circumvent this limitation. As African populations are largely understudied, there is only a minimal number of samples and data (GWAS data and some low-coverage WGS data) available within these databases, highlighting this as a limited resource in African based genetic studies.

A second limitation to consider is associated to the cost estimation. Due to import costs (reagents and consumables), fluctuating exchange rates (Habanabakize, T., 2020), and local demand the cost estimate identified in this project may increase and/or decrease drastically with time. The absence of local manufacturing means this limitation cannot easily be overcome, however, exchange rates and import costs can continuously be monitored to predict costing fluctuation.

4.8 Future Directions

Research reports will be generated and provided to the referring patient's clinician in order to feedback on the results found in this project. These patients form a subset of the DDD-Africa project, as such feedback of results is done as part of a research report. At the Johannesburg, South Africa site of the project, genetic counsellors will form part of the team that provide feedback to patients. Feedback of findings takes place for multiple reasons such as when a significant associated risk to disease is

identified; when the identified disease has future implications (co-morbidities, reproductive implications, premature death); or when therapeutic and preventative interventions are available (Bookman et al., 2006; Budin-Ljøsne et al., 2016). Genetic counsellors are included in the feedback of findings as their unique skill set allows them to provide adequate risk assessment, and education and support to the patients and families receiving this feedback (Skirton et al., 2015; Joseph, F.M., 2018; Papaz et al., 2019).

Once permission has been granted to access the data available through the AGVP we will investigate the allele frequencies of the variants identified in this project and add them to the research reports being generated for the clinicians. This is necessary as allele frequencies contribute to the categorisation of variants and population-specific frequencies allow for better categorisation (Shearer et al., 2014). We will then also input this avenue of investigation into the recommended data analysis pipeline of investigating WES data of DD patients in South Africa which was developed in this project (Chapter 3, Figure 8).

The *KMT2D* variant that was not validated, will be investigated to identify possible ways of further optimising the assay. New PCR primers may be designed or alternative validation methods will be sought out, such as making use of long-range PCR. In current literature, the recommendations for Sanger sequencing to validate NGS results varies among researchers, with some suggesting it is not necessary (Beck et al., 2016); some suggesting it is necessary (Lier et al., 2018); and others suggesting the necessity in certain circumstances, such as in regions of poor base coverage (Mu et al., 2016; Celestino-Soper et al., 2017; Zheng et al., 2019). As this was a research project, validation was necessary before providing results, even though the specific variants showed high coverage during WES. In a diagnostic environment, it would be necessary to validate variants until a threshold of variant accuracy is identified; this would likely only be established by the sequencing of thousands of samples.

In patients where no possible putative disease-causing mutations were identified (such as Patient 2), reanalysis would be possible at a later date. Reanalysis of this specific sample five months after the initial analysis identified four new variants for investigation. Although none of these mutations were found to be causative, their emergence and other reports in literature (Wright et al., 2018) hold positive

expectations for later reanalysis. It is unknown how many of the patients recruited in this project as well as patients outside of this project that request exome sequencing, have monogenic aetiologies for their disorders. While this testing method drastically improves the diagnostic yield of these patients, there is no guarantee that a diagnosis is possible. As such, clinicians and counsellors will need to be prepared to counsel and support families in which no diagnosis is identified. Following this, alternative testing methods are available such as whole genome sequencing, RNA analysis (transcriptomics), long-read sequencing, and investigating alternative cell-types. These methods have been shown to provide some diagnoses in cases of 'negative' exome results (Meienberg et al., 2016; Byron et al., 2016; Frésard and Montgomery 2018; Merker et al., 2018).

We have successfully investigated the potential of locally implementing WES to detect putative disease-causing SNVs in South African DD patients however, we did not investigate the detection and analysis of CNVs. In the near future, a CNV baseline will need to be generated for this laboratory's Ion Torrent system to efficiently call CNVs and then set up a pipeline for their analysis. Generating this baseline in-house creates a unique opportunity of using ethnically and racially matched controls to allow for better analysis of African data. This would be a step towards the implementation of personalised precision medicine in South Africa.

An additional consideration arising from the results of this project was to implement a clinical meeting with medical geneticists to describe and discuss the phenotypes of Patients' 3 and 7 and the corresponding variants identified in the *EFTUD2* gene. This would provide clinicians with a better indication of the phenotypical presentation of Mandibulofacial Dysostosis with Microcephaly in African patients. Additionally, this data and the other patient data/results will contribute to the accurate clinical phenotyping of DD patients in South Africa which is a larger aim of the DDD-Africa study under which this project operated.

4.9 Conclusion

This project aimed to implement a WES workflow and evaluate the financial and computational costs involved in the technique. The workflow was achieved and refined by performing WES on seven DD patients on the Ion Torrent S5™ sequencing platform and analysing the generated data across three comparative programs.

Positive diagnoses were identified in 70% of patients involved in this small sub-study, showing the benefits of WES in increasing the diagnostic yield of patients with DD in comparison to current locally available testing methods. This study identified the first reported case (to our knowledge) of Takenouchi-Kosaki syndrome in Africa and extended the phenotype of two additional rare genetic disorders.

The financial and computational costs were estimated to be approximately R13,000.00 per patient which was a cheaper alternative to the current practice of outsourcing sequencing services internationally. Additionally, the costing appeared comparative to that of local research testing facilities highlighting the feasibility of implementing this test at this local laboratory.

The resources necessary to establish and implement a custom in-house bioinformatic pipeline for data analysis in a diagnostic setting were investigated. Absent resources were identified and quoted to better understand the requirements of possibly implementing this pipeline in the future with the expansion of local diagnostic NGS testing.

To conclude, it was determined that WES would be a highly beneficial tool in the investigation of DD within this department. The feasibility of local implementation would allow steps towards personalised precision medicine in South Africa, job creation, and local skill development.

5. References

Adzhubei, I.A., Schmidt, S., Peshkin, L., Raemnsky, V.E., Gerasimova, A., Bork, P., et al., 2010. A method and server for predicting damaging missense mutations. Nature Methods 7, 248-249. https://doi.org/10.1038/nmeth0410-248

Ajay, S.S., Parker, S.C., Abaan, H.O., Fajardo, K.V., Margulies, E.H., 2011. Accurate and Comprehensive Sequencing of Personal Genomes. Genome Research 21, 1498-1505. https://doi.org/10.1101/gr.123638.111

Akasaka-Manya, K., Manya, H., Nakajima, A., Kawakita, M., Endo, T., 2006. Physical and functional association of human protein O-mannosyltransferases 1 and 2. Journal of Biological Chemistry 281, 19339-19345. https://doi.org/10.1074/jbc.M601091200

Ambardar, S., Gupta, R., Trakroo, D., Lal, R., Vakhlu, J., 2016. High Throughput Sequencing: An Overview of Sequencing Chemistry. Indian Journal of Microbiology 56, 394–404. https://doi.org/10.1007/s12088-016-0606-4

Arts, P., Simons, A., AlZharani, M.S., Yilmaz, E., Alldrissa, E., van Aerde, K.J., et al., 2019. Exome sequencing in routine diagnostics: a generic test for 254 patients with primary immunodeficiencies. Genome Med 11, 38. https://doi.org/10.1186/s13073-019-0649-3

Baker, S.W., Murrell, J.R., Nesbitt, A.I., Pechter, K.B., Balciuniene, J., Zhao, X., et al., 2019. Automated clinical exome reanalysis reveals novel diagnosis. The Journal of Molecular Diagnostics 21, 38-48. https://doi.org/10.1016/j.jmoldx.2018.07.008

Ball, M., List, A.F., Padron, E., 2016. When clinical heterogeneity exceeds genetic heterogeneity: thinking outside the genomic box in chronic myelomonocytic leukemia. Blood 128, 2381-2387. https://doi.org/10.1111/cge.13365

Batzir, N.A., Shohat, M., Maya, I., 2015. Chromosomal Microarray Analysis (CMA) a Clinical Diagnostic Tool in the Prenatal and Postnatal Setting. Pediatric Endocrinology Reviews 13, 448-454.

Beauchamp, M., Djedid, A., Daupin, K., Clokie, K., Kumar, S., Majewski, J., et al., 2019. Loss of function mutation of *Eftud2*, the gene responsible for mandibulofacial dysostosis with microcephaly (MFDM), leads to pre-implantation arrest in mouse. PLOS One 14, e0219280. https://doi.org/10.1371/journal.pone.0219280

Beck, T.F., Mullikin, J.C., Biesecker, L.G., 2016. Systematic evaluation of Sanger validation of NextGen sequencing variants. Clinical Chemistry 62, 647-654. https://doi.org/10.1373/clinchem.2015.249623

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456, 53-59.

BGI- Exome Seq Fall Pricing Promotion [WWW Document], 2015. BGI - Global. URL https://www.bgi.com/global/company/news/exome-sequencing-fall-promotion/ (Accessed 3 November 2019).

Bone, W.P., Washington, N.L., Buske, O.J., Adams, M.D., Davis, J., Draper, D., et al., 2016. Computational evaluation of exome sequence data using human and model organism phenotypes improves diagnostic efficiency. Genetics in Medicine 18, 608-617. https://doi.org/10.1038/gim.2015.137

Bookman, E.B., Langehorne, A.A., Eckfeldt, J.H., Glass, K.C., Jarvic, G.P., Klag, M., et al., 2006. Reporting genetic results in research studies: summary and recommendations of an NHLBI working group. American Journal of Medical Genetics Part A 140, 1033-1040. https://doi.org/10.1002/ajmg.a.31195

Boyle, C.A., Boulet, S., Schieve, L.A., Cohen, R.A., Blumberg, S.J., Yeargin-Allsopp, M., et al., 2011. Trends in the Prevalence of Developmental Disabilities in US Children, 1997–2008. Pediatrics 127, 1034–1042. https://doi.org/10.1542/peds.2010-2989

Bragg, L.M., Stone, G., Butler, M.K., Hugenholtz, P., Tyson, G.W., 2013. Shining a Light on Dark Sequencing: Characterising Errors in Ion Torrent PGM Data. PLoS Computational Biology 9, e1003031. https://doi.org/10.1371/journal.pcbi.1003031

Brnich, S.E., Tayoun, A.N.A., Couch, F.J., Cutting, G.R., Greenblatt, M.S., Heinen, C.D., et al., 2020. Recommendations for application of the functional evidence PS3/BS3 criterion using ACMG/AMP sequence variant interpretation framework. Genome Medicine 12, 3. https://doi.org/10.1186/s13073-019-0690-2

Budin-Ljøsne, I., Mascalzoni, D., Soini, S., Machado, H., Kaye, J., Bentzen, H.B., et al., 2016. Feedback of individual genetic results to research participants: is it feasible in Europe? Biopreservation and Biobanking 14, 241-248. https://doi.org/10.1089/bio.2015.0115

Byron, S.A., Van Keuren-Jensen, K.R., Engelthaler, D.M., Carpten, J.D., Craig, D.W., 2016. Translating RNA sequencing into clinical diagnostics: oppurtunities and challenges. Nature Reviews Genetics 17, 257-271. https://doi.org/ 10.1038/nrg.2016.10

Campbell, M.C., and Tishkoff, S.A., 2010. African Genetic Diversity: Implications for human demographic history, modern human origins, and complex disease mapping. Annual Review of Genomics and Human Genetics 9, 403-433. https://doi.org/10.1146/annurev.genom.9.081307.164258

Cao, Y., Fanning, S., Proos, S., Jordan, K., Srikumar, S., 2017. A review on the applications of next generation sequencing technologies as applied to food-related microbiome studies. Frontiers in Microbiology 8, 1829. https://doi.org/10.3389/fmicb.2017.01829

Castellana, S., Ronai, J., Mazza, T., 2015. MitImpact: and exhaustive collection of pre-computed pathogenicity prediction of human mitochondrial and non-synonymous variants. Human Mutation 36, E2413-E2422. https://doi.org/10.1002/humu.22720

Celestino-Soper, P.B.S., Gao, H., Lynnes, T.C., Lin, H., Liu, Y., Spoonamore, K.G., et al., 2017. Validation and utilisation of a clinical next generation sequencing panel for selected cardiovascular disorders. Frontiers in Cardiovascular Medicine 4, 11. https://doi.org/10.3389/fcvm.2017.00011

Cheon, C., and Ko, J.M., 2015. Kabuki syndrome: clinical and molecular characteristics. Korean Journal of Pediatrics 58, 317-324. https://doi.org/10.3345/kjp.2015.58.9.317

Clark, M.M., Hildreth, A., Batalov, S., Ding, Y., Chowdhury, S., Watkins, K., et al., 2019. Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation. Science Translational Medicine 11. https://doi.org/10.1126/scitranslmed.aat6177

Córdoba, M., Rodriguez-Quiroga, S.A., Vega, P.A., Salinas, V., Perez-Maturo, J., Amartino, H., et al., 2018. Whole exome sequencing in neurogenetic odysseys: An effective, cost- and time-saving diagnostic approach. PLOS ONE 13, e0191228. https://doi.org/10.1371/journal.pone.0191228

Cunningham, F., Achuthan, P., Akanni, W., Allen, J., Amode, M.R., Armean, I.M., et al., 2018. Ensembl 2019. Nucleic Acids Research 47, D745-D751. https://doi.org/10.1093/nar/gky1113

Daber, R., Sukhadia, S., Morrissette, J.J.D., 2013. Understanding the limitations of next generation sequencing informatics, an approach to clinical pipeline validation using artificial data sets. Cancer Genetics 206, 441–448. https://doi.org/10.1016/j.cancergen.2013.11.005

Damiati, E., Borsani, G., Giacopuzzi, E., 2016. Amplicon-based semiconductor sequencing of human exomes: performance evaluation and optimization strategies. Human Genetics 135, 499-511. https://doi.org/10.1007/s00439-016-1656-8

De Haan, M.C., Nio, C.Y., Thomeer, M., de Vries, A.H., Bossuyt, P.M., Kuipers, E.J., et al., 2012. Comparing the Diagnostic Yields of Technologists and Radiologists in an Invitational Colorectal Cancer Screening Program Performed with CT Colonography. Radiology 264, 771-778. https://doi.org/10.1148/radiol.12112486

De Menil, V., Hoogenhout, M., Kipkemoi, P., Kamuya, D., Eastman, E., Galvin, A., et al., 2019. The NeuroDev Study: Phenotypic and Genetic Characterisation of Neurodevelopmental Disorders in Kenya and South Africa. Neuron 101, 15-19. https://doi.org/10.1016/j.neuron.2018.12.016.

Deepti, G., Carstensen, T., Fatumo, S., Chen, G., Franklin, C.S., Prado-Martinez, J., et al., 2019. Uganda genome resource enables insights into population history and genomic diversity in Africa. Cell 179, 984-1002. https://doi.org/10.1016/j.cell.2019.10.004

Desvignes, J., Bartoli, M., Delague, V., Krahn, M., Miltgen, M., Béroud, C., et al., 2018. VarAFT: a variant annotation and filtration system for human next generation sequencing data. Nucleic Acids Research 46, W545-W553. https://doi.org/10.1093/nar/gky471

Dunn, N.A., Unni, D.R., Diesh, C., Munoz-Torres, M., Harris, N.L., Yao, E., et al., 2019. Apollo: Democratizing genome annotation. PLoS Computational Biology 15, e1006790. https://doi.org/10.1371/journal.pcbi.1006790

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., et al., 2019. The Pfam protein families database in 2019. Nucleic Acids Research 47, D427-D432. https://doi.org/10.1093/nar/gky995

Essop, F.B., Krause, A., 2013. Diagnostic, carrier and prenatal genetic testing for fragile X syndrome and other FMR-1-related disorders in Johannesburg, South Africa: A 20-year review. South African Medical Journal 103, 994-998–998. https://doi.org/10.7196/SAMJ.7144

Ewans, L.J., Schofield, D., Shrestha, R., Zhu, Y., Gayevskiy, V., Ying, K., et al., 2018. Whole-exome sequencing reanalysis at 12 months boosts diagnosis and is cost-effective when applied early in Mendelian disorders. Genetics in Medicine 20, 1564–1574. https://doi.org/10.1038/gim.2018.39

Ewing, B., Hillier, L., Wendl, M.C., Green, P., 1998. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. Genome Research 8, 175–185.

Fieggen, K.J., Lambie, L.A., Donald, K.A., 2019. Investigating developmental delay in South Africa: A pragmatic approach. South African Journal of Medicine 109, 210-213.

Firth, H.V., Richards, S.M., Bevan, A.P., Clayton, S., Corpas, M., Rajan, D., et al., 2009. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using ENSEMBL Resources. American Journal of Human Genetics 84, P524-533. https://doi.org/10.1016/j.ajhg.2009.03.010

Flores-Juárez, C.R., González-Jasso, E., Antaramian, A., Pless, R.C., 2016. PCR amplification of GC-rich DNA regions using the nucleotide analog N4-methyl-2´-deoxycytidine 5´-triphosphate. Bio Techniques 61, 175-182. https://doi.org/10.2144/000114457

Fogel, B.L., Satya-Murti, S., Cohen, B.H., 2016. Clinical exome sequencing in neurologic disease. Neurology: Clinical Practice 6, 164–176. https://doi.org/10.1212/CPJ.0000000000000239

Frésard, L. and Montgomery, S.B., 2018. Diagnosing rare disease after the exome. Cold Spring Harbor Molecular Case Studies 4, a003392. doi: 10.1101/mcs.a003392

Genomenon., 2020. Mastermind Genomic Search Engine. Available at: https://www.genomenon.com/mastermind (Accessed 15 April 2020)

Gibson G., 2011. Rare and Common Variants: Twenty Arguments. Nature Review Genetics 13, 135-145. https://doi.org/10.1038/nrg3118

Gilissen, C., Hehir-Kwa, J.Y., Thung, D.T., van de Vorst, M., van Bon, B.W.M., Willemsen, M.H., et al., 2014. Genome sequencing identifies major causes of severe intellectual disability. Nature 511, 344–347. https://doi.org/10.1038/nature13394

Goodwin, S., McPherson, J.D., McCombie, W.R., 2016. Coming of Age: Ten Years of Next-Generation Sequencing Technologies. Nature Reviews 17, 333-351.

Gomez, F., Hirbo, J., Tishkoff, S.A., 2014. Genetics variation and adaptation in Africa: implications for human evolution and disease. Cold Spring Harbor Perspectives in Biology 6, a008524. https://doi.org/10.1101/cshperspect.a008524

Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., et al., 2015. The African Genome Variation Project Shapes Medical Genetics in Africa. Nature 517, 327-332. https://doi.org/10.1038/nature13997

Habanabakize, T., 2020. The effect of economic growth and exchange rate on imports and exports: the South African post-2008 financial crisis case. International Journal of Economics and Finance Studies 12, 223-238. https://doi.org/10.34109/ijefs.202012114

Haft, D.H., DiCuccio, M., Badretdin, A., Brover, V., Chetvernin, V., O'Neil, K., et al., 2018. RefSeq: an update on prokaryotic genome annotation and curation. Nucleic Acids Research 46, D851-D860. https://doi.org/10.1093/nar/gkx1068

Harel, T. and Lupski, J.R., 2017. Genomic disorders 20 years on- mechanisms for clinical manifestations. Clinical Genetics 93, 439-449. https://doi.org/10.1111/cge.13146

Hayes, A. and Batshaw, M.L., 1993. Down syndrome. Pediatric Clinics of North America 40, 523-535. https://doi.org/10.1016/S0031-3955(16)38548-0

Homer, N., Merriman, B., Nelson, S.F., 2009. Local alignment of two-base encoded DNA sequence. BMC Bioinformatics 10, 175. http://dx.doi.org/10.1186/1471-2105-10-175

Huang, L., Vanstone, M.R., Hartley, T., Osmond, M., Barrowman, N., Allanson, J., et al., 2016. Mandibulofacial Dysostosis with Microcephaly: Mutation and Database Update. Human Mutation 37, 148-154. https://doi.org/10.1002/humu.22924

IBM., 2020. IBM Tape Drives. Available at: https://www.ibm.com/za-en/it-infrastructure/storage/tape/drives (Accessed: 3 July 2020).

Iglesias, A., Anyane-Yeboa, K., Wynn, J., Wilson, A., Cho, M.T., Guzman, E., et al., 2014. The usefulness of whole-exome sequencing in routine clinical practice. Gen Med 16, 922-931. https://doi.org/10.1038/gim.2014.58

Illumina Inc., 2020. Specifications for the MiSeq system. Available at: https://emea.illumina.com/systems/sequencing-platforms/miseq/specifications.html (Accessed: 29 April 2020).

Institute of Medicine (US) Committee on Nervous System Disorders in Developing Countries., 2001. Neurological, Psychiatric, and Developmental Disorders: Meeting the Challenge in the Developing World. National Academic Press, Washington (DC) USA. Chapter 5, Developmental Disabilities. Available from: https://www.ncbi.nlm.nih.gov/books/NBK223473/

Ion Torrent., 2014. Torrent Suite-Signal Processing and Base Calling Application Note Torrent Suite Software Analysis Pipeline. Technical Note. Available online: http://coolgenes.cahe.wsu.edu/ion-docs/Technical-Note-\--Analysis-Pipeline_6455567.html (Accessed: 10 April 2020).

Johnson, D.I. and Pringle, J.R., 1990. Molecular Characterisation of *CDC42*, a *Saccharomyces cerevisiae* Gene Involved in the Development of Cell Polarity. The Journal of Cell Biology 111, 143-152. https://doi.org/10.1083/jcb.111.1.143

Joseph, F.M., 2018. A key role for the genetic counsellor in the genomics era. F1000 Research 7, 420. https://doi.org/10.12688/f1000research.14222.2

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alfoldi, J., Wang, Q., et al., 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. bioRxiv. https://doi.org/10.1101/531210

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., et al., 2002. The human genome browser at UCSC. Genome Res. 12, 996–1006. https://doi.org/10.1101/gr.229102

Kharbanda, M., Tolmie, J., Joss, S., 2015. How to use… microarray comparative genomic hybridisation to investigate developmental disorders. Archives of disease in childhood - Education & practice edition 100, 24–29. https://doi.org/10.1136/archdischild-2014-306022

Kieleczawa, J., 2006. Fundamentals of Sequencing Difficult Templates- An Overview. Journal of Biomolecular Techniques 17, 207-217.

Kohler, S., Carmody, L., Vasilevsky, N., Jacobsen, J.O.B., Danis, D., Gourdine, J., et al., 2018. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Research 47, D1018-D1027. https://doi.org/10.1093/nar/gky1105

Kopanos, C., Tsiolkas, V., Kouris, A., Chapple, C.E., Aguilera, M.A., Meyer, R., et al., 2018. VarSome: The Human Genomic Variant Search Engine. Oxford Bioinformatics 35, 1978-1980. https://doi.org/10.1093/bioinformatics/bty897

Kõressaar, T., Lepamets, M., Kaplinski, L., Raime, K., Andreson, R., Remm, M., 2018. Primer3_masker: integrating masking of template sequence with primer design software. Bioinformatics 34, 1937–1938. https://doi.org/10.1093/bioinformatics/bty036

Kulkarni, P. and Frommolt, P., 2017. Challenges in the setup of large-scale next-generation sequencing analysis workflows. Computational and Structural Biotechnology Journal 15, 471-477. https://doi.org/10.1016/j.csbj.2017.10.001

Lahens, N.F., Ricciotti, E., Smirnova, O., Toorens, E., Kim, E.J., Baruzzo, G., et al., 2017. A comparison of Illumina and Ion Torrent sequencing platforms in the context of differential gene expression. BMC Genomics 18, 602. https://doi.org/10.1186/s12864-017-4011-0

Lai, D., and Love, D. R. 2012. Automation of a primer design and evaluation pipeline for subsequent sequencing of the coding regions of all human Refseq genes. Bioinformation 8, 365–368. https://doi.org/10.6026/97320630008365

Landrum, M.J., Kattman, B.L., 2018. ClinVar at five years: Delivering on the promise. Human Mutation 39, 1623-1630. https://doi.org/10.1002/humu.23641

Lappalailen, I., Lopez, J., Skipper, L., Hefferon, T., Spalding, J.D., Garner, J., et al., 2013. dbVar and DGVa: public archives for genomic structural variation. Nucleic Acids Research 41, D936-D941. https://doi.org/10.1093/nar/gks1213

Larizza, L. and Finelli, P., 2018. Developmental disorders with intellectual disability driven by chromatin dysregulation: Clinical overlaps and molecular mechanisms. Clinical Genetics 1-10. https://doi.org/10.1111/cge.13365

Lee, I., Lee, K., Hsing, M., Choe, Y., Park, H., Kim, S.H., et al., 2014. Prioritizing disease-linked variants, genes, and pathways with an interactive whole genome analysis pipeline. Human Mutation 35, 537-547. https://doi.org/10.1002/humu.22520

Legius, E., Schollen, E., Matthijs, G., Fryns, P., 1998. Fine mapping of Noonan/cardio-facio cutaneous syndrome in a large family. European Journal of Human Genetics 6, 32-37.

Lesko, L.J., Zineh, I., Huang, S.-M., 2010. What Is Clinical Utility and Why Should We Care? Clinical Pharmacology & Therapeutics 88, 729–733. https://doi.org/10.1038/clpt.2010.229

Lier, A., Penzel, R., Heining, C., Horak, P., Fröhlich, M., Uhrig, S., et al., 2018. Validating comprehensive next-generation sequencing results for precision oncology: the NCT/DKTK molecularly aided stratification for tumor eradication research experience. JCO Precision Oncology 2, 1-13.

Lines, M.A., Huang, L., Schwartzentruber, J., Douglas, S.L., Lynch, D.C., Beaulieu, C., et al., 2012. Haploinsufficiency of a spliceosomal GTPase encoded by EFTUD2 causes mandibulofacial dysostosis with microcephaly. American Journal of Human Genetics 90, 369-377. https://doi.org/10.1016/j.ajhg.2011.12.023

Liu, X., Wu, C., Li, C., Boerwinkle, E., 2016. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Non-synonymous and Splice Site SNVs. Human Mutation 37, 235-241.

Lott, M.T., Leipzig, J.N., Derbeneva, O., Xie, H.M., Chalkia, D., Sarmady, M., et al., 20113. mtDNA variation and analysis using MITOMAP and MITOMASTER. Current Protocols in Bioinformatics 1, 1.23.1-26.

Lumaka, A., Race, V., Peeters, H., Corveleyn, A., Coban-Akdemir, Z., Jhangiani, S.N., et al., 2018. A comprehensive clinical and genetic study in 127 patients with ID in Kinshasa, DR Congo. American Journal of Medical Genetics 176, 1897-1909. https://doi.org/10.1002/ajmg.a.40382

Luquetti, D.V., Hing, A.V., Rieder, M.J., Nickerson, D.A., Turner, E.H. et al., 2013. “Madibulofacial Dysostosis With Microcephaly” Caused by EFTUD2 Mutations: Expanding the Phenotype. America Journal of Medical Genetics 161A, 108-113. https://doi.org/10.1002/ajmg.a.35696

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 Research 42, D986-D992. https://doi.org/10.1093/nar/gkt958

Mammedov, T.G., Pienaar, E., Whitney, S.E., TerMaat, J.R., Carvill, G., Subramanian, A., et al., 2008. A Fundamental Study of the PCR Amplification of GC-Rich DNA Templates. Computational Biological Chemistry 32, 452-457. https://doi.org/10.1016/j.compbiolchem.2008.07.021

Marine, R.L., Magaña, L.C., Castro, C.J., Zhao, K., Montmayeur, A.M., Schmidt, A., et al., 2020. Comparison of Illumina MiSeq and the Ion Torrent PGM and S5 platforms for whole-genome sequencing of picornaviruses and caliciviruses. Journal of Virological Methods 280, 113865. https://doi.org/10.1016/j.jviromet.2020.113865

Martinelli, S., Krumbach, O.H.F., Pantaleoni, F., Copolla, S., Amin, E., Pannone, L., et al., 2018. Functional Dysregulation of *CDC42* Cause Diverse Developmental Phenotypes. The American Journal of Human Genetics 102, 309-320. https://doi.org/10.1016/j.ajhg.2017.12.015

Matsuo, M., Yamauchi, A., Ito, Y., Sakauchi, M., Yamamoto, T., Okamoto, N., et al., 2017. Mandibulofacial Dysostosis With Microcephaly: A Case Presenting With Seizures. Brain & Development 39, 177-181. https://doi.org//10.1016/j.braindev.2016.08.008

McKusick, V.A., 1998. Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders. Baltimore: Johns Hopkins University Press, 12.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., et al., 2016. The ensembl variant effect predictor. Genome Biology 17, 1-14. https://doi.org/10.1186/s13059-016-0974-4

Meienberg, J., Bruggman, R., Oexle, K., Matyas, G., 2016. Clinical sequencing: is WGS better than WES? Human Genetics 135, 359-362. https://doi.org/10.1007/s00439-015-1631-9

Merker, J.D., Wenger, A.M., Sneddon, T., Grove, M., Zappala, Z., Frésard, L., et al., 2018. Long-read genome sequencing identifies causal structural variation in a Mendelian disease. Genetics in Medicine 20, 159-163. https://doi.org/10.1038/gim.2017.86

Merriman, B., Ion Torrent R&D Team, Rothberg, J.M., 2012. Progress on Ion Torrent semiconductor chip based sequencing. Electrophoresis 33, 3397-3417.

Miclea, D., Peca, L., Cuzmici, Z., Pop, I.V., 2015. Genetic testing in patients with global developmental delay/intellectual disabilities. A review. Clujul Medical 88, 288–292. https://doi.org/10.15386/cjmed-461

Miller, D.T., Adam, M.P., Aradhya, S., Biesecker, L.G., Brothman, A.R., Carter, N.P., et al., 2010. Consensus Statement: Chromosomal Microarray Is a First-Tier Clinical Diagnostic Test for Individuals with Developmental Disabilities or Congenital Anomalies. The American Journal of Human Genetics 86, 749–764. https://doi.org/10.1016/j.ajhg.2010.04.006

Miller, S.A., Dykes, D.D., Polesky, H.F., 1988. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Research 11, 1215. https://doi.org/10.1093/nar/16.3.1215

Monroe, G.R., Frederix, G.W., Savelberg, S.M.C., de Vries, T.I., Duran, K.J., van der Smagt, J.J., et al., 2016. Effectiveness of whole-exome sequencing and costs of the traditional diagnostic trajectory in children with intellectual disability. Genetics in Medicine 18, 949–956. https://doi.org/10.1038/gim.2015.200

Moorthie, S., Hall, A., Wright, C.F., 2013. Informatics and clinical genome sequencing: Opening the black box. Genetics in Medicine 15, 165-171. https://doi.org/10.1038/gim.2012.116

Mu, W., Lu, H., Chen, J., Li, S., Elliot, A.M., 2016. Sanger confirmation is required to achieve optimal sensitivity and specificity in next generation sequencing panel testing. The Journal of Molecular Diagnostics 18, 923-932. https://doi.org/10.1016/j.jmoldx.2016.07.006

Nakayama, Y., Yamaguchi, H., Einaga, N., Esumi, M., 2016. Pitfalls of DNA Quantification using DNA-binding fluorescent dyes and suggested solutions. PLoS One 11, e0150528. https://doi.org/10.1371/journal.pone.0150528

NCBI Resource Coordinators., 2018. Database resources of the National Center for Biotechnology Information. Nucleic acids research 46(D1), D8–D13. https://doi.org/10.1093/nar/gkx1095

Need, A.C., Shashi, V., Hitomi, Y., Schoch, K., Shianna, K.V., McDonald, M.T., et al., 2012. Clinical application of exome sequencing in undiagnosed genetic conditions. Journal of Medical Genetics 49, 353–361. https://doi.org/10.1136/jmedgenet-2012-100819

Ng, S.B., Bigham, A.W., Buckingham, K.J., Hannibal, M.C., McMillin, M.J., Gildersleeve, H.I., et al., 2010. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nature Genetic 42, 790-793. https://doi.org/10.1038/ng.646

Ochs, H.D., 2009. Mutations of the Wiskott-Aldrich Syndrome Protein affect protein expression and dictate the clinical phenotype. Immunology Research 44, 84-88. https://doi.org/10.1007/s12026-008-8084-3

Papaz, T., Liston, E., Zahavich, L., Stavropoulus, D.J., Jobling, R.K., Kim, R.H., et al., 2019. Return of genetic and genomic research findings: experience of a pediatric biorepository. BMC Medical Genomics 12, 173. https://doi.org/10.1186/s12920-019-0618-0

Pereira, R., Oliverira, J., Sousa, M., 2020. Bioinformatics and Computational Tools for Next-Generation Sequencing Analysis in Clinical Genetics. Journal of Clinical Medicine 9, 132. https://doi.org/10.3390/jcm9010132

Qiagen., 2020. CLC Main Workbench User Guide. Denmark: Qiagen, 368-392.

Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Connor, S.R., Betroni, A., et al., 2012. A tale of three next generation sequencing platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13, 341. https://doi.org/10.1186/1471-2164-13-341

Rajaratnam, A., Shergill, J., Salcedo-Arellane, M., Saldarriaga, W., Duan, X., Hagerman, R., 2017. Fragile X syndrome and fragile X-associated disorders. F1000Research 6, 2112. https://doi.org/10.12688/f1000research.11885.1

Rehm, H.L., Bale, S.J., Bayrak-Toydemir, P., Berg, J.S., Brown, K.K., Deignan, J.L., et al., 2013. ACMG clinical laboratory standards for next generation sequencing. Genetics in Medicine 15, 733-747. https://doi.org/10.1038/gim.2013.92

Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J., Kircher, M., 2018. CADD: Predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Research 47, D886-894. https://doi.org/10.1093/nar/gky1016

Richards, S., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., et al., 2015. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genetics in Medicine 17, 405–423. https://doi.org/10.1038/gim.2015.30

Robin, J. D., Ludlow, A. T., LaRanger, R., Wright, W. E., Shay, J. W., 2016. Comparison of DNA Quantification Methods for Next Generation Sequencing. Scientific Reports 6, 1-10. https://doi.org/10.1038/srep24067

Rosell, A.M., Pena, L.D.M., Schoch, K., Spillmann, R., Sullivan, J., Hooper, S.R., et al., 2016. Not the End of the Odyssey: Parental Perceptions of Whole Exome Sequencing (WES) in Pediatric Undiagnosed Disorders. Journal of Genetic Counseling 25, 1019–1031. https://doi.org/10.1007/s10897-016-9933-1

Rotimi, C.N., Bentley, A.R., Doumatey, A.P., Chen, G., Shriner, S., Adeyemo, A., 2017. The genomic landscape of African populations in health and disease. Human Molecular Genetics 26, R225-R236. https://doi.org/10.1093/hmg/ddx253

Salgado, D., Desvignes, J., Rai, G., Blanchard, A., Miltgen, M., Pinard, A., et al., 2016. UMD-Predictor: A High-Throughput Sequencing Compliant System for Pathogenicity Prediction of Any Human cDNA Substitution. Human Mutation 37, 439-446. https://doi.org/10.1002/humu.22965

Sanger, F., Coulson, A.R., 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. Journal of Molecular Biology 94, 441–448. https://doi.org/10.1016/0022-2836(75)90213-2

Sarkar, A., Emrick, L.T., Smith, E.M., Austin, E.G., Yang, Y., Hunter, J.V., et al., 2014. Novel *De novo* Mutations in EFTUD2 Detected by Exome Sequencing in Mandibulofacial Dysostosis with Microcephaly Syndrome. American Journal of Medical Genetics Part A, 914-918. https://doi.org/10.1002/ajmg.a.36948

Sarkozy, P., Enyedi, M.Z., Antal, P., 2014. Flow index based characterization of next generation sequencing errors. Visualizing pyrosequencing and semiconductor sequencing to cope with homopolymer errors. In the proceedings of the international conference on bioinformatic models, methods and algorithms, 271-277. https://doi.org/10.5220/0004924902710277

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. Genetics in Medicine 20, 1122–1130. https://doi.org/10.1038/gim.2017.247

Shabani, M., Vears, D., Borry, P., 2018. Raw genomic data: storage, access, and sharing. Trends in Genetics 34, 8-10. https://doi.org/10.1016/j.tig.2017.10.004

Shearer, A.E., Eppsteiner, R.W., Booth, K.T., Ephraim, S.S., Gurrola, J., Simpson, A., et al., 2014. Utilizing ethnic-specific differences in minor allele frequency to recategorize reported pathogenic deafness variants. The American Journal of Human Genetics 95, 445-453. http://dx.doi.org/10.1016/j.ajhg.2014.09.001

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., et al., 2001. dbSNP: the NCBI dataset of genetic variation. Nucleic Acids Research 29, 308-311.

Shevell, M., Ashwal, S., Donley, D., Flint, J., Gingold, M., Hirtz, D., et al., 2003. CME Practice parameter: Evaluation of the child with global developmental delay. American Academy of Neurology 60, 367–380.

Sim, N., Kumar, P., Hu, J., Henikoff, S., Schneider, G., Ng, P.C., 2012. SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Research 40, W452–W457, https://doi.org/10.1093/nar/gks539

Skirton, H., Cordier, C., Ingvoldstad, C., Taris, N., Benjamin, C., 2015. The role of the genetic counsellor: a systematic review of research evidence. European Journal of Human Genetics 23, 452-458. https://doi.org/10.1038/ejhg.2014.116

Smith, T.F., and Waterman, M.S., 1981. Identification of common molecular subsequences. Journal of Molecular Biology 147, 195–197.

Soden, S.E., Saunders, C.J., Willig, L.K., Farrow, E.G., Smith, L.D., Petrikin, J.E., et al., 2014. Effectiveness of exome and genome sequencing guided by acuity of illness for diagnosis of neurodevelopmental disorders. Science Translational Medicine 6, 265ra168. https://doi.org/10.1126/scitranslmed.3010076

SolarWinds Worldwide., 2019. What is server management? Definition, best practices, and best software 2020. Available online: https://www.dnsstuff.com/server-management (Accessed: 1 July 2020).

Song, L., Huang, W., Kang, J., Huang, Y., Ren, H., Ding, K., 2017. Comparison of error correction algorithms for Ion Torrent PGM data: application to hepatitis B virus. Scientific Reports 7, 8106. https://doi.org/10.1038/s41598-017-08139-y

Srivastava, S., Cohen, J.S., Vernon, H., Barañano, K., McClellan, R., Jamal, L., et al., 2014. Clinical whole exome sequencing in child neurology practice. Annals of Neurology 76, 473-483. https://doi.org/10.1002/ana.24251

Srivastava, S., Love-Nichols, J.A., Dies, K.A., Ledbetter, D.H., Martin, C.L., Chung, W.K., et al., 2019. Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders. Genetics in Medicine 21, 2413-2421. https://doi.org/10.1038/s41436-019-0554-6

Srour, M., Shevell, M., 2015. Chapter 14 - Global Developmental Delay and Intellectual Disability, in: Rosenberg, R.N., Pascual, J.M. (Eds.), Rosenberg's Molecular and Genetic Basis of Neurological and Psychiatric Disease (Fifth Edition). Academic Press, Boston, pp. 151–161. https://doi.org/10.1016/B978-0-12-410529-4.00014-0

Steenhuysen, J., 2015. Personal Genome Test Will Sell at New Low Price of $250. Scientific America, 22 September. Available at: https://www.scientificamerican.com/article/craig-venter-s-company-in-deal-for-whole-exome-tests-at-new-low-cost/ (Accessed: 9 April 2019)

Strand Life Sciences, 2016. Guide to storage and computation requirements. StrandNGS, viewed 7 January 2020. https://www.strand-ngs.com/support/ngs-data-storage-requirements

Suwinski, P., Ong, C., Ling, M.H.T., Poh, Y.M., Khan, A.M., Ong, H.S., 2019. Advancing Personalised Medicine Through the Application of Whole Exome Sequencing and Big Data Analytics. Frontiers in Genetics 10, 49. https://doi.org/10.3389/fgene.2019.00049

Swaminathan, G.J., Bragin, E., Chatzimichali, E.A., Corpas, M., Bevan, A.P., Wright, C.F., et al., 2012. DECIPHER: web-based, community resource for clinical interpretation of rare variants in developmental disorders. Human Molecular Genetics 21, R37–R44. https://doi.org/10.1093/hmg/dds362

Takenouchi, T., Kosaki, R., Niizuma, T., Hata, K., Kosaki, K., 2015. Macrothrombocytopenia and Developmental Delay with a *de novo CDC42* Mutation: Yet Another Locus for Thrombocytopenia and Developmental Delay. American Journal of Medical Genetics Part A, 2822-2825. https://doi.org/10.1002/ajmg.a.37275

Tekendo-Ngongang, C., Agenbag, G., Bope, C.D., Esterhuizen, A. I., Wonkam, A., 2019. Noonan Syndrome in South Africa: Clinical and Molecular Profiles. Frontiers in Genetics 10, 333. https://doi.org/10.3389/fgene.2019.00333

Thermo Fisher Scientific Inc., 2018. Ion AmpliSeq™ Exome RDY Library Preparation User Guide. Carlsbad: Life Technologies Holdings, 1-39.

Thermo Fisher Scientific Inc., 2019. Ion Reporter™ Software 5.12 User Guide. Carlsbad: Life Technologies Holdings, 230-276.

Thevenon, J., Duffourd, Y., Masurel-Paulet, A., Lefebvre, M., Feillet, F., Chehadeh-Djebbar, A., et al., 2016. Diagnostic odyssey in severe neurodevelopmental disorders: toward clinical whole-exome sequencing as a first-line diagnostic test. Clinical Genetics 89, 700-707. https://doi.org/10.1111/cge.12732

Thien, C.B.F., Walker, F., Langdon, W.Y., 2001. RING finger mutations that abolish c-Cbl-directed polyubiquitination and downregulation of the EGF receptor are insufficient for cell transformation. Molecular Cell 7, 355-365. doi:10.1016/s1097-2765(01)00183-6

Thormann, A., Halachev, M., McLaren, W., Moore, D.J., Svinti, V., Campbell, A., et al., 2019. Flexible and scalable diagnostic filtering of genomic variants using G2P with Ensemble VEP. Nature Communications 10, 2373-2382. https://doi.org/10.1038/s41467-019-10016-3

Uehara, T., Suzuki, H., Okamoto, N., Kondoh, T., Ahmad, A., O'Connor, B.C., et al., 2019. Pathogenetic basis of Takenouchi-Kosaki syndrome: Electron microscopy study using platelets in patients and functional studies in a *Caenorhabditis elegans* model. Scientific Reports 9, 4418. https://doi.org/10.1038/s41598-019-40988-7

UNAIDS., 2020. The Global HIV/AIDS Epidemic. HIV.gov, 07 July. Available at: https://www.hiv.gov/hiv-basics/overview/data-and-trends/global-statistics (Accessed: 16 October 2020)

Untergasser, A., Cutcutach, I., Koressaar, T., Ye, J., Faircloth, BC., Remm, M., et al., 2012. Primer3—new capabilities and interfaces. Nucleic Acids Research 40, e115.

Voight, C., Mègarbanè, A., Neveling, K., Czeschik, J.C., Albrecht, B., Calleweart, B., et al., 2013. Oto-facial syndrome and esophageal atresia, intellectual disability and zygomatic anomalies - expanding the phenotypes associated with *EFTUD2* mutations. Orphanet Journal of Rare Diseases 8, 110. https://doi.org/10.1186/1750-1172-8-110

Wang, K., Li, M., Hakonarson, H., 2010. ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data. Nucleic Acids Research 38, e164. https://doi.org/10.1093/nar/gkq603

Wang, Y., Xu, N., Wang, J., Wang, X., 2019. Kabuki syndrome: Review of the Clinical Features, Diagnosis and Epigenetic Mechanisms. World Journal of Pediatrics 15, 525-535. https://doi.org/10.1007/s12519-019-00309-4

Wenger, A.M., Guturu, H., Bernstein, J.A., Nejerano, G., 2017. Systematic reanalysis of clinical exome data yields additional diagnoses: implications for providers. Genetics in Medicine 19, 209-214. https://doi.org/10.1038/gim.2016.88

Wildeman, M., van Ophuizen, E., den Dunnen, J.T., Taschner, P.E.M., 2007. Improving sequence variant descriptions in mutation databases and literature using Mutalyzer sequence variation nomenclature checker. Human Mutation 29, 6-13. https://doi.org/10.1002/humu.20654

Wright, C.F., Clayton, S., Gallone, G., Aitken, S., FitzGerald, T.W., Jones, P., et al., 2018. Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. Genetics in Medicine 20, 1216–1223. https://doi.org/10.1038/gim.2017.246

Wright, C.F., Fitzgerald, T.W., Jones, W.D., Clayton, S., McRae, J.F., van Kogelenberg, M., et al., 2015. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. The Lancet 385, 1305–1314. https://doi.org/10.1016/S0140-6736(14)61705-0

Xiao, B., Qiu, W., Ji, X., Liu, X., Huang, Z., Liu, H., et al., 2018. Marked yield of re-evaluating phenotype and exome/target sequencing data in 33 individuals with intellectual disabilities. American Journal of Medical Genetics Part A 176, 107–115. https://doi.org/10.1002/ajmg.a.38542

Yanagisawa, A., Bouchet, C., Quijano-Roy, S., Vuillaumier-Barrot, S., Clarke, N., Odent, S., et al., 2009. POMT2 intragenic deletions and splicing abnormalities causing congenital muscular dystrophy with mental retardation. European Journal of Medical Genetics 52, 201-206. https://doi.org/10.1016/j.ejmg.2008.12.004

Yeo, Z.X., Chan, M., Yap, Y.S., Ang, P., Rozen, S., Lee, A.S.G.L., 2012. Improving Indel Detection Specificity of the Ion Torrent PGM Benchtop Sequencer. PLoS One 7, e45798. https://doi.org/10.1371/journal.pone.0045798

Zheng, J., Zhang, H., Banerjee, S., Li, Y., Zhou, J., Yang, Q., 2019. A comprehensive assessment of next-generation sequencing variants validation using a secondary technology. Molecular Genetics and Genomic Medicine 7, e:00748. https://doi.org/10.1002/mgg3.748

Zhong, K., Zhu, G., Jing, X., Hendriks, A.E.J., Drop, A.L.S., Ikram, M.A., et al., 2017. Genome-wide compound heterozygote analysis highlights alleles associated with adult height in Europeans. Human Genetics 136, 1407-1417. https://doi.org/10.1007/s00439-017-1842-3

6. Appendices

Appendix 1

UNIVERSITY OF THE WITWATERSRAND JOHANNESBURG

R14/49 Ms K Flynn

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
CLEARANCE CERTIFICATE NO. M190578

NAME: (Principal Investigator)	Ms K Flynn
DEPARTMENT:	School of Pathology Department of Human Genetics National Health Laboratory Service
PROJECT TITLE:	Evaluating whole exome sequencing on the Ion Torrent S5TM as a potential diagnostic tool for developmental disorders
DATE CONSIDERED:	Ad hoc
DECISION:	Approved unconditionally
CONDITIONS:	Sub-study under M180678
SUPERVISOR:	Professor Z Lombard
APPROVED BY:	Dr CB Penny, Chairperson, HREC (Medical)
DATE OF APPROVAL:	2019/06/10

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

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary on the 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 will therefore reports and re-certification will be due early in the month of **May** each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature

Date 10/06/2019

PLEASE QUOTE THE CLEARANCE CERTIFICATE NUMBER IN ALL ENQUIRIES

Appendix 2

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

R14/49 Prof. A Krause & Drs N Carstens & Z Lombard

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

NAME: (Principal Investigator)	Prof. A Krause & Drs N Carstens & Z Lombard
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 NIH Study No. UO1MH115483
SUPERVISOR:	Not applicable
APPROVED BY:	Professor CB Penny, Chairperson, HREC (Medical)
DATE OF APPROVAL:	06/07/2018

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

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary on 3rd floor, Phillip V Tobias Building, Parktown, University of the Witwatersrand, Johannesburg.
I/We fully understand the conditions under which I am/we are authorised to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated from the research protocol as approved, I/we undertake to resubmit to the Committee. **I agree to submit a yearly progress report**. The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. In this case, the study was initially reviewed in **June** and will therefore be due in the month of **June** each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature _____ Date _____

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

Appendix 3

List of Websites

CADD: https://cadd.gs.washington.edu/

CLC Bio Main Workbench: https://digitalinsights.qiagen.com/products-overview/discovery-insights-portfolio/analysis-and-visualization/qiagen-clc-main-workbench/

ClinVar: https://www.ncbi.nlm.nih.gov/clinvar/

dbSNP: https://www.ncbi.nlm.nih.gov/snp/

DGV: http://dgv.tcag.ca/dgv/app/home

ENSEMBL: https://www.ensembl.org/index.html

G2P: https://www.ebi.ac.uk/gene2phenotype/g2p_vep_plugin

GnomAD: https://gnomad.broadinstitute.org/

HPO: https://hpo.jax.org/app/

Illumina: https://emea.illumina.com/systems/sequencing-platforms/miseq/specifications.html

Invitae: https://www.invitae.com/en/physician/category/CAT000168/#exome-faq-panel-reporting

Ion Reporter: https://ionreporter.thermofisher.com/ir/

Mastermind: https://mastermind.genomenon.com/

Moon: http://www.diploid.com/moon

Mutalyzer: https://mutalyzer.nl/

NEB Tm Calculator: https://tmcalculator.neb.com/#!/main

OMIM: https://omim.org/

Pfam: https://pfam.xfam.org/

PolyPhen-2: http://genetics.bwh.harvard.edu/pph2/

Primer 3: http://primer3.ut.ee/

Primer BLAST: https://www.ncbi.nlm.nih.gov/tools/primer-blast/

PubMed: https://www.ncbi.nlm.nih.gov/pubmed/

Radboudumc, Maatricht UMC+: https://order.radboudumc.nl/en/services/exome-sequencing-diagnostics#grid[%220%22]=eJyrVipITE9VsjLSUUrLzClJLSpWssorzcmpBQBqEQik

RefSeq: https://www.ncbi.nlm.nih.gov/

SIFT: https://sift.bii.a-star.edu.sg/www/code.html

SNPCheck: https://genetools.org/SNPCheck/snpcheck.htm

UCSC *in silico* PCR: https://genome.ucsc.edu/cgi-bin/hgPcr

VarAFT: https://varaft.eu/

Varsome: https://varsome.com/

VCFtools: http://vcftools.sourceforge.net/

VEP: https://www.ensembl.org/info/docs/tools/vep/index.html

Appendix 4

The following tables were taken from the Richards et al., 2015 paper providing variant classification guidelines to allow for ease of access to the content. Please refer to the above-mentioned paper for the original content.

Table 3

Criteria for Classifying Pathogenic Variants

Very strong evidence of pathogenicity		
PVS1	Null variant (nonsense, frameshift, canonical +/−1 or 2 splice sites, initiation codon, single or multi-exon deletion) in a gene where loss of function (LOF) is a known mechanism of disease	
	Caveats:	
	• Beware of genes where LOF is not a known disease mechanism (e.g. *GFAP, MYH7*)	
	• Use caution interpreting LOF variants at the extreme 3' end of a gene	
	• Use caution with splice variants that are predicted to lead to exon skipping but leave the remainder of the protein intact	
	• Use caution in the presence of multiple transcripts	
Strong evidence of pathogenicity		
PS1	Same amino acid change as a previously established pathogenic variant regardless of nucleotide change	
	Example:	Val->Leu caused by either G>C or G>T in the same codon
	Caveat:	Beware of changes that impact splicing rather than at the amino acid/protein level
PS2	*De novo* (both maternity and paternity confirmed) in a patient with the disease and no family history	
	Note: Confirmation of paternity only is insufficient. Egg donation, surrogate motherhood, errors in embryo transfer, *etc*. can contribute to non-maternity	
PS3	Well-established *in vitro* or *in vivo* functional studies supportive of a damaging effect on the gene or gene product	
	Note: Functional studies that have been validated and shown to be reproducible and robust in a clinical diagnostic laboratory setting are considered the most well-established	
PS4	The prevalence of the variant in affected individuals is significantly increased compared to the prevalence in controls	
	Note 1: Relative risk (RR) or odds ratio (OR), as obtained from case-control studies, is >5.0 and the confidence interval around the estimate of RR or OR does not include 1.0. See manuscript for detailed guidance.	
	Note 2: In instances of very rare variants where case-control studies may not reach statistical significance, the prior observation of the variant in multiple unrelated patients with the same phenotype, and its absence in controls, may be used as moderate level of evidence.	

Moderate evidence of pathogenicity

PM1	Located in a mutational hot spot and/or critical and well-established functional domain (*e.g.* active site of an enzyme) without benign variation
PM2	Absent from controls (or at extremely low frequency if recessive) (see Table 6) in Exome Sequencing Project, 1000 Genomes or ExAC Caveat: Population data for indels may be poorly called by next generation sequencing
PM3	For recessive disorders, detected in *trans* with a pathogenic variant Note: This requires testing of parents (or offspring) to determine phase
PM4	Protein length changes due to in-frame deletions/insertions in a non-repeat region or stop-loss variants
PM5	Novel missense change at an amino acid residue where a different missense change determined to be pathogenic has been seen before Example: Arg156His is pathogenic; now you observe Arg156Cys Caveat: Beware of changes that impact splicing rather than at the amino acid/protein level
PM6	Assumed *de novo*, but without confirmation of paternity and maternity

Supporting evidence of pathogenicity

PP1	Co-segregation with disease in multiple affected family members in a gene definitively known to cause the disease Note: May be used as stronger evidence with increasing segregation data
PP2	Missense variant in a gene that has a low rate of benign missense variation and where missense variants are a common mechanism of disease
PP3	Multiple lines of computational evidence support a deleterious effect on the gene or gene product (conservation, evolutionary, splicing impact, etc) Caveat: As many *in silico* algorithms use the same or very similar input for their predictions, each algorithm should not be counted as an independent criterion. PP3 can be used only once in any evaluation of a variant.
PP4	Patient's phenotype or family history is highly specific for a disease with a single genetic etiology
PP5	Reputable source recently reports variant as pathogenic but the evidence is not available to the laboratory to perform an independent evaluation

Table 4

Criteria for Classifying Benign Variants

Stand-Alone evidence of benign impact	
BA1	Allele frequency is above 5% in Exome Sequencing Project, 1000 Genomes, or ExAC
Strong evidence of benign impact	
BS1	Allele frequency is greater than expected for disorder (see table 6)
BS2	Observed in a healthy adult individual for a recessive (homozygous), dominant (heterozygous), or X-linked (hemizygous) disorder with full penetrance expected at an early age
BS3	Well-established *in vitro* or *in vivo* functional studies shows no damaging effect on protein function or splicing
BS4	Lack of segregation in affected members of a family
	Caveat: The presence of phenocopies for common phenotypes (*i.e.* cancer, epilepsy) can mimic lack of segregation among affected individuals. Also, families may have more than one pathogenic variant contributing to an autosomal dominant disorder, further confounding an apparent lack of segregation.
Supporting evidence of benign impact	
BP1	Missense variant in a gene for which primarily truncating variants are known to cause disease
BP2	Observed in *trans* with a pathogenic variant for a fully penetrant dominant gene/disorder; or observed in *cis* with a pathogenic variant in any inheritance pattern
BP3	In-frame deletions/insertions in a repetitive region without a known function
BP4	Multiple lines of computational evidence suggest no impact on gene or gene product (conservation, evolutionary, splicing impact, etc)
	Caveat: As many *in silico* algorithms use the same or very similar input for their predictions, each algorithm cannot be counted as an independent criterion. BP4 can be used only once in any evaluation of a variant.
BP5	Variant found in a case with an alternate molecular basis for disease
BP6	Reputable source recently reports variant as benign but the evidence is not available to the laboratory to perform an independent evaluation
BP7	A synonymous (silent) variant for which splicing prediction algorithms predict no impact to the splice consensus sequence nor the creation of a new splice site AND the nucleotide is not highly conserved

Table 5

Rules for Combining Criteria to Classify Sequence Variants

Pathogenic

1. 1 Very Strong (PVS1) *AND*
   a. ≥1 Strong (PS1–PS4) *OR*
   b. ≥2 Moderate (PM1–PM6) *OR*
   c. 1 Moderate (PM1–PM6) and 1 Supporting (PP1–PP5) *OR*
   d. ≥2 Supporting (PP1–PP5)
2. ≥2 Strong (PS1–PS4) *OR*
3. 1 Strong (PS1–PS4) *AND*
   a. ≥3 Moderate (PM1–PM6) *OR*
   b. 2 Moderate (PM1–PM6) *AND* ≥2 Supporting (PP1–PP5) *OR*
   c. 1 Moderate (PM1–PM6) *AND* ≥4 Supporting (PP1–PP5)

Likely Pathogenic

1. 1 Very Strong (PVS1) *AND* 1 Moderate (PM1–PM6) *OR*
2. 1 Strong (PS1–PS4) *AND* 1–2 Moderate (PM1–PM6) *OR*
3. 1 Strong (PS1–PS4) *AND* ≥2 Supporting (PP1–PP5) *OR*
4. ≥3 Moderate (PM1–PM6) *OR*
5. 2 Moderate (PM1–PM6) *AND* ≥2 Supporting (PP1–PP5) *OR*
6. 1 Moderate (PM1–PM6) *AND* ≥4 Supporting (PP1–PP5)

Benign

1. 1 Stand-Alone (BA1) *OR*
2. ≥2 Strong (BS1–BS4)

Likely Benign

1. 1 Strong (BS1–BS4) and 1 Supporting (BP1–BP7) *OR*
2. ≥2 Supporting (BP1–BP7)

* Variants should be classified as Uncertain Significance if other criteria are unmet or the criteria for benign and pathogenic are contradictory.

Appendix 5: Primers Selected for Variant Validation

Gene	Variant	Transcript	RefSeq	Primer Pair	Primer Length	Annealing Temperature (°C)	Optimised	Samples Tested/ Validated
CDC42	p.Tyr23Cys	ENST400259	NG_047042.2	5'-TCTGAGTGCCTGAACCTGTT-3'	20	51.7	Yes	Sample 1 Maternal Paternal
	c.68A>G			5'-TCACCCCTTCTGACTTTCCA-3'	20			
EFTUD2	p.Arg578*	ENST426333	NG_032674.1	5'-ACTCAGTTCAGATGCGAGGA-3'	20	54.6	Yes	Sample 3 Maternal Paternal
	c.1732C>T			5'-GATAGTGGGGCTTGGCGG-3'	18			
IQSEC2	p.Gly283fs	ENST396435	NG_021296.2	5'-TATCTGTGAGGAAGCTGGGG-3'	20	53.3	Yes	Sample1 Sample 3 Sample 4 Sample 5 Sample7
	c.848_849delGC			5'-ATTCCCAACCAGATGCCTCT-3'	20			
POMT2	p.Arg475Pro	ENST261534	NG_008897.1	5'-TGGCCATTTCCCTTTCTGAC-3'	20	51.7	Yes	Sample 4 Maternal Sample 5
	c.1424G>C			5'-CCCTCCATCTGCCACTAAGA-3'	20			
POMT2	p.Trp744Cys			5'-TGGGGAATTTCTCTGGAGCT-3'	20	53.3	Yes	
	c.2232G>T			5'-CTGTCCTCTCCGTGGTGC-3'	18			
BRAF	p.Gly30fs	ENST288602	NG_007873.3	5'-GCTCTCGGTTATAAGATGGCG-3'	21	53.3	Yes	Sample 6 Maternal Paternal
	c.87_88delCG			5'-CGAGAAGGTGGCTGAGGG-3'	18			
KMT2D	p.Glu748fs	ENST301067	NG_027827.1	5'-GGTATCGCGCCTATCCCC-3'	18	53.8	No	
	c.2243_2244insC			5'-GATGTGGTCCCTCAGCCT-3'	18			
EFTUD2	p.Leu89fs	ENST426333	NG_032674.1	5'-TCCCTCTTGCTTTCAGATGGA-3'	21	54.6	Yes	Sample 7 Maternal
	c.263_264delCT			5'-ACGACAGTCTTCATGCACCT-3'	20			

Appendix 6: Identified possible putative-disease causing variant information

Patient	Gene	Location	dbSNP	Transcript	RefSeq	HGVS Nomenclature	Effect	Frequency	Allele Depth	Genotype Quality	ClinVar
1	*CDC42*	1:22405039	rs797044916	ENST400259	NM_001791.3	c.68A>G p.Tyr23Cys	Missense	0.00% GnomAD	120	225	Pathogenic
2	*CBL*	11:119170290	rs112330156	ENST264033	NM_005188.3	c.2520T>G p.Cys840Trp	Missense	0.0032% GnomAD	83	189	VUS
3	*EFTUD2*	17:42937401	rs1057520673	ENST426333	NM_001258354.1	c.1732C>T p.Arg578*	Stop gained	0.00% GnomAD	52	83	Pathogenic
4	*POMT2*	14:77743740	-	ENST261534	NM_013382.5	c.2232G>T P.Trp744Cys	Missense	0.00% GnomAD	49	113	-
5		14:77751884	-	ENST261534	NM_013382.5	c.1424G>C P.Arg475Pro	Missense	0.00% GnomAD	67	153	-
6	*BRAF*	7:140624415	-	ENST288602	NM_004333.4	c.87_88delCG p. Gly30fs	Frameshift deletion	0.00% GnomAD	10	27	-
	KMT2D	12:49445222	-	ENST301067	NM_003482.3	c.2243_2244insC p.Glu748fs	Frameshift insertion	0.00% GnomAD	185	39	-
7	*EFTUD2*	17:42963959	-	ENST426333	NM_001258354.1	c.263_264delCT p.Leu89fs	Frameshift deletion	0.00% GnomAD	159	99	-
	IQSEC2	X:53285131	-	ENST396435	NM_001111125.2	c.848_849delGC p.Gly283fs	Frameshift deletion	0.00% GnomAD	-	-	-

Appendix 7

Table A7A. Genes in which variants were identified for Patient 1 using the Custom Pipeline.

Gene	**Variant excluded or prioritised**
ABCB7	Excluded: benign supporting evidence identified.
WDR19	Excluded: two variants identified with benign supporting evidence.
moNUP214	Excluded: three variants identified all with benign supporting evidence, all reported in GnomAD (09/12/2019) and phenotype of patient does not match the suggested disease phenotype.
MACF1	Excluded: two variants identified both with benign supporting evidence.
POMT1	Excluded: variant identified with benign supporting evidence and the phenotype of the patient does not match the suggested disease phenotype.
TMPRSS6	Excluded: pathogenic variant identified however; the suggested disease phenotype does not match the phenotype of the patient.
ALMS1	Excluded: two variants identified with benign supporting evidence.
MYO5B	Excluded: variant identified with benign supporting evidence and insufficient phenotype overlap.
TUBGCP6	Excluded: variant not found in BAM file.
LRP5	Excluded: two variants identified with benign supporting evidence and the patient phenotype does not match the suggested disease phenotype.
C5ORF42	Excluded: Likely pathogenic variant within homopolymer, suggested disease phenotype does not match patient phenotype.
ADAR	Excluded: VUS identified.
KMT2D	Prioritised: likely pathogenic variant identified with a somewhat similar phenotype to the suggested disease.
ASXL3	Excluded: variants of uncertain significance
MRE11A	Excluded: Autosomal recessive pattern of inheritance for disease and one likely pathogenic and one benign variant identified.
FAM20C	Excluded: variant not found in BAM file.
ARID1A	Excluded: one VUS and one benign variant identified
ALX3	Excluded: one likely pathogenic variant identified, disease phenotype of does not match patient phenotype. Possible false-positive due to homopolymer region.
FGFR2	Excluded: VUS identified.
HPD	Excluded: Likely benign variant identified.
NBEA	Excluded: VUS identified.
TBX22	Excluded: Phenotype doesn't match, and seen in one healthy African patient (GnomAD 12/04/2019).

Table A7B. Genes in which variants were identified for Patient 2 using the Custom Pipeline.

Gene	Variant excluded or prioritised
HECW2	Excluded: VUS and benign variants identified.
GLI3	Excluded: Likely benign variant identified.
SUV420H1	Excluded: VUS identified.
DCHS1	Excluded: Likely benign variants identified.
CLCNKB	Excluded: Same variant identified in sequencing control.
KANSL1	Excluded: VUS identified.
ALMS1	Excluded: Benign and likely benign variants identified.
GATA6	Excluded: VUS identified, suggested disease phenotype does not match patient phenotype.
ACAN	Excluded: Likely pathogenic variant identified; suggested disease phenotype does not match patient phenotype.
TGIF1	Excluded: Likely benign variant identified.
TRIP12	Excluded: VUS identified.
FREM2	Excluded: VUS and likely benign variants identified.
KCNN3	Excluded: VUS identified.
DNAH9	Excluded: One pathogenic and one VUS (with benign supporting evidence) identified, suggested disease phenotype does not match patient phenotype.
ALX3	Excluded: One likely pathogenic and one likely benign variant identified, with insufficient phenotype overlap of autosomal recessive disease.
ARID1A	Excluded: VUS identified.
COL9A2	Excluded: One benign variant identified.
EHMT1	Excluded: VUS identified with benign supporting evidence.

Table A7C. Genes in which variants were identified in Patient 3 using the Custom Pipeline.

Gene	Variant excluded or prioritised
FANCA	Excluded: Likely benign variants identified.
RECQL4	Excluded: VUS identified.
MED12	Excluded: Likely benign variant identified.
TGIF1	Excluded: Benign variant identified.
SACS	Excluded: VUS identified.
PTF1A	Excluded: VUS identified.
ALDH18A1	Excluded: VUS identified and suggested disease phenotype does not match patient phenotype.
PITX2	Excluded: Suggested disease phenotype does not match patient phenotype.
EHMT1	Excluded: Likely benign variant identified.
NHS	Excluded: VUS identified.
HIVEP2	Excluded: Likely benign and VUS identified.
DNM1	Excluded: Benign and VUS identified, suggested disease phenotype does not match patient phenotype.
CEP290	Excluded: VUS identified.
FANCM	Excluded: Benign and likely benign variants identified; suggested disease phenotype does not match patient phenotype.
LRP2	Excluded: Likely benign and benign variants identified.
GLI2	Excluded: VUS identified.
ALX3	Excluded: Likely pathogenic variant identified; suggested disease phenotype does not match patient phenotype.
COL11A2	Excluded: VUS identified.
DNMT3B	Excluded: VUS identified.
LFNG	Excluded: Benign variant identified.

TOP3A	Excluded: VUS identified.
CLCNKB	Excluded: VUS identified, suggested disease phenotype does not match patient phenotype.
PKD1L1	Excluded: Likely benign variants identified.
FGFR3	Excluded: VUS identified with benign supporting evidence; suggested disease phenotype does match patient phenotype.
DCHS1	Excluded: VUS identified, suggested disease phenotype is similar to patient phenotype.
CPS1	Excluded: VUS identified, suggested disease phenotype does not match patient phenotype.
BIN1	Excluded: variant not found in BAM file.
TCOF1	Excluded: Likely pathogenic variant (indel) identified in homopolymer, suggested disease phenotype does match patient phenotype.
KCNT1	Excluded: VUS and benign variants identified.
ARID1A	Excluded: VUS and intronic variant with no coding or splicing effect identified.
EFTUD2	Prioritised: Pathogenic variant identified; suggested disease phenotype does match patient phenotype.
HPSE2	Excluded: Likely pathogenic variant identified; suggested disease phenotype does not match patient phenotype.
NBAS	Excluded: Likely pathogenic variant identified; suggested disease phenotype does not match patient phenotype.
OTOGL	Excluded: VUS and likely benign variants identified, suggested disease phenotype does not match patient phenotype.
TBX22	Excluded: VUS identified.
LAMC3	Excluded: VUS identified.
PKHD1	Excluded: VUS identified, suggested disease phenotype does not match patient phenotype.

Table A7D. Genes in which variants where identified in Patient 4 using the Custom Pipeline.

Gene	**Variant excluded or prioritised Sample 4**
SON	Excluded: Likely pathogenic and VUS identified, suggested phenotype does not match patient phenotype.
TMPRSS6	Excluded: Likely pathogenic variant and VUS identified, suggested phenotype does not match patient phenotype.
POLR3B	Excluded: Likely pathogenic and VUS identified for suggested autosomal recessive disease.
ALX3	Excluded: Likely pathogenic and 3 prime UTR variant identified; suggested disease phenotype does not match patient phenotype.
POMT2	Prioritised: two VUS identified with patient phenotype highly specific for gene.
CLTC	Excluded: Benign variants identified.
UVSSA	Excluded: Benign variants and VUS identified.
PKD1L1	Same variants identified in patient 5.
CACNA1A	Excluded: Benign variants identified.
RYR1	Excluded: Benign and likely benign variants identified.
SALL4	Excluded: Likely benign variants identified.
DCHS1	Excluded: VUS identified with benign supporting evidence.
RLIM	Excluded: VUS identified.
BRAF	Excluded: Benign variants identified.
ARIDA1	Excluded: VUS identified.

KIF7	Excluded: Pathogenic variant identified in a homopolymer region. Variant not found in affected twin brother. Suggested disease phenotype does not match patient phenotype.
SCN1A	Excluded: Pathogenic variant identified in a homopolymer region. Variant not found in affected twin brother. Suggested disease phenotype does not match patient phenotype.

Table A7E. Genes in which variant were identified in Patient 5 using the Custom Pipeline.

Gene	**Variant(s) excluded/prioritised Sample 5**
KMT2D	Excluded: Benign variants identified.
GRIN2A	Excluded: VUS identified, did not co-segregate with sibling.
MRE11A	Excluded: One likely pathogenic and 3 VUS identified for a suggested autosomal recessive disorder.
TH	Excluded: Benign variant identified.
AMPD2	Excluded: Benign variant identified.
TGIF1	Excluded: Synonymous variant identified.
GPSM2	Excluded: Benign variant identified.
CUX2	Excluded: Benign and VUS identified.
NOTCH1	Excluded: Benign and likely benign variants identified.
LFNG	Excluded: variant not found in BAM file.
ATM	Excluded: Benign and likely benign variants identified.
RIN2	Excluded: Likely benign variant identified.
POMGNT2	Excluded: VUS and likely benign variants identified.
HPSE2	Excluded: One likely pathogenic variant and one intronic variant identified, suggested disease phenotype does not match patient phenotype.
POMT2	Prioritised: two VUS identified with patient phenotype highly specific for gene.
MMAB	Excluded: VUS identified.
ACAN	Excluded: Likely benign variants identified.
GAS2L2	Excluded: Likely benign variants identified.

Table A7F. Genes in which variants were identified in Patient 6 using the Custom Pipeline.

Gene	**Variant(s) excluded/prioritised**
LAMA2	Excluded: Intronic, benign and VUS identified.
EXT2	Excluded: VUS identified.
EHMT1	Excluded: Likely benign and VUS identified.
RLIM	Excluded: Synonymous variant and VUS identified.
PTCH1	Excluded: VUS identified.
COL11A2	Excluded: Intronic and VUS identified.
KAT6B	Excluded: Benign variant identified.
NOTCH2	Excluded: Likely benign variant identified.
GRIA3	Excluded: Likely benign variant identified.
COL18A1	Excluded: VUS and intronic variants identified, suggested disease phenotype does not match patient phenotype.
DNAH9	Excluded: VUS and intronic variants identified.
SON	Excluded: Likely benign variant identified.
FRAS1	Excluded: Intronic, likely benign and VUS identified.
CCDC40	Excluded: Likely benign variant identified.
KMT2D	Prioritised: Likely pathogenic variants identified; suggested disease phenotype does match patient phenotype.

ATRX	Excluded: VUS identified.
MYOB5	Excluded: Benign variant identified.
DNMT3B	Excluded: VUS identified.
CTSA	Excluded: Intronic variant identified.
TCF20	Excluded: VUS identified.
ALG9	Excluded: VUS identified.
MTR	Excluded: Intronic and likely benign variants identified.
BRAF	Prioritised: Likely pathogenic variant identified; suggested disease phenotype does match patient phenotype.
TELO2	Excluded: VUS identified.
TGIF1	Excluded: VUS identified.
SNC11A	Excluded: Synonymous, intronic and VUS identified.

Table A7G. Genes in which variants were identified for Patient 7 using the Custom Pipeline.

Gene	**Variant(s) excluded/prioritised**
MED13L	Excluded: Likely benign variant identified.
SMPD1	Excluded: Benign variant identified.
FLT4	Excluded: VUS identified.
SLC17A5	Excluded: Likely pathogenic, suggested disease phenotype does not match patient phenotype.
SYNGAP1	Excluded: VUS identified.
NEB	Excluded: Benign and likely benign variants identified.
MRE11A	Excluded: Likely pathogenic variant identified; suggested disease phenotype does not match patient phenotype.
LGI4	Excluded: VUS identified.
MYT1L	Excluded: Benign variant identified.
FBN2	Excluded: VUS identified.
BCOR	Excluded: VUS identified.
PACS1	Excluded: Likely benign variant identified.
SLC6A17	Excluded: VUS identified.
EFTUD2	Prioritised: Likely pathogenic variant identified; suggested disease phenotype does match patient phenotype.

Appendix 8

KMT2D assay optimisation:

An initial temperature gradient was performed with a control DNA sample using the designed *KMT2D* primers seen in Appendix 5. The PCR protocol shown in Chapter 2 (Table's 3 and 4) were followed. 2µl of PCR product and 3µl of agarose gel loading buffer were run on a 3% (w/v) agarose gel at 5V/cm and produced the following result:

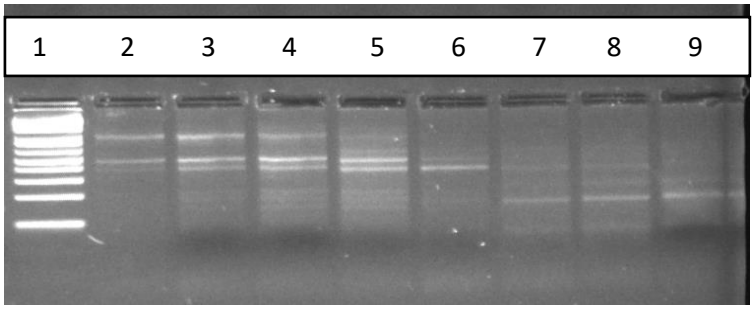


Figure A7.1 3% (w/v) agarose gel image of *KMT2D* temperature gradient. Lane 1 indicates the 1kb+ marker. Lanes 2-9 show the PCR product run at temperatures: 60.0°C; 59.2°C; 58.0°C; 56.1°C; 53.8°C; 51.9°C; 50.7°C; 50°C respectively. Non-specific binding was identified at all temperatures.

The PCR product in Lane 6 showed the desired amplification product with minimal non-specific binding. The protocol was followed with the Patient and parental samples at this annealing temperature (53.8°C) and Sanger sequencing was then performed. The sequencing showed a product of the correct size however, there was too much 'background noise' to interpret the results attributed to the minimal non-specific binding seen in the gel image and further assay optimisation was undertaken.

An alternative website was used to obtain the annealing temperatures of the primer pair namely NEB Tm Calculator (see Appendix 3 for URL), which suggested an annealing temperature of 65°C. A second PCR reaction was set-up and the new annealing temperature was used during thermal cycling; however, the non-specific binding was not resolved. The next intervention applied was to extend the annealing and elongation time during thermal cycling; however, this was unsuccessful. Following this, an additive known as Dimethyl Sulfoxide (DMSO, Merck, USA) was introduced into the PCR set-up to reduce secondary structures that may result due to the high GC content of the area targeted for amplification (Kieleczawa, J., 2006; Mammedov et al., 2008). The PCR protocol with added DMSO can be seen below:

Table A81. PCR protocol with additive DMSO

PCR Component	**Volume (µl)**
DNA [100ng/µl]	1
Forward Primer [10ng/µl]	1
Reverse Primer [10ng/µl]	1
DMSO [100%]	1.25
dNTP Mix [10mM per dNTP]	2.5
AmpliTaq Gold Buffer II [10X]	2.5
AmpliTaq Gold MgCl2 [25nM]	2.5
ddH2O	13.05
AmpliTaq Gold DNA Polymerase [5U/ µl]	0.2
TOTAL	25

The incorporation of the additive reduced majority of the non-specific binding, however, some non-specific product was still identified. This protocol was run at four different temperature namely: 53.8°C**;** 58.5°C**;** 61°C**;** 64°C**,** however, the non-specific binding was not resolved. The additive was then increased within the PCR protocol to 2.5µl DMSO per 25µl reaction, however this prevented amplification of the target region. The protocol showed above was then used in a second temperature gradient where higher annealing temperatures were used ranging from 56°C to 66°C and yielded the following result:

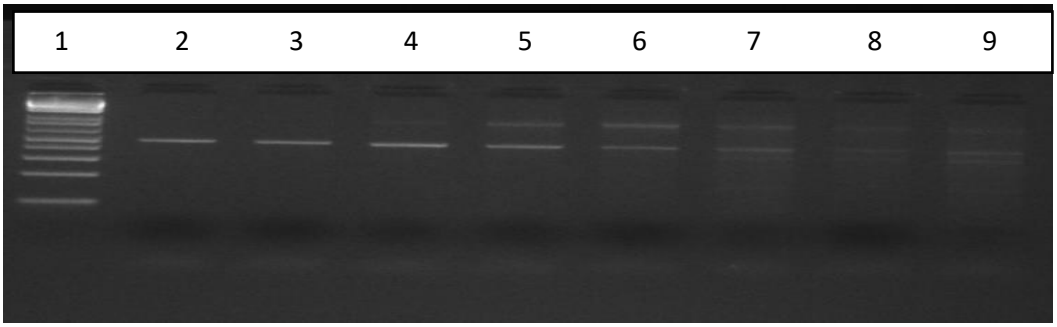


Figure A7.2 3% (w/v) agarose gel image of *KMT2D* temperature gradient with DMSO additive. Lane 1 shows the 1kb+ marker. Lanes 2-9 show the PCR product run at temperatures: 66.0°C; 65.3°C; 64.0°C; 62.1°C; 59.8°C; 57.9°C; 56.7°C; 56.0°C respectively. Lanes 2 and 3 show the amplification of the desired PCR product in the control DNA sample.

This protocol was then run on the Patient and his parent's samples at an annealing temperature of 66°C, however, no amplification was seen except for a faint band indicating PCR amplification in the Patient's paternal sample, seen below.

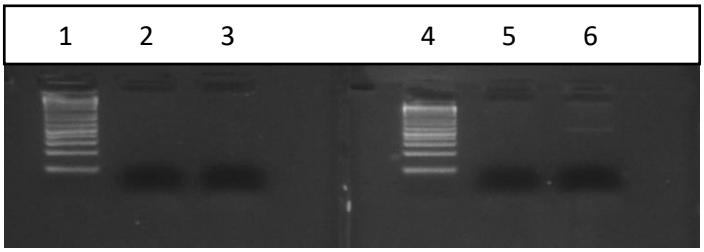
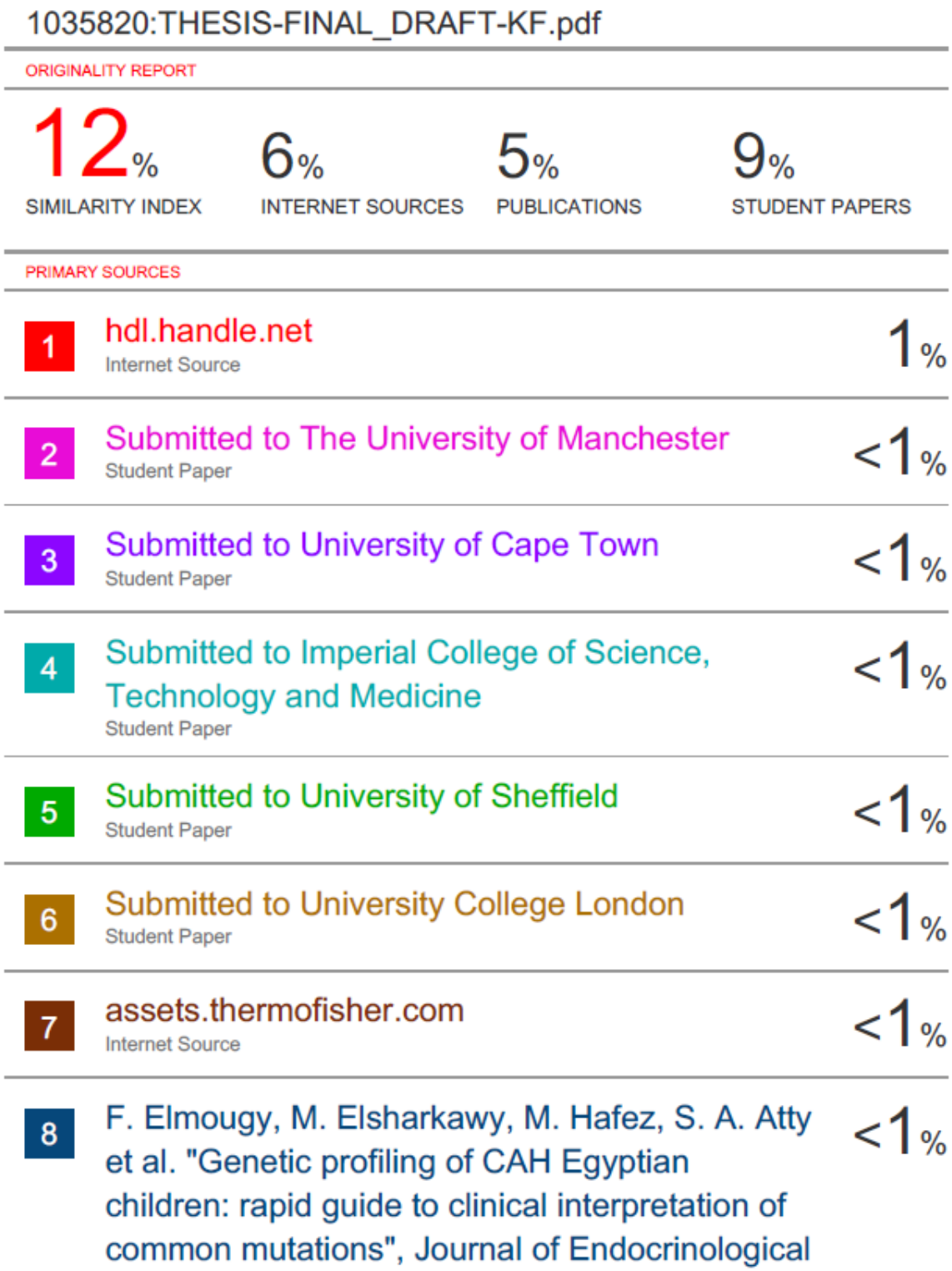


Figure A7.3 3% (w/v) agarose gel image of *KMT2D* PCR product on Patient 6 and parental samples. Lanes 1 and 4 show the 1kb+ marker. Lane 2 is the PCR product showing no amplification of the Patient's DNA. Lane 3 shows a non-template control. Lane 5 shows the maternal sample with no amplification and Lane 6 shows the paternal sample with a faint band indicating amplification.

This same protocol was then repeated a second time at an annealing temperature of 65.5°C and faint bands showing amplification were seen in all three samples. This PCR product was then cleaned and subjected to cycle sequencing, post-sequencing clean-up, and then sequenced on an Applied Biosystems 3130 Genetic Analyzer (Thermo Fisher, USA). However, no sequencing data was produced, most likely as a result of inefficient product amplification during PCR. The final optimisation intervention introduced was to increase the concentration of the DNA within the PCR reaction. A constant concentration of 100ng/µl DNA was used throughout the optimisation procedures for both the control sample and patient samples used. The final PCR set-up used the patient samples stock DNA concentrations run comparatively with the 100ng/µl diluted samples to see if amplification improved and the control DNA sample was also included in the set-up. These seven PCR reactions were run with an annealing temperature of 65.5°C and showed no amplification except for a faint band present in the control sample reaction.

Appendix 9



1035820:THESIS-FINAL_DRAFT-KF.pdf

ORIGINALITY REPORT

12%	6%	5%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	hdl.handle.net Internet Source	1%
2	Submitted to The University of Manchester Student Paper	<1%
3	Submitted to University of Cape Town Student Paper	<1%
4	Submitted to Imperial College of Science, Technology and Medicine Student Paper	<1%
5	Submitted to University of Sheffield Student Paper	<1%
6	Submitted to University College London Student Paper	<1%
7	assets.thermofisher.com Internet Source	<1%
8	F. Elmougy, M. Elsharkawy, M. Hafez, S. A. Atty et al. "Genetic profiling of CAH Egyptian children: rapid guide to clinical interpretation of common mutations", Journal of Endocrinological	<1%