MUTATION PROFILING IN SOUTH AFRICAN PATIENTS WITH CORNELIA DE LANGE SYNDROME PHENOTYPE USING TARGETED NEXT GENERATION SEQUENCING

Heather Seymour

Student number: 543629

Supervisor: Dr Nadia Carstens

Co-supervisors: Dr Candice Feben and Prof Zané Lombard

UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

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

Johannesburg, 2019

Declaration

I, Heather Jessica Seymour, 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)

_________________day of______________________20_________________in______________

Abstract

Cornelia de Lange Syndrome (CdLS) is a monogenic, heterogeneous, congenital disorder. Its main features are intellectual and developmental delay, failure to thrive and skeletal abnormalities. It displays a wide phenotypic range and has a significant phenotypic overlap with other conditions resulting in multiple differential diagnoses.

To date, mutations in five genes have been reported to cause CdLS, accounting for 70% of clinically diagnosed patients. These genes are involved in the cohesin pathway. The remaining 30% of cases could either harbour a mutation in another gene, potentially also in the cohesin pathway, in a regulatory region or patients with a less classical phenotype could have been misdiagnosed owing to the broad phenotypic spectrum. Currently no molecular studies have been carried out on patients with CdLS in South Africa and thus the molecular cause of this disease is unknown in this population. This study aimed to use a targeted next generation sequencing technique to study the molecular aetiology of CdLS in South African patients and families. By adopting this technique, we were able to study multiple genes simultaneously to elucidate the mutation profile of this condition in a South African setting. Our gene panel included the five known causal genes as well as genes implicated in the differential diagnoses and other genes involved in the cohesin pathway.

Out of the 14 patients that underwent targeted sequencing, putative disease-causing mutations were identified in eight. These were classified as pathogenic using the ACMG guidelines in addition to other bioinformatic tools and databases. Four of these mutations were small deletions, one was a single base pair duplication, one was a splice site mutation and two were missense mutations. The phenotypes of these eight patients correlated in severity in accordance with other genotype-phenotype studies that have been conducted in the past. Seven of the mutations were identified in the *NIPBL* gene, the most commonly mutated gene in CdLS. The remaining mutation was identified in *STAG1*. At the time of designing the targeted gene panel, no mutations had been identified in *STAG1* in humans. This gene is involved in the cohesin pathway and was only suspected to be involved in CdLS.

This study provides novel insights into the mutation profile of CdLS in South African patients. Studies such as this can inform the development of diagnostic techniques involving next generation sequencing panels as well as offering testing options for patients with CdLS in South Africa.

Acknowledgements

I would like to thank my main supervisor, Dr Nadia Carstens, for the opportunity to carry out the current Master's project and for her exceptional support and guidance throughout the entire process. I would like to thank her and the National Research Foundation (NRF) for the funding that was provided through the NRF Thuthuka grant that was awarded to her for this project. I would also like to thank my two co-supervisors, Dr Candice Feben and Prof Zané Lombard for their guidance and support over the last two years. Dr Feben has offered invaluable direction and clinical insight and Prof Lombard has been an excellent mentor, particularly in the writing process. I would like to thank the remainder of everyone involved in the FRASC research group: Prof Amanda Krause, Dr Careni Spencer, Dr Robyn Kerr, Ms Maria Mudau and Ms Patracia Nevondwe. Thank you to Mr Dean Harris and Dr Michelle Bronze from Agilent Technologies for the technical support and guidance as well as Dr Adriaan Basson and the University of the Witwatersrand Medical School for the use of their Illumina MiSeq® sequencing machine. I would like to thank my parents for their unwavering support and confidence in me. Lastly, I would like to thank Ms Michaella Hulley for the invaluable help, support and guidance throughout the process and write up of the current project.

Contents

List of Abbreviations and Symbols

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

ACMG: American College of Medical Genetics

BAM: Binary Alignment Map

BED file: Browser Extensible Data file

CADD: Combined Annotation-Dependent Depletion

CdLS: Cornelia de Lange syndrome

cDNA: complementary Deoxyribonucleic Acid

CHOPS: Cognitive impairment and coarse facies, heart defects, obesity, pulmonary involvement and short stature ad Skeletal dysplasia

COSMIC: Catalogue of Somatic Mutations in Cancer

DANN: Deleterious Annotation of genetic variants using Neural Networks

dbSNP: Single Nucleotide Polymorphism Database

ddH20: Double Distilled Water

DNA: Deoxyribonucleic Acid

dNTP: deoxyribonucleotide triphosphate

ExAC: Exome Aggregation Consortium

FATHMM: Functional Analysis Through Hidden Markov Models

FRASC: Facial dysostosis syndrome, RASopathies and Cornelia de Lange syndrome working group

gDNA: genomic Deoxyribonucleic Acid

GERP++: Genomic Evolutionary Rate Profiling

gnomAD: The Genome Aggregation Database

GRCh37: Genome Reference Consortium Human Build 37

GWAS: Genome Wide Association Studies

HGVS: Human Genome Variation Society

IGV: Integrative Genomics Viewer program

LOD: Logarithm Of the Odds

LOF: Loss Of Function

MAF: Minor Allele Frequency

MLPA: Multiplex Ligation-dependant Probe Amplification

NGS: Next Generation Sequencing

NHLBI: National Heart, Lung and Blood Institute

NHLS: National Health Laboratory Services

PCR: Polymerase Chain Reaction

PhiX: Phi X 174 bacteriophage

PolyPhen-2: Polymorphism Phenotyping version 2

PROVEAN: Protein Variation Effect Analyzer

QC: Quality Control

Q-score: Quality Score

RNA: Ribonucleic Acid

SAV: Sequence Analysis Viewer

SIFT: Sorting Intolerant from Tolerant

SNP: Single Nucleotide Polymorphism

SWI/SNF: SWItch/sucrose NonFermenting

UK: United Kingdom

UTR: Untranslated Region

VCF: Variant Call File

wANNOVAR: web server of ANNOtate VARiation software

Wits: University of the Witwatersrand

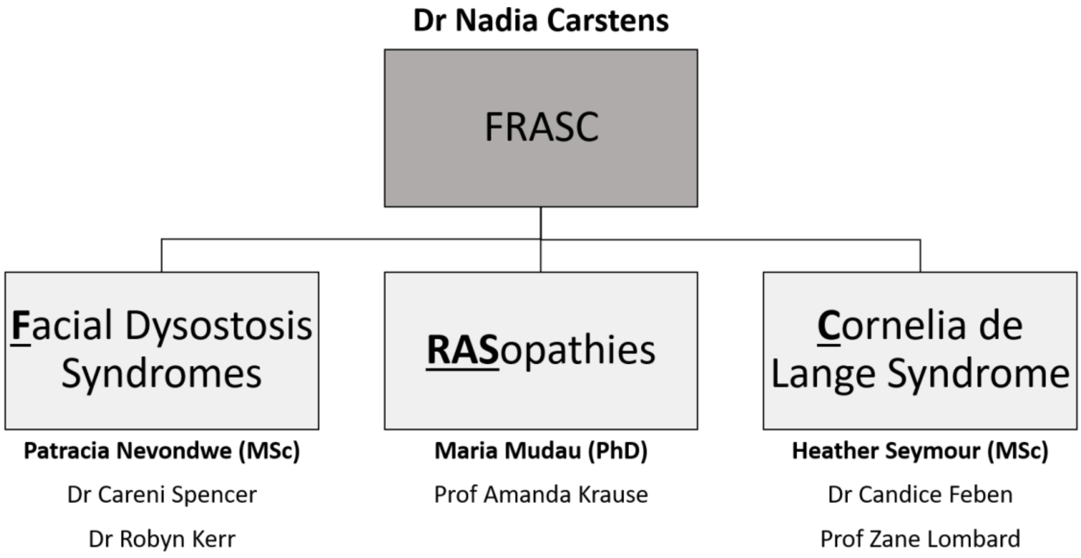
WSS: Wiedemann-Steiner Syndrome

List of Figures

List of Tables

Preface

The initial project design was a PhD looking into mutation profiles of Kabuki Syndrome, Rasopathies and Cohesinopathies as national genetic testing is not currently available for these conditions. However, after intra-departmental discussions, it was decided to separate the single project into three projects falling under a single working group. Subsequent changes were made to the study design and the resulting projects consisted of two MSc projects looking at mutation profiles of Facial Dysostosis Syndromes and Cornelia de Lange-like phenotypes respectively, and one PhD project examining the mutation profile of RASopathies. Thus, the FRASC working group was formed consisting of clinical geneticists, senior researchers and students. The clinical geneticists included Prof Amanda Krause, Dr Candice Feben and Dr Careni Spencer. The senior researchers included Prof Zané Lombard, Dr Robyn Kerr and Dr Nadia Carstens. The students consisted of Ms Maria Mudau, Ms Patracia Nevondwe and myself. The diagram below summarises the FRASC working group and how everyone was involved – Dr Nadia Carstens headed up the working group, each student is indicated in bold underneath their respective disease of interest and the supervisors for each student are indicated below the students' names. Dr Nadia Carstens was also a supervisor to each student.



Each student utilized the same methodology and therefore one gene panel was used to incorporate the genes of interest for all three projects. Library preparation was carried out independently by each student, however, library pooling and subsequent sequencing was occasionally completed as a group to reduce the costs involved in the FRASC research group.

1. Introduction

1.1. Cornelia de Lange Syndrome

Cornelia de Lange syndrome (CdLS), also known as Brachmann de Lange syndrome, is a heterogeneous, monogenic condition. It was first described in reports by Dutch anatomists, Willem and Gerardus Vrolik (Vrolik, 1849), then by the German physician, Winfried Robert Clemens Brachmann (Brachmann, 1916), and finally by the Dutch paediatrician, Cornelia de Lange (De Lange, 1933), after whom the condition was named. Currently, the prevalence of CdLS is estimated at 1.6-2.2:100 000 in European populations (Barisic *et al.*, 2008); however, there are no studies reporting the prevalence of this disorder in other population groups.

CdLS presents with a variable phenotype with features ranging from mild to severe developmental- and intellectual delay and physical abnormalities. These physical abnormalities include both craniofacial and limb abnormalities (Jackson *et al.*, 1993) (Figure 1.1.). The diverse nature of this heterogeneous phenotype can be attributed to inter- and intragenic variability (Mannini *et al.*, 2013), which will be discussed in the sections that follow. The most commonly observed phenotypes in patients with CdLS are summarized in the following categories: facial features; sensory development; skeletal anomalies; genitalia; neurodevelopment; cardiac anomalies and the gastrointestinal system. Table 1.1 has been adapted from an inhouse clinical tick sheet used to diagnose patients with CdLS (Appendix A.1.). Patients with CdLS will have some but not necessarily all the following phenotypes (the most common being coarse facial features, synophrys, hirsutism, intellectual delay and limb deformities).

Table 1.1. Summary of the typical phenotypic features of patients presenting with Cornelia de Lange syndrome.

Facial features	Coarse facial features	**Skeletal anomalies**	Micromelia (upper limbs)
	Brachycephaly		Oligodactyly
	Down slanted palpebral fissures		Clinodactyly
	Ptosis		Elbow contractures
	Synophrys		Transverse palmar crease
	Curly eyelashes		Kyphosis (spine)
	Down turned mouth		Scoliosis
	Long philtrum	**Genitalia**	Undescended testes
	Cleft palate		Small penis
	Widely spaced teeth		Hypospadias
	Depressed nasal bridge		Hypoplastic labia minora
	Anteverted nares	**Neurodevelopment**	Mild to severe intellectual delay
	Short neck		Hypertonia
	Low posterior hairline		Autism spectrum disorder
	Low set ears	**Cardiac anomalies**	Atrial septal defect
	Posteriorly rotated ears		Ventricular septal defect
	Hirsutism		Aortic stenosis
Sensory development	Myopia		Pulmonary stenosis
	Strabismus		
	Hearing loss – conductive and sensorineural	**Gastrointestinal system**	Malrotation
			Gastro-oesophageal reflux

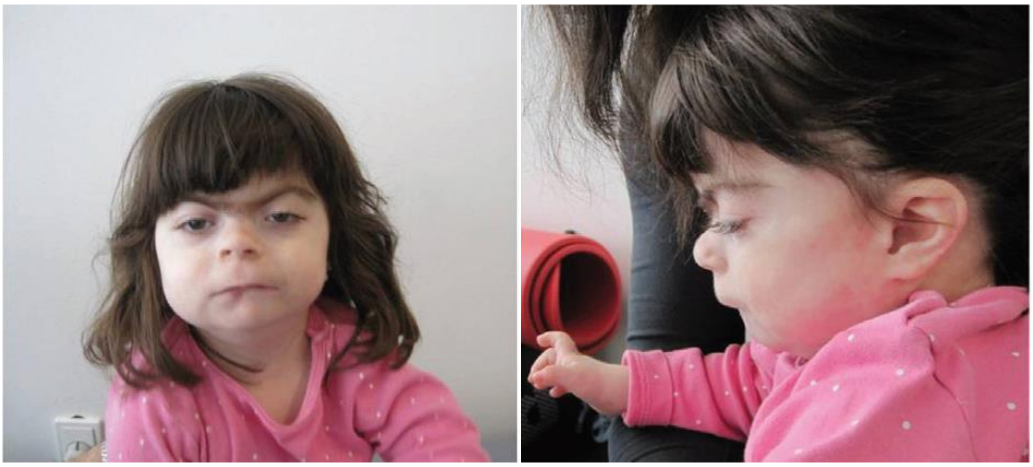


Figure 1.1. A patient with a classical Cornelia de Lange phenotype (Boyle et al. 2014). The classical facial features and upper limb abnormalities can be clearly seen in the figure.

1.2.**Genetics and pathology of Cornelia de Lange Syndrome**

CdLS expresses wide phenotypic variability and as such, a number of differential diagnoses exist. This presents a need for molecular confirmation by means of genetic testing, particularly in those patients with an atypical phenotype. It is therefore imperative to understand the genetics and molecular biology of CdLS.

1.2.1. **Cohesin pathway / complex**

CdLS falls within the spectrum of a larger group of diseases termed ‘cohesinopathies’. These disorders arise as a result of mutations in genes forming part of the cohesin protein complex (Figure 1.2.). This protein complex forms a ring-like structure that encircles the DNA strand and is comprised of four subunits: SMC1 and SMC 3 (Structural Maintenance of Chromosome 1 and 3), and SCC1 and SCC3 (Sister Chromatid Cohesion 1 and 3, SCC1 - also referred to as RAD21 cohesin complex component and SCC3 - also referred to as Stromal antigen 1 and 2 (STAG1 and STAG2). There is also a large number of regulatory proteins that aid in the functioning and organisation of the cohesin protein complex. These include, NIPBL (Nipped B Like) and SCC4 (Sister Chromatid Cohesion 4), which aid in loading the cohesin complex onto the DNA. The proteins ESCO2 (Establishment of Sister Chromatid Cohesion N-Acetyltransferase 2) and HDAC8 (Histone Deacetylase 8) assist the loading and dissociation of the complex through acetylation and deacetylation respectively. The PDS5 (Precocious Dissociation of Sisters 5) complex, made up of PDS5A and PDS5B, is responsible for the maintenance of the cohesin complex, as well as assisting in dissociation (reviewed by Barbero, 2013).

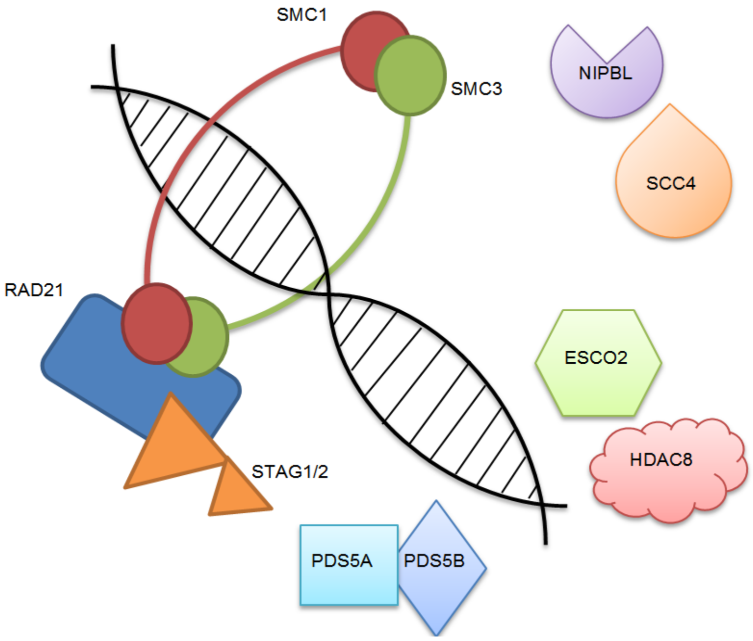


Figure 1.2. Diagram representing the cohesin protein complex and selected regulatory proteins (adapted from Barbero, 2013). SMC1, SMC3, RAD21 and STAG1/2 comprise the main protein complex which encircles the DNA strand. The regulatory proteins include NIPBL and SCC4 (which aid in the loading of the cohesin onto the chromosomes), ESCO2 and HDAC8 (which facilitate the loading and dissociation of the complex via acetylation or deacetylation respectively) and the PDS5 complex (which is responsible for the maintenance of the protein complex and also assists in its dissociation).

The cohesin complex has a wide range of roles including the cohesion of sister chromatids during cell division, assisting in DNA repair and regulating gene expression (reviewed by Pezic, Weeks and Hadjur, 2017). These are the processes which are deleteriously affected in CdLS.

1.2.1.1. Sister chromatid cohesion

Sister chromatid cohesion is the process whereby sister chromatids are held together for various phases of the cell cycle allowing for processes to take place where close proximity is needed between sister chromatids. A study carried out in 1998 showed that the cohesion of sister chromatids remained unaffected at interphase in an environment depleted of cohesin proteins (SMC family) (Losada, Hirano and Hirano, 1998). However, upon entry of the mitotic phase of the cell cycle, the sister chromatids no longer remained tightly associated with each other. Additionally, they observed that the most common chromosomal abnormality in a cohesin-depleted environment was double stranded DNA breaks, alluding to an additional role of cohesin in the cell. This study grouped SMC1 and SMC3 as cohesin proteins responsible for the cohesion of sister chromatids after DNA replication during the mitotic phase. They then grouped SMC2 and SCM4 as condensins, as they are the key proteins from the cohesin protein complex involved in the condensation of chromatids in preparation for mitosis (Hirano and Mitchison, 1994; Kimura and Hirano, 1997).

1.2.1.2. DNA repair

Sjögren and Nasmyth, (2001) decided to test, and subsequently accept, the hypothesis that the cohesion of sister chromatids is essential for post-replicative double stranded DNA repair. They utilised the idea that double stranded DNA breaks require an undamaged copy to be used as a template for repair. They then tested this by subjecting cells with mutated cohesin components to γ-radiation to inflict double stranded breaks. The results showed that poorly associated sister chromatids showed markedly less DNA repair, whereas the sister chromatids that were associated strongly due to the presence of cohesin were able to repair the double stranded breaks. They concluded that it is the association of sister chromatids facilitated by cohesin and not cohesin itself that aids in double stranded DNA break repair.

Another study by Lightfoot *et al.*, (2011) demonstrated how the lack of cohesin not only impaired joining of sister chromatids, and thus lack of double stranded DNA repair, but also that in the absence of cohesin, the cell no longer releases an apoptotic response at the pachytene checkpoint during meiosis. This implicates cohesin in DNA damage checkpoints as well.

1.2.1.3. Regulation of gene expression

A study carried out by Rollins, Morcillo and Dorsett, (1999) on *Drosophila melanogaster* (fruit flies) aimed to investigate long range gene activation i.e. enhancer-promoter communication. By studying the inhibitory effects of Nipped-B mutants (*Drosophila* homologue of NIPBL in humans) on promoter-enhancer communication, they showed the essential structural role Nipped-B plays. Further studies have elucidated the same role of NIPBL and the cohesin pathway in yeast and humans (Dorsett, 1999; Liu *et al.*, 2009;

Dorsett, 2007). Genome-wide analysis of transcription in 16 patients with CdLS has shown dysregulated gene expression and, after validating these findings in an additional 101 patients, can even correlate this dysregulation to phenotypic severity (Liu *et al.*, 2009).

These roles are critical in cell function and development and it is evident that mutations in any one of the genes involved, affecting protein function, will have serious downstream consequences.

1.2.2. Genes involved in Cornelia de Lange Syndrome

Due to the condition's heterogeneous nature, CdLS can be inherited in either an autosomal dominant or X-linked manner depending in which gene a mutation is present. Causal mutations have been identified in five genes thus far: *NIPBL*, *RAD21*, *SMC1A*, *SMC3* and *HDAC8* (reviewed by Barbero, 2013). The mutations identified in these genes have predominantly been small mutations i.e. frameshift, nonsense or splice site mutations which are predicted to lead to truncated proteins or a loss of function (reviewed by Mannini *et al.*, 2013) (Table 1.2.). Somatic mosaicism in *NIPBL*, although it is rare, has been described to contribute to the disease as well (reviewed by Kline *et al.*, 2018).

1.2.2.1. *NIPBL*

NIPBL, the most commonly mutated gene in CdLS cases (approximately 60% of cases in European populations have a mutation in this gene (reviewed by Mannini *et al.*, 2013), was first reported to cause CdLS in 2004 (Krantz *et al.*, 2004; Tonkin *et al.*, 2004). Krantz *et al.*, (2004) used linkage exclusion mapping to identify five regions of interest with a positive logarithm of the odds (LOD) score in nine families presenting with CdLS. Further fine mapping and multipoint linkage analysis with more markers narrowed the region down to chromosome 5p13.1-13.3. By carrying out mutation analysis and identifying two mutated overlapping transcripts, they defined the gene in which these mutations were found as being the homologue of the Nipped-B gene in Drosophila and named it *NIPBL*. Simultaneously, Tonkin *et al.*, (2004) carried out chromosome breakage analysis on patients diagnosed with CdLS to narrow down the region for the CdLS-causing gene. Once *NIPBL* had been identified, further screening was carried out to elucidate additional point mutations causing CdLS.

1.2.2.2. *SMC1A* and *SMC3*

In a study by Musio *et al.*, (2006), it was hypothesized that *SMC1A* (located in an inactivated region of the X chromosome) was a potential CdLS causing gene based on its involvement in the cohesin pathway (it forms part of the cohesin core ring complex). *SMC1A* sequencing was used to identify a three base pair (bp) deletion and a missense mutation in two families. One of the families had three males diagnosed with CdLS and the other was a single *de novo* case. In a later study by Deardorff *et al.*, (2007), both *SMC1A* and *SMC3* were screened for mutations in a cohort of 115 CdLS patients. Similar to the previous study, the hypothesis of *SMC3*'s involvement was based on its function in the cohesin's core ring structure. An additional 14 *SMC1A* mutations and a single *SMC3* mutation were identified. Deardorff *et al.*, (2007) observed that the patients who tested positive for mutations in *SMC1A* and *SMC3* presented with a milder phenotype compared with patients who tested positive for a mutation in *NIPBL*, and in some cases, patients may even present with intellectual disabilities exclusively with little to no physical abnormalities. Gil-Rodriguez *et al.*, (2015) screened patients for *SMC3* mutations and found of the 16 patients who tested positive, they too presented with a milder phenotype compared with typical CdLS cases. However, Hoppman-Chaney *et al.*, (2011) presented a case study where a female patient had a multi-exon deletion in *SMC1A* and presented with a severe form of CdLS. These studies present evidence for inter- as well as intragenic phenotypic variability while also identifying additional CdLS causal genes. Mutations in the *SMC1A* gene accounts for 5% of CdLS-causing mutations in European populations (reviewed by Mannini *et al.*, 2013).

1.2.2.3. *RAD21*

Deardorff *et al.*, (2012) carried out a study whereby they utilised genome wide copy number analysis and subsequent sequencing in a cohort of 290 patients, 101 of which presented with classical CdLS and 189 were defined as only having an overlapping phenotype. At the time of this study, only mutations in *NIPBL*, *SMC1A* and *SMC3* had been identified as CdLS causal genes and the cohort of 290 patients all tested negative for mutations in these three genes. They identified *RAD21*, a known cohesin gene at the time, to be a potentially disease-causing gene through their copy number analysis, where they found one patient to have a deletion of 8q24.1 (which includes *RAD21*). They then sequenced *RAD21*'s exons to identify possible point mutations in their cohort. They observed that patients who tested positive for *RAD21* mutations appeared to have milder intellectual- and physical abnormalities compared with other CdLS patients, further suggesting a genotype-phenotype correlation.

1.2.2.4. *HDAC8*

Mutations in the *HDAC8* gene (which is also located on the X chromosome)were first reported by Harakalova *et al.*, (2012) in a family with multiple members presenting with intellectual disabilities amongst other phenotypes. It was not until Deardorff *et al.*, (2012) reported the role *HDAC8* plays in the cohesin pathway via deacetylation of *SMC3,* and the subsequent sequencing of patients for *HDAC8* mutations, that this gene was implicated in CdLS. Kaiser *et al.*, (2014) published a study that followed on from the one published by Deardorff *et al.*, (2012) and presented an additional 38 patients with *HDAC8* mutations and their phenotypic variability ranging from mild to severe.

Table 1.2. Types of mutations identified in CdLS cases in the 5 known causal genes as of 2013 (adapted from Mannini *et al.*, (2013), *n=311*

Type of mutation	**Number of mutations identified**					**% each type of mutation accounts for**
	NIPBL	*SMC1A*	*SMC3*	*RAD21*	*HDAC8*	
Missense	67	19	-	2	4	29.5%
Nonsense	43	-	-	-	1	14%
Small deletions	71	5	1	-	-	25%
Small insertions	33	-	-	-	-	10%
Small Indels	2	-	-	-	-	1%
Splicing	43	-	-	-	-	14%
Regulatory	2	-	-	-	-	1%
Gross deletions	16	-	-	1	-	5%
Translocations	1	-	-	-	-	0.5%

As is evident, inter- and intragenic variability, and possibly mosaicism, plays an important role on the phenotypic outcome of patients with CdLS. To date, a wide variety of types of mutations have been reported including nonsense, missense and splice site mutations; small and large insertions and deletions as well as large genomic rearrangements (Table 1.2.). Upon carrying out a phenotype-genotype correlation study, Mannini *et al.*, (2013) concluded that truncating mutations found in *NIPBL* result in a more severe phenotype while missense and in-frame deletions found in *NIPBL* and *SMC1A/SMC3* result in a milder form of the disease. Interestingly, mutations found in the HEAT domain of the NIPBL protein result in a severe phenotype as well, including missense and in-frame deletions thus proving not only are the type of mutations and genes important, but also the protein domain in which they are found.

The vast majority of mutations within the five causal genes occurs *de novo*. The severity of the disease clearly results in reduced reproductive fitness (Jackson *et al.*, 1993). There are, however, a few familial cases where a parent is very mildly affected and harbours a disease-causing mutation which is unique to each family. (Krantz *et al.*, 2004; Zhang *et al.*, 2009).

Mutations in the five genes only account for approximately 70% of CdLS cases and the causative genes and mutations contributing to the remaining 30% are yet unknown (reviewed by Mannini *et al.*, 2013).

1.2.3. **Other candidate cohesin genes implicated in Cornelia de Lange Syndrome**

Considering only approximately 70% of CdLS cases receive a molecular diagnosis, many studies have been carried out to elucidate the disease-causing variants in the remainder of patients. Studies by Zhang *et al.*, (2007, 2009) implicate *PDS5A* and *PDS5B* (Cohesin Associated Factor A and B respectively) in CdLS in mice showing a CdLS-like phenotype. They initially created a *PDS5B*-deficient mouse model and observed a phenotype resembling features of CdLS: including a cleft lip, short limbs and congenital heart defects (2007). In 2009, they went on to create a *PDS5A*-deficient mouse model and also observed features similar to those observed in human patients with CdLS and similar to the *PDS5B* mouse model: congenital heart defects, cleft palate and growth retardation. Based on these two mouse models, they concluded that *PDS5A* and *B* could be potential candidate genes for the 30% of molecularly undiagnosed CdLS patients. They then carried out *PDS5A* and *B* sequencing on 114 patients with CdLS who had previously tested negative for mutations in *NIPBL*, *SMC1A* and *SMC3*. They identified a familial CdLS case where three siblings were clinically affected with CdLS (two of whom had passed away, - DNA was only available for one of the deceased siblings) while the parents and the other sibling were clinically unaffected. Although it was noted that three siblings and the father shared the same deleterious mutation R1292Q, only two of the siblings who shared the same maternal *PDS5B* allele were affected with CdLS (Figure 1.3.). Since the mutation was found in the unaffected parent and unaffected sibling, they suggested that the CdLS phenotype was inherited in an autosomal recessive manner. There was potentially another deleterious mutation on the maternal allele shared by the affected siblings that went undetected in the sequencing to account for the unique inheritance pattern.

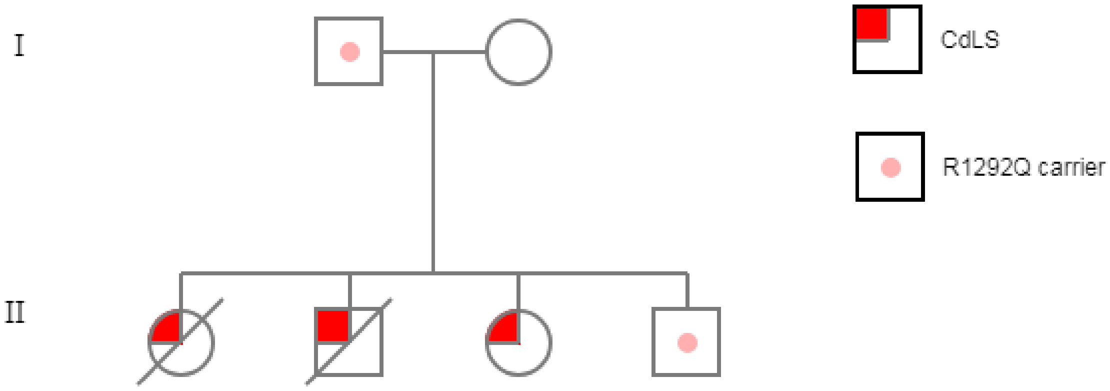


Figure 1.3. Family pedigree from Zhang et al (2009) study where they screened 114 patients for PDS5A and B mutations. One familial case was observed to have a PDS5B mutation present (represented by the red fraction or red dot in the square or circle). Affected siblings are represented by red fractions while carriers of the mutation are represented by a red dot.

Another gene thought to be implicated in CdLS is *STAG1*. This gene’s protein forms part of the core cohesin ring. In a study carried out by Remeseiro *et al.*, (2012), lack of STAG1 protein in mouse models result in a phenotype similar to CdLS: reduced body size (short stature), skeletal abnormalities and impaired lipid metabolism. Upon analysis of the function of *STAG1*, it was found that *STAG1* regulates the distribution of the cohesin complex at promoter sites and affects cohesin’s gene regulatory functioning. It was also postulated that a decrease in NIPBL may influence the loading of STAG1 and thus *STAG1*, even if not mutated itself, may have a role in the aetiology of CdLS. Two studies published within the last two years have identified the first *STAG1* mutations in humans with a phenotype similar to CdLS (Lehalle *et al.*, 2017; Yuan *et al.*, 2018). The evidence from the studies by Remeseiro *et al.* (2012); Lehalle *et al.* (2017) and Yuan *et al.* (2018) shows that *STAG1* is a good candidate gene to investigate as a potential CdLS-causing gene inpatients who do not have a molecular diagnosis.

SCC4 (also known as *MAU2*) was studied in depth by Watrin *et al.*, (2006) and it was reported that *SCC4* was essential in the loading of the cohesin complex onto chromatin. It was also observed that *SCC4* is conserved from yeast to humans, indicating its essential role in the cell. Another study further elucidated that *SCC4* interacts with *NIPBL* to facilitate the loading of cohesin onto chromatids (Braunholz *et al.*, 2012). In this study, the specific *SCC4*-*NIPBL* interaction domain was mapped and mutations were induced in the *SCC4*-interacting domain of *NIPBL* to determine the importance of this interaction. These mutations resulted in reduced binding and subsequently reduced loading of cohesin. In the same study, Braunholz *et al.*, (2012) tested 184 patients with CdLS who tested negative for the known CdLS-causing genes. No *SCC4* mutations were found in this cohort but this would not exclude *SCC4* from being a causal gene, it may possibly just be very rare.

Although studies have identified mutations in the four aforementioned genes, they have not been confirmed to cause CdLS in human patients to date (Oliver *et al.*, 2010; Braunholz *et al.*, 2012; Remeseiro *et al.*, 2012). These genes are potential causal genes in molecularly undiagnosed patients based on mouse models or hypotheses deduced from functional studies. Alternatively, these patients could potentially be misdiagnosed with CdLS due to the wide range of phenotypic variability as well as CdLS having significant phenotypic overlap with other conditions i.e. differential diagnoses.

1.3.<u>Differential Diagnoses</u>

Currently, a number of differential diagnoses exist for CdLS, some of which are discussed below (reviewed by Deardorff *et al.*, 2016). Although cases of these conditions can be clinically distinguished from classic cases of CdLS, a wide phenotypic spectrum exists which makes milder or atypical cases of CdLS more difficult to differentiate from other conditions. Several diagnostic gene panels exist which test for CdLS mutations that also include some of the below mentioned differential diagnoses. Each of these disorders have convincing evidence of overlapping phenotypic features with CdLS and are thus good candidates to include on a diagnostic or research gene panel (Table 1.3.).

1.3.1. CHOPS syndrome (*AFF4*)

In 2015, Izumi *et al.* described a previously unknown condition termed 'CHOPS' which they named based on aspects of the phenotype. This is described in the paper as "C for Cognitive impairment and Coarse facies, H for Heart defects, O for Obesity, P for Pulmonary involvement and S for Short stature and Skeletal dysplasia". As is evident from the description, there is a large phenotypic overlap with CdLS and thus is considered a differential diagnosis. CHOPS is a result of a genetic mutation in the *AFF4* gene (AF4/FMR2 Family Member 4). *AFF4* forms part of the super elongation complex which is responsible for elongation during transcription in the cell (Lin *et al.*, 2011). Interruptions in this process may lead to dysregulation of gene expression during critical stages in embryogenesis and thus result in the observed phenotype.

1.3.2. KBG syndrome (*ANKRD11*)

KBG Syndrome, named after the families originally diagnosed and reported in the literature (Herrmann *et al.*, 1975), is another example of a differential diagnosis for CdLS. Patients affected with this condition present with macrodontia, intellectual and developmental delay, short stature and skeletal and craniofacial abnormalities (Herrmann *et al.*, 1975). KBG is a result of mutations in the ankyrin repeat domain 11 gene (*ANKRD11*) which is involved in the regulation of transcription (Zhang *et al.*, 2004). In a study that sequenced 163 patients clinically diagnosed with CdLS, three patients who tested negative for mutations in the five known CdLS causal genes were seen to carry *ANKRD11* mutations thus demonstrating the phenotypic overlap between the two conditions (Ansari *et al.*, 2014).

1.3.3. Roberts syndrome (*ESCO2*)

Roberts Syndrome, also known as SC Phocomelia (Herrmann and Opitz, 1977), is another example of a cohesinopathy. It is a result of mutations in the *ESCO2* gene (Vega *et al.*, 2005) which is involved in establishment of the cohesin complex via acetylation. The clinical phenotype and its variability were defined in 2010 by Vega *et al.* This includes pre- and postnatal growth retardation, symmetrical limb malformations (predominantly and more severely in the upper limbs compared with the lower limbs), microcephaly, cleft lip and palate as well as distinct facial features not unlike those observed in patients with CdLS.

1.3.4. Wiedemann-Steiner Syndrome (*KMT2A)*

The gene found to be mutated in Wiedemann-Steiner Syndrome (WSS) is *KMT2A (*histone-lysine N-methyltransferase 2A*)* (also known as *MLL (*myeloid/lymphoid or mixed-lineage leukemia*)*) (Jones *et al.*, 2012). As the name suggests, this gene is involved in gene regulation by means of histone methylation. WSS, like CdLS, displays a wide phenotypic range but the most commonly shared feature is excessive hair growth (Yuan *et al.*, 2015). Some of the other features of WSS include intellectual and developmental disability, short stature and coarse facial features. In a study carried out by Yuan *et al.*, (2015), patients that were clinically diagnosed with WSS were found to have mutations in the CdLS-causing gene, *SMC1A*. Additionally, they found *KMT2A* mutations in a patient clinically diagnosed with CdLS and mutations in *SMC3* and *SMC1A* in patients with combined features of WSS and CdLS. This evidence makes *KMT2A* a likely candidate for a proportion of molecularly undiagnosed CdLS patients who may have been clinically misdiagnosed.

1.3.5. X-linked syndromic mental retardation syndrome (*TAF1*)

Males presenting with intellectual disability, distinct facial features (some shared with CdLS – long philtrum, low set ears, down slanted palpebral fissures to name a few), hypotonia and various neurological abnormalities were grouped together as having a common condition (O'Rawe *et al.*, 2015). A study on 9 families with individuals with the above-mentioned phenotype identified mutations in the *TAF1* (TBP associated factor 1) gene found on the X chromosome (O'Rawe *et al.*, 2015). *TAF1* plays a role by associating with TATA binding proteins to make up the transcription factor II D complex which is responsible for the initiation of transcription.

1.3.6. Alazami-Yuan syndrome (*TAF6)*

Alazami-Yuan syndrome was described simultaneously by Alazami *et al.*, (2015) and Yuan *et al.*, (2015) in consanguineous families from Saudi Arabia and Turkey respectively. Some of the phenotypes described by these two papers included synophrys, short stature, microcephaly, hirsutism, long philtrum and dysmorphic facial features, all similar to CdLS. Yuan *et al.*, (2015) reported on Alazami-Yuan syndrome based on its similarity with CdLS alongside *KMT2A* mutations. The gene independently reported by both Alazami *et al.*, (2015) and Yuan *et al.*, (2015) was the *TAF6* gene (TBP associated factor 6) which, like *TAF1*, assists the transcription factor II D complex in initiating transcription.

1.3.7. Autosomal dominant mental retardation syndrome *(SETD5)*

A study carried out by Grozeva *et al.*, (2014) identified mutations in the *SETD5* gene (SET domain-containing 5) in patients with a 3p25 microdeletion syndrome – autosomal dominant mental retardation syndrome. The main phenotype of this syndrome is intellectual disability with additional phenotypes that overlap with CdLS: synophrys, low set ears, skeletal abnormalities and developmental delay (Grozeva *et al.*, 2014). A further study carried out by Parenti *et al.*, (2017) looked at patients with a CdLS phenotype who tested negative for mutations in the five known causal genes. They identified mutations in the *KMT2A* (WSS) and *SETD5* genes, both of which are classified as methyltransferases.

1.3.8. Coffin-Siris Syndrome (*SMARCB1* and *ARID1B*)

Mutations in components in the SWItch/sucrose nonfermenting (SWI/SNF) complex have been found to cause Coffin Siris Syndrome (Santen *et al.,* 2012; Tsurusaki *et al.,* 2012). The resulting phenotype is intellectual and developmental delay, coarse facial features and microcephaly, which shows an overlap with the CdLS phenotype (Parenti *et al.*, 2017). Four patients in a cohort with a CdLS overlapping phenotype were found to harbour mutations in two of the genes making up the SWI/SNF complex: *ARID1B* (AT-rich interaction domain 1B) and *SMARCB1* (SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily B, member 1) (Parenti *et al.,* 2017). Additionally, in the same study, Parenti *et al.*, (2017) examined patients clinically diagnosed with Coffin Siris syndrome who tested negative for mutations in genes comprising the SWI/SNF complex. They sequenced the five CdLS causal genes and found one patient with a mutation in the *NIPBL* gene thus strengthening the argument that Coffin Siris syndrome is a differential diagnosis of CdLS.

Table 1.3. Differential diagnoses of CdLS, the genes involved in each condition and overlapping phenotypes with CdLS.

Differential Diagnosis	**Gene**	**Some common features with CdLS**
CHOPS	*AFF4*	Coarse facial features, skeletal anomalies, short stature, intellectual delay
KBG	*ANKRD11*	Intellectual and developmental delay, coarse facial features, skeletal anomalies
Roberts Syndrome	*ESCO2*	Upper limb malformations, coarse facial features, microcephaly
Wiedemann-Steiner	*KMT2A*	Intellectual and developmental delay, hirsutism, coarse facial features
X-linked syndromic mental retardation	*TAF1*	Long philtrum, low set ears, down slanted palpebral fissures
Alazami-Yuan	*TAF6*	Short stature, microcephaly, hirsutism, long philtrum
Autosomal dominant mental retardation syndrome	*SETD5*	Intellectual and developmental delay, synophrys, low set ears, skeletal abnormalities
Coffin-Siris Syndrome	*SMARCB1* and *ARID1B*	Intellectual and developmental delay, coarse facial features, microcephaly

1.4. Next Generation sequencing (NGS)

Before the era of next generation sequencing (NGS), Sanger sequencing was mainly used to identify disease-causing mutations and molecularly diagnose patients with CdLS. However, due to the genetically heterogeneous nature of CdLS, with five causal genes being identified to date, in combination with the multiple differential diagnoses, it is not always the most cost-effective approach to perform Sanger sequencing as a first line test in the diagnostic and research process anymore. In cases like CdLS when there are multiple genes that could potentially harbour a mutation, a diagnostic test that can examine multiple genes at once is preferred to a single gene approach like Sanger sequencing. Sanger sequencing however, is still the gold standard for molecular diagnostic testing examining individual genes as it has a lower error rate than NGS (Mu *et al.*, 2016).

Over the past few years, NGS (a high-throughput sequencing technique) has moved from the research to the diagnostic field and has proven to have powerful diagnostic abilities (Yohe *et al.*, 2015; Goodwin *et al.*, 2016). With the price of NGS decreasing, it is replacing older techniques in diagnostic laboratories across the world. One specific application of NGS is sequencing gene panels which are designed to evaluate a specified set of genes simultaneously. The gene panel could potentially span a number of different disorders and can be applied to a number of different patient samples simultaneously. This approach decreases both time and cost involved in obtaining a molecular diagnosis and is an ideal solution for a heterogeneous condition like CdLS. In the South African public health sector, NGS techniques are not yet widely adopted in a diagnostic setting. However, the use of techniques such as gene panel testing has the potential to increase diagnostic yields, thus benefitting the affected patient and their families, who may otherwise have had limited testing options.

1.5. Cornelia de Lange syndrome in a South African context

The prevalence of CdLS is approximately 1.6-2.2:100 000 in European populations (Barisic *et al.*, 2008); however, this estimate is likely to underrepresent the number of CdLS cases as the phenotype ranges from mild to severe and a number of differential diagnoses exist. No epidemiological study has been carried out on South African patients suspected to have CdLS and only a handful of case reports have previously been published on patients with African ancestry (Ptacek *et al.*, 1963; Cicoria, 1974; Begeman and Duggan, 1976). There is currently no molecular data on patients with CdLS from Africa, although a vastly different mutation profile is not expected as majority of the mutations are *de novo* (reviewed by Mannini *et al.*, 2013).

In the clinical setting, a diagnosis of CdLS is initially suspected based on the phenotype of the patient. Ideally, molecular genetic testing would then be requested to confirm the diagnosis in order to care for the patient appropriately and to provide them with accurate family planning information. Currently, the state health care system, which caters for approximately 80% of the population, offers no molecular diagnostic test to confirm a clinical diagnosis of CdLS (E. Vorster, personal communication, 16 March 2017). The diagnosis is thus made based on the clinical phenotype alone and often only in the patients who fall on the more severe end of the spectrum (C. Feben, personal communication 6 March 2017). Multiplex ligation-dependant probe amplification (MLPA) is occasionally requested to exclude the chromosome 3q duplication syndrome - a differential diagnosis of CdLS (Deardorff *et al.*, 2016), but thus far has not yielded a positive result (E. Vorster, personal communication, 16 March 2017). For patients in the private health care system and who are able to self-fund testing, clinicians may offer an NGS test performed by an international laboratory (such as CentoGene, Germany or Invitae (for website URL, see list of websites at the end of dissertation). These tests screen several genes implicated in CdLS causation and/or a differential diagnosis for CdLS. However, this approach is not feasible for the vast majority of patients, owing to the limited financial resources available.

1.6.Current study

1.6.1. Rationale

The current project utilised a specifically designed gene panel which incorporated the known CdLS causal genes, as well as the suspected genes and genes from the differential diagnoses mentioned above, to perform NGS on patients with a CdLS-like phenotype. It was envisioned that this would generate novel data on our diverse local population and address the current lack of African CdLS data. Further implications could include the provision of accurate and useful genetic counselling to the study participants which may have an influence on patient care and recurrence risk counselling. The generation of the data and the designing of the panel may have utility in the diagnostic field in the future, as simultaneous multi-gene testing could become more cost effective than sequential gene testing.

1.6.2. Aims and objectives

In the present project, the aim was to investigate the genetics of CdLS in the South African context as the majority of molecular research has been conducted on European populations thus far. We proposed that by designing a gene panel, we would be able to evaluate the relevant genes known to be involved in CdLS, as well as selected plausible causal genes simultaneously, in order to establish a mutation profile in South African patients with a CdLS phenotype.

The specific objectives of the present study were to:

- Recruit patients presenting with CdLS or CdLS-like phenotypes, and their immediate family members where possible, and to extract DNA from blood samples of these individuals.
- Determine known and suspected causal genes for CdLS and CdLS-like syndromes to include on the targeted NGS gene panel.
- Sequence the selected genes using the aforementioned gene panel and analyse the results using Agilent SureCall software.
- Validate any putative pathogenic mutations identified by the NGS technique via Sanger sequencing and segregation analysis where possible.

2. Materials and methods

2.1. Patient recruitment and ethics

The cohort for the study comprised 14 patients presenting with CdLS or a CdLS-like phenotype. Patients were recruited from genetic clinics that were held in and around the Johannesburg and Pretoria areas. These clinics are staffed by medical geneticists and genetic counsellors from the National Health Laboratory Services (NHLS) in association with the University of the Witwatersrand and the University of Pretoria. This research is a sub-study of a larger project on developmental disorders in the Division of Human Genetics (certificate number M160830) and was approved by the Human Research Ethics Committee (Medical) at the University of the Witwatersrand (certificate number M170761; Appendix B.1.) and the University of Pretoria (certificate number 80/2018).

Most patients in the study had previously been assessed at one of the genetic clinics staffed by the NHLS, including those held at Charlotte Maxeke Johannesburg Academic Hospital, Chris Hani Baragwanath Hospital and Rahima Moosa Mother and Child Hospital. An additional recruitment clinic was held at the University of Pretoria for patients who had previously been assessed at Steve Biko Pretoria Academic Hospital. Patients presenting with suspected CdLS were identified and recruited to the study by a medical geneticist. For the present study, the only inclusion criterion was that the patients met the diagnosis/suspected diagnosis criteria as set forth by the medical geneticist. At the time of recruitment, formal consensus criteria were not available; these have now been published in an international consensus statement (Kline *et al.*, 2018).

At the clinic appointment, information sheets were given to the patients and/or their guardian and written informed consent was obtained after discussing the study with the patient or family. The consent form covered aspects of the project and clinic appointment: clinical phenotyping by examination, photographs and blood sampling and storage as well as family members' blood sampling and storage (Table 2.1.). It was also discussed that if a putative disease-causing mutation were identified and validated in a diagnostic setting that the family may request to receive this information in a follow up genetic counselling session.

Table 2.1. List of demographics and the procedures each patient consented to.

Patient number	**Date of birth**	**Gender**	**Clinical examination**	**Blood sampling and storage**	**Photograph**	**Relative's blood sampling and storage (number of relatives)**
FRASC8	3/1/2017	F	X	X	X	1
FRASC29	14/6/2011	F	X	X		1
FRASC30	1/4/2014	F	X	X		1
FRASC32	24/10/2003	F	X	X		0
FRASC41	5/4/2016	F	X	X		0
FRASC47			X	X		0
FRASC49	28/4/2016	F	X	X	X	0
FRASC51	16/1/2011	M	X	X	X	2
FRASC72	19/8/2016	F	X	X	X	1
FRASC75	22/8/2016	F	X	X		0
FRASC76	15/5/2018	M	X	X		1
FRASC77	13/8/2017	F	X	X		1
FRASC78	4//6/2011	M	X	X		1
FRASC79	5/7/2018	F	X	X		1

A clinical examination was conducted, specifically to document relevant growth and physical features. An inhouse clinical tick sheet (Appendix A.1.) has previously been specifically designed by the medical geneticists based on commonly reported CdLS phenotypes as well as their own experience with CdLS. Additionally, the clinic file was reviewed to document any known medical diagnoses or congenital abnormalities in the patient and other family members as appropriate.

2.2.Sample collection and DNA extraction

Blood samples of between 1-10ml were collected from each patient and their family members where possible, at the clinic appointment. A modified version of the salting out protocol that is routinely used in the diagnostic department in the Division of Human Genetics was used to extract DNA (Miller, Dykes and Polesky, 1988). This was carried out by a diagnostic staff member, assisted by the students of the FRASC team. The process involved lysing whole blood samples with a triton-X sucrose buffer, degrading nuclear membranes and proteins with a proteinase mix and precipitating out the DNA from the solution with a saturated NaCl solution followed by the addition of ethanol. The final step was assessing DNA quality and quantity using the Nanodrop 2000 Spectrophotometer (Thermo-Fisher Scientific, California USA). The A260/A280 and A260/A230 ratios were examined to ensure the quality of the DNA was sufficient and no RNA or protein interference would occur downstream. The genomic DNA was then further assessed using the Qubit v3.0 (Invitrogen by Thermo-Fisher Scientific, Johannesburg, South Africa) to determine the concentration of double stranded DNA. Agarose gel electrophoresis was utilised as another quality control step to analyse the molecular weight of the DNA and was visualised using the Omega Fluor™ Gel Documentation System, (Vacutec, Johannesburg, South Africa). This process is achieved by running an electrical current through a tank containing a buffer. The DNA that has been inserted in the agarose gel will then move at a speed that is determined by the size of each DNA fragment.

2.3.Candidate gene selection and assay design

A targeted gene panel was designed specifically for this project using the Agilent SureDesign software (for website URL, see list of websites at the end of dissertation). The gene panel incorporated 18 genes that were identified in the literature to be related to CdLS. This included the five known causative genes (*NIPBL, SMC3, SMC1A, RAD21* and *HDAC8*), four genes suspected to account for the 30% of molecularly undiagnosed patients with CdLS (based on mouse-model studies) were also included (*PDS5A, PDS5B, STAG1* and *SCC4*). Lastly, nine genes involved in CdLS differential diagnoses were added to the gene panel (*SETD5, AFF4, ARID1B, TAF6, ESCO2, KMT2A, ANKRD11, SMARCB1* and *TAF1*). Only exonic regions were included, covering an additional 10 intronic bases on either side of each exon to account for splice site mutations.

2.4.Library preparation

Library preparation was carried out using the Agilent^{QXT} target enrichment system protocol (Agilent Technologies, CA, USA). An overview of the protocol is summarised below (Figure 2.1.).

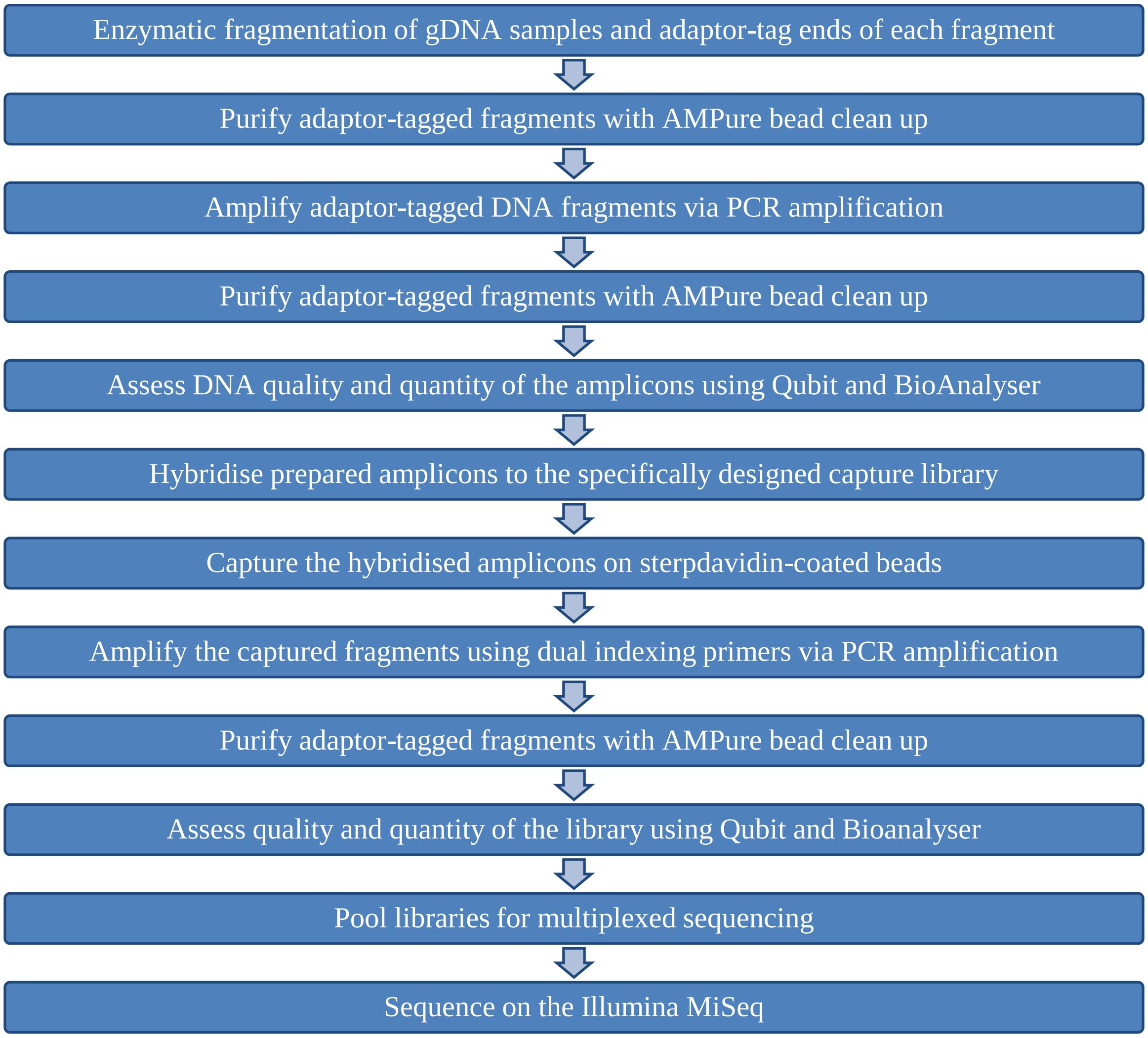

Enzymatic fragmentation of gDNA samples and adaptor-tag ends of each fragment
Purify adaptor-tagged fragments with AMPure bead clean up
Amplify adaptor-tagged DNA fragments via PCR amplification
Purify adaptor-tagged fragments with AMPure bead clean up
Assess DNA quality and quantity of the amplicons using Qubit and BioAnalyser
Hybridise prepared amplicons to the specifically designed capture library
Capture the hybridised amplicons on sterpdavidin-coated beads
Amplify the captured fragments using dual indexing primers via PCR amplification
Purify adaptor-tagged fragments with AMPure bead clean up
Assess quality and quantity of the library using Qubit and Bioanalyser
Pool libraries for multiplexed sequencing
Sequence on the Illumina MiSeq

Figure 2.1. Flow diagram depicting an overview of the library preparation protocol (Adapted from SureSelect^{QXT} Target Enrichment for Illumina Multiplexed Sequencing, Agilent Technologies, CA, USA).

Transposons were used to fragment 50ng of gDNA and to simultaneously add adaptors to the ends of the DNA fragments during the first step. The fragmented, adaptor-tagged DNA was then washed twice using AMPure XP beads (Beckman Coulter, CA, USA) as per the protocol (using a ratio of 1:1.8). The DNA was then amplified using a limited cycle PCR (on the Agilent Technologies SureCycler 8800 G8800A, Germany) and washed twice again using AMPure XP beads (Beckman Coulter, CA, USA) (using a ratio of 1:1).

The quantity and quality of each DNA sample was then assessed by diluting a 1µl aliquot of each library by 1:100 and analysing it on the Agilent Technologies 2100 Bioanalyser (Agilent

Technologies, Waldbronn, Germany) using the High Sensitivity DNA Reagents kit (Agilent Technologies, Waldbronn Germany) as per the manufacturer's instructions. If the libraries fell within the acceptable limits (fragment sizes between 245 and 325bp and with a minimum concentration of 50ng/µl) the protocol could proceed to target capture. The fragment sizes were assessed to ensure under- or over-fragmentation had not occurred as this will negatively affect library preparation and sequencing downstream. A concentration of at least 50ng/ul had to be established to ensure optimal library preparation downstream, particularly regarding DNA to bead ratios. Libraries that did not achieve this quality metric were then discarded.

The samples that did pass the quality control step were then hybridised to the specifically designed capture probes targeting the genes of interest previously described. The probes are biotinylated RNA molecules that were designed to be complementary to the regions of interest and would therefore bind only to the regions to be sequenced. The hybridised fragments were subsequently captured on Dynabeads^{TM} MyOne^{TM} Strepdavidin T1 beads (Invitrogen, ThermoFisher Scientific, Baltics, UAB, Norway) (with a ratio of 1:6.7) where they were amplified using pairs of sample-specific dual indexing primers to enable multiplexing. Samples were then washed twice using AMPure XP beads (Beckman Coulter, California, USA) (using a ratio of 1:1.2) and underwent another quality control step using the 2100 Bioanalyser High Sensitivity DNA Reagents kit (Agilent Technologies, Lithuania) to determine fragment length and concentration. This quality control step was essential to gain measurements needed in equimolar pooling calculations requiring fragment lengths and concentrations of each sample. Once the samples passed this quality control step, they were either stored at 4°C if sequencing was to commence the following day or at -20°C if the samples were to be sequenced in the next month.

2.5.<u>Next Generation Sequencing</u>

Once the index-tagged libraries passed all quality control steps, the molarity was calculated for each sample and equimolar pooling was carried out according to the Illumina MiSeq System Denature and Dilute Libraries Guide (Illumina, California, USA). Sequencing was then carried out using the MiSeq® Reagent Nano Kit V2 and Micro Kit V3 (Illumina, California, USA) on an

Illumina MiSeq NGS system (Illumina, California, USA) in three separate runs. The first run was conducted using a MiSeq® Reagent Nano V2 kit (Illumina, CA, USA) and the last two runs were performed using a MiSeq® Reagent Micro V2 kits (Illumina, CA, USA). The Nano kit was used initially to assess the utility of the MiSeq® kits in the present study design. Once this was shown to yield good quality data, a larger kit, the MiSeq® Reagent Micro V2 kit, was utilized. Compared with the Nano kit, the Micro kit has a larger sample capacity and yields more data, particularly when fewer samples are added. Each kit comes with a cartridge with Illumina TruSeq primers designed for libraries prepared with Illumina kits. However, a SureSelect library preparation kit was used and therefore the TruSeq primers were spiked with 3µl of SureSelect^{QXT} read and indexing primers to correctly identify the unique identifiers of our prepared libraries as per the SureSelect^{QXT} Target Enrichment for Illumina Multiplexed Sequencing protocol (Agilent Technologies, California, USA). As an additional quality control check, 12.5pM of a PhiX FC-110-3001 control was also added (Illumina, California, USA) to our pooled library. The PhiX control is a bacteriophage with a genome of 5386 bases (Sanger *et al.*, 1977). Once the run had finished, the data on the PhiX control was an indicator of how well or how poorly a run went since the data on these outcomes are known.

2.6.<u>Sequencing output quality control and variant calling</u>

The sequencing output files, in FASTQ and FASTQC format, generated from the MiSeq underwent quality assessment using Illumina Sequence Analysis Viewer (SAV) (for website URL, see list of websites at the end of dissertation). This provided information on a variety of quality control metrics including equimolar pooling accuracy, clustering densities and the amount of data generated that is usable, i.e. eliminating the probability of incorrect base calling. Cluster density was measured in K/mm^2 (the number of clusters per tile measured in thousands per mm^2). This measurement determined how many clusters of libraries were present on the flow cell provided by the sequencing kit. The cluster density is important because over- or under-clustering could have negative effects downstream regarding quality data and/or output. Over-clustering could cause the sequencing run to crash or negatively affect data output whereas under-clustering affects the amount of data output but not data quality. The quality score (Q-score) indicates the probability of an incorrect base being called. A Q-score of 30 indicates a one in 1 000 chance of a base being called incorrectly. The percentage of data with a Q-score of above 30 were considered, the higher the percentage, the more bases were called correctly indicating the amount of usable data (according to the Illumina technical note: sequencing, Quality Scores for Next Generation Sequencing, for website URL, see list of websites at the end of dissertation). These quality metrics were also analysed for the PhiX control and when compared with the sequence data, the quality in comparison could be extrapolated.

The FASTQ files were used as the input for Agilent's SureCall software which aligned reads to the hg19 human genome assembly (Genome Reference Consortium Human Build 37 (GRCh37)) and produced quality control (QC) reports for each sample. This provided information on the coverage of each sample overall as well as per exon. It also indicated exons that were not covered sufficiently (exons with less than 20X coverage). This information was used to identify regions that were covered well and to also identify regions where Sanger sequencing may be necessary to fill any coverage gaps of the genes sequenced. The Agilent SureCall software was used to generate a BAM file (a binary format of a Sequence Alignment Map). The BAM files were inputted into the Integrative Genomics Viewer program (IGV) (for website URL, see list of websites at the end of dissertation) for further quality assessment. The loading of BAM files onto IGV allows for the visualisation of coverage per exon and to determine if any specific exon had particularly poor coverage and needed Sanger sequencing to fill in any potentially vital gaps created in the targeted gene panel sequencing. This was done with the aid of a browser extensible data file (BED) file which was specifically generated to define the regions that were sequenced as opposed to focusing on the entire genome. It provided information such as chromosome number, start and end positions of sequenced regions as well as exon numbers. The variants were then called using Agilent SureCall software to generate a Variant Call File (VCF). The VCF files are used to identify variants detected during sequencing. VCF files were used later in the data analysis part of the methodology. Once the quality of the data was determined to be sufficient, data analysis was carried out.

2.7.NGS Data Analysis

The variants were analysed by inputting the VCF files generated by Agilent's SureCall software into the online tool wANNOVAR (Chang and Wang, 2012; for website URL, see list of websites at the end of dissertation). This generated a table of all variants identified during sequencing and data about each variant (e.g. population frequency, prediction scores from multiple variant effect prediction tools described in table 2.2., HGVS nomenclature etc.) (Table 2.3.; Appendix C.1., for Appendix C.1. file URL, see list of websites at the end of dissertation – available via Google Drive link). These prediction tools included Sorting Intolerant From Tolerant (SIFT) (Sim *et al.*, 2012), Polymorphism Phenotyping (PolyPhen-2) (Adzhubei *et al.*, 2010), Protein Variation Effect Analyzer (PROVEAN) (Choi and Chan, 2015), MutationTaster (Schwarz *et al.*, 2014), Functional Analysis Through Hidden Markov Models (FATHMM) (Shihab *et al.*, 2012), Combined Annotation-Dependant Depletion (CADD) (Rentzsch *et al.*, 2019), Deleterious Annotation of genetic variants using Neural Networks (DANN) (Quang, Chen and Xie, 2015) and Genomic Evolutionary Rate Profiling (GERP++) (Davydov *et al.*, 2010). Each tool has a specific scoring system to represent predictions on a spectrum of deleterious to benign (Table 2.2.). ClinVar predictions were also considered if any information was available in the ClinVar database.

Table 2.2. Prediction tools and interpretations of each scoring system on the deleterious or benign nature of a variant.

<table>
<tr><th>Prediction tool</th><th colspan="12">Scoring system (from benign to deleterious)</th></tr>
<tr><td>SIFT*</td><td colspan="4">Scale of 0-1</td><td colspan="4">Benign: 0.05-1</td><td colspan="4">Deleterious: 0.05-0</td></tr>
<tr><td>PolyPhen2*</td><td colspan="4">Scale of 0-1</td><td colspan="4">Benign: 0-0.15</td><td colspan="4">Deleterious: 0.15-1</td></tr>
<tr><td>Provean*</td><td colspan="4">Indefinite scale</td><td colspan="4">Benign: >-2.5</td><td colspan="4">Deleterious: ≤-2.5</td></tr>
<tr><td>MutationTaster</td><td colspan="3">P: Polymorphism automatic</td><td colspan="3">N: Polymorphism</td><td colspan="3">D: Disease causing</td><td colspan="3">A: Disease causing automatic</td></tr>
<tr><td>FATHMM*</td><td colspan="4">Indefinite scale</td><td colspan="4">Benign: >0</td><td colspan="4">Deleterious: <0</td></tr>
<tr><td>CADD*</td><td colspan="3">Benign: <10</td><td colspan="3">Top 10% of deleterious mutations: >10</td><td colspan="3">Top 1% of deleterious mutations: >20</td><td colspan="3">Top 0.1% of deleterious mutations: >30</td></tr>
<tr><td>DANN*</td><td colspan="4">Scale of 0-1</td><td colspan="4">Benign: <0.98</td><td colspan="4">Deleterious: >0.98</td></tr>
<tr><td>GERP++*</td><td colspan="4">Measure of conservation</td><td colspan="4">Not conserved region: <4 (mutations more likely to be benign)</td><td colspan="4">Conserved regions: >4 (mutations more likely to be pathogenic)</td></tr>
</table>

**SIFT: Sorting Intolerant From Tolerant; PolyPhen: Polymorphism Phenotyping; Provean: Protein Variation Effect Analyzer; FATHMM: Functional Analysis Through Hidden Markov Models; CADD: Combined Annotation-Dependant Depletion; DANN: Deleterious Annotation of genetic variants using Neural Networks; GERP++: Genomic Evolutionary Rate Profiling.*

Table 2.3. Various data reported by the online tool wANNOVAR for each variant identified (Chang and Wang, 2012).

Type of information	Datasets used	Utility for data analysis
Chromosomal positions and variant nomenclature	Human hg19 genome build	Useful for searches on other genomic search engines or online tools (e.g. Ensembl and VarSome)
Effect of variant	Human hg19 genome build	This helps in assessing a variant e.g. is it exonic or intronic or in the 3’ UTR region. It also assists with variant effect e.g. nonsynonymous, synonymous, frameshift, splice site mutations etc.
Population frequencies	1000 Genomes Project (1000G), ExAC dataset, NHLBI Exome Sequencing Project (ESP6500si), gnomAD	This provides the minor allele frequency (MAF) which is the first step in the variant filtering process
Public database references	dbSNP, COSMIC, ClinVar, GWAS	If a variant had been reported previously, these tools provided their database reference number and some basic information available on each variant e.g. was it classified as benign or pathogenic
Prediction tools (functional and evolutionary)	SIFT, PolyPhen, MutationTaster, FATHMM, PROVEAN, CADD, DANN, GERP++	These tools give a range of scores to determine if a variant is pathogenic or benign or falls somewhere in between. It ranges from functional protein predictions to conservation of a region and predicting effects should that region contain a variant. This helps when forming an opinion of whether a variant is pathogenic

Variants were then filtered sequentially based on the information produced by wANNOVAR (Figure 2.2.). The top candidates for potential mutations in each sample were then examined in depth according to the codes provided by the American College of Medical Genetics and Genomics (ACMG) guidelines for interpreting variant pathogenicity (Richards *et al.,* 2015) and assigned one of five statuses: benign, likely benign, unknown significance, likely pathogenic or pathogenic. The ACMG codes are explained in depth in tables taken from the Richards et al. paper (2015) which can be accessed in Appendices D.1. – D.3. If sufficient evidence was presented to indicate that a variant was putatively disease-causing, various tools were used to better visualise the effect of the variant. IGV was used to visualise the variant at the genomic level and to also determine if the region was covered sufficiently to call the variant confidently.

The online tool Mutalyzer (Wildeman *et al.*, 2009) was then used to visualise the effect a variant had on the amino acid sequence (i.e. if there was a truncation or a change in the sequence).

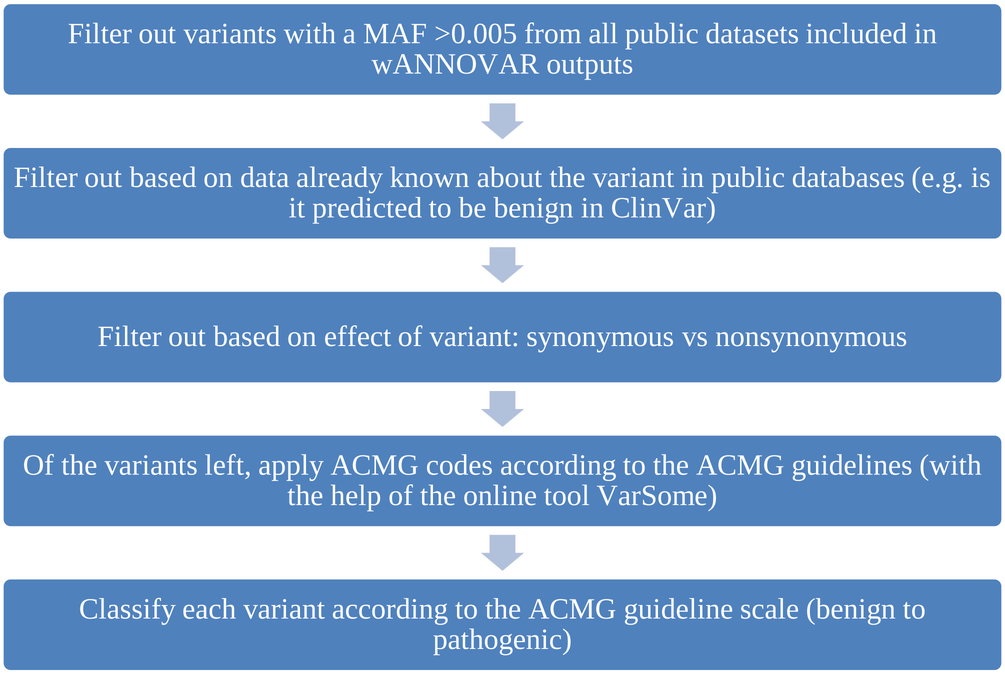

Filter out variants with a MAF >0.005 from all public datasets included in wANNOVAR outputs
Filter out based on data already known about the variant in public databases (e.g. is it predicted to be benign in ClinVar)
Filter out based on effect of variant: synonymous vs nonsynonymous
Of the variants left, apply ACMG codes according to the ACMG guidelines (with the help of the online tool VarSome)
Classify each variant according to the ACMG guideline scale (benign to pathogenic)

Figure 2.2. Flow diagram depicting the variant filtering pipeline of all variants to produce a list of potential causative mutations for each patient.

The first step in filtering out the benign variants was to assess MAF; any variants with a MAF of >0.005 were excluded as CdLS is a rare dominant condition and therefore causative mutations are not expected in a healthy population. There are exceptions to this rule as outlined by Ghosh *et al.*, (2018) where a variant that has a MAF >0.05 may still be pathogenic. The second step was to analyse all the known data on each variant that had been reported in tools and public databases. This type of information included the types of variants present (synonymous, nonsynonymous, frameshift etc.) as well as where the mutation occurred (exonic or intronic regions, splice sites or 3' UTR regions etc.). Once all the mutations deemed to be benign were excluded, a list of putatively disease-causing mutations were generated for each sample. These were then analysed in more depth using the ACMG guidelines and assigning various ACMG codes (Appendices D.1. and D.2.) to classify them on the scale of benign to pathogenic (Appendix D.3.). These variants were also inputted into the online tool, VarSome (Kopanos *et al.*, 2018) (which also adopts the ACMG guidelines (Richards *et al.*, 2015)) as an additional tool for variant interpretation to ensure no data points were missed. The five genes known to cause CdLS were examined first and if no putative disease-causing mutation was identified, the remaining genes in the candidate lists were then analysed. If any putative disease-causing mutations were identified after the filtering had been completed, the variants were then validated by means of Sanger sequencing.

2.8.Variant validation with Sanger sequencing

If any putative disease-causing mutations were classified according to the ACMG guidelines as being either likely pathogenic or pathogenic or if they had a status of unknown significance but with other evidence strongly indicating pathogenicity, validation would be carried out by means of Sanger sequencing. Sanger sequencing was carried out for the eight patients with putative disease-causing mutations identified. Sequences were obtained from Ensembl Release 94 (build GRCh37) (Zerbino *et al.*, 2018) and primers for the sequencing were designed using the online program Primer3 v.0.4.0 (Koressaar and Remm, 2007; Untergasser *et al.*, 2012) (for website URL, see list of websites at the end of dissertation). A PCR was carried out using the reagents, specifically designed primers and thermocycler conditions depicted in tables 2.4. and 2.5. The products were then cleaned using ddH$_2$O on a MultiScreen® plate (Merck Millipore, Cork, Ireland) with the vacuum pump Millipore Millivac Maxi SD1P014M04 (Merck Millipore, Cork, Ireland). Cycle sequencing was then carried out using a BigDye^{TM} Terminator v.3.1 Cycle Sequencing Kit (Applied Biosystems, ThermoFisher Scientific, Austin, USA) using the reagents and thermal cycler conditions specified in tables 2.6. and 2.7. The samples were cleaned using Injection Solution (Merck Millipore, Cork, Ireland) again with the use of the MultiScreen® plates on the Millipore Millivac Maxi SD1P014M04 vacuum pump system (Merck Millipore, Cork, Ireland). Finally, the products were denatured, and Sanger sequencing was carried out on the Genetic Analyser 3500xL (622-0015) (Applied Biosystems, HITACHI). Sanger sequencing electropherograms were then analysed using the biological sequence alignment editor BioEdit Version 7.2 (Hall, 1999).

Table 2.4. Primers designed to validate mutations identified in the targeted gene panel.

Sample number	**Forward primer (5'-3')**	**Reverse primer (5'-3')**	**PCR thermocycler conditions**					
			Initial denaturation	**Denaturation ***	**Annealing ***	**Elongation***	**Final elongation**	**Hold**
FRASC8	AGACTCTGACAATAAAGGTGTGA	AGTGAGAATGTGGTTGACGC	95°C: 10 minutes	95°C: 15 seconds	51.5°C: 30 seconds	72°C: 30 seconds	72°C: 5 minutes	4°C
FRASC30	AGGAGGGATTCTGGAAAGCC	CGAACCCTAGACTGATCCCC	95°C: 10 minutes	95°C: 15 seconds	58°C: 30 seconds	72°C: 30 seconds	72°C: 5 minutes	4°C
FRASC75	TGGTATCAGTGTCAGGAAAAGAG	CCTCTTCATCATTGACTCTGCG	95°C: 10 minutes	95°C: 15 seconds	61.7°C: 30 seconds	72°C: 30 seconds	72°C: 5 minutes	4°C
FRASC76	TTGGCAGTGATGACCCAGAA	AGGCATAAACATCGCATTCCT	95°C: 10 minutes	95°C: 15 seconds	51.2°C: 30 seconds	72°C: 30 seconds	72°C: 5 minutes	4°C
FRASC72	GGTGCTCCAGTGCTTTCT	TGTTCCGCATAGCAGGTTCT	95°C: 10 minutes	95°C: 15 seconds	51.2°C: 30 seconds	72°C: 30 seconds	72°C: 5 minutes	4°C
FRASC47	TGCTGCATTGTGAAAGGACC	GGATACGGTAATTACACACCCT	95°C: 10 minutes	95°C: 15 seconds	52.5°C: 30 seconds	72°C: 30 seconds	72°C: 5 minutes	4°C

*Cycle steps carried out 30 times

Table 2.5. Reagents added per sample in PCR set up.

Reagent	**Volume (µl)**	**Final concentration**
AmpliTaq Gold Buffer II (10X)	2.5	1X
dNTP Mix (10mM per dNTP)	1.0	200µM
AmpliTaq Gold $MgCl_2$ (25mM)	1.5	1.5mM
Forward primer (10µM)	0.5	0.2µM
Reverse primer (10µM)	0.5	0.2µM
DNA (50ng/µl)	1.0	2ng/µl
AmpliTaq Gold DNA polymerase (5U/µl)	0.13	1.25U
ddH_2O	17.9	N/A
Total volume	25.0	

Table 2.6. Thermocycler conditions for cycle sequencing performed on all samples to undergo Sanger sequencing.

Initial denaturation	**Denaturation***	**Annealing***	**Elongation***	**Hold**
96°C: 1 minute	96°C: 10 seconds	50°C: 5 seconds	60°C: 4 minutes	4°C

*Cycle steps carried out 25 times

Table 2.7. Reagents added per sample for cycle sequencing PCR set up.

Reagent	**Volume (µl)**
Cleaned PCR product	2.0
BigDye^{TM} Terminator 3.1 Ready Reaction Mix	1.0
BigDye^{TM} Terminator 5X Sequencing Buffer	1.5
Primer (forward or reverse) (10µM)	1.0
ddH_2O	4.5
Total volume	10.0

3. Results

3.1. Patient cohort

Fourteen patients were recruited for the present study. Of these, the phenotype varied from classical to mild and atypical. The most frequently observed clinical phenotypes in the patients are summarized in Table 3.1. A review of the patients' phenotypes revealed that there were seven patients who presented with classical CdLS, three who presented with a mild CdLS phenotype and one who presented with an atypical phenotype of CdLS based on the internally developed tick sheet used at the recruitment clinics (Table 3.1. and Appendix A.1.). The recent publication of the CdLS consensus guidelines outlines a clinical scoring system for the diagnosis of CdLS. Retrospective analysis of the patient cohort indicates that nine of the 14 patients had a score indicative of the diagnosis of classical or non-classical CdLS (Kline *et al.*, 2018)

Table 3.1. Most frequent occurring phenotypes in a South African CdLS patient cohort.

Phenotype	**FRASC study number**													
	8	**29**	**30**	**32**	**41**	**47**	**49**	**51**	**72**	**75**	**76**	**77**	**78**	**79**
Failure to thrive	X		X		X	X		X	X			X	X	X
Microcephaly	X		X	X	X		X	X	X	X	X	X	X	X
Hirsutism	X	X	X	X	X	X	X		X	X	X	X	X	X
Smooth philtrum	X				X		X	X	X	X	X	X	X	X
Upper limb abnormalities	X		X	X	X		X		X	X	X	X		X
Lower limb abnormalities	X													
Structural malformations	X		X	X				X	X	X	X	X		
Mild intellectual disability	N/A		X		X					N/A	N/A			N/A
Moderate/severe intellectual disability	N/A	X					X	X	X	N/A	N/A	X	X	N/A
Clinical score according to the recent consensus guidelines*	13	6	9	8	12	5	10	5	12	12	9	12	10	7

*Blue shading indicates a classical phenotype, pink shading indicates a mild phenotype and green shading indicates an atypical CdLS phenotype. No shading indicates no information was available if a classical, mild or atypical phenotype was observed. *Scoring system: ≥11=classical CdLS, 9-10=non-classic CdLS, 4-8=molecular testing for CdLS indicated, <4=insufficient clinical information to indicate molecular testing.*

Three patients consented to photographs being published. These are shown below to demonstrate the phenotype of CdLS (Figure 3.1.).

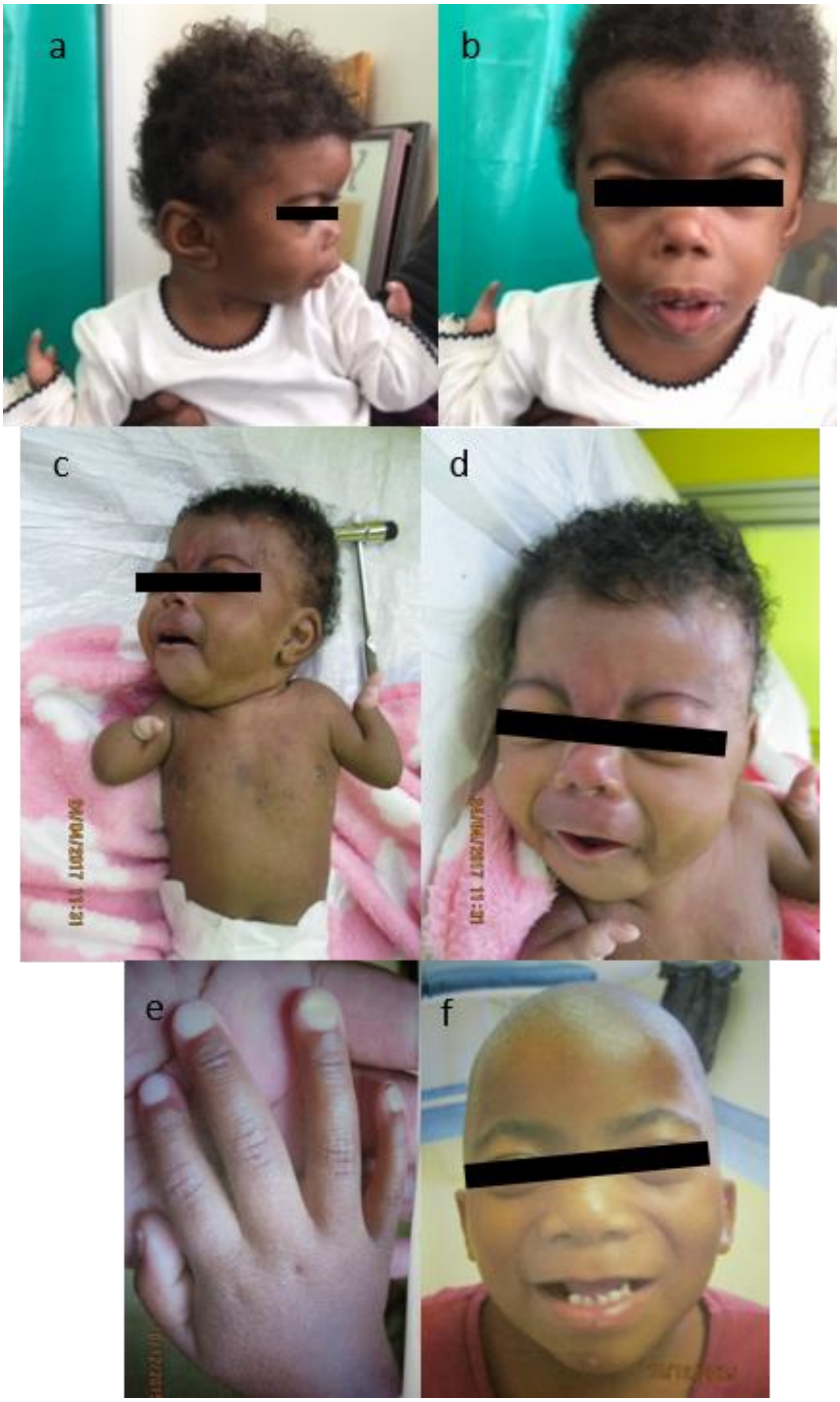


Figure 3.1. Photos showing the spectrum of phenotypes observed in FRASC72: a and b; FRASC8: c and d; and FRASC51: e and f. The classical lower limb abnormalities can be seen in a, b and c whereas there was a lack of hand malformations observed in e. Patients FRASC72 and 8 share a classical CdLS phenotype with coarse facial features, smooth philtrum and hirsutism clearly shown in a, b, c and d. However, FRASC51 presented with an atypical phenotype; mild synophrys and a smooth philtrum can be observed in f.

3.2.Sequencing data quality control

Samples were sequenced in three sequential runs (Table 3.2.).

Table 3.2. Samples included in each NGS run.

Run 1	**Run 2**	**Run 3**
FRASC32	FRASC29	FRASC49
FRASC41	FRASC30	FRASC72
	FRASC8	FRASC51
	FRASC75	FRASC78
	FRASC76	
	FRASC77	
	FRASC47	
	FRASC79	

The samples were run on the Agilent Technologies 2100 Bioanalyzer (Waldbronn, Germany) using the High Sensitivity DNA Reagents kit (Agilent Technologies, Lithuania) as a quality control step during library preparation. The Bioanalyzer traces obtained at the end of the library preparation are shown below (Figure 3.2.). Samples were prepped in batches and some had to be repeated due to low quality, however, all samples were eventually at a sufficient quality to continue on to sequencing. A fragment size between 200 and 400 bp was expected.

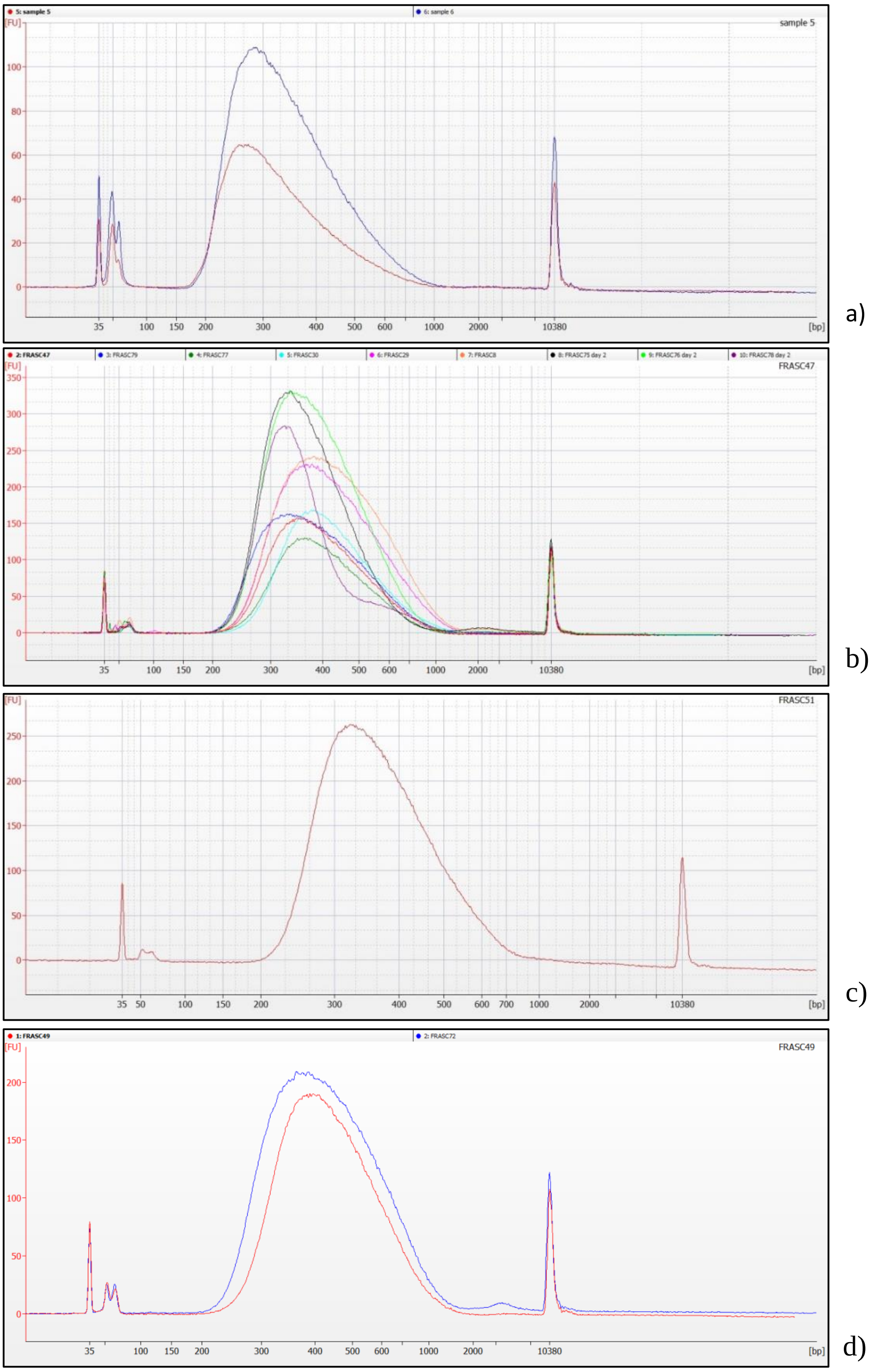


Figure 3.2. Bioanalyzer traces at the end of library preparation indicating each sample was in the range of the acceptable fragment length to proceed with sequencing. The traces were overlapped where possible to make a comparison of the samples fragment sizes: a) FRASC32 and FRASC41; b) FRASC47, 79, 77, 30, 29, 8, 75, 76, 78; c) FRASC51 and d) FRASC49 and FRASC72.

All three runs exhibited quality of an acceptable standard. Quality is based on the Q scores and cluster density obtained from the Illumina SAV software and the coverage obtained from the QC reports generated by the Agilent SureCall software (Table 3.3.).

Table 3.3. Summary of the quality control scores of the three Illumina MiSeq runs.

	Run one	**Run two**	**Run three**
%>Q30	94.0	84.6	90.5
Cluster density (K/mm^2)*	628	646	742
Average coverage for all patients in run	57X	100.75X	180X

*Clusters on flow cell measured in thousands per mm^2

As can be seen, the percentage of samples with a Q-score above 30 was sufficiently high to ensure the confidence of the base calling i.e. very low probability of incorrect base calling (according to the Illumina technical note: sequencing, Quality Scores for Next Generation Sequencing - for website URL, see list of websites at the end of dissertation). The cluster density, which should range from between 865-965 (Genohub, 2019) was below average in each run. Under-clustering does not affect the quality of the data but does affect the total data output. These values, however, are not low enough to have too large an impact on data output. The recommended average read depth coverage is 35X for targeted re-sequencing (Ajay *et al.*, 2011). This study yielded a higher average read depth coverage for all three runs falling between 57X and 180X (Table 3.3.).

The coverage of each individual sample was analysed using the Agilent SureCall QC reports. This indicates the overall coverage of each patient as well as the percentage of specific bases covered in each patient (Table 3.4.).

Table 3.4. Coverage for each patient sequenced on the Illumina MiSeq platform.

Patient number	**Overall coverage**	**% of bases with at least 20 reads**
FRASC8	101X	97.28
FRASC29	105X	96.07
FRASC30	118X	98.22
FRASC32	50X	88.59
FRASC41	64X	93.84
FRASC47	113X	97.19
FRASC49	169X	98.36
FRASC51	184X	99.25
FRASC72	164X	98.80
FRASC75	64X	95.88
FRASC76	106X	98.09
FRASC77	79X	95.95
FRASC78	203X	99.16
FRASC79	120X	97.92

As seen in Table 3.4., FRASC32 had the lowest overall coverage of 50X which is still higher than the recommended 35X. It is important to note that the samples included in the first run had the lowest coverage of between 50 and 64X whilst the samples included in the last run have the highest coverage of between 164 and 203X. The percentage of bases with at least 20 reads is sufficiently high, further indicating the good quality of the sequencing runs.

Upon further analysis of the coverage data using the Agilent SureCall QC reports, it was observed that there were exons that were not optimally covered in each patient. Exons were classified as being not optimally covered if more than 40% of that exon had below 20X coverage (Table 3.5.). The reasons for the decreased coverage and potential remedial action will be discussed in the final chapter. This could be due to an error that occurred while sequencing, or the specifically designed probes could not bind sufficiently to that region of DNA (either due to a high GC content or a homopolymer region). There is an overlap in some of the poorly covered exons between samples, the most commonly observed ones being *NIPBL*: exon 33 in 10 patients, *PDS5A*: exon 4 in all 14 patients, *STAG1*: exon 9 in 5 patients and *PDS5B*: exon 27 in 5 patients.

Table 3.5. Exons with <20X coverage in at least 40% of that exon for each individual patient.

Patient number	**Exons poorly covered**
FRASC8	*PDS5A**: exon 4 and 5
FRASC29	*NIPBL**: exon 33 and 13, *SMC3**: exon 25, *STAG1**: exon 9, *PDS5A**: exon 4, *PDS5B**: exon 14 and 27
FRASC30	*STAG1**: exon 9, *PDS5A**: exon 4
FRASC32	*AFF4**: exon 15 and 6, *NIPBL**: exon 38, 33, 20 and 13, *PDS5B**: exon 27 and 16, *SMC3**: exon 10, 3 and 25, *TAF1**: exon 5, *ESCO2**: exon 4, *STAG1**: exon 9 and 2, *KMT2A**: exon 1, *PDS5A**: exon 4, 29, 5 and 7
FRASC41	*PDS5B**: exon 27, *NIPBL**: exon 38 and 33, *SMC3**: exon 26 and 25, *STAG1**: exon 9, *PDS5A**: exon 29, 4 and 5
FRASC47	*NIPBL**: exon 33, *KMT2A**: exon 1, *PDS5A*:* exon 4
FRASC49	*PDS5A**: exon 4
FRASC51	*NIPBL**: exon 33, *PDS5A**: exon 4
FRASC72	*PDS5A**: exon 4
FRASC75	*NIPBL**: exon 33 and 39, *STAG1**: exon 9, *PDS5A**: exon 4 and 29, *PDS5B**: exon 27
FRASC76	*NIPBL**: exon 33, *PDS5A**: exon 4 and 5, *PDS5B**: exon 27
FRASC77	*NIPBL**: exon 33, *PDS5A**: exon 4, *KMT2A**: exon 1
FRASC78	*NIPBL**: exon 33, *PDS5A**: exon 4
FRASC79	*NIPBL**: exon 33, *PDS5A**: exon 4

**PDS5A*: Cohesin Associated Factor A; *NIPBL*: Nipped B Like; *SMC3*: Structural Maintenance of Chromosome 3; *STAG1*: Stromal antigen 1; *PDS5B*: Cohesin Associated Factor B; *AFF4*: AF4/FMR2 Family Member 4; *TAF1*: TBP associated factor 1; *ESCO2*: Establishment of Sister Chromatid Cohesion N-Acetyltransferase 2; *KMT2A*: histone-lysine N-methyltransferase 2A.

Overall, the quality of the data obtained was of a sufficient standard to proceed to variant analysis.

3.3.<u>Variant analysis</u>

Upon annotating each VCF file generated by the Agilent SureCall software, using the online wANNOVAR tool (for website URL, see list of websites at the end of dissertation), a list of variants was obtained (Appendix C.1., for Appendix C.1. file URL, see list of websites at the end of dissertation – available via Google Drive link). Variants were initially filtered based on population MAF from the gnomAD and 1000 Genomes projects (1000G) data (Lek *et al.*, 2016; The 1000 Genomes Project Consortium, 2015) produced by wANNOVAR with a threshold of <0.005. The threshold recommended by the ACMG guidelines (Richards *et al.*, 2015) is 0.05, with a few exceptions (Ghosh *et al.*, 2018), however considering this is a dominant condition and a large portion of variants passed the 0.05 MAF threshold, the threshold was made more stringent at 0.005. This threshold was applied to the full datasets as well as to the African subset where available. Benign variants/polymorphisms were then further filtered out by examining scores produced by prediction tools.

The top candidates for potential disease-causing variants were narrowed down for each patient (Table 3.6.-3.19.). Not every variant had prediction scores available from each tool used in the analysis as some tools only make predictions for specific types of variants e.g. missense and splice site variants (Liu *et al.*, 2016). Scores predicting the variant to be deleterious are highlighted in red and scores predicting a benign effect are highlighted in green in the following summary tables. Additional filtering was carried out according to the ACMG guidelines (Richards *et al.*, 2015; Appendices D.1. - D.3.).

3.3.1. FRASC8

A total of 45 variants were identified in FRASC8. Out of these 45 variants, three were selected as top candidates for potential disease-causing mutations after MAF filtering had occurred (Table 3.6.).

Table 3.6. Summary of top candidate variants of FRASC8 with results and information from various bioinformatic tools and databases.

	Variant 1	**Variant 2**	**Variant 3**
Gene	*NIPBL*	*HDAC8*	*PDS5B*
Variant	c.6027_6030del GTTC	c.*5 A>T	c.*43_*44insGCT
GnomAD freq ALL	0.000	0.001	0.000
GnomAD freq AFR	0.000	0.002	0.000
1000G freq ALL	0.000	<0.001	0.000
1000G freq AFR	0.000	0.001	0.000
SIFT			1.000
PolyPhen2			
PROVEAN			0.000
MutationTaster	D (Disease causing)		
FATHMM			
CADD	35.000	6.863	5.846
DANN		0.459	
GERP++	4.950	4.690	4.830
dbSNP ID		rs782509754	
ClinVar			
Benign ACMG codes*		BP7, BP2	BP4, BP2
Pathogenic ACMG codes*	PP4, PP3, PVS1, PM2	PM1	
ACMG classification*	Pathogenic	VUS	VUS

Cells highlighted in pink indicate a deleterious prediction, cells highlighted in green indicate a benign prediction.

*See Appendices D.1. – D.3. for detailed descriptions of ACMG codes and classifications

The *HDAC8* c.*5 A>T and *PDS5B* c.*43_*44insGCT variants were both classified as variants of unknown significance. The *HDAC8* c.*5 A>T is a synonymous (silent) variant for which splicing prediction algorithms predict no impact to the splice site nor the creation of a new splice site (ACMG code BP7). Although this variant is seen to occur within a mutation hotspot of *HDAC8* which is found on the X chromosome (ACMG code PM1), it has been seen in 24 apparently healthy individuals (including males) from gnomAD. Additionally, it passes this study's MAF threshold of 0.005 but is still observed in the general population which is highly unlikely for a condition that is predominantly caused by de novo mutations. This variant is therefore unlikely to be pathogenic. The *PDS5B* c.*43_*44insGCT is located at the 3' end of the gene and it does not alter a splice site or have a predicted impact on the protein according to the Mutalyzer website (for website URL, see list of websites at the end of dissertation), even though this nucleotide is conserved according to the GERP++. Both variants have been identified in the presence of a variant that has convincing evidence to suggest pathogenicity and therefore the BP62 ACMG code is applied.

The following ACMG codes could be applied to the *NIPBL* c.6027_6030del GTTC frameshift variant (p.Leu2009PhefsTer6): PM2, PP3, PP4 and PVS1. This variant is therefore classified as a pathogenic variant. This variant is present in the *NIPBL* gene which provides more convincing evidence of pathogenicity as it is the most commonly mutated gene in patients with CdLS. The variant occurs in exon 34 of the *NIPBL* gene (Figure 3.3.) and is predicted to result in a truncated protein by Mutalyzer (Figure 3.4.). The frameshift leads to a considerable portion of the protein being truncated; and from initial speculation, without functional studies being carried out, it is strongly suggestive of the functional loss of one copy of the *NIPBL* gene.

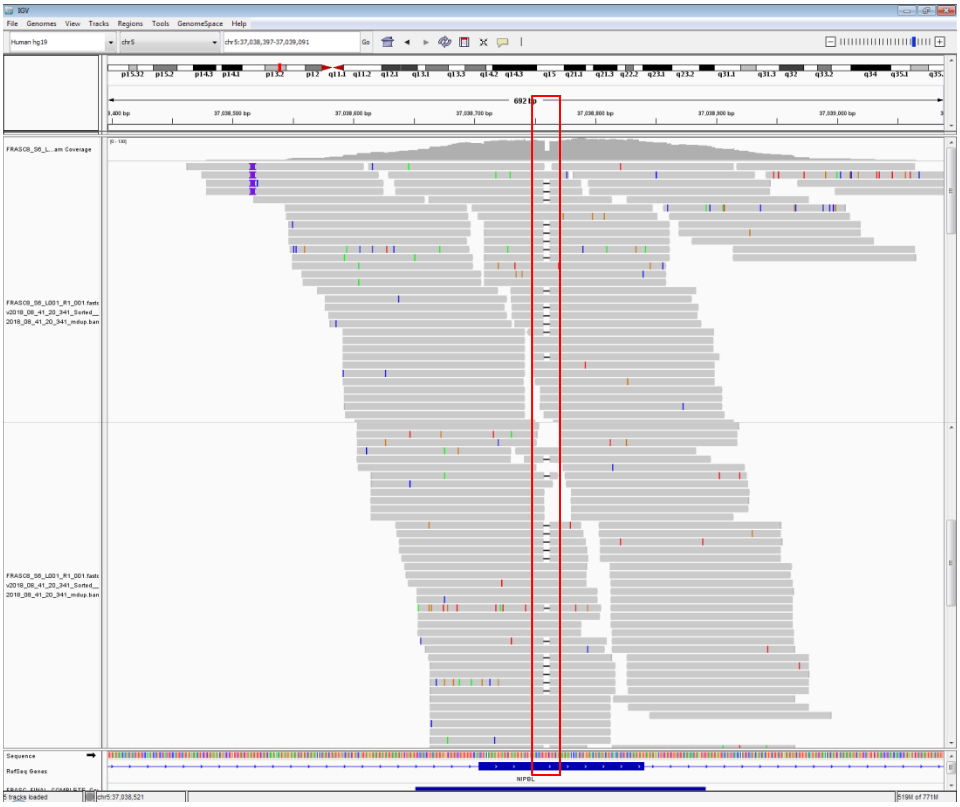
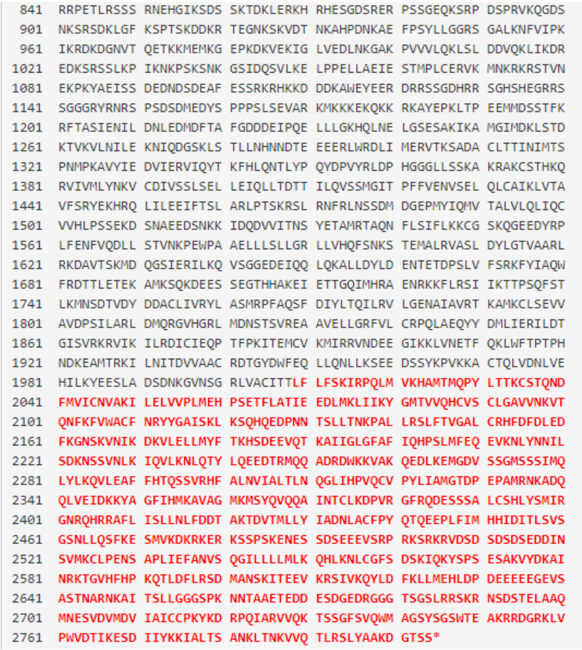
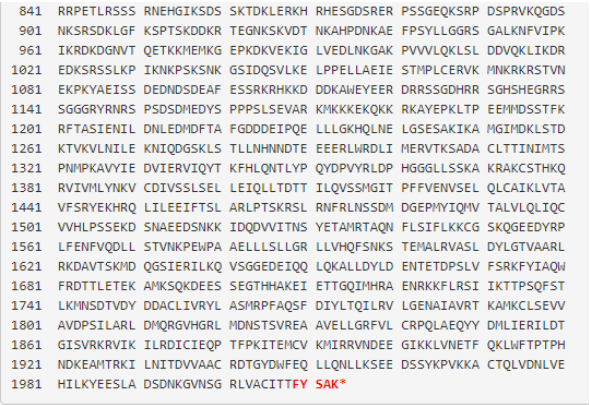

IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
Human hg19
chr5
chr5:37,038,397-37,039,091
Go
p15.32 p15.2 p14.3 p14.1 p13.2 p12 q11.1 q11.2 q12.1 q13.1 q13.3 q14.2 q14.3 q15 q21.1 q21.3 q22.2 q23.1 q23.2 q31.1 q31.3 q32 q33.2 q34 q35.1 q35.
692 bp
,400 bp
37,038,500 bp
37,038,600 bp
37,038,700 bp
37,038,800 bp
37,038,900 bp
37,039,000 bp
FRASC8_S6_L...am Coverage
[0 - 130]
FRASC8_S6_L001_R1_001.fastc
v2018_08_41_20_341_Sorted_
2018_08_41_20_341_mdup.ban
FRASC8_S6_L001_R1_001.fastc
v2018_08_41_20_341_Sorted_
2018_08_41_20_341_mdup.ban
Sequence
RefSeq Genes
NIPBL
5 tracks loaded
chr5:37,038,521
519M of 771M

Figure 3.3. IGV screenshot depicting the 4 bp deletion in exon 34 in the NIPBL gene (outlined in red). It is a heterozygous mutation as it occurs in only 52 of the 104 aligned reads as indicated by the IGV software. There are sufficient reads containing the deletion to call this mutation with confidence.



```
 841  RRPETLRSSS RNEHGIKSDS SKTDKLERKH RHESGDSRER PSSGEQKSRP DSPRVKQGDS
 901  NKSRSDKLGF KSPTSKDDKR TEGNKSKVDT NKAHPDNKAE FPSYLLGGRS GALKNFVIPK
 961  IKRDKDGNVT QETKKMEMKG EPKDKVEKIG LVEDLNKGAK PVVVLQKLSL DDVQKLIKDR
1021  EDKSRSSLKP IKNKPSKSNK GSIDQSVLKE LPPELLAEIE STMPLCERVK MNKRKRSTVN
1081  EKPKYAEISS DEDNDSDEAF ESSRKRHKKD DDKAWEYEER DRRSSGDHRR SGHSHEGRRS
1141  SGGGRYRNRS PSDSDMEDYS PPPSLSEVAR KMKKKEKQKK RKAYEPKLTP EEMMDSSTFK
1201  RFTASIENIL DNLEDMDFTA FGDDDEIPQE LLLGKHQLNE LGSESAKIKA MGIMDKLSTD
1261  KTVKVLNILE KNIQDGSKLS TLLNHNNDTE EEERLWRDLI MERVTKSADA CLTTINIMTS
1321  PNMPKAVYIE DVIERVIQYT KFHLQNTLYP QYDPVYRLDP HGGGLLSSKA KRAKCSTHKQ
1381  RVIVMLYNKV CDIVSSLSEL LEIQLLTDTT ILQVSSMGIT PFFVENVSEL QLCAIKLVTA
1441  VFSRYEKHRQ LILEEIFTSL ARLPTSKRSL RNFRLNSSDM DGEPMYIQMV TALVLQLIQC
1501  VVHLPSSEKD SNAEEDSNKK IDQDVVITNS YETAMRTAQN FLSIFLKKCG SKQGEEDYRP
1561  LFENFVQDLL STVNKPEWPA AELLLSLLGR LLVHQFSNKS TEMALRVASL DYLGTVAARL
1621  RKDAVTSKMD QGSIERILKQ VSGGEDEIQQ LQKALLDYLD ENTETDPSLV FSRKFYIAQW
1681  FRDTTLETEK AMKSQKDEES SEGTHHAKEI ETTGQIMHRA ENRKKFLRSI IKTTPSQFST
1741  LKMNSDTVDY DDACLIVRYL ASMRPFAQSF DIYLTQILRV LGENAIAVRT KAMKCLSEVV
1801  AVDPSILARL DMQRGVHGRL MDNSTSVREA AVELLGRFVL CRPQLAEQYY DMLIERILDT
1861  GISVRKRVIK ILRDICIEQP TFPKITEMCV KMIRRVNDEE GIKKLVNETF QKLWFTPTPH
1921  NDKEAMTRKI LNITDVVAAC RDTGYDWFEQ LLQNLLKSEE DSSYKPVKKA CTQLVDNLVE
1981  HILKYEESLA DSDNKGVNSG RLVACITTLF LFSKIRPQLM VKHAMTMQPY LTTKCSTQND
2041  FMVICNVAKI LELVVPLMEH PSETFLATIE EDLMKLIIKY GMTVVQHCVS CLGAVVNKVT
2101  QNFKFVWACF NRYYGAISKL KSQHQEDPNN TSLLTNKPAL LRSLFTVGAL CRHFDFDLED
2161  FKGNSKVNIK DKVLELLMYF TKHSDEEVQT KAIIGLGFAF IQHPSLMFEQ EVKNLYNNIL
2221  SDKNSSVNLK IQVLKNLQTY LQEEDTRMQQ ADRDWKKVAK QEDLKEMGDV SSGMSSSIMQ
2281  LYLKQVLEAF FHTQSSVRHF ALNVIALTLN QGLIHPVQCV PYLIAMGTDP EPAMRNKADQ
2341  QLVEIDKKYA GFIHMKAVAG MKMSYQVQQA INTCLKDPVR GFRQDESSSA LCSHLYSMIR
2401  GNRQHRRAFL ISLLNLFDDT AKTDVTMLLY IADNLACFPY QTQEEPLFIM HHIDITLSVS
2461  GSNLLQSFKE SMVKDKRKER KSSPSKENES SDSEEEVSRP RKSRKRVDSD SDSDSEDDIN
2521  SVMKCLPENS APLIEFANVS QGILLLLMLK QHLKNLCGFS DSKIQKYSPS ESAKVYDKAI
2581  NRKTGVHFHP KQTLDFLRSD MANSKITEEV KRSIVKQYLD FKLLMEHLDP DEEEEEGEVS
2641  ASTNARNKAI TSLLGGGSPK NNTAAETEDD ESDGEDRGGG TSGSLRRSKR NSDSTELAAQ
2701  MNESVDVMDV IAICCPKYKD RPQIARVVQK TSSGFSVQWM AGSYSGSWTE AKRRDGRKLV
2761  PWVDTIKESD IIYKKIALTS ANKLTNKVVQ TLRSLYAAKD GTSS*
```

a)



```
 841  RRPETLRSSS RNEHGIKSDS SKTDKLERKH RHESGDSRER PSSGEQKSRP DSPRVKQGDS
 901  NKSRSDKLGF KSPTSKDDKR TEGNKSKVDT NKAHPDNKAE FPSYLLGGRS GALKNFVIPK
 961  IKRDKDGNVT QETKKMEMKG EPKDKVEKIG LVEDLNKGAK PVVVLQKLSL DDVQKLIKDR
1021  EDKSRSSLKP IKNKPSKSNK GSIDQSVLKE LPPELLAEIE STMPLCERVK MNKRKRSTVN
1081  EKPKYAEISS DEDNDSDEAF ESSRKRHKKD DDKAWEYEER DRRSSGDHRR SGHSHEGRRS
1141  SGGGRYRNRS PSDSDMEDYS PPPSLSEVAR KMKKKEKQKK RKAYEPKLTP EEMMDSSTFK
1201  RFTASIENIL DNLEDMDFTA FGDDDEIPQE LLLGKHQLNE LGSESAKIKA MGIMDKLSTD
1261  KTVKVLNILE KNIQDGSKLS TLLNHNNDTE EEERLWRDLI MERVTKSADA CLTTINIMTS
1321  PNMPKAVYIE DVIERVIQYT KFHLQNTLYP QYDPVYRLDP HGGGLLSSKA KRAKCSTHKQ
1381  RVIVMLYNKV CDIVSSLSEL LEIQLLTDTT ILQVSSMGIT PFFVENVSEL QLCAIKLVTA
1441  VFSRYEKHRQ LILEEIFTSL ARLPTSKRSL RNFRLNSSDM DGEPMYIQMV TALVLQLIQC
1501  VVHLPSSEKD SNAEEDSNKK IDQDVVITNS YETAMRTAQN FLSIFLKKCG SKQGEEDYRP
1561  LFENFVQDLL STVNKPEWPA AELLLSLLGR LLVHQFSNKS TEMALRVASL DYLGTVAARL
1621  RKDAVTSKMD QGSIERILKQ VSGGEDEIQQ LQKALLDYLD ENTETDPSLV FSRKFYIAQW
1681  FRDTTLETEK AMKSQKDEES SEGTHHAKEI ETTGQIMHRA ENRKKFLRSI IKTTPSQFST
1741  LKMNSDTVDY DDACLIVRYL ASMRPFAQSF DIYLTQILRV LGENAIAVRT KAMKCLSEVV
1801  AVDPSILARL DMQRGVHGRL MDNSTSVREA AVELLGRFVL CRPQLAEQYY DMLIERILDT
1861  GISVRKRVIK ILRDICIEQP TFPKITEMCV KMIRRVNDEE GIKKLVNETF QKLWFTPTPH
1921  NDKEAMTRKI LNITDVVAAC RDTGYDWFEQ LLQNLLKSEE DSSYKPVKKA CTQLVDNLVE
1981  HILKYEESLA DSDNKGVNSG RLVACITTFY SAK*
```

b)

Figure 3.4. Mutalyzer screenshots indicating a part of the truncated NIPBL protein resulting from the 4bp deletion. A) The reference protein is displayed on the left depicting the amino acid sequence. Amino acids shown in red are those that are not present in the truncated protein. B) The resulting amino acid sequence encoded by the mutated NIPBL *gene. The amino acids shown in red is where the protein has been truncated.*

3.3.2. FRASC29

Three candidate variants were selected from the 46 variants identified in FRASC29 after MAF filtering (Table 3.7.).

Table 3.7. Summary of top candidate variants of FRASC29 with results and information from various bioinformatic tools and databases.

	Variant 1	**Variant 2**	**Variant 3**
Gene	*PDS5A*	*ARID1B*	*SMC3*
Variant	c.3086+13_3086+15delGTT	c.3270 C>T	c.1410-4T>G
GnomAD freq ALL	<0.001	<0.001	0.000
GnomAD freq AFR	0.000	<0.001	0.000
1000G freq ALL	0.000	0.001	0.000
1000G freq AFR	0.000	0.002	0.000
SIFT		1	
PolyPhen2			
PROVEAN		0	
MutationTaster		D	
FATHMM			
CADD	4.097	10.950	16.520
DANN		0.740	0.745
GERP++	5.047	5.940	5.250
dbSNP ID	rs559982388	rs111368751	
ClinVar		Likely benign	
Benign ACMG codes		BP6, BP4	
Pathogenic ACMG codes		PM1	PM2, PP3
ACMG classification	VUS	Likely benign	VUS

Cells highlighted in pink indicate a deleterious prediction, cells highlighted in green indicate a benign prediction.
*See Appendices D.1. – D.3. for detailed descriptions of ACMG codes and classifications

The *ARID1B* c.3270 C>T (p.Leu1090=) variant was classified as likely benign, both by ClinVar and according to the ACMG guidelines (Richards *et al.*, 2015). This variant is synonymous and has recently been reported as benign by the Genetic Services Laboratory, University of Chicago (ClinVar submission accession number: SCV000246513).

Both the *PDS5A* c.3086+13_3086+15delGTT and *SMC3* c.1410-4T>G variants have been classified as variants of unknown significance. The *PDS5A* c.3086+13_3086+15delGTT variant occurs within a homopolymer region within an intron, it is therefore likely not to have a deleterious effect as sequencing systems commonly have higher error rates within these regions.

The *SMC3* c.1410-4T>G variant occurs near a splice site and according to the Human Splicing Finder (Desmet *et al.*, 2009), this variant has the potential to affect splicing by either altering the splice site or by creating a new splice site. There is still not enough evidence, however, to predict this variant to be the disease-causing mutation in this patient.

3.3.3. FRASC30

A total of 61 variants were identified in FRASC30, of which six were presented as candidates for potential disease-causing mutations after MAF filtering (Table 3.8.).

Table 3.8. Summary of top candidate variants of FRASC30 with results and information from various bioinformatic tools and databases.

	Variant 1	**Variant 2**	**Variant 3**	**Variant 4**	**Variant 5**	**Variant 6**
Gene	*NIPBL*	*STAG1*	*ARID1B*	*ESCO2*	*PDS5B*	*PDS5B*
Variant	c.2479_2480delAG	c.30-26C>T	c.3270 C>T	c.247 A>G	c.*43_*44insTTTTT	c.*45_*46insT
GnomAD freq ALL	0.000	<0.001	<0.001	0.000	0.000	0.000
GnomAD freq AFR	0.000	<0.001	<0.001	0.000	0.000	0.000
1000G freq ALL	0.000	<0.001	0.001	0.000	0.000	0.000
1000G freq AFR	0.000	<0.001	0.002	0.000	0.000	0.000
SIFT			1	0.173		
PolyPhen2				0.015		
PROVEAN			0	-3.510		
MutationTaster	D		D	D		
FATHMM				0.410		
CADD	29.400	15.120	10.950	<0.001	5.731	0.432
DANN		0.855	0.740	0.783		
GERP++	5.990	5.290	5.940	-1.270	4.830	5.250
dbSNP ID	rs398124465	rs200992196	rs111368751	rs113305862		
ClinVar	Pathogenic		likely benign			
Benign ACMG codes		BP4, BP2	BP6, BP4, BP2	BP4, BP2	BP4, BP2	BP4, BP2
Pathogenic ACMG codes	PP4, PP3, PM2, PVS1, PP5, PM1		PM1	PM2		
ACMG classification	Pathogenic	VUS	Likely benign	VUS	VUS	VUS

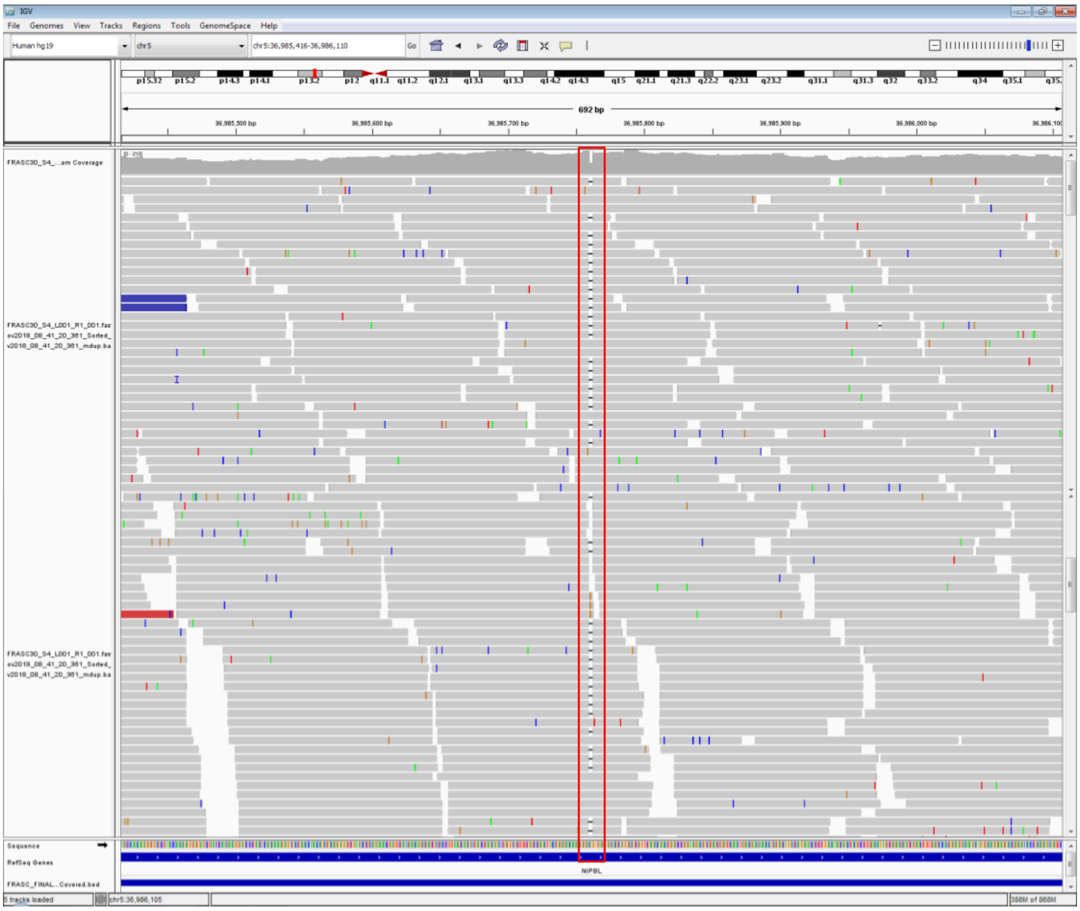
Cells highlighted in pink indicate a deleterious prediction, cells highlighted in green indicate a benign prediction.

*See Appendices D.1. – D.3. for detailed descriptions of ACMG codes and classifications

The *ARID1B* c.3270 C>T (p.Leu1090=) synonymous variant, classified as likely benign by ACMG and ClinVar (ClinVar submission accession number: SCV000246513), was seen in a previous patient – FRASC29. Four of the six candidate variants were classified as being a VUS: *STAG1* c.30-26C>T, *ESCO2* c.247 A>G (p.Thr83Ala), *PDS5B* c.*43_*44insTTTTT and *PDS5B* c.*45_*46insT. *STAG1* c.30-26C>T is an intronic variant that, according to

Human Splicing Finder, is unlikely to affect splicing and therefore not likely to be deleterious. *ESCO2* c.247 A>G is an exonic missense mutation that was predicted to be benign by six out of the eight prediction tools used. This also makes it unlikely to be the disease-causing mutation in this patient. Both *PDS5B* variants occur in the 3' end of the gene and have no apparent effect on splicing despite both being present in a conserved region (according to the GERP++ scores).

All five variants were identified in the presence of a known pathogenic mutation (ACMG code PB2): *NIPBL* c.2479_2480delAG. The *NIPBL* c.2479_2480delAG (p.Arg827GlyfsTer2) variant has been reported in the literature before (Gillis *et al.*, 2004; Yuan *et al.*, 2018) and has already been classified as pathogenic according to ClinVar. This matches the classification according to the ACMG guidelines. The following scores have been assigned to this variant: PP4, PP3, PM2, PVS1, PM1 and PP5, all contributing to the evidence that this is the disease-causing variant in this patient. The variant occurs within a NIPBL hotspot, exon 10 (in alignment with the ACMG code PM1) (Figure 3.5.). The resulting effect on the protein sequence is a truncation of more than half the amino acids (Figure 3.6.). This truncation most likely leads to a loss of function of the protein and, as mentioned previously, this is a common disease mechanism for patients with CdLS.



Human hg19

chr5

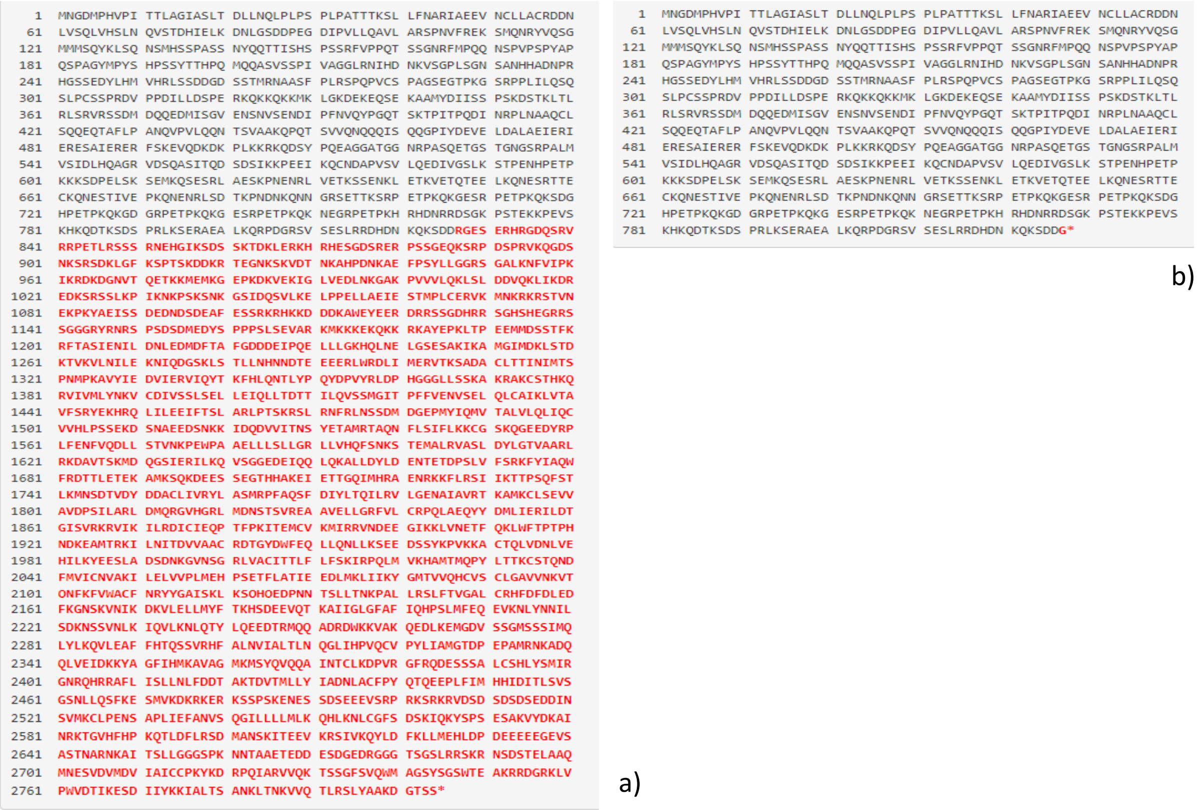
chr5:36,985,416-36,986,110

5 tracks loaded

chr5:36,986,105

388M of 868M

Figure 3.5. IGV screenshot depicting the 2 bp deletion in exon 10 in the NIPBL gene (outlined in red). Exon 10 of the NIPBL gene is a known hotspot for CdLS causative mutations to occur. The deletion is heterozygous and is present in 91 of the 188 reads.



```
   1  MNGDMPHVPI TTLAGIASLT DLLNQLPLPS PLPATTTKSL LFNARIAEEV NCLLACRDDN
  61  LVSQLVHSLN QVSTDHIELK DNLGSDDPEG DIPVLLQAVL ARSPNVFREK SMQNRYVQSG
 121  MMMSQYKLSQ NSMHSSPASS NYQQTTISHS PSSRFVPPQT SSGNRFMPQQ NSPVPSPYAP
 181  QSPAGYMPYS HPSSYTTHPQ MQQASVSSPI VAGGLRNIHD NKVSGPLSGN SANHHADNPR
 241  HGSSEDYLHM VHRLSSDDGD SSTMRNAASF PLRSPQPVCS PAGSEGTPKG SRPPLILQSQ
 301  SLPCSSPRDV PPDILLDSPE RKQKKQKKMK LGKDEKEQSE KAAMYDIISS PSKDSTKLTL
 361  RLSRVRSSDM DQQEDMISGV ENSNVSENDI PFNVQYPGQT SKTPITPQDI NRPLNAAQCL
 421  SQQEQTAFLP ANQVPVLQQN TSVAAKQPQT SVVQNQQQIS QQGPIYDEVE LDALAEIERI
 481  ERESAIERER FSKEVQDKDK PLKKRKQDSY PQEAGGATGG NRPASQETGS TGNGSRPALM
 541  VSIDLHQAGR VDSQASITQD SDSIKKPEEI KQCNDAPVSV LQEDIVGSLK STPENHPETP
 601  KKKSDPELSK SEMKQSESRL AESKPNENRL VETKSSENKL ETKVETQTEE LKQNESRTTE
 661  CKQNESTIVE PKQNENRLSD TKPNDNKQNN GRSETTKSRP ETPKQKGESR PETPKQKSDG
 721  HPETPKQKGD GRPETPKQKG ESRPETPKQK NEGRPETPKH RHDNRRDSGK PSTEKKPEVS
 781  KHKQDTKSDS PRLKSERAEA LKQRPDGRSV SESLRRDHDN KQKSDDRGES ERHRGDQSRV
 841  RRPETLRSSS RNEHGIKSDS SKTDKLERKH RHESGDSRER PSSGEQKSRP DSPRVKQGDS
 901  NKSRSDKLGF KSPTSKDDKR TEGNKSKVDT NKAHPDNKAE FPSYLLGGRS GALKNFVIPK
 961  IKRDKDGNVT QETKKMEMKG EPKDKVEKIG LVEDLNKGAK PVVVLQKLSL DDVQKLIKDR
1021  EDKSRSSLKP IKNKPSKSNK GSIDQSVLKE LPPELLAEIE STMPLCERVK MNKRKRSTVN
1081  EKPKYAEISS DEDNDSDEAF ESSRKRHKKD DDKAWEYEER DRRSSGDHRR SGHSHEGRRS
1141  SGGGRYRNRS PSDSDMEDYS PPPSLSEVAR KMKKKEKQKK RKAYEPKLTP EEMMDSSTFK
1201  RFTASIENIL DNLEDMDFTA FGDDDEIPQE LLLGKHQLNE LGSESAKIKA MGIMDKLSTD
1261  KTVKVLNILE KNIQDGSKLS TLLNHNNDTE EEERLWRDLI MERVTKSADA CLTTINIMTS
1321  PNMPKAVYIE DVIERVIQYT KFHLQNTLYP QYDPVYRLDP HGGGLLSSKA KRAKCSTHKQ
1381  RVIVMLYNKV CDIVSSLSEL LEIQLLTDTT ILQVSSMGIT PFFVENVSEL QLCAIKLVTA
1441  VFSRYEKHRQ LILEEIFTSL ARLPTSKRSL RNFRLNSSDM DGEPMYIQMV TALVLQLIQC
1501  VVHLPSSEKD SNAEEDSNKK IDQDVVITNS YETAMRTAQN FLSIFLKKCG SKQGEEDYRP
1561  LFENFVQDLL STVNKPEWPA AELLLSLLGR LLVHQFSNKS TEMALRVASL DYLGTVAARL
1621  RKDAVTSKMD QGSIERILKQ VSGGEDEIQQ LQKALLDYLD ENTETDPSLV FSRKFYIAQW
1681  FRDTTLETEK AMKSQKDEES SEGTHHAKEI ETTGQIMHRA ENRKKFLRSI IKTTPSQFST
1741  LKMNSDTVDY DDACLIVRYL ASMRPFAQSF DIYLTQILRV LGENAIAVRT KAMKCLSEVV
1801  AVDPSILARL DMQRGVHGRL MDNSTSVREA AVELLGRFVL CRPQLAEQYY DMLIERILDT
1861  GISVRKRVIK ILRDICIEQP TFPKITEMCV KMIRRVNDEE GIKKLVNETF QKLWFTPTPH
1921  NDKEAMTRKI LNITDVVAAC RDTGYDWFEQ LLQNLLKSEE DSSYKPVKKA CTQLVDNLVE
1981  HILKYEESLA DSDNKGVNSG RLVACITTLF LFSKIRPQLM VKHAMTMQPY LTTKCSTQND
2041  FMVICNVAKI LELVVPLMEH PSETFLATIE EDLMKLIIKY GMTVVQHCVS CLGAVVNKVT
2101  ONFKFVWACF NRYYGAISKL KSOHOEDPNN TSLLTNKPAL LRSLFTVGAL CRHFDFDLED
2161  FKGNSKVNIK DKVLELLMYF TKHSDEEVQT KAIIGLGFAF IQHPSLMFEQ EVKNLYNNIL
2221  SDKNSSVNLK IQVLKNLQTY LQEEDTRMQQ ADRDWKKVAK QEDLKEMGDV SSGMSSSIMQ
2281  LYLKQVLEAF FHTQSSVRHF ALNVIALTLN QGLIHPVQCV PYLIAMGTDP EPAMRNKADQ
2341  QLVEIDKKYA GFIHMKAVAG MKMSYQVQQA INTCLKDPVR GFRQDESSSA LCSHLYSMIR
2401  GNRQHRRAFL ISLLNLFDDT AKTDVTMLLY IADNLACFPY QTQEEPLFIM HHIDITLSVS
2461  GSNLLQSFKE SMVKDKRKER KSSPSKENES SDSEEEVSRP RKSRKRVDSD SDSDSEDDIN
2521  SVMKCLPENS APLIEFANVS QGILLLLMLK QHLKNLCGFS DSKIQKYSPS ESAKVYDKAI
2581  NRKTGVHFHP KQTLDFLRSD MANSKITEEV KRSIVKQYLD FKLLMEHLDP DEEEEEGEVS
2641  ASTNARNKAI TSLLGGGSPK NNTAAETEDD ESDGEDRGGG TSGSLRRSKR NSDSTELAAQ
2701  MNESVDVMDV IAICCPKYKD RPQIARVVQK TSSGFSVQWM AGSYSGSWTE AKRRDGRKLV
2761  PWVDTIKESD IIYKKIALTS ANKLTNKVVQ TLRSLYAAKD GTSS*
```

```
   1  MNGDMPHVPI TTLAGIASLT DLLNQLPLPS PLPATTTKSL LFNARIAEEV NCLLACRDDN
  61  LVSQLVHSLN QVSTDHIELK DNLGSDDPEG DIPVLLQAVL ARSPNVFREK SMQNRYVQSG
 121  MMMSQYKLSQ NSMHSSPASS NYQQTTISHS PSSRFVPPQT SSGNRFMPQQ NSPVPSPYAP
 181  QSPAGYMPYS HPSSYTTHPQ MQQASVSSPI VAGGLRNIHD NKVSGPLSGN SANHHADNPR
 241  HGSSEDYLHM VHRLSSDDGD SSTMRNAASF PLRSPQPVCS PAGSEGTPKG SRPPLILQSQ
 301  SLPCSSPRDV PPDILLDSPE RKQKKQKKMK LGKDEKEQSE KAAMYDIISS PSKDSTKLTL
 361  RLSRVRSSDM DQQEDMISGV ENSNVSENDI PFNVQYPGQT SKTPITPQDI NRPLNAAQCL
 421  SQQEQTAFLP ANQVPVLQQN TSVAAKQPQT SVVQNQQQIS QQGPIYDEVE LDALAEIERI
 481  ERESAIERER FSKEVQDKDK PLKKRKQDSY PQEAGGATGG NRPASQETGS TGNGSRPALM
 541  VSIDLHQAGR VDSQASITQD SDSIKKPEEI KQCNDAPVSV LQEDIVGSLK STPENHPETP
 601  KKKSDPELSK SEMKQSESRL AESKPNENRL VETKSSENKL ETKVETQTEE LKQNESRTTE
 661  CKQNESTIVE PKQNENRLSD TKPNDNKQNN GRSETTKSRP ETPKQKGESR PETPKQKSDG
 721  HPETPKQKGD GRPETPKQKG ESRPETPKQK NEGRPETPKH RHDNRRDSGK PSTEKKPEVS
 781  KHKQDTKSDS PRLKSERAEA LKQRPDGRSV SESLRRDHDN KQKSDDG*
```

Figure 3.6. Mutalyzer screen shots indicating the reference and truncated protein sequences. A) This indicates the reference amino acid sequence without any alterations. B) This is the resulting amino acid sequence from the NIPBL c.2479_2480 del AG mutation. More than half the protein sequence has been lost.

3.3.4. FRASC47

FRASC47 produced 57 variants, of which, four were classified as candidate variants after MAF filtering (Table 3.9.).

Table 3.9. Summary of top candidate variants of FRASC47 with results and information from various bioinformatic tools and databases.

	Variant 1	**Variant 2**	**Variant 3**	**Variant 4**
Gene	*STAG1*	*SETD5*	*PDS5B*	*ANKRD11*
Variant	c.17 T>G	c.4317 G>A	c.*43_*44insCA	c.4884 C>T
GnomAD freq ALL	0.000	0.000	0.000	<0.001
GnomAD freq AFR	0.000	0.000	0.000	0.003
1000G freq ALL	0.000	0.000	0.000	<0.001
1000G freq AFR	0.000	0.000	0.000	0.002
SIFT		1.000		0.519
PolyPhen2				
PROVEAN		0.000		0.000
MutationTaster	A	D		D
FATHMM				
CADD	40.000	10.180	5.902	0.106
DANN	0.988	0.663		0.423
GERP++	5.650	5.010	5.510	5.070
dbSNP ID		rs539395910		rs144721281
ClinVar				
Benign ACMG codes		BP7, BP2	BP4, BP2	BP7, BP2
Pathogenic ACMG codes	PP4, PP3, PM2, PVS1	PM2		
ACMG classification	Pathogenic	VUS	VUS	VUS

Cells highlighted in pink indicate a deleterious prediction, cells highlighted in green indicate a benign prediction.

*See Appendices D.1. – D.3. for detailed descriptions of ACMG codes and classifications

PDS5B c.*43_*44insCA, like some of the other PDS5B mutations identified previously, is found in a conserved region at the 3’ end of the gene. However, it has no implication in splicing and can be ruled out as a putative disease-causing mutation. Both *SETD5* c.4317 G>A (p.Thr1439=) and *ANKRD11* c.4884 C>T (p.Asp1628=) variants are synonymous and have been assigned an ACMG classification of VUS. In both cases, four of the eight prediction tools used indicate the variants are benign. These three variants were all assigned the ACMG code: BP2 as a more convincingly pathogenic variant was identified in this patient.

The c.17 T>G (p.Leu6Ter) (p.Leu6Ter) stop-gain variant identified in the *STAG1* gene has four pathogenic ACMG codes assigned to it: PP4, PP3, PM2 and PVS1. PVS1 (null variant where LOF is the main mechanism of disease) is the strongest line of evidence to suggest pathogenicity in the ACMG guidelines and according to a more recent paper (Abou Tayoun *et al.*, 2018), a more accurate interpretation of the PVS1 code could be achieved by looking at the type of LOF mutation, it's exact position and whether it is a true null effect or not. This variant occurs in exon 2 of the *STAG1* gene (Figure 3.7.) and results in a truncation six positions into the amino acid sequence (Figure 3.8.). Mutations in the *STAG1* gene haven't been reported to cause CdLS, instead, it has been reported that *STAG1* mutations lead to a CdLS-like phenotype (Lehalle *et al.*, 2017; Yuan *et al.*, 2018).

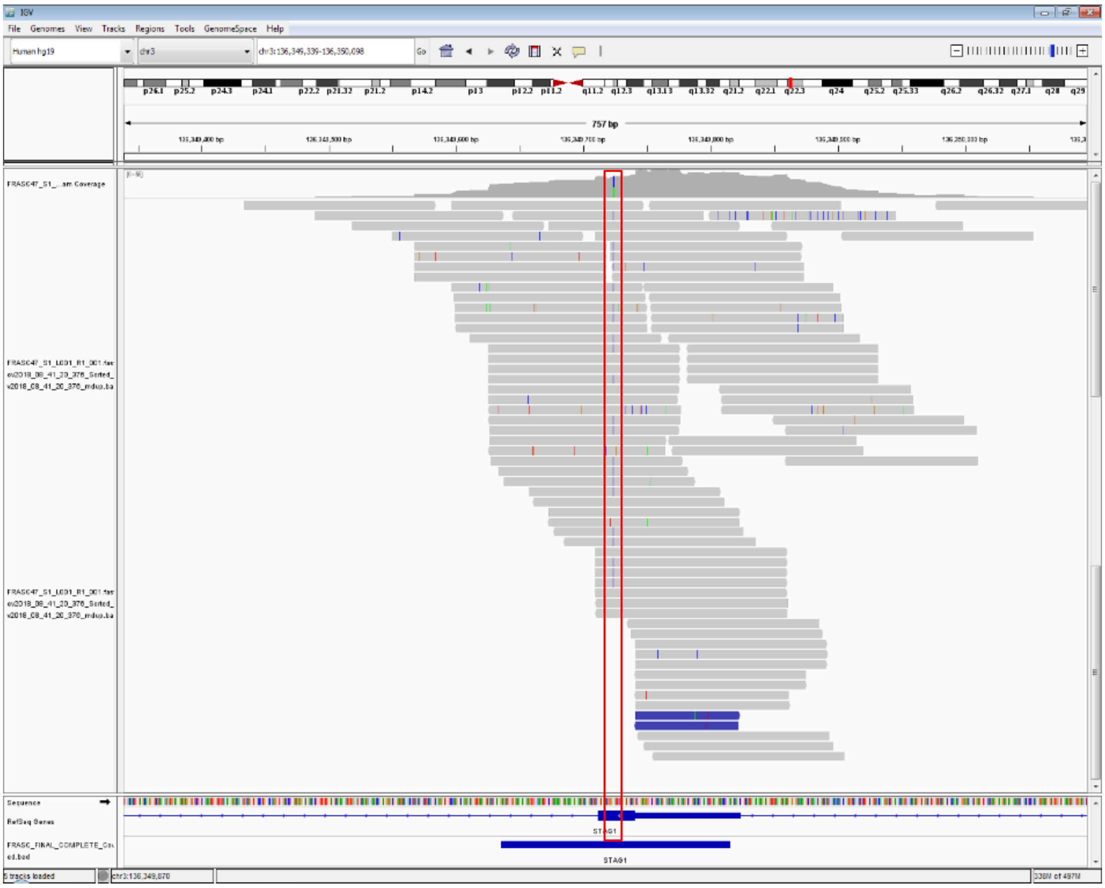
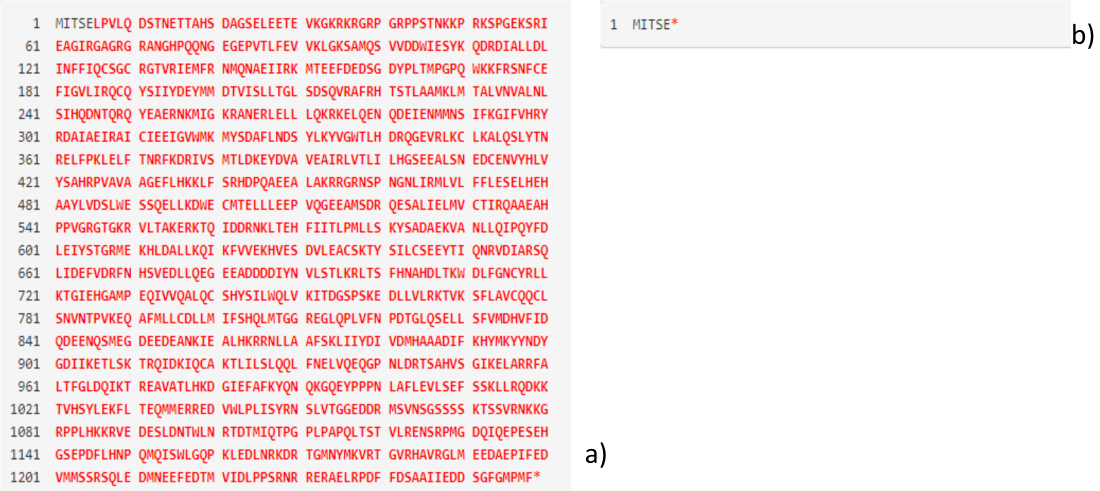

IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
Human hg19
chr3
chr3:136,349,339-136,350,098
Go
p26.1 p25.2 p24.3 p24.1 p22.2 p21.32 p21.2 p14.2 p13 p12.2 p11.2 q11.2 q12.3 q13.13 q13.32 q21.2 q22.1 q22.3 q24 q25.2 q25.33 q26.2 q26.32 q27.1 q28 q29
757 bp
136,349,400 bp
136,349,500 bp
136,349,600 bp
136,349,700 bp
136,349,800 bp
136,349,900 bp
136,350,000 bp
FRASC47_S1_...am Coverage
[0-56]
FRASC47_S1_L001_R1_001.fas ev2018_08_41_20_376_Sorted_ v2018_08_41_20_376_rmdup.ba
FRASC47_S1_L001_R1_001.fas ev2018_08_41_20_376_Sorted_ v2018_08_41_20_376_rmdup.ba
Sequence
RefSeq Genes
STAG1
FRASC_FINAL_COMPLETE_Cov ed.bed
STAG1
5 tracks loaded
chr3:136,349,870
338M of 497M

Figure 3.7. IGV screenshot showing the T>G SNP in the second exon of the STAG1 gene (outlined in red). This is evidence for the first STAG1 putative disease-causing mutation found in a human. The screenshot depicts that this mutation occurs in a heterozygous state.



```
   1  MITSELPVLQ DSTNETTAHS DAGSELEETE VKGKRKRGRP GRPPSTNKKP RKSPGEKSRI
  61  EAGIRGAGRG RANGHPQQNG EGEPVTLFEV VKLGKSAMQS VVDDWIESYK QDRDIALLDL
 121  INFFIQCSGC RGTVRIEMFR NMQNAEIIRK MTEEFDEDSG DYPLTMPGPQ WKKFRSNFCE
 181  FIGVLIRQCQ YSIIYDEYMM DTVISLLTGL SDSQVRAFRH TSTLAAMKLM TALVNVALNL
 241  SIHQDNTQRQ YEAERNKMIG KRANERLELL LQKRKELQEN QDEIENMMNS IFKGIFVHRY
 301  RDAIAEIRAI CIEEIGVWMK MYSDAFLNDS YLKYVGWTLH DRQGEVRLKC LKALQSLYTN
 361  RELFPKLELF TNRFKDRIVS MTLDKEYDVA VEAIRLVTLI LHGSEEALSN EDCENVYHLV
 421  YSAHRPVAVA AGEFLHKKLF SRHDPQAEEA LAKRRGRNSP NGNLIRMLVL FFLESELHEH
 481  AAYLVDSLWE SSQELLKDWE CMTELLLEEP VQGEEAMSDR QESALIELMV CTIRQAAEAH
 541  PPVGRGTGKR VLTAKERKTQ IDDRNKLTEH FIITLPMLLS KYSADAEKVA NLLQIPQYFD
 601  LEIYSTGRME KHLDALLKQI KFVVEKHVES DVLEACSKTY SILCSEEYTI QNRVDIARSQ
 661  LIDEFVDRFN HSVEDLLQEG EEADDDDIYN VLSTLKRLTS FHNAHDLTKW DLFGNCYRLL
 721  KTGIEHGAMP EQIVVQALQC SHYSILWQLV KITDGSPSKE DLLVLRKTVK SFLAVCQQCL
 781  SNVNTPVKEQ AFMLLCDLLM IFSHQLMTGG REGLQPLVFN PDTGLQSELL SFVMDHVFID
 841  QDEENQSMEG DEEDEANKIE ALHKRRNLLA AFSKLIIYDI VDMHAAADIF KHYMKYYNDY
 901  GDIIKETLSK TRQIDKIQCA KTLILSLQQL FNELVQEQGP NLDRTSAHVS GIKELARRFA
 961  LTFGLDQIKT REAVATLHKD GIEFAFKYQN QKGQEYPPPN LAFLEVLSEF SSKLLRQDKK
1021  TVHSYLEKFL TEQMMERRED VWLPLISYRN SLVTGGEDDR MSVNSGSSSS KTSSVRNKKG
1081  RPPLHKKRVE DESLDNTWLN RTDTMIQTPG PLPAPQLTST VLRENSRPMG DQIQEPESEH
1141  GSEPDFLHNP QMQISWLGQP KLEDLNRKDR TGMNYMKVRT GVRHAVRGLM EEDAEPIFED
1201  VMMSSRSQLE DMNEEFEDTM VIDLPPSRNR RERAELRPDF FDSAAIIEDD SGFGMPMF*
```

```
1  MITSE*
```

Figure 3.8. Mutalyzer screenshot depicting a wild type and mutated amino acid sequence of STAG1. A) The STAG1 protein sequence is considerably shorter than the NIPBL protein sequence with only 1259 amino acid residues. B) This is the amino acid sequence produced by the mutated STAG1 gene. The truncation occurs very early on the protein sequence at residue 6.

3.3.5. FRASC75

There were 27 variants identified in FRASC75, however only two presented themselves as candidate variants after filtering based on MAF (Table 3.10.).

Table 3.10. Summary of top candidate variants of FRASC75 with results and information from various bioinformatic tools and databases.

	Variant 1	**Variant 2**
Gene	*NIPBL*	*SCC4*
Variant	c.5639_5642 del CAAC	c.1156-30G>A
GnomAD freq ALL	0.000	0.000
GnomAD freq AFR	0.000	0.000
1000G freq ALL	0.000	<0.001
1000G freq AFR	0.000	0.000
SIFT		
PolyPhen2		
PROVEAN		
MutationTaster	D	
FATHMM		
CADD	35.000	0.073
DANN		0.453
GERP++	5.620	2.000
dbSNP ID		rs199982797
ClinVar		
Benign ACMG codes		BP4, BP2
Pathogenic ACMG codes	PP4, PP3, PM2, PVS1	
ACMG classification	Pathogenic	VUS

Cells highlighted in pink indicate a deleterious prediction, cells highlighted in green indicate a benign prediction.

*See Appendices D.1. – D.3. for detailed descriptions of ACMG codes and classifications

The *SCC4* c.1156-30G>A variant is intronic and occurs 30bp away from an intron-exon boundary and is therefore not likely to be deleterious or affect splicing. It does not occur in a conserved region according to the GERP++ score and is predicted to be benign by both CADD and DANN (Table 17). It also has the BP2 ACMG code assigned to it as a variant that is likely to be disease-causing was identified in the patient.

The variant identified in the *NIPBL* gene, c.5639_5642 del CAAC (p.Pro1880HisfsTer10), is a four bp deletion in exon 30 of the NIPBL gene and results in a frameshift variant (Figure 3.9.). It is predicted to be pathogenic by the ACMG guidelines according to the codes: PP4, PP3, PM2 and PVS1. It results in a truncated protein (Figure 3.10.) and is likely to cause loss of function of the gene and therefore this mutation can be predicted to be the putative disease-causing variant in this patient.

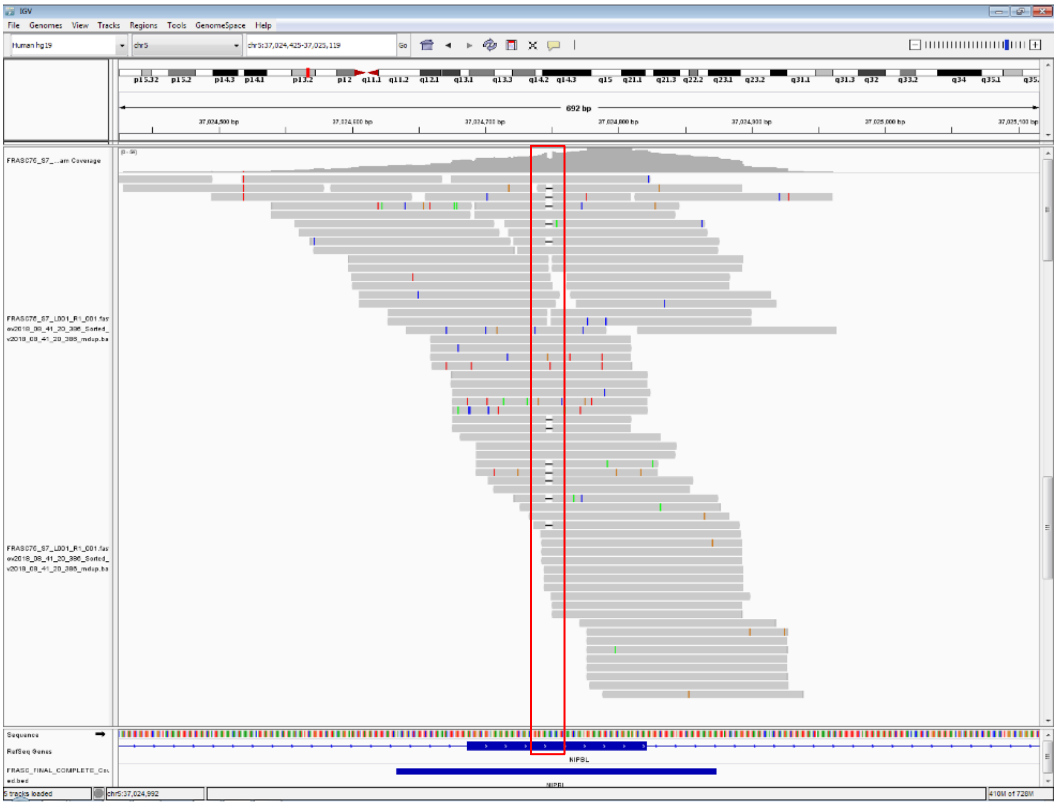
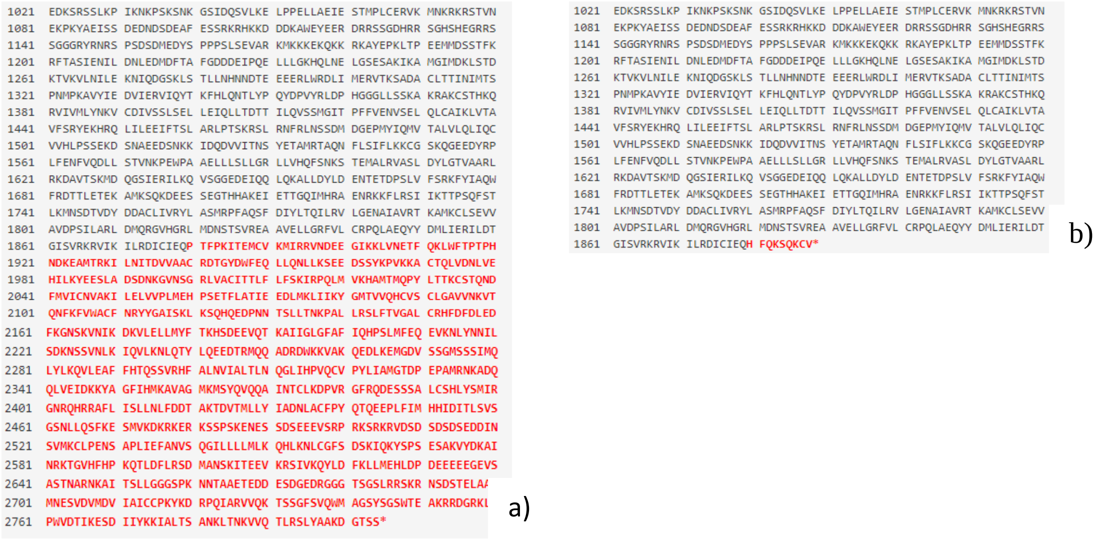

IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
Human hg19
chr5
chr5:37,024,425-37,025,119
Go
p15.32 p15.2 p14.3 p14.1 p13.2 p12 q11.1 q11.2 q12.1 q13.1 q13.3 q14.2 q14.3 q15 q21.1 q21.3 q22.2 q23.1 q23.2 q31.1 q31.3 q32 q33.2 q34 q35.1 q35.
692 bp
37,024,500 bp
37,024,600 bp
37,024,700 bp
37,024,800 bp
37,024,900 bp
37,025,000 bp
37,025,100 bp
[0 - 58]
FRAS076_S7_...am Coverage
FRAS076_S7_L001_R1_001.fastqv2018_08_41_20_386_Sorted_v2018_08_41_20_386_mdup.ba
FRAS076_S7_L001_R1_001.fastqv2018_08_41_20_386_Sorted_v2018_08_41_20_386_mdup.ba
Sequence
RefSeq Genes
NIPBL
FRASC_FINAL_COMPLETE_Cored.bed
NIPBL
5 tracks loaded
chr5:37,024,992
410M of 728M

Figure 3.9. IGV screenshot depicting the 4 bp deletion in exon 30 of the NIPBL gene (outlined in red). The heterozygous deletion occurs in sufficient reads to call the variant with confidence.



```
1021  EDKSRSSLKP IKNKPSKSNK GSIDQSVLKE LPPELLAEIE STMPLCERVK MNKRKRSTVN
1081  EKPKYAEISS DEDNDSDEAF ESSRKRHKKD DDKAWEYEER DRRSSGDHRR SGHSHEGRRS
1141  SGGGRYRNRS PSDSDMEDYS PPPSLSEVAR KMKKKEKQKK RKAYEPKLTP EEMMDSSTFK
1201  RFTASIENIL DNLEDMDFTA FGDDDEIPQE LLLGKHQLNE LGSESAKIKA MGIMDKLSTD
1261  KTVKVLNILE KNIQDGSKLS TLLNHNNDTE EEERLWRDLI MERVTKSADA CLTTINIMTS
1321  PNMPKAVYIE DVIERVIQYT KFHLQNTLYP QYDPVYRLDP HGGGLLSSKA KRAKCSTHKQ
1381  RVIVMLYNKV CDIVSSLSEL LEIQLLTDTT ILQVSSMGIT PFFVENVSEL QLCAIKLVTA
1441  VFSRYEKHRQ LILEEIFTSL ARLPTSKRSL RNFRLNSSDM DGEPMYIQMV TALVLQLIQC
1501  VVHLPSSEKD SNAEEDSNKK IDQDVVITNS YETAMRTAQN FLSIFLKKCG SKQGEEDYRP
1561  LFENFVQDLL STVNKPEWPA AELLLSLLGR LLVHQFSNKS TEMALRVASL DYLGTVAARL
1621  RKDAVTSKMD QGSIERILKQ VSGGEDEIQQ LQKALLDYLD ENTETDPSLV FSRKFYIAQW
1681  FRDTTLETEK AMKSQKDEES SEGTHHAKEI ETTGQIMHRA ENRKKFLRSI IKTTPSQFST
1741  LKMNSDTVDY DDACLIVRYL ASMRPFAQSF DIYLTQILRV LGENAIAVRT KAMKCLSEVV
1801  AVDPSILARL DMQRGVHGRL MDNSTSVREA AVELLGRFVL CRPQLAEQYY DMLIERILDT
1861  GISVRKRVIK ILRDICIEQP TFPKITEMCV KMIRRVNDEE GIKKLVNETF QKLWFTPTPH
1921  NDKEAMTRKI LNITDVVAAC RDTGYDWFEQ LLQNLLKSEE DSSYKPVKKA CTQLVDNLVE
1981  HILKYEESLA DSDNKGVNSG RLVACITTLF LFSKIRPQLM VKHAMTMQPY LTTKCSTQND
2041  FMVICNVAKI LELVVPLMEH PSETFLATIE EDLMKLIIKY GMTVVQHCVS CLGAVVNKVT
2101  QNFKFVWACF NRYYGAISKL KSQHQEDPNN TSLLTNKPAL LRSLFTVGAL CRHFDFDLED
2161  FKGNSKVNIK DKVLELLMYF TKHSDEEVQT KAIIGLGFAF IQHPSLMFEQ EVKNLYNNIL
2221  SDKNSSVNLK IQVLKNLQTY LQEEDTRMQQ ADRDWKKVAK QEDLKEMGDV SSGMSSSIMQ
2281  LYLKQVLEAF FHTQSSVRHF ALNVIALTLN QGLIHPVQCV PYLIAMGTDP EPAMRNKADQ
2341  QLVEIDKKYA GFIHMKAVAG MKMSYQVQQA INTCLKDPVR GFRQDESSSA LCSHLYSMIR
2401  GNRQHRRAFL ISLLNLFDDT AKTDVTMLLY IADNLACFPY QTQEEPLFIM HHIDITLSVS
2461  GSNLLQSFKE SMVKDKRKER KSSPSKENES SDSEEEVSRP RKSRKRVDSD SDSDSEDDIN
2521  SVMKCLPENS APLIEFANVS QGILLLLMLK QHLKNLCGFS DSKIQKYSPS ESAKVYDKAI
2581  NRKTGVHFHP KQTLDFLRSD MANSKITEEV KRSIVKQYLD FKLLMEHLDP DEEEEEGEVS
2641  ASTNARNKAI TSLLGGGSPK NNTAAETEDD ESDGEDRGGG TSGSLRRSKR NSDSTELAA
2701  MNESVDVMDV IAICCPKYKD RPQIARVVQK TSSGFSVQWM AGSYSGSWTE AKRRDGRKL
2761  PWVDTIKESD IIYKKIALTS ANKLTNKVVQ TLRSLYAAKD GTSS*
```

a)

```
1021  EDKSRSSLKP IKNKPSKSNK GSIDQSVLKE LPPELLAEIE STMPLCERVK MNKRKRSTVN
1081  EKPKYAEISS DEDNDSDEAF ESSRKRHKKD DDKAWEYEER DRRSSGDHRR SGHSHEGRRS
1141  SGGGRYRNRS PSDSDMEDYS PPPSLSEVAR KMKKKEKQKK RKAYEPKLTP EEMMDSSTFK
1201  RFTASIENIL DNLEDMDFTA FGDDDEIPQE LLLGKHQLNE LGSESAKIKA MGIMDKLSTD
1261  KTVKVLNILE KNIQDGSKLS TLLNHNNDTE EEERLWRDLI MERVTKSADA CLTTINIMTS
1321  PNMPKAVYIE DVIERVIQYT KFHLQNTLYP QYDPVYRLDP HGGGLLSSKA KRAKCSTHKQ
1381  RVIVMLYNKV CDIVSSLSEL LEIQLLTDTT ILQVSSMGIT PFFVENVSEL QLCAIKLVTA
1441  VFSRYEKHRQ LILEEIFTSL ARLPTSKRSL RNFRLNSSDM DGEPMYIQMV TALVLQLIQC
1501  VVHLPSSEKD SNAEEDSNKK IDQDVVITNS YETAMRTAQN FLSIFLKKCG SKQGEEDYRP
1561  LFENFVQDLL STVNKPEWPA AELLLSLLGR LLVHQFSNKS TEMALRVASL DYLGTVAARL
1621  RKDAVTSKMD QGSIERILKQ VSGGEDEIQQ LQKALLDYLD ENTETDPSLV FSRKFYIAQW
1681  FRDTTLETEK AMKSQKDEES SEGTHHAKEI ETTGQIMHRA ENRKKFLRSI IKTTPSQFST
1741  LKMNSDTVDY DDACLIVRYL ASMRPFAQSF DIYLTQILRV LGENAIAVRT KAMKCLSEVV
1801  AVDPSILARL DMQRGVHGRL MDNSTSVREA AVELLGRFVL CRPQLAEQYY DMLIERILDT
1861  GISVRKRVIK ILRDICIEQH FQKSQKCV*
```

b)

Figure 3.10. Mutalyzer screenshots of the effect the 4bp deletion identified in FASC75 has on the amino acid sequence. Only a portion of the protein sequence is shown here. A) This is the reference amino acid sequence. The amino acids in red are the ones not included in the mutant sequence. B) The resulting mutant amino acid sequence has a truncation of 924 amino acid residues.

3.3.6. FRASC76

A total of 45 variants were identified in FRASC76, four of which passed the MAF filtering (Table 3.11.).

Table 3.11. Summary of top candidate variants of FRASC76 with results and information from various bioinformatic tools and databases.

	Variant 1	**Variant 2**	**Variant 3**	**Variant 4**
Gene	*NIPBL*	*TAF6*	*KMT2A*	*PDS5B*
Variant	c.302_311delCAAGGAGTCC	c.243+30_243+31delTG	c.664A>C	c.*43_*44insACTTT
GnomAD freq ALL	0.000	0.000	0.000	0.000
GnomAD freq AFR	0.000	0.000	0.000	0.000
1000G freq ALL	0.000	0.000	0.000	0.000
1000G freq AFR	0.000	0.000	0.000	0.000
SIFT			1.000	
PolyPhen2				
PROVEAN			0.000	
MutationTaster	D			
FATHMM				
CADD	33.000	4.014	12.160	5.731
DANN			0.589	
GERP++	5.840	4.660	5.590	4.830
dbSNP ID			rs782212726	
ClinVar				
Benign ACMG codes		BP2	BP7, BP4, BP2	BP4, BP2
Pathogenic ACMG codes	PP4, PP3, PM2, PVS1	PM2		
ACMG classification	Pathogenic	VUS	Likely benign	VUS

Cells highlighted in pink indicate a deleterious prediction, cells highlighted in green indicate a benign prediction.

*See Appendices D.1. – D.3. for detailed descriptions of ACMG codes and classifications

Two of these variants have been classified according to the ACMG guidelines as being VUS's. *TAF6* c.243+30_243+31delTG is an intronic variant and is 30bp away from an intron-exon boundary and is therefore unlikely to affect splicing. Similar variants have been found in other patients at the *PDS5B* c.*43_*44insACTTT site. This is because this site is a homopolymer region and is likely to be a highly variable region within the African population. The *KMT2A* c.664A>C (p.Arg222=) variant is a synonymous variant and is not predicted to have any effect according to the ACMG code BP7. It was classified as likely benign according to the ACMG guidelines. All of the above mentioned mutations were assigned the BP2 ACMG code as the mutation identified in the *NIPBL* gene is convincingly pathogenic.

A deletion of 10bp was identified in *NIPBL:* c.302_311delCAAGGAGTCC (p.Ala101ValfsTer18) (Figure 3.11.). This frameshift mutation truncates the amino acid sequence 100 positions into the 2805 amino acid sequence resulting in the loss of function of the NIPBL protein (Figure 3.12.). It was classified as pathogenic based on the ACMG scores PP4, PP3, PM2 and PVS1 (Abou Tayoun *et al.*, 2018). Another deletion that overlaps the current mutation has been reported on the online tool VarSome – *NIPBL* c.310_317delCCTAATGT (for website URL, see list of websites at the end of dissertation). It was classified as pathogenic according to the ACMG codes: PVS1, PM2 and PP5, although no publications have been released detailing additional information about this mutation.

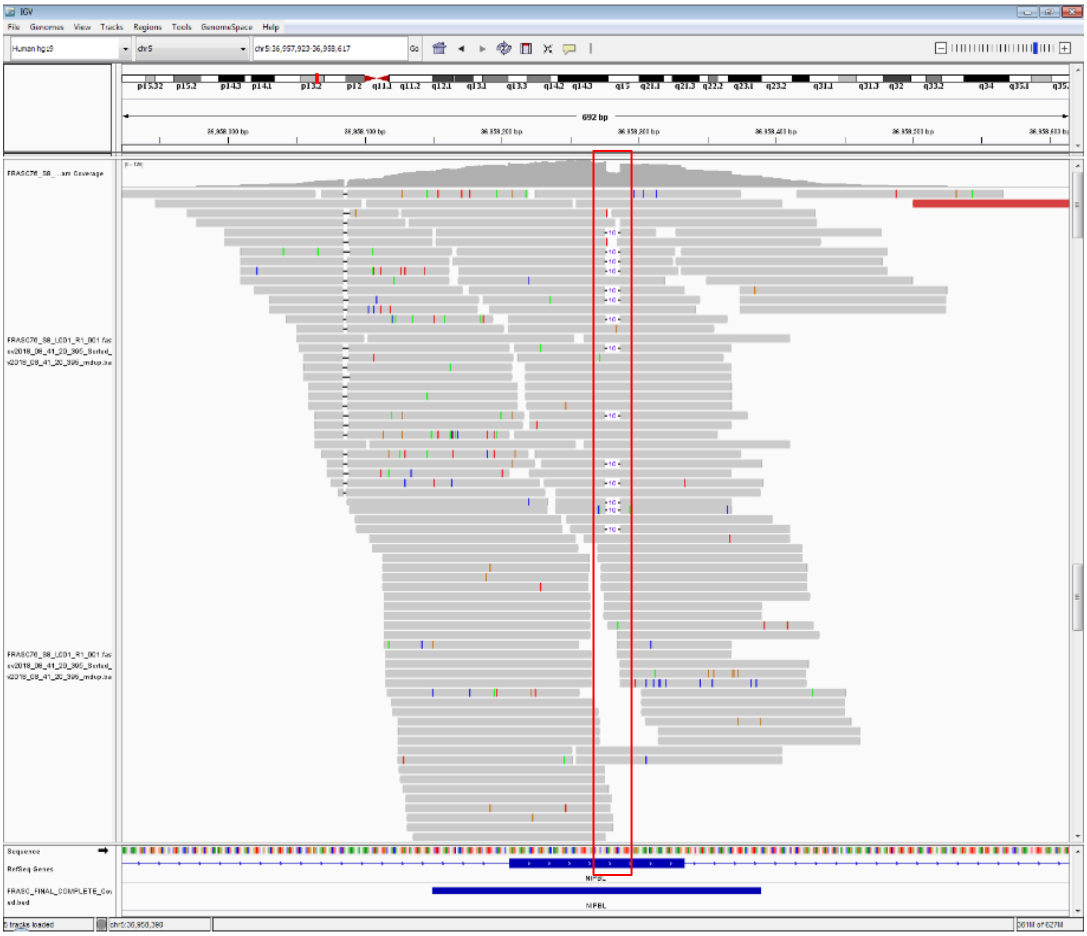
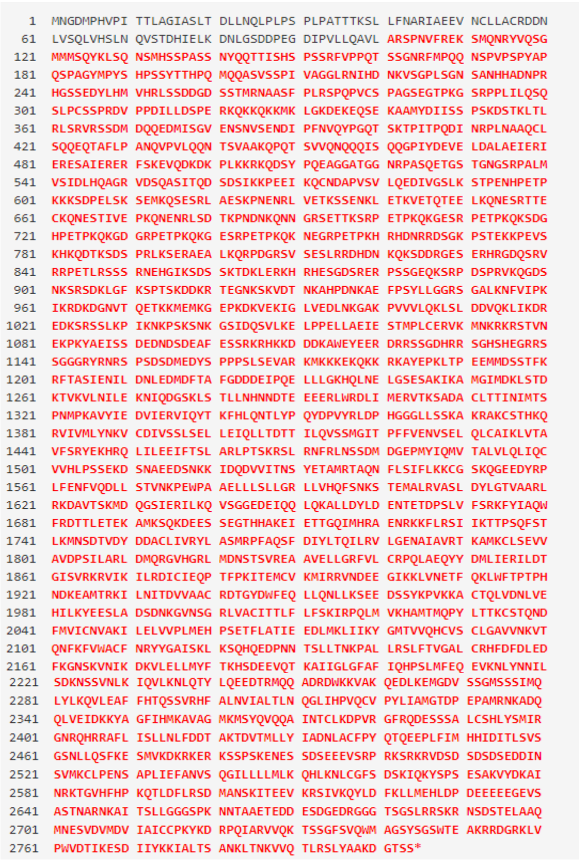
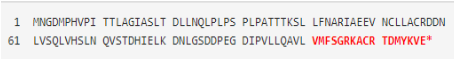

IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
Human hg19
chr5
chr5:36,957,923-36,958,617
692 bp
Sequence
RefSeq Genes
NIPBL
NIPBL

Figure 3.11. IGV screenshot visualising the 10 bp deletion in exon 4 of the NIPBL gene (outlined in red). This mutation occurs in enough of the reads generated by the MiSeq to be called with confidence.



```
   1  MNGDMPHVPI TTLAGIASLT DLLNQLPLPS PLPATTTKSL LFNARIAEEV NCLLACRDDN
  61  LVSQLVHSLN QVSTDHIELK DNLGSDDPEG DIPVLLQAVL ARSPNVFREK SMQNRYVQSG
 121  MMMSQYKLSQ NSMHSSPASS NYQQTTISHS PSSRFVPPQT SSGNRFMPQQ NSPVPSPYAP
 181  QSPAGYMPYS HPSSYTTHPQ MQQASVSSPI VAGGLRNIHD NKVSGPLSGN SANHHADNPR
 241  HGSSEDYLHM VHRLSSDDGD SSTMRNAASF PLRSPQPVCS PAGSEGTPKG SRPPLILQSQ
 301  SLPCSSPRDV PPDILLDSPE RKQKKQKKMK LGKDEKEQSE KAAMYDIISS PSKDSTKLTL
 361  RLSRVRSSDM DQQEDMISGV ENSNVSENDI PFNVQYPGQT SKTPITPQDI NRPLNAAQCL
 421  SQQEQTAFLP ANQVPVLQQN TSVAAKQPQT SVVQNQQQIS QQGPIYDEVE LDALAEIERI
 481  ERESAIERER FSKEVQDKDK PLKKRKQDSY PQEAGGATGG NRPASQETGS TGNGSRPALM
 541  VSIDLHQAGR VDSQASITQD SDSIKKPEEI KQCNDAPVSV LQEDIVGSLK STPENHPETP
 601  KKKSDPELSK SEMKQSESRL AESKPNENRL VETKSSENKL ETKVETQTEE LKQNESRTTE
 661  CKQNESTIVE PKQNENRLSD TKPNDNKQNN GRSETTKSRP ETPKQKGESR PETPKQKSDG
 721  HPETPKQKGD GRPETPKQKG ESRPETPKQK NEGRPETPKH RHDNRRDSGK PSTEKKPEVS
 781  KHKQDTKSDS PRLKSERAEA LKQRPDGRSV SESLRRDHDN KQKSDDRGES ERHRGDQSRV
 841  RRPETLRSSS RNEHGIKSDS SKTDKLERKH RHESGDSRER PSSGEQKSRP DSPRVKQGDS
 901  NKSRSDKLGF KSPTSKDDKR TEGNKSKVDT NKAHPDNKAE FPSYLLGGRS GALKNFVIPK
 961  IKRDKDGNVT QETKKMEMKG EPKDKVEKIG LVEDLNKGAK PVVVLQKLSL DDVQKLIKDR
1021  EDKSRSSLKP IKNKPSKSNK GSIDQSVLKE LPPELLAEIE STMPLCERVK MNKRKRSTVN
1081  EKPKYAEISS DEDNDSDEAF ESSRKRHKKD DDKAWEYEER DRRSSGDHRR SGHSHEGRRS
1141  SGGGRYRNRS PSDSDMEDYS PPPSLSEVAR KMKKKEKQKK RKAYEPKLTP EEMMDSSTFK
1201  RFTASIENIL DNLEDMDFTA FGDDDEIPQE LLLGKHQLNE LGSESAKIKA MGIMDKLSTD
1261  KTVKVLNILE KNIQDGSKLS TLLNHNNDTE EEERLWRDLI MERVTKSADA CLTTINIMTS
1321  PNMPKAVYIE DVIERVIQYT KFHLQNTLYP QYDPVYRLDP HGGGLLSSKA KRAKCSTHKQ
1381  RVIVMLYNKV CDIVSSLSEL LEIQLLTDTT ILQVSSMGIT PFFVENVSEL QLCAIKLVTA
1441  VFSRYEKHRQ LILEEIFTSL ARLPTSKRSL RNFRLNSSDM DGEPMYIQMV TALVLQLIQC
1501  VVHLPSSEKD SNAEEDSNKK IDQDVVITNS YETAMRTAQN FLSIFLKKCG SKQGEEDYRP
1561  LFENFVQDLL STVNKPEWPA AELLLSLLGR LLVHQFSNKS TEMALRVASL DYLGTVAARL
1621  RKDAVTSKMD QGSIERILKQ VSGGEDEIQQ LQKALLDYLD ENTETDPSLV FSRKFYIAQW
1681  FRDTTLETEK AMKSQKDEES SEGTHHAKEI ETTGQIMHRA ENRKKFLRSI IKTTPSQFST
1741  LKMNSDTVDY DDACLIVRYL ASMRPFAQSF DIYLTQILRV LGENAIAVRT KAMKCLSEVV
1801  AVDPSILARL DMQRGVHGRL MDNSTSVREA AVELLGRFVL CRPQLAEQYY DMLIERILDT
1861  GISVRKRVIK ILRDICIEQP TFPKITEMCV KMIRRVNDEE GIKKLVNETF QKLWFTPTPH
1921  NDKEAMTRKI LNITDVVAAC RDTGYDWFEQ LLQNLLKSEE DSSYKPVKKA CTQLVDNLVE
1981  HILKYEESLA DSDNKGVNSG RLVACITTLF LFSKIRPQLM VKHAMTMQPY LTTKCSTQND
2041  FMVICNVAKI LELVVPLMEH PSETFLATIE EDLMKLIIKY GMTVVQHCVS CLGAVVNKVT
2101  QNFKFVWACF NRYYGAISKL KSQHQEDPNN TSLLTNKPAL LRSLFTVGAL CRHFDFDLED
2161  FKGNSKVNIK DKVLELLMYF TKHSDEEVQT KAIIGLGFAF IQHPSLMFEQ EVKNLYNNIL
2221  SDKNSSVNLK IQVLKNLQTY LQEEDTRMQQ ADRDWKKVAK QEDLKEMGDV SSGMSSSIMQ
2281  LYLKQVLEAF FHTQSSVRHF ALNVIALTLN QGLIHPVQCV PYLIAMGTDP EPAMRNKADQ
2341  QLVEIDKKYA GFIHMKAVAG MKMSYQVQQA INTCLKDPVR GFRQDESSSA LCSHLYSMIR
2401  GNRQHRRAFL ISLLNLFDDT AKTDVTMLLY IADNLACFPY QTQEEPLFIM HHIDITLSVS
2461  GSNLLQSFKE SMVKDKRKER KSSPSKENES SDSEEEVSRP RKSRKRVDSD SDSDSEDDIN
2521  SVMKCLPENS APLIEFANVS QGILLLLMLK QHLKNLCGFS DSKIQKYSPS ESAKVYDKAI
2581  NRKTGVHFHP KQTLDFLRSD MANSKITEEV KRSIVKQYLD FKLLMEHLDP DEEEEEGEVS
2641  ASTNARNKAI TSLLGGGSPK NNTAAETEDD ESDGEDRGGG TSGSLRRSKR NSDSTELAAQ
2701  MNESVDVMDV IAICCPKYKD RPQIARVVQK TSSGFSVQWM AGSYSGSWTE AKRRDGRKLV
2761  PWVDTIKESD IIYKKIALTS ANKLTNKVVQ TLRSLYAAKD GTSS*
```

a)



```
 1  MNGDMPHVPI TTLAGIASLT DLLNQLPLPS PLPATTTKSL LFNARIAEEV NCLLACRDDN
61  LVSQLVHSLN QVSTDHIELK DNLGSDDPEG DIPVLLQAVL VMFSGRKACR TDMYKVE*
```

b)

Figure 3.12. Mutalyzer screenshots depicting the reference protein sequence and the mutated protein sequence in FRASC 76. A) The entire reference protein sequence is shown, the amino acids in red are the ones that have been truncated as a result of the 10bp deletion identified in this patient. B) The amino acid sequence as a result of the 10 bp deletion. It occurs in exon 4 and therefore has an effect on the vast majority of the protein.

3.3.7. FRASC77

There were five candidate variants out of 49 total variants identified in FRASC77 after MAF filtering (Table 3.12.).

Table 3.12. Summary of top candidate variants of FRASC77 with results and information from various bioinformatic tools and databases.

	Variant 1	**Variant 2**	**Variant 3**	**Variant 4**	**Variant 5**
Gene	*ARID1B*	*PDS5B*	*PDS5B*	*ANKRD11*	*SMC1A*
Variant	c.5606 C>G	c.*43_*44insTTT	c.*45_*46insT	c.7134C>T	c.1545+4A>C
GnomAD freq ALL	<0.001	0.000	0.000	0.000	<0.001
GnomAD freq AFR	<0.001	0.000	0.000	0.000	0.003
1000G freq ALL	0.000	0.000	0.000	0.000	0.001
1000G freq AFR	0.000	0.000	0.00000	0.00000	0.004
SIFT	0.086			1.000	
PolyPhen2	0.006				
PROVEAN	-1.430			0.000	
MutationTaster	D			D	
FATHMM	4.670				
CADD	21.000	5.846	0.432	0.648	19.560
DANN	0.540			0.740	0.928
GERP++	2.040	4.830	5.250	4.860	4.880
dbSNP ID	rs113818462	rs373003953			rs377270943
ClinVar					VUS/benign
Benign ACMG codes	BP1, BP4	BP4	BP4	BP7	BS2
Pathogenic ACMG codes				PM1	PP3
ACMG classification	Likely benign	VUS	VUS	VUS	VUS

Cells highlighted in pink indicate a deleterious prediction, cells highlighted in green indicate a benign prediction.

*See Appendices D.1. – D.3. for detailed descriptions of ACMG codes and classifications

None of these candidates seem to provide convincing evidence of being the putative disease-causing variant in this patient. The *ARID1B* c.5606 C>G (p.Ser1869Cys) missense variant was predicted to be likely benign according to the ACMG guidelines as missense variants are not the typical disease mechanism in this gene.

The two *PDS5B* mutations: c.*43_*44insTTT and c.*45_*46insT both occur in an intronic homopolymer region indicating they may due to an error in sequencing. Similar mutations in this region have been identified in several other patients in this cohort and they are therefore unlikely to be pathogenic.

The *ANKRD11* c.7134C>T (p.Asp2378=) variant is a synonymous variant as per the ACMG code, BP7. However, it does fall within an exonic splicing enhancer site and may alter splicing within this gene according to Human Splicing Finder. There is little other evidence suggesting this variant is pathogenic.

The *SMC1A* c.1545+4A>C variant is the only variant occurring in a known CdLS causal gene. It occurs near a splice site and the online Human Splicing Finder tool predicts this variant to affect the splice site. However, this variant has been identified in healthy individuals before and is predicted to be benign by three out of four ClinVar entries, therefore this *SMC1A* mutation can be excluded as the putative disease-causing mutation in this patient.

3.3.8. FRASC79

A total of 65 variants were identified in FRASC79. Of these variants, there were four candidate variants that passed the MAF filtering (Table 3.13.).

Table 3.13. Summary of top candidate variants of FRASC79 with results and information from various bioinformatic tools and databases.

	Variant 1	**Variant 2**	**Variant 3**	**Variant 4**
Gene	*NIPBL*	*NIPBL*	*PDS5B*	*HDAC8*
Variant	c.7831dupA	c.1078 T>C	c.*43_*44insTTTTT	c.*5A>T
GnomAD freq ALL	0.000	0.000	0.000	0.001
GnomAD freq AFR	0.000	0.000	0.000	0.002
1000G freq ALL	0.000	0.000	0.000	<0.001
1000G freq AFR	0.000	0.000	0.000	0.001
SIFT				1.000
PolyPhen2				
PROVEAN	0.000			0.000
MutationTaster	D	D		
FATHMM				
CADD	35.000	12.360	5.731	6.863
DANN		0.783		0.459
GERP++	4.890	5.240	4.830	4.690
dbSNP ID				rs782509754
ClinVar				
Benign ACMG codes		BP2	BP4, BP2	BP7, BP2
Pathogenic ACMG codes	PP4, PP3, PM2, PVS1	PM2		PM1
ACMG classification	Pathogenic	VUS	VUS	VUS

Cells highlighted in pink indicate a deleterious prediction, cells highlighted in green indicate a benign prediction.

*See Appendices D.1. – D.3. for detailed descriptions of ACMG codes and classifications

One of two variants identified in the *NIPBL* gene: c.1078 T>C (p.Leu360=) is a synonymous variant and is predicted to have no effect on splicing according to Human Splicing Finder and therefore unlikely to be deleterious. Another *PDS5B* variant was observed at the site *43_*44, this variant is a five bp insertion. Given the number of times mutations have been identified in this region, it is very unlikely to be pathogenic. The *HDAC8* c.*5A>T variant has been identified in FRASC8 and is a synonymous variant which is not predicted to alter splicing by either changing the sequence or creating a new splice site.

The other variant identified in *NIPBL*: c.7831dupA (p.Arg2612LysfsTer20) is a single bp duplication occurring in exon 45 (Figure 3.13.). This results in a truncation of 193 amino acids at the end of the protein sequence (Figure 3.14.). It is predicted to be pathogenic according to the ACMG guidelines codes: PP4, PP3, PM2 and PVS1. There was, however, a prediction tool implying this mutation has a benign effect: Provean (score = 0). This was the first putative disease-causing mutation that had evidence against pathogenicity, it also truncates a much smaller portion of the amino acid sequence than previous mutations identified in this study and the exact effect of this truncation is not known.

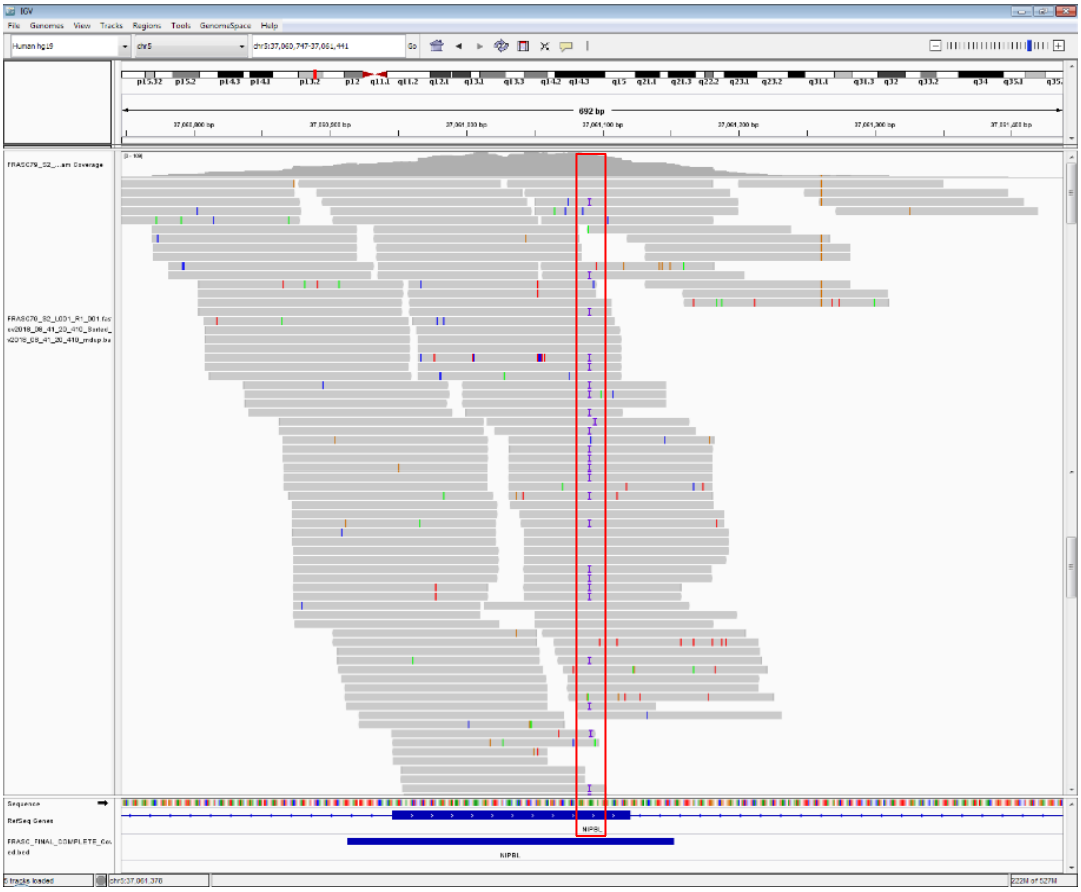
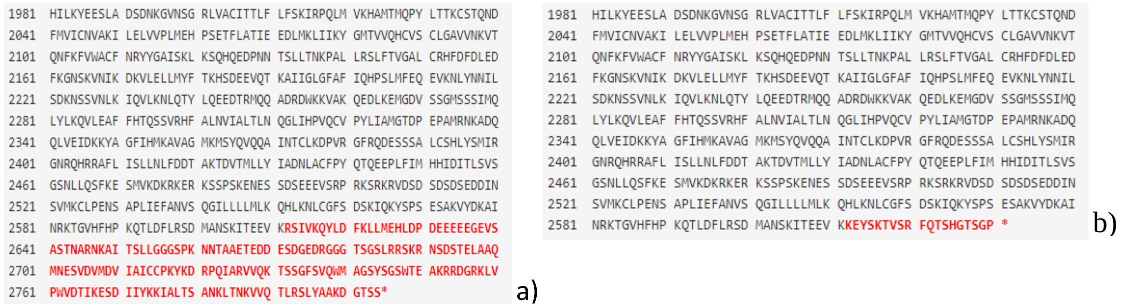

IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
Human hg19
chr5
chr5:37,060,747-37,061,441
692 bp
Sequence
RefSeq Genes
NIPBL

Figure 3.13. IGV screenshot depicting the insertion of a single 'A' nucleotide in exon 45 of the NIPBL gene (outlined in red). The insertion occurs in 40 of the 98 reads and can therefore be called with confidence.



```
1981  HILKYEESLA DSDNKGVNSG RLVACITTLF LFSKIRPQLM VKHAMTMQPY LTTKCSTQND
2041  FMVICNVAKI LELVVPLMEH PSETFLATIE EDLMKLIIKY GMTVVQHCVS CLGAVVNKVT
2101  QNFKFVWACF NRYYGAISKL KSQHQEDPNN TSLLTNKPAL LRSLFTVGAL CRHFDFDLED
2161  FKGNSKVNIK DKVLELLMYF TKHSDEEVQT KAIIGLGFAF IQHPSLMFEQ EVKNLYNNIL
2221  SDKNSSVNLK IQVLKNLQTY LQEEDTRMQQ ADRDWKKVAK QEDLKEMGDV SSGMSSSIMQ
2281  LYLKQVLEAF FHTQSSVRHF ALNVIALTLN QGLIHPVQCV PYLIAMGTDP EPAMRNKADQ
2341  QLVEIDKKYA GFIHMKAVAG MKMSYQVQQA INTCLKDPVR GFRQDESSSA LCSHLYSMIR
2401  GNRQHRRAFL ISLLNLFDDT AKTDVTMLLY IADNLACFPY QTQEEPLFIM HHIDITLSVS
2461  GSNLLQSFKE SMVKDKRKER KSSPSKENES SDSEEEVSRP RKSRKRVDSD SDSDSEDDIN
2521  SVMKCLPENS APLIEFANVS QGILLLLMLK QHLKNLCGFS DSKIQKYSPS ESAKVYDKAI
2581  NRKTGVHFHP KQTLDFLRSD MANSKITEEV KRSIVKQYLD FKLLMEHLDP DEEEEEGEVS
2641  ASTNARNKAI TSLLGGGSPK NNTAAETEDD ESDGEDRGGG TSGSLRRSKR NSDSTELAAQ
2701  MNESVDVMDV IAICCPKYKD RPQIARVVQK TSSGFSVQWM AGSYSGSWTE AKRRDGRKLV
2761  PWVDTIKESD IIYKKIALTS ANKLTNKVVQ TLRSLYAAKD GTSS*
```

```
1981  HILKYEESLA DSDNKGVNSG RLVACITTLF LFSKIRPQLM VKHAMTMQPY LTTKCSTQND
2041  FMVICNVAKI LELVVPLMEH PSETFLATIE EDLMKLIIKY GMTVVQHCVS CLGAVVNKVT
2101  QNFKFVWACF NRYYGAISKL KSQHQEDPNN TSLLTNKPAL LRSLFTVGAL CRHFDFDLED
2161  FKGNSKVNIK DKVLELLMYF TKHSDEEVQT KAIIGLGFAF IQHPSLMFEQ EVKNLYNNIL
2221  SDKNSSVNLK IQVLKNLQTY LQEEDTRMQQ ADRDWKKVAK QEDLKEMGDV SSGMSSSIMQ
2281  LYLKQVLEAF FHTQSSVRHF ALNVIALTLN QGLIHPVQCV PYLIAMGTDP EPAMRNKADQ
2341  QLVEIDKKYA GFIHMKAVAG MKMSYQVQQA INTCLKDPVR GFRQDESSSA LCSHLYSMIR
2401  GNRQHRRAFL ISLLNLFDDT AKTDVTMLLY IADNLACFPY QTQEEPLFIM HHIDITLSVS
2461  GSNLLQSFKE SMVKDKRKER KSSPSKENES SDSEEEVSRP RKSRKRVDSD SDSDSEDDIN
2521  SVMKCLPENS APLIEFANVS QGILLLLMLK QHLKNLCGFS DSKIQKYSPS ESAKVYDKAI
2581  NRKTGVHFHP KQTLDFLRSD MANSKITEEV KKEYSKTVSR FQTSHGTSGP *
```

Figure 3.14. Mutalyzer screenshots depicting a portion of the NIPBL protein sequence. A) A portion of the reference protein sequence is shown with the amino acid residues to be truncated indicated in red. B) The NIPBL protein sequence that is generated as a result of the single nucleotide insertion. A portion of the protein sequence is truncated towards the end of the protein.

3.3.9. FRASC49

A total of 65 variants were identified in FRASC79. Of these variants, there were five candidate variants that passed the MAF filtering (Table 3.14.).

Table 3.14. Summary of top candidate variants of FRASC49 with results and information from various bioinformatic tools and databases.

	Variant 1	**Variant 2**	**Variant 3**	**Variant 4**	**Variant 5**
Gene	*ARID1B*	*TAF6*	*TAF6*	*PDS5B*	*PDS5B*
Variant	c.339_340ins CAGCAGCA GCAGCAGC AGCAGCAA	c.354+30_354+ 31delTG	c.354+19delC	c.*43_*44ins ATTTTTTTT	c.*45_*46insT
GnomAD freq ALL	0.001	0.000	0.000	0.000	0.000
GnomAD freq AFR	0.001	0.000	0.000	0.000	0.000
1000G freq ALL	0.000	0.000	0.000	0.000	0.000
1000G freq AFR	0.000	0.000	0.000	0.000	0.000
SIFT					
PolyPhen2					
PROVEAN	-0.250				
MutationTaster	D				
FATHMM					
CADD	20.500	4.014	1.862	5.497	0.432
DANN					
GERP++	2.480	4.660	5.280	4.830	5.250
dbSNP ID	rs770869529		rs370164610		rs368080614
ClinVar	Likely benign				
Benign ACMG codes	BP3, BP6		BP4	BP4	BP4
Pathogenic ACMG codes	PP3	PM2	PM2		
ACMG classification	Likely Benign	VUS	VUS	VUS	VUS

Cells highlighted in pink indicate a deleterious prediction, cells highlighted in green indicate a benign prediction.

*See Appendices D.1. – D.3. for detailed descriptions of ACMG codes and classifications

The *ARID1B* c.339_340insCAGCAGCAGCAGCAGCAGCAGCAA (p.Gln124_Gln131dup) variant is an in-frame repeat variant and was classified as likely benign by both ClinVar and according to ACMG guidelines. The other four candidate variants were classified as VUS's. The *TAF6* c.354+30_354+31delTG intronic variant was also identified in FRASC76 and was determined to not be pathogenic as it occurs too far away from a splice site to have any deleterious effect. The other mutation identified in *TAF6* was a one bp deletion: c.354+19delC. It is unlikely pathogenic due to the lack of convincing evidence as well as it occurring 19bp from an intron-exon boundary. Again, two mutations were identified in the homopolymer region of PDS5B (c.*43_*44insATTTTTTTT and c.*45_*46insT). These are both unlikely to be the causative mutation in this patient.

3.3.10. FRASC72

A total of 58 variants were identified in FRASC72 of which five were identified to be candidate variants after MAF filtering (Table 3.15.).

Table 3.15. Summary of top candidate variants of FRASC72 with results and information from various bioinformatic tools and databases.

	Variant 1	**Variant 2**	**Variant 3**	**Variant 4**	**Variant 5**
Gene	*NIPBL*	*TAF6*	*TAF6*	*PDS5B*	*PDS5B*
Variant	c.6955-2A>C	c.354+30_354+31delTG	c.354+19delC	c.*43_*44insACT	c.*45_*46insT
GnomAD freq ALL	0.000	0.000	0.000	0.000	0.000
GnomAD freq AFR	0.000	0.000	0.000	0.000	0.000
1000G freq ALL	0.000	0.000	0.000	0.000	0.000
1000G freq AFR	0.000	0.000	0.000	0.000	0.000
SIFT					
PolyPhen2					
PROVEAN					
MutationTaster	D				
FATHMM					
CADD	33.000	4.014	1.862	5.846	0.432
DANN	0.995				
GERP++	5.690	4.660	5.280	4.830	5.250
dbSNP ID			rs370164610		rs368080614
ClinVar					
Benign ACMG codes		BP2	BP4	BP4	BP4
Pathogenic ACMG codes	PP4, PP3, PM2, PVS1	PM2	PM2		
ACMG classification	Pathogenic	VUS	VUS	VUS	VUS

Cells highlighted in pink indicate a deleterious prediction, cells highlighted in green indicate a benign prediction.

*See Appendices D.1. – D.3. for detailed descriptions of ACMG codes and classifications

Four of the variants were classified according to the ACMG guidelines as having a VUS status. The two variants identified in the *TAF6* gene (c.354+30_354+31delTG and c.354+19delC) have been identified previously in other patients in this cohort (FRASC76 and

FRASC49 for the c.354+30_354+31delTG variant and FRASC76 for the c.354+19delC variant). Both were determined not to be pathogenic as they both occur within an intron sufficiently far away from a splice site to not have any negative effect. The two variants identified in the *PDS5B* gene again occur in the highly variable homopolymer region commonly mutated in many previous patients suggesting that these are not pathogenic.

A SNV (c.6955-2A>C) was identified in the *NIPBL* gene of FRASC72 and classified to be pathogenic using the ACMG guidelines. This mutation occurs at a splice site just before exon 41 (Figure 3.15.) and thus has no effect on the amino acid sequence but could play a role in the splicing of the protein product. According to the online tool Human Splicing Finder, this mutation alters the wild type acceptor site for splicing and would therefore have a likely deleterious effect on splicing at this position (Desmet *et al.,* 2009). The ACMG codes assigned to this variant include PP4, PP3, PM2 and PVS1. This is an essential splice site and therefore the PVS1 ACMG code can be applied with stronger evidence for pathogenicity compared with other applications of PVS1 (Abou Tayoun *et al.*, 2018).

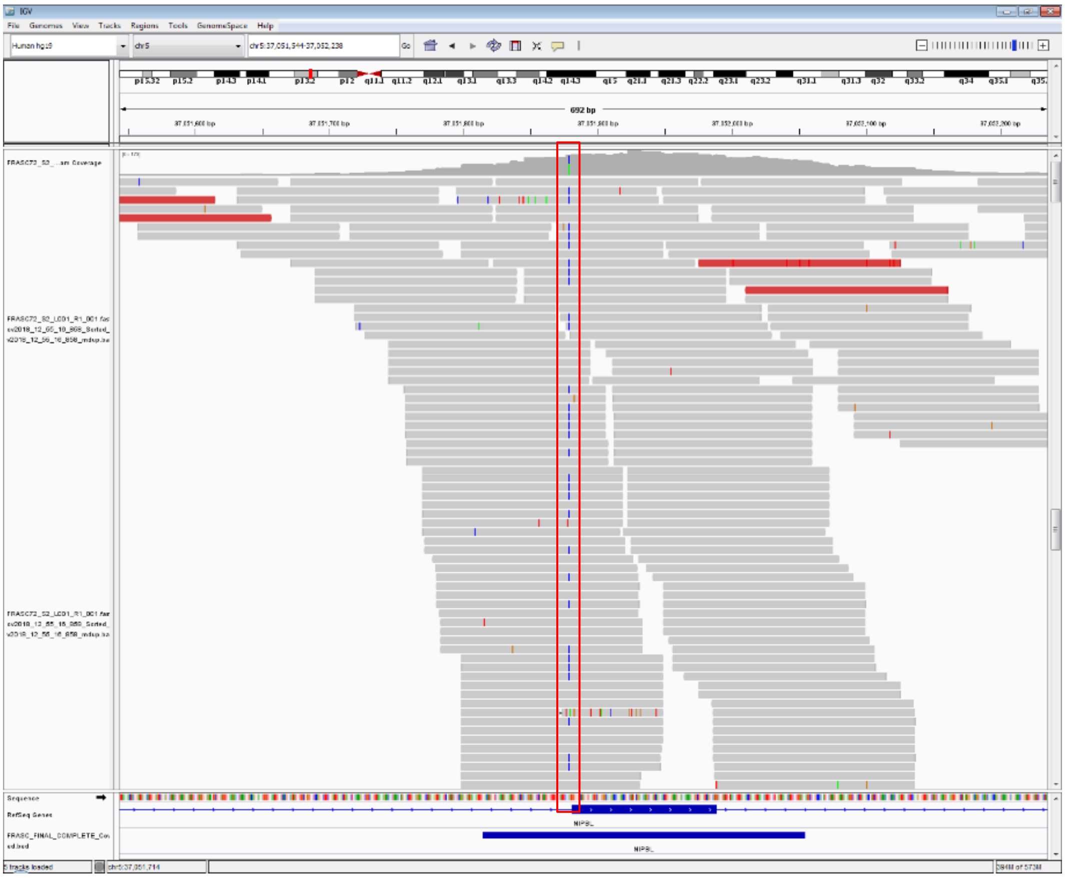

IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
Human hg19
chr5
chr5:37,051,544-37,052,238
Go
p15.32 p15.2 p14.3 p14.1 p13.2 p12 q11.1 q11.2 q12.1 q13.1 q13.3 q14.2 q14.3 q15 q21.1 q21.3 q22.2 q23.1 q23.2 q31.1 q31.3 q32 q33.2 q34 q35.1 q35.
692 bp
37,051,600 bp
37,051,700 bp
37,051,800 bp
37,051,900 bp
37,052,000 bp
37,052,100 bp
37,052,200 bp
FRASC72_S2_...am Coverage
Sequence
RefSeq Genes
NIPBL
FRASC_FINAL_COMPLETE_Co
ed.bed
NIPBL
5 tracks loaded
chr5:37,051,714

Figure 3.15. IGV screenshot depicting the SNP occurring at a splice site just before exon 41 of the NIPBL gene (outlined in red). This mutation would not necessarily affect the resulting amino acid sequence but may affect the splicing of the cDNA post transcription. The mutant 'A' allele occurs within 60 of the 136 reads generated.

3.3.11. FRASC78

FRASC 78 had 42 variants identified and only three of these passed the MAF filtering step (Table 3.16.).

Table 3.16. Summary of top candidate variants of FRASC78 with results and information from various bioinformatic tools and databases.

	Variant 1	**Variant 2**	**Variant 3**
Gene	*SETD5*	*AFF4*	*PDS5B*
Variant	c.2346+15_2346+16insGT	c.1389+21T>G	c.*43_*44ins ACTTTT
GnomAD freq ALL	<0.001	0.001	0.000
GnomAD freq AFR	0.001	0.000	0.000
1000G freq ALL	0.001	0.000	0.000
1000G freq AFR	0.003	0.000	0.000
SIFT			
PolyPhen2			
PROVEAN			
MutationTaster			
FATHMM			
CADD	7.054	5.115	5.673
DANN		0.551	
GERP++	5.320	4.380	4.830
dbSNP ID	rs559066423	rs1480073724	
ClinVar			
Benign ACMG codes	BP4	BP4	BP4
Pathogenic ACMG codes	PM2		
ACMG classification	VUS	VUS	VUS

Cells highlighted in pink indicate a deleterious prediction, cells highlighted in green indicate a benign prediction.

*See Appendices D.1. – D.3. for detailed descriptions of ACMG codes and classifications

The *SETD5* c.2346+15_2346+16insGT variant occurs within an intron and is not predicted to have a deleterious effect according to the online program Human Splicing Finder. The *AFF4* c.1389+21T>G variant also occurs within an intron and, when run through the same online splicing tool as above, is not predicted to have a deleterious effect on splicing. Both variants occur far enough away from an intron-exon boundary to be benign variants. As per majority of the patients so far, a variant was detected in the 3' UTR homopolymer region in *PDS5B* (c.*43_*44ins ACTTTT) which does not result in a pathogenic effect. Although no potentially pathogenic mutation was identified in this patient, the coverage was sufficient and no Sanger sequencing was required for further investigation.

3.3.12. FRASC51

A total of 43 variants were identified in FRASC51 and five of these were considered candidate variants after MAF filtering (Table 3.17.).

Table 3.17. Summary of top candidate variants of FRASC51 with results and information from various bioinformatic tools and databases.

	Variant 1	**Variant 2**	**Variant 3**	**Variant 4**	**Variant 5**
Gene	*PDS5A*	*PDS5B*	*PDS5B*	*ANKRD11*	*ANKRD11*
Variant	c.1499+24C>G	c.*43_*44ins ACTT	c.*45_*46insT	c.6668C>T	c.5185G>A
GnomAD freq ALL	<0.001	0.000	0.000	0.000	0.000
GnomAD freq AFR	<0.001	0.000	0.000	0.000	0.000
1000G freq ALL	0.000	0.000	0.000	0.000	0.000
1000G freq AFR	0.000	0.000	0.000	0.000	0.000
SIFT				0.007	0.339
PolyPhen2				0.231	0.007
PROVEAN				-0.650	-0.700
MutationTaster				N	N
FATHMM				1.020	1.150
CADD	0.928	5.788	0.432	11.420	7.280
DANN	0.563			0.984	0.988
GERP++	4.990	4.830	5.250	1.080	0.093
dbSNP ID	rs371975259		rs368080614	rs537338393	rs368667754
ClinVar					
Benign ACMG codes	BP4	BP4	BP4	BP1, BP4	BP1, BP4
Pathogenic ACMG codes					
ACMG classification	VUS	VUS	VUS	Likely benign	Likely benign

Cells highlighted in pink indicate a deleterious prediction, cells highlighted in green indicate a benign prediction.

*See Appendices D.1. – D.3. for detailed descriptions of ACMG codes and classifications

The *PDS5A* c.1499+24C>G variant was classified as benign according to the ACMG guidelines. It occurs in an intronic region but is predicted to have no effect on a splice site or the creation of a new splice site according to the online tool Human Splicing Finder. FRASC51 is the tenth patient where variants in the 3' homopolymer region of *PDS5B* have been identified (c.*43_*44insACTT and c.*45_*46insT). It has been shown in all the previous cases that variants in this region do not alter splicing and therefore these variants identified in FRASC51 are not deleterious mutations either.

The two mutations identified in *ANKRD11*: c.6668C>T (p.Ala2223Val) and c.5185G>A (p.Ala1729Thr) are both missense mutations. According to the ACMG code BP1, missense variants in the *ANKRD11* gene are not a common mechanism of disease and therefore both have been classified as likely benign according to the ACMG guidelines. Very few exons in FRASC51 have been covered insufficiently and there is no need for further investigation by means of Sanger sequencing.

3.3.13. FRASC32

There were only 20 variants identified in FRASC32, the least number of variants identified per patient within this cohort. Of these variants, six were considered candidate variants after MAF filtering (Table 3.18.).

Table 3.18. Summary of top candidate variants of FRASC32 with results and information from various bioinformatic tools and databases.

	Variant 1	**Variant 2**	**Variant 3**	**Variant 4**	**Variant 5**	**Variant 6**
Gene	*NIPBL*	*PDS5A*	*ARID1B*	*ARID1B*	*KMT2A*	*SMC1A*
Variant	c.3932G>A	c.567T>C	c.3270C>T	c.5766G>A	c.3176C>T	c.1545+4A >C
GnomAD freq ALL	0.000	0.000	<0.001	<0.001	<0.001	0.001
GnomAD freq AFR	0.000	0.000	<0.001	0.000	0.000	0.003
1000G freq ALL	0.000	0.000	<0.001	0.000	0.000	0.001
1000G freq AFR	0.000	0.000	0.002	0.000	0.000	0.004
SIFT	0.001	1.000	1.000	0.884	0.013	
PolyPhen2	0.994				0.987	
PROVEAN	-5.100	0.000	0.000	0.000	-2.550	
MutationTaster	D	D	D	D	D	
FATHMM	0.430				-1.700	
CADD	28.600	10.650	10.950	1.300	32.000	19.560
DANN	0.998	0.507	0.740	0.640	0.999	0.928
GERP++	5.610	5.750	5.940	5.080	5.840	4.880
dbSNP ID		rs563432054	rs111368751	rs372334858		rs37727094 3
ClinVar			likely benign			VUS/ benign
Benign ACMG codes		BP7, BP4, BP2	BP4, BP6, BP2	BP4, BP2	BP2	BP6, BP2
Pathogenic ACMG codes	PP4, PP3, PM2, PM5, PP2		PM1	PM2	PP3	
ACMG classification	Likely Pathogenic	Likely benign	Likely benign	VUS	VUS	VUS

Cells highlighted in pink indicate a deleterious prediction, cells highlighted in green indicate a benign prediction.
*See Appendices D.1. – D.3. for detailed descriptions of ACMG codes and classifications

The *PDS5A* c.567T>C (p.Asp189=) variant is a synonymous variant and is not predicted to alter splicing according to the ACMG code BP7. It has also been classified as likely benign according to the ACMG guidelines. The *ARID1B* mutations (c.3270C>T and c.5766G>A) are also both synonymous variants. *ARID1B* c.3270C>T (p.Leu1090=) has been identified in two previous patients – FRASC29 and FRASC30 and has been reported to be likely benign according to the ACMG guidelines and by ClinVar (ClinVar submission accession number: SCV000246513). *ARID1B* c.5766G>A (p.Glu1922=) is classified as a VUS according to the ACMG guidelines. As mentioned previously, it is a synonymous variant and according to Human Splicing Finder this variant does not occur within a significant splicing motif and therefore is most likely not deleterious.

The *SMC1A* c.1545+4A>C variant has been seen previously in FRASC77. As mentioned before, this variant is predicted to alter the splice site sequence according to the Human Splicing Finder. However, it has been classified as likely benign by multiple contributions on the online ClinVar database (for website URL, see list of websites at the end of dissertation). It has also been reported 63 times in the gnomAD database further suggesting it is not pathogenic. The *KMT2A* c.3176C>T (p.Ser1059Leu) variant has been classified as a VUS according to the ACMG guidelines, although there are seven out of the eight prediction tools used that are predicting this variant has a deleterious effect. This variant has been identified in four apparently healthy individuals from the Latino population from the gnomAD dataset and is therefore unlikely to be a disease-causing variant since the condition associated with this gene is dominant. However, all of the above mutations were identified in the presence of a variant that is likely to be pathogenic according to the ACMG guidelines, therefore these mutations were assigned the ACMG code: BP2.

A missense variant was identified, *NIPBL* c.3932G>A (Figure 3.16.) and was classified as likely pathogenic according to the ACMG guidelines. The SNV results in an amino acid change at position 1311 in the amino acid sequence: changing a cysteine to a tyrosine (p.Cys1311Tyr) (Figure 3.17.). According to ACMG guidelines, the codes PP4, PP3, PM2, PP2 and PM5 were assigned, suggesting this variant is pathogenic. A previous mutation was identified by Tonkin *et al.*, (2004) (*NIPBL* c.3931T>C) and resulted in an amino acid change reported by UniProt to be causative of CdLS (cysteine to arginine) (p.Cys1311Arg) (Tonkin *et al.*, 2004). PM5 was therefore assigned to the *NIPBL* c.3932G>A variant as it occurs within the same codon of the previously reported mutation and results in an amino acid change.

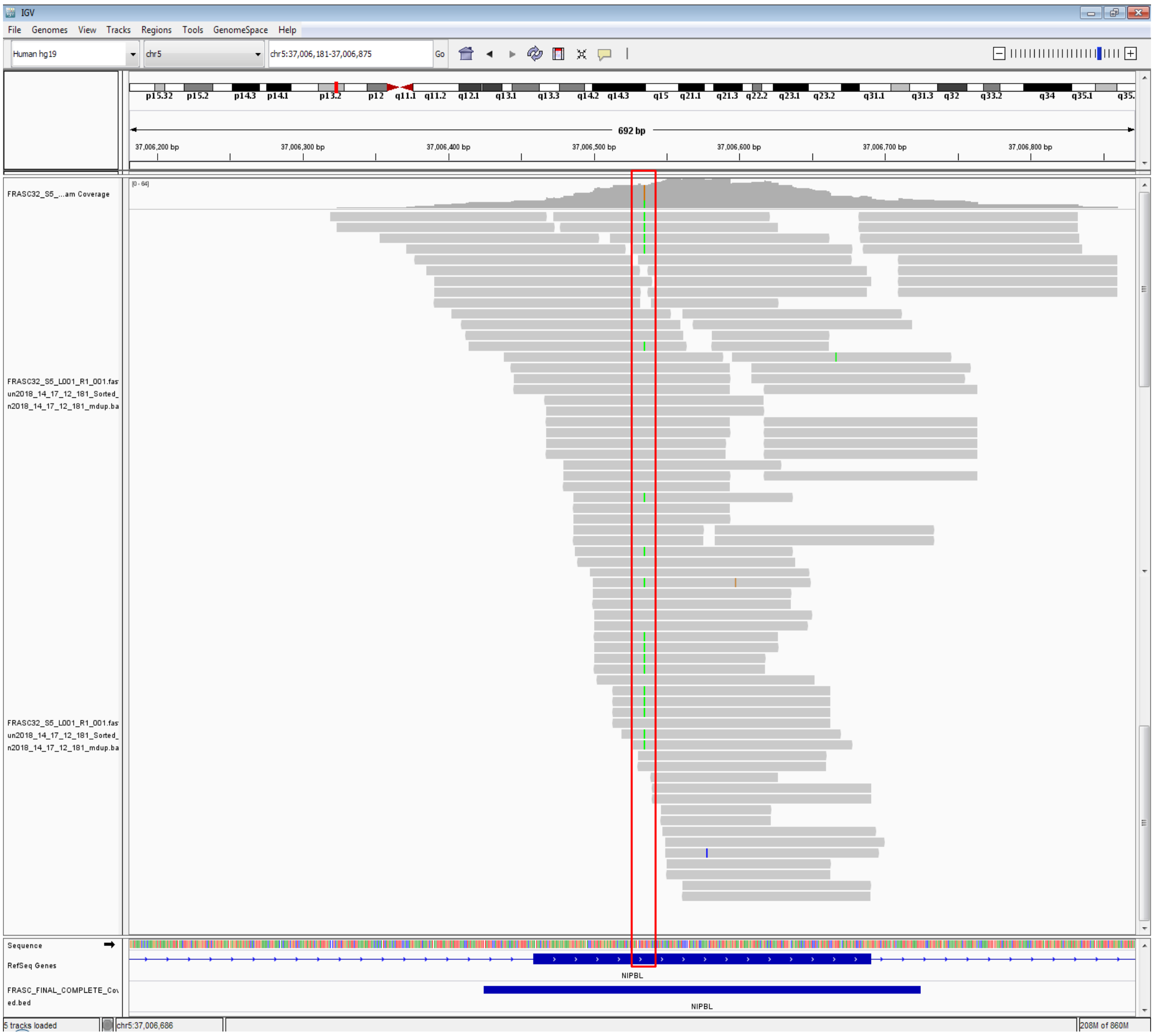
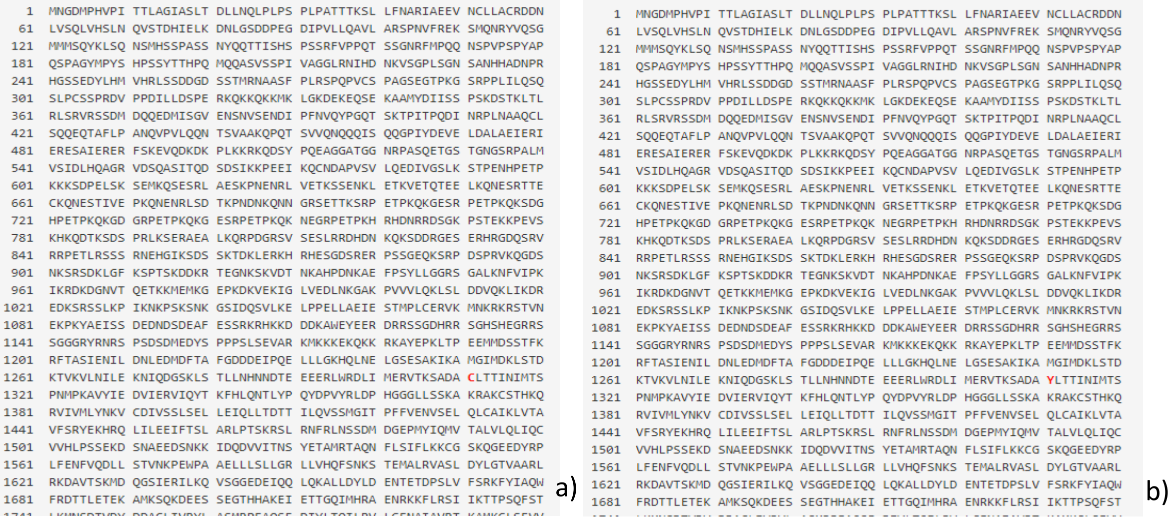

IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
Human hg19
chr5
chr5:37,006,181-37,006,875
Go
p15.32 p15.2 p14.3 p14.1 p13.2 p12 q11.1 q11.2 q12.1 q13.1 q13.3 q14.2 q14.3 q15 q21.1 q21.3 q22.2 q23.1 q23.2 q31.1 q31.3 q32 q33.2 q34 q35.1
692 bp
37,006,200 bp
37,006,300 bp
37,006,400 bp
37,006,500 bp
37,006,600 bp
37,006,700 bp
37,006,800 bp
FRASC32_S5_...am Coverage
[0 - 64]
FRASC32_S5_L001_R1_001.fas un2018_14_17_12_181_Sorted_ n2018_14_17_12_181_mdup.ba
FRASC32_S5_L001_R1_001.fas un2018_14_17_12_181_Sorted_ n2018_14_17_12_181_mdup.ba
Sequence
RefSeq Genes
NIPBL
FRASC_FINAL_COMPLETE_Cov ed.bed
NIPBL
5 tracks loaded
chr5:37,006,686
208M of 860M

Figure 3.16. IGV screenshot depicting the SNP in exon 17 of the NIPBL gene (outlined in red). The A allele is present in 17 of the 49 reads, indicating it is also a heterozygous mutation. Again, there are enough reads with the 'A' allele present to call this mutation confidently.



```
   1  MNGDMPHVPI TTLAGIASLT DLLNQLPLPS PLPATTTKSL LFNARIAEEV NCLLACRDDN
  61  LVSQLVHSLN QVSTDHIELK DNLGSDDPEG DIPVLLQAVL ARSPNVFREK SMQNRYVQSG
 121  MMMSQYKLSQ NSMHSSPASS NYQQTTISHS PSSRFVPPQT SSGNRFMPQQ NSPVPSPYAP
 181  QSPAGYMPYS HPSSYTTHPQ MQQASVSSPI VAGGLRNIHD NKVSGPLSGN SANHHADNPR
 241  HGSSEDYLHM VHRLSSDDGD SSTMRNAASF PLRSPQPVCS PAGSEGTPKG SRPPLILQSQ
 301  SLPCSSPRDV PPDILLDSPE RKQKKQKKMK LGKDEKEQSE KAAMYDIISS PSKDSTKLTL
 361  RLSRVRSSDM DQQEDMISGV ENSNVSENDI PFNVQYPGQT SKTPITPQDI NRPLNAAQCL
 421  SQQEQTAFLP ANQVPVLQQN TSVAAKQPQT SVVQNQQQIS QQGPIYDEVE LDALAEIERI
 481  ERESAIERER FSKEVQDKDK PLKKRKQDSY PQEAGGATGG NRPASQETGS TGNGSRPALM
 541  VSIDLHQAGR VDSQASITQD SDSIKKPEEI KQCNDAPVSV LQEDIVGSLK STPENHPETP
 601  KKKSDPELSK SEMKQSESRL AESKPNENRL VETKSSENKL ETKVETQTEE LKQNESRTTE
 661  CKQNESTIVE PKQNENRLSD TKPNDNKQNN GRSETTKSRP ETPKQKGESR PETPKQKSDG
 721  HPETPKQKGD GRPETPKQKG ESRPETPKQK NEGRPETPKH RHDNRRDSGK PSTEKKPEVS
 781  KHKQDTKSDS PRLKSERAEA LKQRPDGRSV SESLRRDHDN KQKSDDRGES ERHRGDQSRV
 841  RRPETLRSSS RNEHGIKSDS SKTDKLERKH RHESGDSRER PSSGEQKSRP DSPRVKQGDS
 901  NKSRSDKLGF KSPTSKDDKR TEGNKSKVDT NKAHPDNKAE FPSYLLGGRS GALKNFVIPK
 961  IKRDKDGNVT QETKKMEMKG EPKDKVEKIG LVEDLNKGAK PVVVLQKLSL DDVQKLIKDR
1021  EDKSRSSLKP IKNKPSKSNK GSIDQSVLKE LPPELLAEIE STMPLCERVK MNKRKRSTVN
1081  EKPKYAEISS DEDNDSDEAF ESSRKRHKKD DDKAWEYEER DRRSSGDHRR SGHSHEGRRS
1141  SGGGRYRNRS PSDSDMEDYS PPPSLSEVAR KMKKKEKQKK RKAYEPKLTP EEMMDSSTFK
1201  RFTASIENIL DNLEDMDFTA FGDDDEIPQE LLLGKHQLNE LGSESAKIKA MGIMDKLSTD
1261  KTVKVLNILE KNIQDGSKLS TLLNHNNDTE EEERLWRDLI MERVTKSADA CLTTINIMTS
1321  PNMPKAVYIE DVIERVIQYT KFHLQNTLYP QYDPVYRLDP HGGGLLSSKA KRAKCSTHKQ
1381  RVIVMLYNKV CDIVSSLSEL LEIQLLTDTT ILQVSSMGIT PFFVENVSEL QLCAIKLVTA
1441  VFSRYEKHRQ LILEEIFTSL ARLPTSKRSL RNFRLNSSDM DGEPMYIQMV TALVLQLIQC
1501  VVHLPSSEKD SNAEEDSNKK IDQDVVITNS YETAMRTAQN FLSIFLKKCG SKQGEEDYRP
1561  LFENFVQDLL STVNKPEWPA AELLLSLLGR LLVHQFSNKS TEMALRVASL DYLGTVAARL
1621  RKDAVTSKMD QGSIERILKQ VSGGEDEIQQ LQKALLDYLD ENTETDPSLV FSRKFYIAQW
1681  FRDTTLETEK AMKSQKDEES SEGTHHAKEI ETTGQIMHRA ENRKKFLRSI IKTTPSQFST
```

```
   1  MNGDMPHVPI TTLAGIASLT DLLNQLPLPS PLPATTTKSL LFNARIAEEV NCLLACRDDN
  61  LVSQLVHSLN QVSTDHIELK DNLGSDDPEG DIPVLLQAVL ARSPNVFREK SMQNRYVQSG
 121  MMMSQYKLSQ NSMHSSPASS NYQQTTISHS PSSRFVPPQT SSGNRFMPQQ NSPVPSPYAP
 181  QSPAGYMPYS HPSSYTTHPQ MQQASVSSPI VAGGLRNIHD NKVSGPLSGN SANHHADNPR
 241  HGSSEDYLHM VHRLSSDDGD SSTMRNAASF PLRSPQPVCS PAGSEGTPKG SRPPLILQSQ
 301  SLPCSSPRDV PPDILLDSPE RKQKKQKKMK LGKDEKEQSE KAAMYDIISS PSKDSTKLTL
 361  RLSRVRSSDM DQQEDMISGV ENSNVSENDI PFNVQYPGQT SKTPITPQDI NRPLNAAQCL
 421  SQQEQTAFLP ANQVPVLQQN TSVAAKQPQT SVVQNQQQIS QQGPIYDEVE LDALAEIERI
 481  ERESAIERER FSKEVQDKDK PLKKRKQDSY PQEAGGATGG NRPASQETGS TGNGSRPALM
 541  VSIDLHQAGR VDSQASITQD SDSIKKPEEI KQCNDAPVSV LQEDIVGSLK STPENHPETP
 601  KKKSDPELSK SEMKQSESRL AESKPNENRL VETKSSENKL ETKVETQTEE LKQNESRTTE
 661  CKQNESTIVE PKQNENRLSD TKPNDNKQNN GRSETTKSRP ETPKQKGESR PETPKQKSDG
 721  HPETPKQKGD GRPETPKQKG ESRPETPKQK NEGRPETPKH RHDNRRDSGK PSTEKKPEVS
 781  KHKQDTKSDS PRLKSERAEA LKQRPDGRSV SESLRRDHDN KQKSDDRGES ERHRGDQSRV
 841  RRPETLRSSS RNEHGIKSDS SKTDKLERKH RHESGDSRER PSSGEQKSRP DSPRVKQGDS
 901  NKSRSDKLGF KSPTSKDDKR TEGNKSKVDT NKAHPDNKAE FPSYLLGGRS GALKNFVIPK
 961  IKRDKDGNVT QETKKMEMKG EPKDKVEKIG LVEDLNKGAK PVVVLQKLSL DDVQKLIKDR
1021  EDKSRSSLKP IKNKPSKSNK GSIDQSVLKE LPPELLAEIE STMPLCERVK MNKRKRSTVN
1081  EKPKYAEISS DEDNDSDEAF ESSRKRHKKD DDKAWEYEER DRRSSGDHRR SGHSHEGRRS
1141  SGGGRYRNRS PSDSDMEDYS PPPSLSEVAR KMKKKEKQKK RKAYEPKLTP EEMMDSSTFK
1201  RFTASIENIL DNLEDMDFTA FGDDDEIPQE LLLGKHQLNE LGSESAKIKA MGIMDKLSTD
1261  KTVKVLNILE KNIQDGSKLS TLLNHNNDTE EEERLWRDLI MERVTKSADA YLTTINIMTS
1321  PNMPKAVYIE DVIERVIQYT KFHLQNTLYP QYDPVYRLDP HGGGLLSSKA KRAKCSTHKQ
1381  RVIVMLYNKV CDIVSSLSEL LEIQLLTDTT ILQVSSMGIT PFFVENVSEL QLCAIKLVTA
1441  VFSRYEKHRQ LILEEIFTSL ARLPTSKRSL RNFRLNSSDM DGEPMYIQMV TALVLQLIQC
1501  VVHLPSSEKD SNAEEDSNKK IDQDVVITNS YETAMRTAQN FLSIFLKKCG SKQGEEDYRP
1561  LFENFVQDLL STVNKPEWPA AELLLSLLGR LLVHQFSNKS TEMALRVASL DYLGTVAARL
1621  RKDAVTSKMD QGSIERILKQ VSGGEDEIQQ LQKALLDYLD ENTETDPSLV FSRKFYIAQW
1681  FRDTTLETEK AMKSQKDEES SEGTHHAKEI ETTGQIMHRA ENRKKFLRSI IKTTPSQFST
```

Figure 3.17. Mutalyzer screenshots depicting a portion of the amino acid sequence of the NIPBL gene. A) The reference sequence with an amino acid residue of cysteine at position 1311 (highlighted in red). B) the mutated amino acid sequence showing the tyrosine amino acid residue at position 1311 (also highlighted in red).

3.3.14. FRASC41

There was only a single candidate identified in FRASC41 after MAF filtering out of the 57 variants identified during sequencing (Table 3.19.).

Table 3.19. Summary of the top candidate variant of FRASC41 with results and information from various bioinformatic tools and databases.

	Variant 1
Gene	*PDS5A*
Variant	c.2153+15T>C
GnomAD freq ALL	0.000
GnomAD freq AFR	0.000
1000G freq ALL	0.000
1000G freq AFR	0.002
SIFT	
PolyPhen2	
PROVEAN	
MutationTaster	
FATHMM	
CADD	4.480
DANN	0.455
GERP++	4.840
dbSNP ID	rs181065196
ClinVar	
Benign ACMG codes	BP4
Pathogenic ACMG codes	
ACMG classification	VUS

Cells highlighted in pink indicate a deleterious prediction, cells highlighted in green indicate a benign prediction.

*See Appendices D.1. – D.3. for detailed descriptions of ACMG codes and classifications

The *PDS5A* c.2153+15T>C variant occurs within an intron and according to the online tool Human Splicing Finder it is not predicted to affect the splicing sequence or to add a new splice site. This variant has been identified in an African population before and is likely not a disease-causing mutation. There were a few exons that were not covered sufficiently (Table 3.5.) and therefore it is recommended that Sanger sequencing be carried out on those exons for further variant identification.

3.4.Possible disease-causing mutations for validation

Possible disease-causing mutations were identified in eight of the 14 patients. Seven of these were identified in the *NIPBL* gene and one was identified in the *STAG1* gene (Table 3.20.). Seven mutations were classified as pathogenic and one was classified as likely pathogenic according to the ACMG guidelines. These classifications are obtained by assigning the ACMG codes described in Appendix D.1. - D.3.

Table 3.20. Potential disease-causing mutations identified in eight CdLS patients.

Patient number	**Gene**	**Mutation**	**ACMG classification**	**ACMG classification codes***							
				PP4	**PP3**	**PM2**	**PM5**	**PVS1**	**PP5**	**PP2**	**PM1**
FRASC8	*NIPBL*	c.6027_6030 del GTTC	Pathogenic	X	X	X		X			
FRASC30	*NIPBL*	c.2479_2480 del AG	Pathogenic	X	X	X		X	X		X
FRASC32	*NIPBL*	c.3932 G>A	Likely pathogenic	X	X	X	X			X	
FRASC47	*STAG1*	c.17 T>G	Pathogenic	X	X	X		X			
FRASC72	*NIPBL*	c.6955-2 A>C	Pathogenic	X	X	X		X			
FRASC75	*NIPBL*	c.5639_5642 del CAAC	Pathogenic	X	X	X		X			
FRASC76	*NIPBL*	c.302_311 del CAAGGAGTCC	Pathogenic	X	X	X		X			
FRASC79	*NIPBL*	c.7831 dup A	Pathogenic	X	X	X		X			

*Refer to Appendices D.1. and D.2. for a detailed explanation of the ACMG codes

Validation of seven of the eight possibly disease-causing mutations was carried out using Sanger sequencing. The DNA for FRASC32 was depleted and therefore validation could not be carried out at this stage. The patients (and immediate family members where possible) were sequenced and the results were analysed using BioEdit Version 7.2. Although results were obtained for all those sequenced, the chromatograms were not clear and background noise was visible, interfering with interpretation. These samples will be sequenced again, however there was one sample (FRASC8) with a chromatogram that clearly depicted the identified deletion (Figure 3.18.).

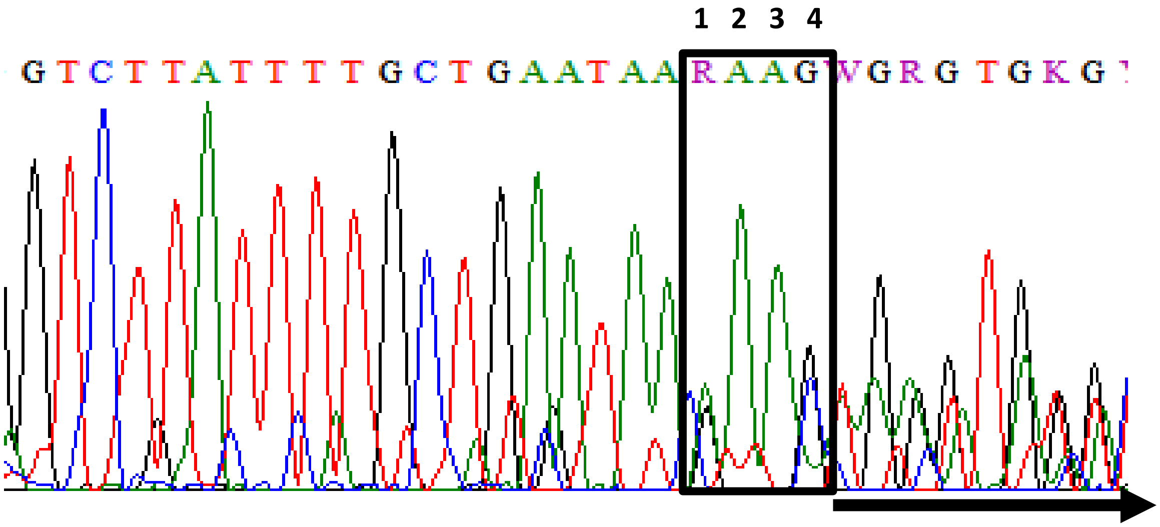


Figure 3.18. Chromatogram showing a portion of the reverse sequence including the deletion identified in FRASC 8 (NIPBL c.6027_6030del GTTC): GCTGAATAAGAACAAAGTGGTTATG. The reference sequence would be GAAC and the sequence at the same position with the deletion would be AAAG. Position 1 is heterozygous for a G and an A, position 2 and 3 are both A's and position 4 is heterozygous for a C and a G. A comparison of the sequence preceding and succeeding the deletion shows the latter's peaks are indistinguishable thus supporting the presence of the deletion.

4. Discussion

Literature shows that only 70% of patients that are clinically diagnosed with Cornelia de Lange syndrome also receive a molecular diagnosis (Braunholz *et al.*, 2015). The remaining 30% of patients either harbour a mutation in a yet unknown CdLS causative gene or have been clinically misdiagnosed due to the broad phenotypic spectrum of the disease. However, these are global statistics and may not reflect our South African CdLS population, as no molecular study has been carried out on patients with South African ancestry. The present study aimed to determine the molecular basis of CdLS using targeted NGS in a South African cohort clinically diagnosed with a CdLS phenotype.

4.1.Putative mutations identified

Fourteen patients with a Cornelia de Lange syndrome phenotype underwent NGS by means of a specifically designed gene panel. Putative disease-causing mutations were identified in eight of these patients. Seven of the eight mutations identified occurred within the *NIPBL* gene, and the other mutation was found in the *STAG1* gene.

NIPBL is the most commonly mutated gene amongst patients with CdLS. Globally, approximately 89% of patients with a molecular diagnosis have mutations in this gene (Mannini *et al.*, 2013). The same trend can be seen in the present cohort as 87.5% of the identified putative disease-causing mutations occurred in the *NIPBL* gene. The most common types of mutations occurring in the *NIPBL* gene in patients with CdLS are small deletions and missense mutations (Mannini *et al.*, 2013). Similarly, the most common type of mutation identified in *NIPBL* within the present cohort was small deletions (occurring in four patients). A missense variant, splice site variant and a duplication were also identified. Considering CdLS is most often an autosomal dominant condition caused by unique, de novo mutations, it is not expected for there to be differences in the mutation profile of South African patients with CdLS compared to those from other countries.

At the time of designing the gene panel, mutations in the *STAG1* gene had not been found to cause CdLS in humans. In the search for genes that may account for the 30% of patients without a molecular diagnosis, many mouse-model studies had been conducted on different genes. One such study was carried out by Remeseiro *et al.*, (2012); they observed *STAG1* knock-out mice and compared the phenotypes to wild type mice. There was a significant overlap in the phenotype observed in the *STAG1* knock-out mice and that of CdLS. The phenotype included reduced stature and skeletal abnormalities occurring through delayed osteogenesis. This observation lead Remeseiro *et al.*, (2012) to believe that *STAG1* plays an essential role in the cohesin pathway and, when mutated in a human, could potentially lead to a CdLS phenotype. Since then, however, there have been reports in the literature of *STAG1* mutations identified in humans, first by Lehalle *et al.*, (2017) and then by Yuan *et al.*, (2018). Lehalle *et al.*, (2017) reported *STAG1* mutations in 17 patients tested from around the world and described a novel cohesinopathy with nonspecific syndromic intellectual disability. In 2018, Yuan *et al*., carried out a retrospective study on patients who had undergone clinical exome sequencing to identify mutations in cohesinopathy genes. They subsequently identified three patients with a *STAG1* mutation who presented with developmental and intellectual delay as well as dysmorphic facial features and skeletal anomalies and concluded that these patients had a phenotype overlapping with that of CdLS.

Each putative disease-causing mutation identified in these two genes within the present cohort has compelling evidence to suggest pathogenicity. This evidence was gathered and evaluated following the ACMG guidelines.

4.1.1. **Small deletion mutations**

Four small deletions were identified within the *NIPBL* gene in the present cohort. These occurred in the following patients:

- FRASC8 (c.6027_6030 del GTTC),
- FRASC30 (c.2479_2480 del AG),
- FRASC75 (c.5639_5642 del CAAC) and
- FRASC76 (c.302_311 del CAAGGAGTCC).

Each of these patients presented with a classical CdLS phenotype (as assessed by the geneticists at the Division of Human Genetics) including shared features of microcephaly, hirsutism and upper limb abnormalities. The phenotypes in these patients correlate with the type of putative disease-causing mutation identified in each patient. Previous studies highlight that small deletions cause a classical and severe form of CdLS compared with other mutations, such as missense mutations. (Mannini *et al.*, 2013). None of the deletions identified in the patients listed above have been reported in public databases as per the ACMG code PM2. Additionally, multiple prediction tools (which included CADD, MutationTaster and GERP++) provide supporting evidence that these mutations are pathogenic or occur within a conserved region. Lastly, the strongest line of evidence suggesting pathogenicity is the assigning of the ACMG code PVS1 to all four of the above mutations. The small deletions identified all resulted in a truncation at varying points within the protein sequence resulting in a loss of function mutation. Loss of function mutations are the most common mechanism of disease in patients with CdLS. A study investigating cohesin and β-globin expression showed that even a slight decrease in NIPBL affects the binding of the cohesin complex, including to CTCF sites, resulting in decreased gene expression (Chien *et al.*, 2011). This was supported by a study investigating the genome-wide effect of *NIPBL* haploinsufficiency (Newkirk *et al.*, 2017). Reduced binding of the cohesin complex to its targets and CTCF sites, and the subsequent reduced gene expression, is the proposed underlying molecular cause of CdLS.

The two-bp deletion identified in FRASC30 has additional evidence suggesting pathogenicity. This mutation occurred in a hotspot within the *NIPBL* gene – exon 10 and was therefore assigned the ACMG code PM1. Exon 10 is very large in comparison to the other *NIPBL* exons and transcribes multiple functional regions including CTCF sites, promoters, enhancers and other transcription factor binding sites, according to Ensembl (Zerbino *et al.*, 2018). Additionally, the ACMG code PP5 was assigned to this mutation as it has previously been identified in the literature (Gillis *et al.*, 2004; Yuan *et al.*, 2015). The mutation was previously identified in a patient with CdLS and was reported to be pathogenic on the ClinVar database (ClinVar ID: RCV000146547.3).

It is recommended that functional studies or RNA and cDNA sequencing be completed to obtain further knowledge on these mutations and the precise consequences or deleterious effects, although, on some occasions functional studies are not routinely performed in the face of strong evidence of pathogenicity. Given the current evidence, it can be surmised with reasonable certainty that these four deletion mutations are the disease-causing mutations in these patients.

4.1.2. **Missense mutation identified**

One missense mutation was identified within the present cohort, in FRASC32 (*NIPBL* c.3932 G>A) and was classified as likely pathogenic using the ACMG guidelines. The patient presented with a mild CdLS phenotype including microcephaly, hirsutism and clinodactyly. The described phenotype is in accordance with multiple genotype-phenotype studies previously conducted that conclude that missense mutations correlate to a milder CdLS phenotype (Mannini *et al.*, 2013). Seven prediction tools provided evidence suggesting the mutation in FRASC32 was pathogenic, including SIFT, PolyPhen2, PROVEAN, DANN, MutationTaster, CADD and GERP++.

While the ACMG code PM2 was assigned indicating that this mutation has not been identified in public datasets, the ACMG code PM5 was also assigned. This code indicates that another missense mutation (predicted to be pathogenic) has been identified previously that occurs within the same amino acid residue. FRASC32 had a missense mutation identified at position c.3932 in the *NIPBL* gene, one bp away from the previously reported mutation. The previously identified mutation (*NIPBL* c.3931T>C) resulted in an amino acid change from cysteine to arginine (p.Cys1311Arg) (Tonkin *et al.*, 2004), while the amino acid change in FRASC32 was a cysteine to a tyrosine (p.Cys1311Tyr). The patient reported by Tonkin *et al.*, (2004) also presented with a mild CdLS phenotype similar to FRASC32 in the present cohort.

Finally, the ACMG code PP2 was assigned to this missense variant. This indicates that benign missense mutations are not common in this gene but rather they are a common mechanism of disease in *NIPBL* (Mannini *et al.*, 2013). When examining *NIPBL* in the ExAC database (Lek *et al.*, 2016) missense variants were assigned a positive Z-score. This indicates that there is an increased intolerance to variation in *NIPBL* when compared to missense variants in other genes. This missense mutation does not result in a truncated protein but rather a single amino acid change in the protein sequence. This could shed light on the mild phenotype of the patient as the mutation does not result in haploinsufficiency, but potentially a protein where the function is somewhat altered (Mannini *et al.*, 2013). Functional studies are recommended to elucidate the true effect this mutation has on the protein's function in the cohesin pathway. However, it is plausible that this is the disease-causing mutation in this patient.

4.1.3. Splice site mutation identified

A splice site variant was identified in FRASC72 (*NIPBL* c.6955-2 A>C) and was classified as pathogenic according to the ACMG guidelines. The patient presented with a classical CdLS phenotype with failure to thrive, microcephaly, hirsutism, a smooth philtrum, upper limb abnormalities, cardiac defects and moderate to severe intellectual disability. The prediction tools MutationTaster, CADD, DANN and GERP++ gave evidence suggesting that the mutation is deleterious as per the ACMG code PP3. This mutation also has not been reported in public datasets to date.

The ACMG code PVS1 can be applied to this mutation as it occurs at an essential splice site according to the online tool Human Splicing Finder (Desmet *et al.*, 2009). In order to elucidate the exact mechanism of disease, it is recommended that RNA and cDNA sequencing should be carried out. This could reveal phenomena such as altered reading frames or exon skipping as the mechanism of disease in this patient.

4.1.4. Duplication mutation identified

The only duplication mutation identified in the present cohort was seen in FRASC79 (*NIPBL* c.7831dupA) and was classified as pathogenic according to the ACMG guidelines. This patient presented with hirsutism, a smooth philtrum and multiple upper limb abnormalities. As with the previous putative disease-causing mutations identified in *NIPBL*, this mutation has not been reported in a public dataset to date. Additionally, as per the ACMG code PP3, multiple prediction tools, including MutationTaster, CADD and GERP++, gave evidence to suggest that this duplication is pathogenic.

The ACMG code PVS1 was applied as the duplication results in a truncation in the NIPBL protein sequence similarly to the small deletion mutations discussed previously. However, this addition of another adenine in the gene sequence occurs within a homopolymer region. It has been suggested in the literature that Sanger sequencing is essential in validating any putative disease-causing mutations identified by NGS, particularly in homopolymer regions where the NGS error rate may be higher than normal (Mu *et al.*, 2016). It is therefore very important to pursue additional lines of testing to confirm that this is a true mutation and not an NGS error in a homopolymer region. If this is a true duplication, then the resulting truncation is likely to cause a loss of function in the protein and this subsequent haploinsufficiency is the probable disease mechanism in this patient.

4.1.5. ***STAG1* missense mutation identified**

The mutation identified in FRASC47 was a missense variant in the *STAG1* gene (*STAG1* c.17 T>G). The mutation was classified as pathogenic according to the ACMG guidelines.

FRASC47 was clinically diagnosed with CdLS and presented with a mild phenotype including failure to thrive and hirsutism. A potentially causative mutation was not identified in the five known CdLS causal genes. Considering this, either FRASC47 should be diagnosed with the new cohesinopathy that was described by Lehalle *et al.*, (2017) or *STAG1* mutations cause a mild form of CdLS. It is unclear at this point whether STAG1 mutations will fall within the spectrum of genes implicated in CdLS (and account for the 30% of unknown cases) or whether they will be classified as a new cohesinopathy. In order to solve discrepancies such as this, a multidisciplinary approach should be taken. More in depth phenotyping could be carried out, biochemical functional studies should be done and perhaps a genome-wide sequencing and segregation approach should also be adopted to identify additional interacting genes. This could potentially differentiate between what is a new cohesinopathy and what is a newly identified

CdLS causal gene. An example of this can be seen in a study by Alesi *et al.*, (2018), where they examined patients with mutations in the *BRD4* gene, which had recently been reported to cause a CdLS-like phenotype in three unrelated patients (Olley *et al.*, 2018). The study by Olley *et al.*, (2018) carried out various biochemical tests to determine the underlying pathogenic mechanisms that lead to the observed phenotype. This showed that BRD4 and NIPBL proteins interacted with each other and had overlapping functions. By studying collated data of patients with a *BRD4* mutation, it was theorized that *BRD4* haploinsufficiency could result in a milder form of CdLS. This approach could potentially be applied to *STAG1*.

4.1.6. **Conclusions on the above patients**

The present evidence shows support for the pathogenic or likely pathogenic classification for all the above-mentioned mutations. Family segregation and functional studies would still be recommended to conclusively classify these mutations as pathogenic. However, classifications presented with as much evidence as these could be sufficient to report back to families once validated in a molecular diagnostic setting. Although in a clinical and counselling setting it is made clear that curative treatments cannot be offered, a clear diagnosis helps with future planning and clinical management. It has been shown to be beneficial to receive a definitive diagnosis even when no treatment is available (Lingen *et al.*, 2016). It is therefore highly recommended that these mutations be validated under diagnostic conditions and reported back to the families in the appropriate clinical and counselling setting.

Of the 14 patients that underwent sequencing, mutations were identified in eight. That gives a mutation pick-up rate of 57%, whereas the global pick-up rate is 70% (Braunholz *et al.*, 2015). Of the patients where a putative disease-causing mutation was identified, seven were found in the *NIPBL* gene and one was found in the *STAG1* gene. Globally, mutations in the *NIPBL* gene account for 89% of CdLS mutations whereas in this study, mutations identified in *NIPBL* accounted for 87.5% of mutations identified.

4.2. **Project design**

4.2.1. **Gene panel design**

The aim of the present study was to identify pathogenic variants in patients with a CdLS phenotype. The chosen method was to design a gene panel with known causative genes, suspected causative genes, as well as genes involved in differential diagnoses. In total 18 genes were included on the panel of which only two were found to harbour putative disease-causing mutations: *NIPBL* and *STAG1*. At the time of designing the gene panel, four suspected causative genes were included: *PDS5A*, *PDS5B*, *STAG1* and *SCC4*. During the study, there have been studies published describing mutations in the cohesin genes: *STAG1* and *STAG2*, *PDS5A*, *WAPL* and *BRD4* in patients with a CdLS like phenotype (Lehalle *et al.*, 2017; Olley *et al.*, 2018; Yuan *et al.*, 2018; Alesi *et al.*, 2018). Considering these new studies, one recommendation would be adding *STAG2*, *WAPL* and *BRD4* to the gene panel that was designed to expand the search for a molecular diagnosis for patients with CdLS in South Africa. However, with new research constantly being carried out, there are challenges with finalising which genes to incorporate onto a gene panel as was outlined by a meeting facilitated by the UK Cancer Group regarding cancer gene panels (Taylor *et al.*, 2018). Therefore, the list of genes to be included on a gene panel need to be regularly checked and current research continually incorporated if there is sufficient supporting evidence.

When examining the efficiency of gene panel, it is important to consider cohort size, specifically for cases of rare diseases. If this gene panel were to be utilised in a diagnostic setting or re-used in another research study, its utility could be optimised by expanding it to include genes involved in other rare genetic conditions. By including genes from multiple genetic conditions onto one panel, it is more cost and time efficient than designing individual panels for each disease separately. This is particularly true when the conditions in question do not have a wide array of genes to be examined.

4.2.2. **NGS and data analysis**

Manual library preparation was carried out in the study and sequencing was performed on the Illumina MiSeq at Wits Medical School Campus. For the purposes of the study, manual library preparation was effective. However, it would be ideal if the panel could be offered as a diagnostic test using automated library preparation in the future, the benefits of which has been outlined in a similar context by (Lundin *et al.*, 2010). They describe an automated library preparation system that could prepare between 36 and 96 libraries in a day which is ideal for small to medium sized sequencing facilities. However, automated library preparation has added equipment costs and a full cost-benefit analysis would be a useful endeavour.

Data analysis was carried out using various bioinformatics tools, as well as following the ACMG guidelines (Richards *et al.*, 2015). These guidelines were designed in an American setting, but they were designed to be utilised in any population and for any Mendelian disease (Richards *et al.*, 2015). However, the PM2 ACMG code (which considers MAF in public datasets) should be applied in the context of the relevant population. South African studies are sometimes pressed to use smaller data sets for MAF interpretations due to the paucity of African data. Overall, the ACMG guidelines were a useful strategy for variant interpretation in this dataset. Additionally, the varying strengths of the PVS1 code as described by Abou Tayoun *et al.*, (2018) was informative.

4.3. Limitations and future directions

4.3.1. Limitations

There were a few limitations in the current study. The cohort size was small due to the rare nature of the disease. By sequencing more patients from across South Africa, a more complete mutation profile of South African patients with CdLS could potentially be produced. The prevalence of CdLS in the country could be investigated if an epidemiologic study was carried out in order to compare it with global statistics.

The targeted gene panel used could detect the most common types of mutations described in patients with CdLS, however, it would not be able to identify intronic or large deletion or duplication mutations that may be present. As mentioned previously in chapter 3.2., there were regions of decreased coverage. Various exons were poorly covered in majority of the patients sequenced and could therefore indicate that the specifically designed probes weren't binding sufficiently to these regions. It is recommended that Sanger sequencing be performed on these regions for patients where no mutation was identified to possibly increase the mutation detection rate.

One of the biggest obstacles encountered in data analysis was the lack of data generated from various prediction tools. The limitations of bioinformatics tools present difficulties as the majority of these tools only give predictions for missense variants. CADD and GERP++ consistently gave prediction scores for every variant identified, whereas SIFT, PolyPhen2, Provean and DANN very rarely provided prediction scores due to majority of them only providing predictions for SNPs. MutationTaster frequently predicted variants to be deleterious even when many other prediction tools gave evidence to the contrary. All this indicates that as many prediction tools as possible should be used when analysing variants to account for lacking or conflicting predictions as recommended in the Richards *et al*., (2015) paper. Additionally, more prediction tools are needed that are able to interpret pathogenicity for variant classes other than missense variants. Alternatively, more evidence is needed to allow these tools to provide a pathogenic or benign prediction.

4.3.2. Future directions/ recommendations

Sanger sequencing was conducted on the samples where a putative disease-causing mutation was identified, however the chromatograms were not clear and impeded interpretation. It is recommended that Sanger sequencing be carried out again with an alternative cleaning up method as this may reduce the background noise that was present. For patients where a putative disease-causing mutation was identified, it is recommended that the results be validated in a diagnostic setting and returned to the patient during a follow up visit with a medical geneticist or genetic counsellor. Functional, biochemical studies should also be carried out on these mutations to confirm the loss of function of these proteins.

To account for the 43% of patients without a molecular diagnosis in the present cohort, other methods of testing should be considered. These include testing for copy number variants (CNV's) (by means of multiplex ligation-dependent probe amplification or arrays), expanding the gene panel to include more genes (e.g. *STAG2, BRD4* and *WAPL*), or taking a non-directive approach (e.g. whole exome or whole genome sequencing). Whole genome sequencing may not be the obvious choice of alternative testing, however, there have been studies in other monogenic diseases identifying causative mutations in intronic regions previously (Ngcungcu *et al.*, 2017) and it cannot be excluded as a course of testing. CNV's have not previously been shown to be a significant cause of CdLS, however, in light of the reduced mutation pick up rate this may be a possible strategy. Investigations into possible mosaicism in *NIPBL* could prove beneficial as well as there have been previous reports of identifying mosaicism in fibroblasts (reveiwed by Kline *et al.*, 2018).

It is recommended to adjust the gene panel designed in the study to include genes from other, completely unrelated genetic conditions to optimise efficiency and cost effectiveness in the diagnostic setting. After a full cost-benefit analysis has been conducted, using an automated library preparation system may prove to be more efficient and reduce turn-around time (which is currently from six to twelve weeks onwards on average, personal communication, E. Vorster, 07 June 2019).

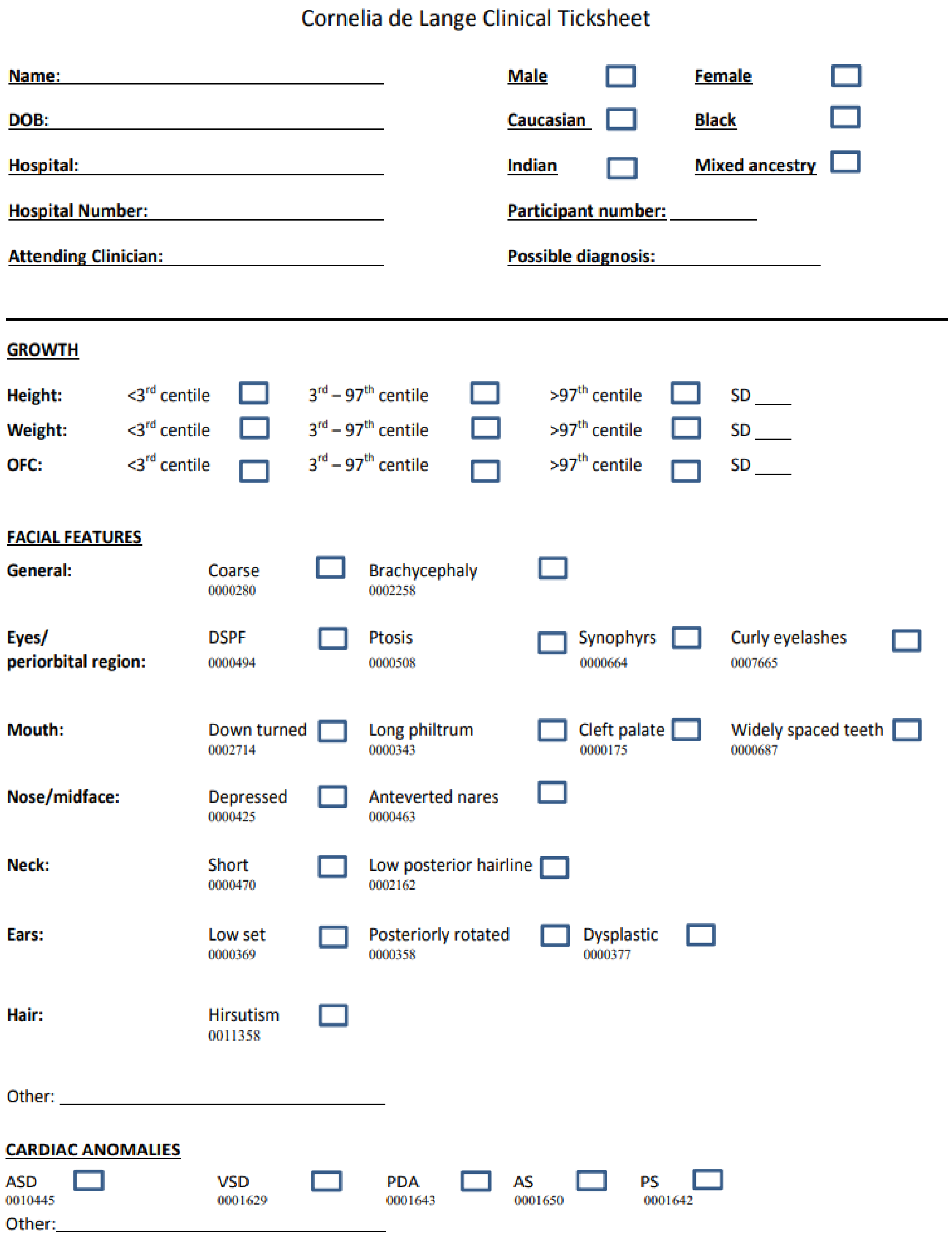
On average, 49 variants per individual patient were identified in the 18 genes included on the panel. With the appropriate filtering methods, the analysis of data for each patient was not considerably longer than the average analysis ongoing in the diagnostic setting currently. A targeted gene panel would therefore be a useful testing strategy in South Africa given our various constraints (e.g. staff shortages, cost of reagents, sample batching). This may not be possible to implement in the near future, so it is recommended that sequencing of the *NIPBL* gene be offered to patients with CdLS seen at our clinics as this was the most commonly mutated gene in the present cohort. This approach for low throughput sequencing could increase diagnostic offerings significantly in the immediate future.

4.4.<u>Conclusion</u>

A cohort of 14 patients of African ancestry with a CdLS like phenotype underwent targeted NGS sequencing by means of a specifically designed gene panel to generate a mutation profile. This is the largest South African cohort to undergo molecular studies for CdLS to date. Seven mutations were identified in the *NIPBL* gene and one was identified in the *STAG1* gene. These consisted of four deletions, two SNV's, one duplication and one splice site mutation. The pick-up rate and results obtained in this study are comparable to what is observed globally. The present study has produced a baseline mutation profile of CdLS in South African patients and has provided direction to improve future genetic testing for this rare disorder.

Appendix

Appendix A.1. Clinical tick sheet used to diagnose patients with suspected CdLS.



Cornelia de Lange Clinical Ticksheet

Name: ______________________	**Male** □	**Female** □
DOB: ______________________	**Caucasian** □	**Black** □
Hospital: ______________________	**Indian** □	**Mixed ancestry** □
Hospital Number: ______________________	**Participant number:** ________	
Attending Clinician: ______________________	**Possible diagnosis:** ______________	

GROWTH

Height:	<3rd centile □	3rd – 97th centile □	>97th centile □	SD ____
Weight:	<3rd centile □	3rd – 97th centile □	>97th centile □	SD ____
OFC:	<3rd centile □	3rd – 97th centile □	>97th centile □	SD ____

FACIAL FEATURES

General:	Coarse □ 0000280	Brachycephaly □ 0002258		
Eyes/ periorbital region:	DSPF □ 0000494	Ptosis □ 0000508	Synophyrs □ 0000664	Curly eyelashes □ 0007665
Mouth:	Down turned □ 0002714	Long philtrum □ 0000343	Cleft palate □ 0000175	Widely spaced teeth □ 0000687
Nose/midface:	Depressed □ 0000425	Anteverted nares □ 0000463		
Neck:	Short □ 0000470	Low posterior hairline □ 0002162		
Ears:	Low set □ 0000369	Posteriorly rotated □ 0000358	Dysplastic □ 0000377	
Hair:	Hirsutism □ 0011358			

Other: ______________________________

CARDIAC ANOMALIES

ASD □ 0010445	VSD □ 0001629	PDA □ 0001643	AS □ 0001650	PS □ 0001642

Other:______________________________

Cornelia de Lange Clinical Ticksheet

SKELETAL ANOMALIES

Upper limbs:	Micromelia 0000171	□	Olidodactyly 0012165	□	Clinodactyly 0001588	□
	Elbow contractures 0002987	□	Transverse palmar crease 0000954	□		
Spine:	Kyphosis 0002808	□	Scoliosis 0002650	□		

GENITALIA

Undescended testes 0000028	□	Small penis 0000054	□	Hypospadias 0000047	□	Hypoplastic labia minora 0000064	□

NEURODEVELOPMENT / CENTRAL NERVOUS SYSTEM

Normal	□	Mild ID 0001256	□	Moderate ID 0002342	□	Severe ID 0010864	□
Hypertonia 0001276	□			Autism spectrum disorder 0000717	□		

SENSORY DEVELOPMENT

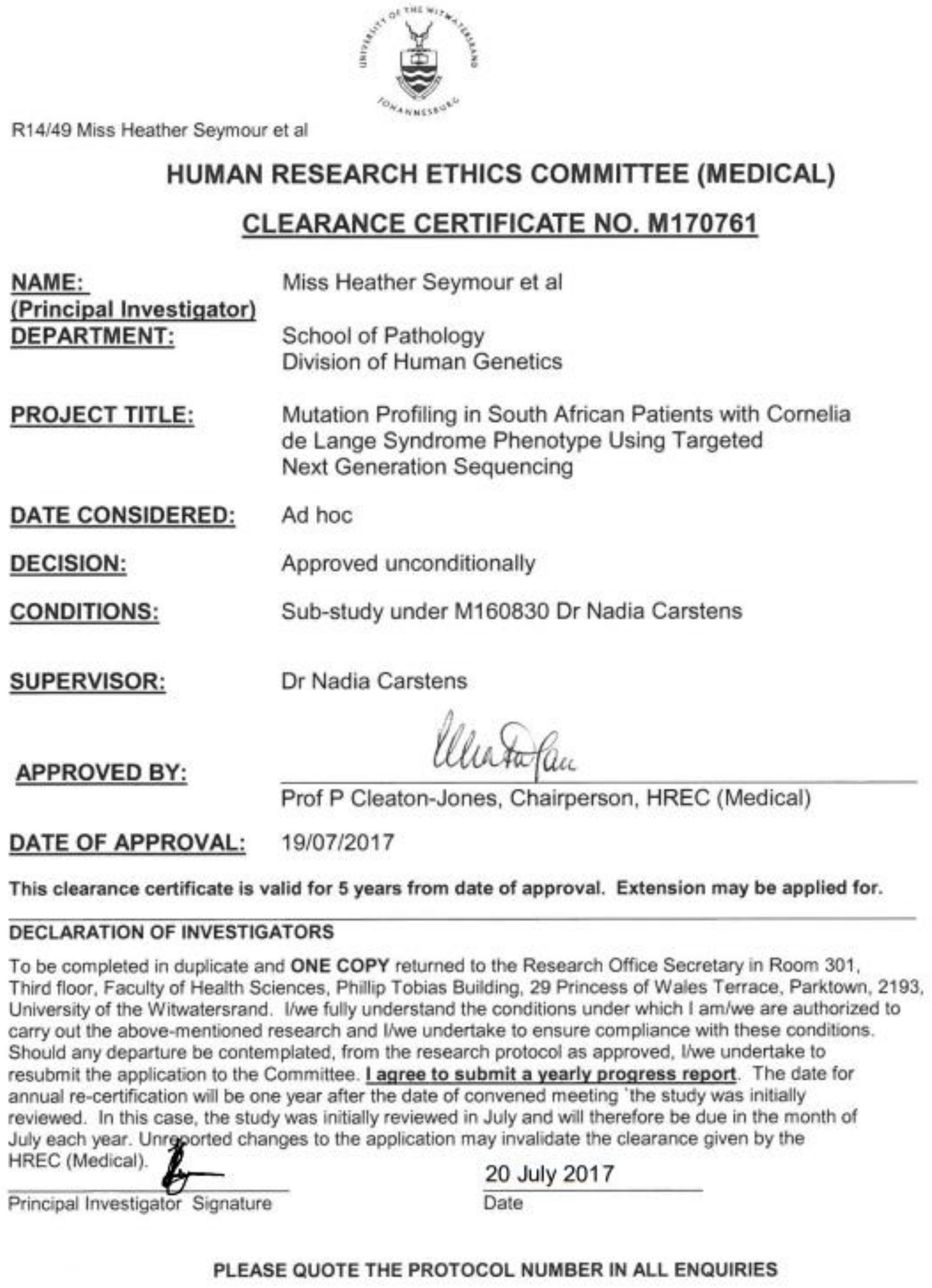
Vision:	Myopia 0000545	□	Strabismus 0000486	□		
Hearing:	Hearing loss	□	Conductive 0000405	□	Sensorineural 0000407	□

GASTROINTESTINAL SYSTEM

Gastro-oesophageal reflux 0002020	□	Malrotation 0002566	□

OTHER SIGNIFICANT ABNORMALITIES

Appendix B.1. Ethics Clearance Certificate obtained by Heather Seymour (M170761).



R14/49 Miss Heather Seymour et al

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M170761

NAME:
(Principal Investigator) Miss Heather Seymour et al

DEPARTMENT: School of Pathology
Division of Human Genetics

PROJECT TITLE: Mutation Profiling in South African Patients with Cornelia de Lange Syndrome Phenotype Using Targeted Next Generation Sequencing

DATE CONSIDERED: Ad hoc

DECISION: Approved unconditionally

CONDITIONS: Sub-study under M160830 Dr Nadia Carstens

SUPERVISOR: Dr Nadia Carstens

APPROVED BY: ______________________
Prof P Cleaton-Jones, Chairperson, HREC (Medical)

DATE OF APPROVAL: 19/07/2017

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

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary in Room 301, Third floor, Faculty of Health Sciences, Phillip Tobias Building, 29 Princess of Wales Terrace, Parktown, 2193, University of the Witwatersrand. I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated, from the research protocol as approved, I/we undertake to resubmit the application to the Committee. **I agree to submit a yearly progress report**. The date for annual re-certification will be one year after the date of convened meeting `the study was initially reviewed. In this case, the study was initially reviewed in July and will therefore be due in the month of July each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

______________________ 20 July 2017
Principal Investigator Signature Date

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

Appendix D.1. Pathogenic ACMG codes and descriptions (Table 3 taken from Richards et al., 2015).

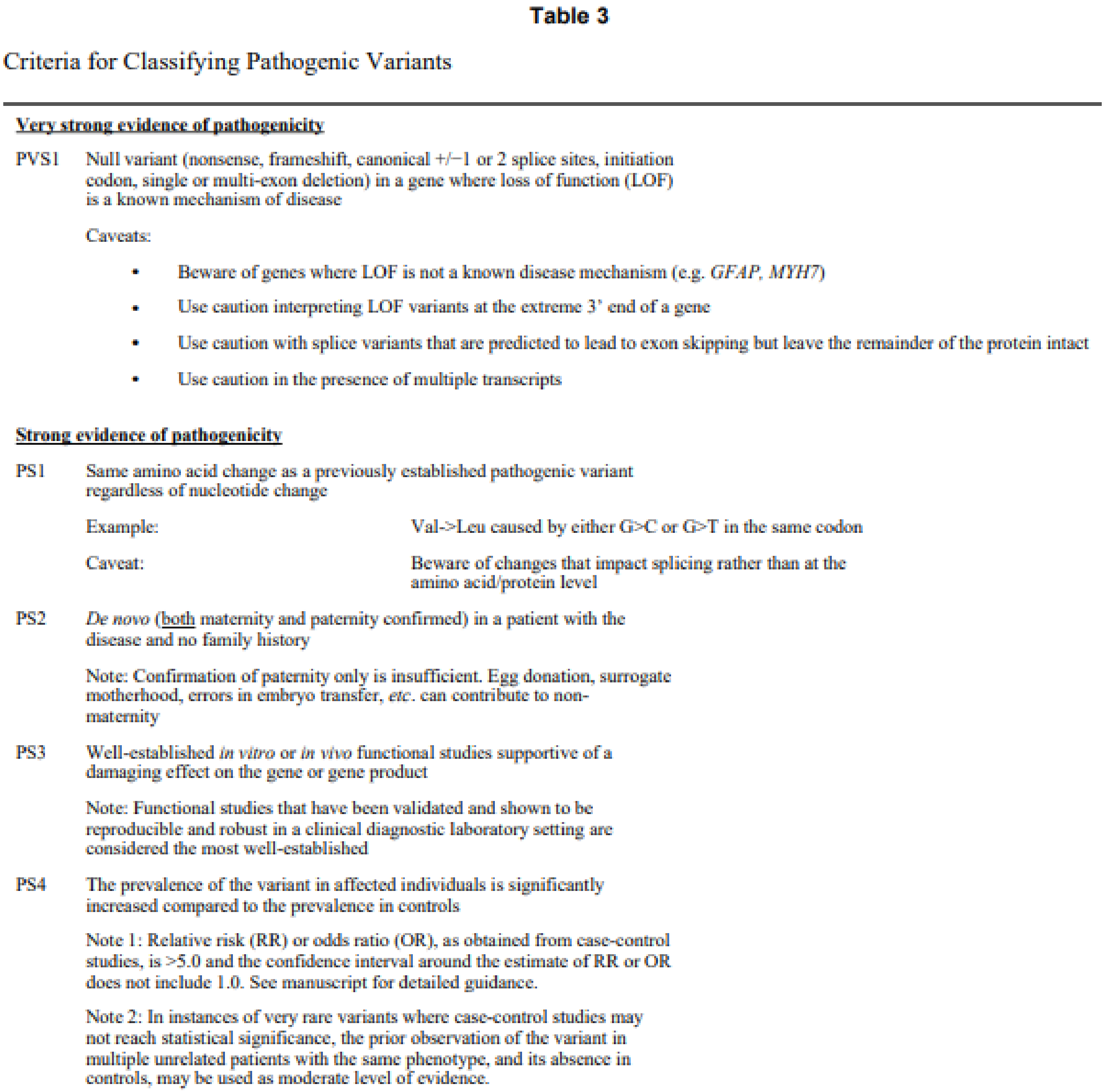


Table 3

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

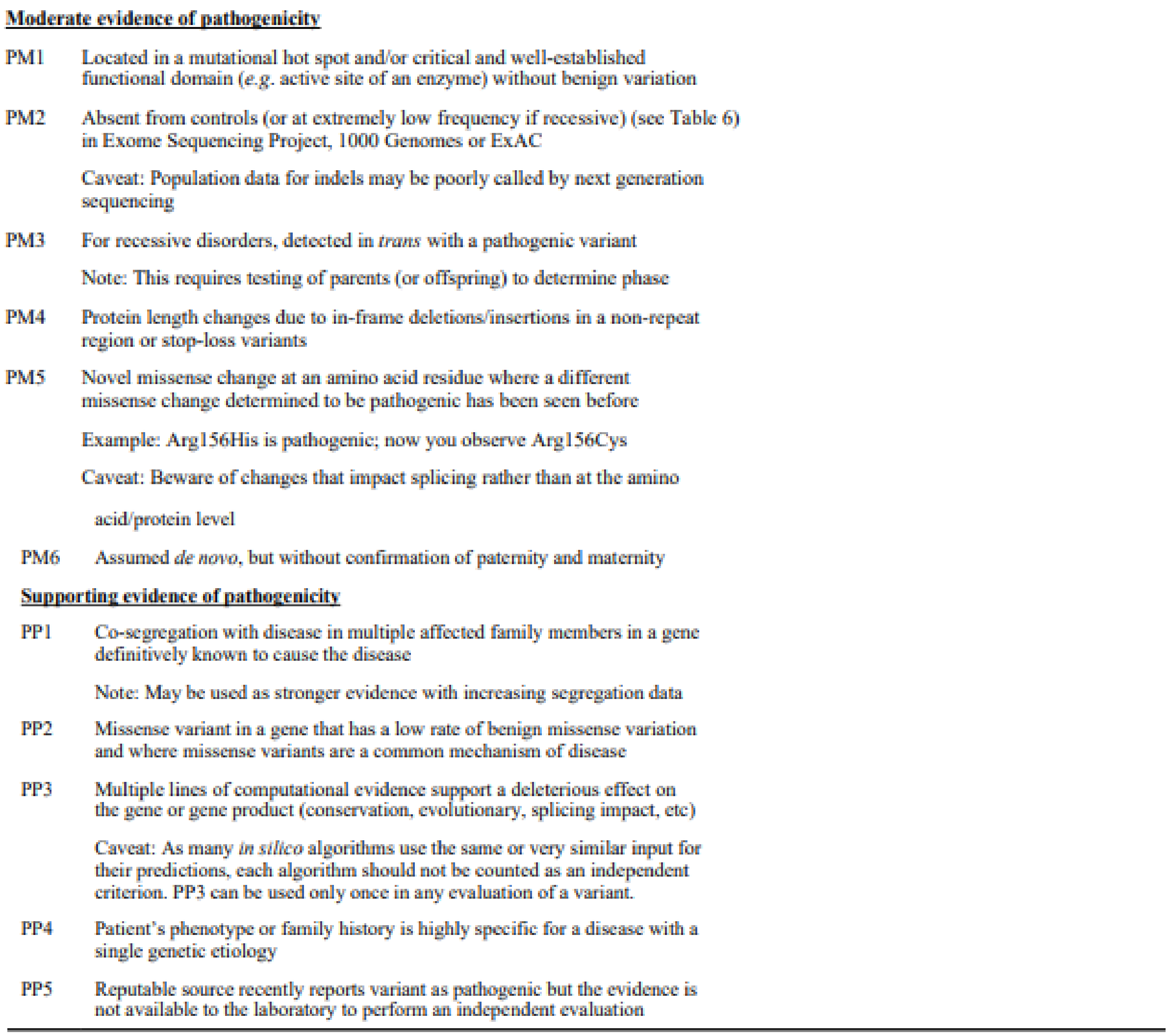
PS3 Well-established *in vitro* or *in vivo* functional studies supportive of a damaging effect on the gene or gene product

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

Appendix D.2. Benign ACMG codes and descriptions (Table 4 taken from Richards et al., 2015).

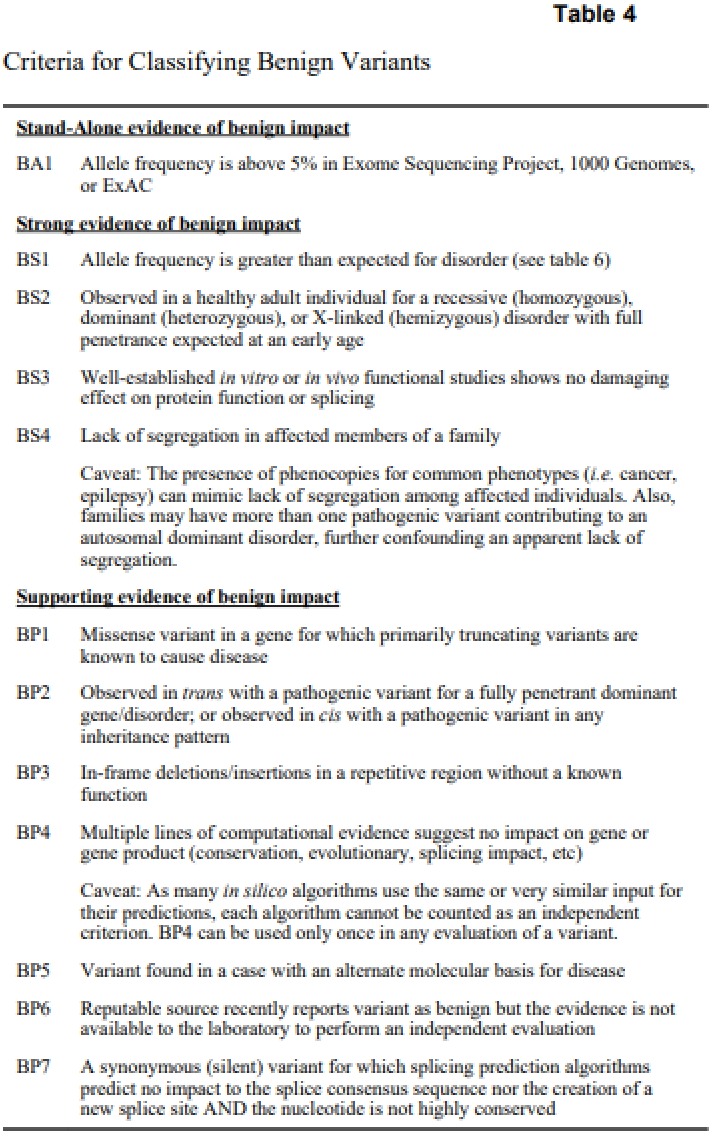


Table 4

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

Appendix D.3. ACMG Classification guidelines based on ACMG codes (Table 5 taken from Richards et al., 2015).

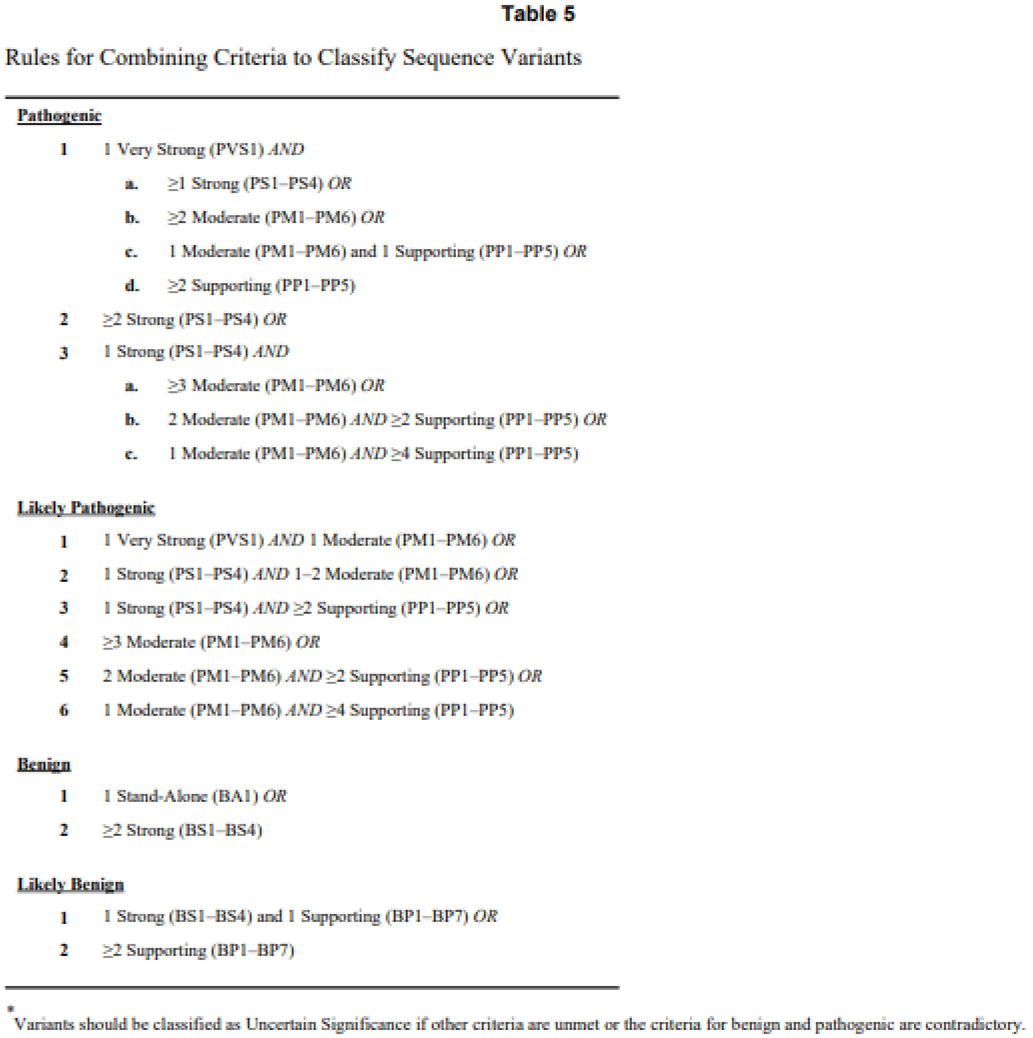


Table 5

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)

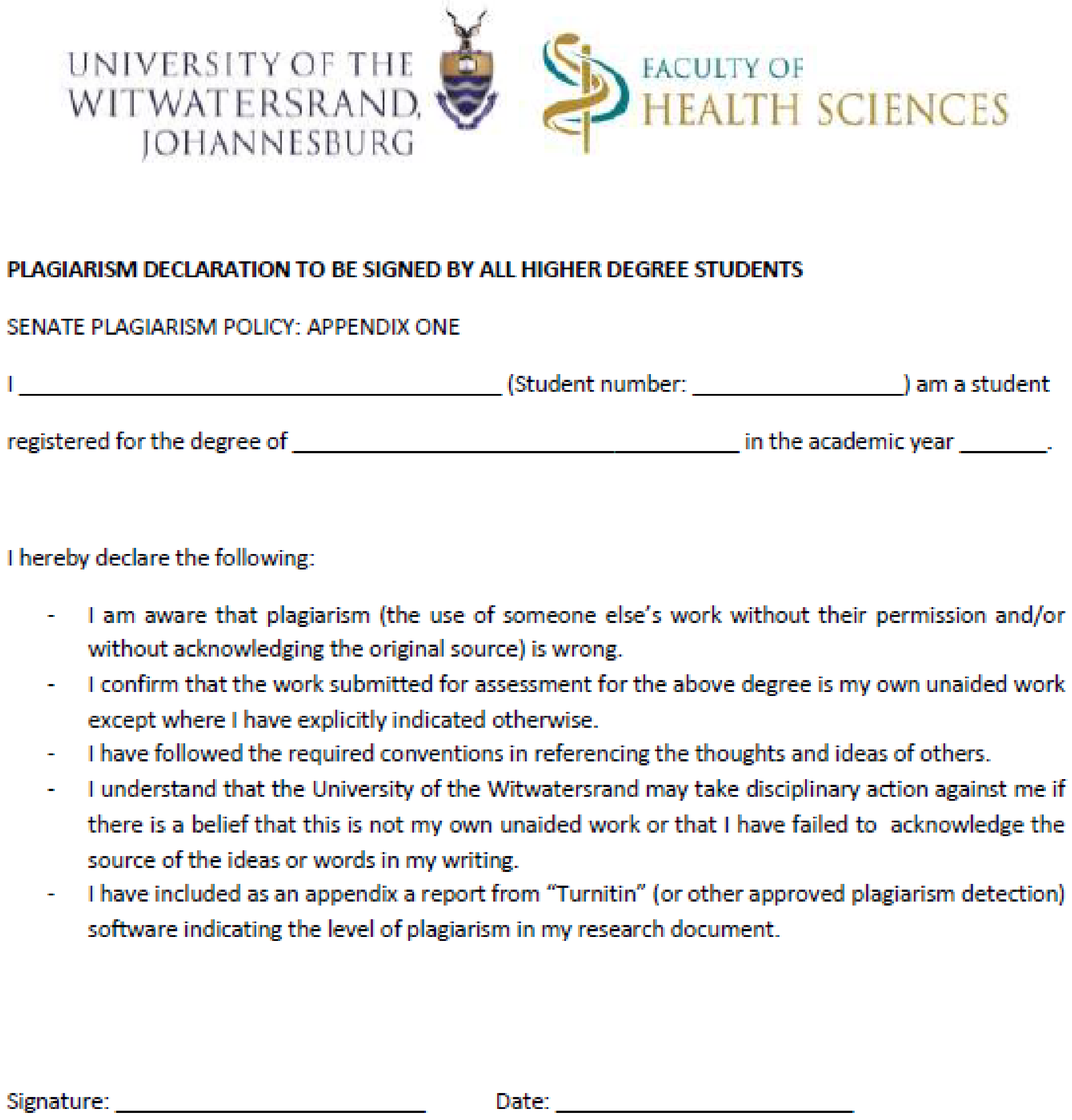
Likely Benign

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 E.1. Plagiarism Declaration



UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

FACULTY OF HEALTH SCIENCES

PLAGIARISM DECLARATION TO BE SIGNED BY ALL HIGHER DEGREE STUDENTS

SENATE PLAGIARISM POLICY: APPENDIX ONE

I ____ (Student number: __________________) am a student

registered for the degree of ______________________________________ in the academic year ________.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work except where I have explicitly indicated otherwise.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.
- I have included as an appendix a report from "Turnitin" (or other approved plagiarism detection) software indicating the level of plagiarism in my research document.

Signature: __________________________ Date: _________________________

List of Websites

Appendix C.1.: https://drive.google.com/file/d/1PgtS1p9CHotIbNceSbyaLkpEZXNL4iPQ/view?usp=sharing

CentoGene: https://www.centogene.com/

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

Illumina Sequence Analysis Viewer (SAV): https://support.illumina.com/sequencing/sequencing_software/sequencing_analysis_viewer_sav.html

Illumina Quality scores: https://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf

Integrative Genomics Viewer program (IGV): http://software.broadinstitute.org/software/igv/

Invitae: https://www.invitae.com/en/

Mutalyzer: https://mutalyzer.nl/

Primer3: http://bioinfo.ut.ee/primer3-0.4.0/

wANNOVAR tool: http://wannovar.wglab.org/index.php

References

Abou Tayoun, A. N. *et al.* (2018) ‘Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion’, *Human Mutation*, 39(11), pp. 1517–1524. doi: 10.1002/humu.23626.

Adzhubei, I. A. *et al.* (2010) ‘A method and server for predicting damaging missense mutations’, *Nature Methods*, 7(4), pp. 248–249. doi: 10.1038/nmeth0410-248.A.

Ajay, S. S. *et al.* (2011) ‘Accurate and comprehensive sequencing of personal genomes’, *Genome Research*, 21, pp. 1498–1505. doi: 10.1101/gr.123638.111.

Alazami, A. M. *et al.* (2015) ‘Accelerating novel candidate gene discovery in neurogenetic disorders via whole-exome sequencing of prescreened multiplex consanguineous families’, *Cell Reports*. The Authors, 10(2), pp. 148–161. doi: 10.1016/j.celrep.2014.12.015.

Alesi, V. *et al.* (2018) ‘Confirmation of BRD4 haploinsufficiency role in Cornelia de Lange–like phenotype and delineation of a 19p13.12p13.11 gene contiguous syndrome’, *Annals of Human Genetics*, (September), pp. 1–10. doi: 10.1111/ahg.12289.

Ansari, M. *et al.* (2014) ‘Genetic heterogeneity in Cornelia de Lange syndrome (CdLS) and CdLS-like phenotypes with observed and predicted levels of mosaicism’, *Journal of Medical Genetics*, 51(10), pp. 659–668. doi: 10.1136/jmedgenet-2014-102573.

Barbero, J. L. (2013) ‘Genetic basis of cohesinopathies’, *Application of Clinical Genetics*, 6, pp. 15–23. doi: 10.2147/TACG.S34457.

Barisic, I. *et al.* (2008) ‘Descriptive Epidemiology of Cornelia de Lange Syndrome in Europe’, *American Journal of Medical Genetics Part A*, 146A, pp. 51–59. doi: 10.1002/ajmg.a.

Begeman, G. and Duggan, R. (1976) ‘The Cornelia de Lange syndrome: a study of 9 affected individuals’, *S Afr Med J*, 50(38), pp. 1475–1478.

Brachmann, W. (1916) ‘En Fall von symmetrischer Monodaktylie durch Ulnadefekt, mit symmetrischer Flughautbildung in den CORNELIA DE LANGE SYNDROME IN EUROPE’, *Ellenbogen sowie anderen Abnormalita¨ten. J B Kinderheilk Phys Erzieh*, 84, pp. 225–235. doi: DOI 10.1002/ajmg.a.

Braunholz, D. *et al.* (2012) ‘Isolated NIBPL missense mutations that cause Cornelia de Lange syndrome alter MAU2 interaction’, *European Journal of Human Genetics*, 20(3), pp. 271–276. doi: 10.1038/ejhg.2011.175.

Braunholz, D. *et al.* (2015) ‘Hidden mutations in CdLS limitations of sanger sequencing in molecular diagnostics’, *Human Mutation*, 36(1), pp. 26–29. doi: https://doi.org/10.1002/humu.22685.

Chang, X. and Wang, K. (2012) ‘wANNOVAR: annotating genetic variants for personal genomes via the web’, *J Med Genet.*, 49(7), pp. 433–436. doi: 10.1136/jmedgenet-2012-100918.wANNOVAR.

Chien, R. *et al.* (2011) ‘Cohesin Mediates Chromatin Interactions That Regulate Mammalian β-globin Expression’, *Journal of Biological Chemistry*, 286(20), pp. 17870–17878. doi: 10.1074/jbc.m110.207365.

Choi, Y. and Chan, A. P. (2015) ‘PROVEAN web server: A tool to predict the functional effect of amino acid substitutions and indels’, *Bioinformatics*, 31(16), pp. 2745–2747. doi: 10.1093/bioinformatics/btv195.

Cicoria, A. (1974) ‘The Brachmann-De Lange Syndrome’, *S Afr Med J*, 48(May), pp. 919–921.

Davydov, E. V. *et al.* (2010) ‘Identifying a high fraction of the human genome to be under selective constraint using GERP++’, *PLoS Computational Biology*, 6(12). doi: 10.1371/journal.pcbi.1001025.

Deardorff, M. A. *et al.* (2007) ‘Mutations in Cohesin Complex Members SMC3 and SMC1A Cause a Mild Variant of Cornelia de Lange Syndrome with Predominant Mental Retardation’, *The American Journal of Human Genetics*, 80(3), pp. 485–494. doi: 10.1086/511888.

Deardorff, M. A., Bando, M., *et al.* (2012) ‘HDAC8 mutations in Cornelia de Lange Syndrome affect the cohesin acetylation cycle’, *Nature*, 489(7415), pp. 313–317. doi: 10.1038/nature11316.HDAC8.

Deardorff, M. A., Wilde, J. J., *et al.* (2012) ‘RAD21 mutations cause a human cohesinopathy’, *American Journal of Human Genetics*, 90(6), pp. 1014–1027. doi: 10.1016/j.ajhg.2012.04.019.

Deardorff, M. A., Noon, S. and Krantz, I. D. (2016) *Cornelia de Lange Syndrome*, *GeneReviews*.

Desmet, F.-O. *et al.* (2009) ‘Human Splicing Finder: an online bioinformatics tool to predict splicing signals’, *Nucleic Acids Research*, 37(9), pp. e67–e67. doi: 10.1093/nar/gkp215.

Dorsett, D. (1999) ‘Distant liaisons: long range enhancer–promoter interactions in Drosophila.’, *Curr Opin Genet Dev*, 9, pp. 505–514.

Dorsett, D. (2007) ‘Roles of the sister chromatid cohesion apparatus in gene expression, development, and human syndromes’, *Chromosoma*, 116(1), pp. 1–13. doi: 10.1158/0008-5472.CAN-10-4002.BONE.

Genohub (2019) *Beginner’s Handbook of Next Generation Sequencing*. Available at: https://genohub.com/next-generation-sequencing-handbook/.

Ghosh, R. *et al.* (2018) ‘Updated recommendation for the benign stand-alone ACMG/AMP criterion’, *Human Mutation*, 39(11), pp. 1525–1530. doi: 10.1002/humu.23642.

Gil-Rodriguez, M. C. *et al.* (2015) ‘De Novo Heterozygous Mutations in SMC3 Cause a Range of Cornelia de Lange Syndrome-Overlapping Phenotypes’, *Human Mutation*, 36(4), pp. 454–462. doi: 10.1002/humu.22761.

Gillis, L. A. *et al.* (2004) ‘NIPBL Mutational Analysis in 120 Individuals with Cornelia de Lange Syndrome and Evaluation of Genotype-Phenotype Correlations’, *American Journal of Human Genetics*, 75, pp. 610–623. Available at: papers2://publication/doi/10.1086/424698.

Goodwin, S., McPherson, J. D. and McCombie, W. R. (2016) ‘Coming of age: ten years of next-generation sequencing technologies’, *Nature Reviews Genetics*. Nature Publishing Group, 17(6), pp. 333–351. doi: 10.1038/nrg.2016.49.

Grozeva, D. *et al.* (2014) ‘De novo loss-of-function mutations in SETD5, encoding a methyltransferase in a 3p25 microdeletion syndrome critical region, cause intellectual disability’, *American Journal of Human Genetics*, 94(4), pp. 618–624. doi: 10.1016/j.ajhg.2014.03.006.

Hall, T. A. (1999) ‘BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT.’, *Nucl. Acids. Symp. Ser.*, 41, pp. 95–98. doi: citeulike-article-id:691774.

Harakalova, M. *et al.* (2012) ‘X-exome sequencing identifies a HDAC8 variant in a large pedigree with X-linked intellectual disability, truncal obesity, gynaecomastia, hypogonadism and unusual face’, *Journal of Medical Genetics*, 49(8), pp. 539–543. doi: 10.1136/jmedgenet-2012-100921.

Herrmann, J. *et al.* (1975) ‘The KBG syndrome-a syndrome of short stature, characteristic facies, mental retardation, macrodontia and skeletal anomalies.’, *Birth Defects Orig. Artic. Ser.*, 11, pp. 7–18.

Herrmann, J. and Opitz, J. M. (1977) ‘The SC Phocomelia and the Roberts Syndrome: Nosologic aspects’, *European Journal of Pediatrics*, 125, pp. 117–134.

Hirano, T. and Mitchison, T. J. (1994) ‘A heterodimeric coiled-coil protein required for mitotic chromosome condensation in vitro’, *Cell*, 79(3), pp. 449–458. doi: 10.1016/0092-8674(94)90254-2.

Hoppman-Chaney, N. *et al.* (2011) ‘In-frame multi-exon deletion of SMC1A in a severely affected female with Cornelia de Lange Syndrome’, *American Journal of Medical Genetics, Part A*, 158 A(1), pp. 193–198. doi: 10.1002/ajmg.a.34360.

Izumi, K. *et al.* (2015) ‘Germline Gain-of-Function Mutations in AFF4 Cause a Developmental Syndrome Functionally Linking the Super Elongation Complex and Cohesin’, *Nature Genetics*, 47(4), pp. 338–344. doi: 10.1055/s-0035-1556872.Free.

Jackson, L. *et al.* (1993) ‘de Lange syndrome: A clinical review of 310 individuals’, *American Journal of Medical Genetics*, 47(7), pp. 940–946. doi: 10.1002/ajmg.1320470703.

Jones, W. D. *et al.* (2012) ‘De novo mutations in MLL cause Wiedemann-Steiner syndrome’, *American Journal of Human Genetics*, 91(2), pp. 358–364. doi: 10.1016/j.ajhg.2012.06.008.

Kaiser, F. J. *et al.* (2014) ‘Loss-of-function HDAC8 mutations cause a phenotypic spectrum of Cornelia de Lange syndrome-like features, ocular hypertelorism, large fontanelle and X-linked inheritance’, *Human Molecular Genetics*, 23(11), pp. 2888–2900. doi: 10.1093/hmg/ddu002.

Kimura, K. and Hirano, T. (1997) ‘ATP-dependent positive supercoiling of DNA by 13S condensin: A biochemical implication for chromosome condensation’, *Cell*, 90(4), pp. 625–634. doi: 10.1016/S0092-8674(00)80524-3.

Kline, A. D. *et al.* (2018) ‘Diagnosis and management of Cornelia de Lange syndrome: first international consensus statement’, *Nature Reviews Genetics*. Springer US, 19(10), pp. 649–666. doi: 10.1038/s41576-018-0031-0.

Kopanos, C. *et al.* (2018) ‘VarSome: the human genomic variant search engine’, *Bioinformatics*, 3, pp. 10–12. doi: 10.1093/bioinformatics/bty897.

Koressaar, T. and Remm, M. (2007) ‘Enhancements and modifications of primer design program Primer3’, *Bioinformatics*, 23(10), pp. 1289–1291. doi: 10.1093/bioinformatics/btm091.

Krantz, I. D. *et al.* (2004) ‘Cornelia de Lange syndrome is caused by mutations in NIPBL, the human homolog of Drosophila melanogaster Nipped-B’, *Nature Genetics*, 36(6), pp. 631–635. doi: 10.1038/ng1364.

De Lange, C. (1933) ‘Sur un type nouveau de´ge´neration (typus Amstelodamensis)’, *Arch Med Enfant*, 36, pp. 713–719.

Lehalle, D. *et al.* (2017) ‘STAG1 mutations cause a novel cohesinopathy characterised by unspecific syndromic intellectual disability’, *Journal of Medical Genetics*, 54(7), pp. 479–488. doi: 10.1136/jmedgenet-2016-104468.

Lek, M. *et al.* (2016) ‘Analysis of protein-coding genetic variation in 60,706 humans’, *Nature*. Nature Publishing Group, 536(7616), pp. 285–291. doi: 10.1038/nature19057.

Lightfoot, J. *et al.* (2011) ‘Loading of meiotic cohesin by SCC-2 is required for early processing of DSBs and for the DNA damage checkpoint’, *Current Biology*. Elsevier Ltd, 21(17), pp. 1421–1430. doi: 10.1016/j.cub.2011.07.007.

Lin, C. *et al.* (2011) ‘Dynamic transcriptional events in embryonic stem cells mediated by the super elongation complex (SEC)’, *Genes and Development*, 25(14), pp. 1486–1498. doi: 10.1101/gad.2059211.

Lingen, M. *et al.* (2016) ‘Obtaining a genetic diagnosis in a child with disability: Impact on parental quality of life’, *Clinical Genetics*, 89(2), pp. 258–266. doi: 10.1111/cge.12629.

Liu, J. *et al.* (2009) ‘Transcriptional Dysregulation in NIPBL and Cohesin Mutant Human Cells’, *PLoS Biology*, 7(5), p. e1000119. doi: 10.1371/journal.pbio.1000119.

Liu, X. *et al.* (2016) ‘dbNSFP v3.0: A one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs’, *Human Mutation*, 37(3), pp. 235–241. doi: 10.1002/humu.22932.

Losada, A., Hirano, M. and Hirano, T. (1998) ‘Identification of Xenopus SMC protein complexes required for sister chromatid cohesion’, *Genes and Development*, 12(13), pp. 1986–1997. doi: 10.1101/gad.12.13.1986.

Lundin, S. *et al.* (2010) ‘Increased throughput by parallelization of library preparation for massive sequencing’, *PLoS ONE*, 5(4), p. e10029. doi: 10.1371/journal.pone.0010029.

Mannini, L. *et al.* (2013) ‘Mutation Spectrum and Genotype-Phenotype Correlation in Cornelia de Lange Syndrome', *Human Mutation*, 34(12), pp. 1589–1596. doi: 10.1002/humu.22430.

Miller, S. A., Dykes, D. D. and Polesky, H. F. (1988) 'A simple salting out procedure for extracting DNA from human nucleated cells', *Nucleic Acids Research*, 16(3), p. 1215. doi: 10.1093/nar/16.3.1215.

Mu, W. *et al.* (2016) 'Sanger Confirmation Is Required to Achieve Optimal Sensitivity and Specificity in Next-Generation Sequencing Panel Testing', *Journal of Molecular Diagnostics*. American Society for Investigative Pathology and the Association for Molecular Pathology, 18(6), pp. 923–932. doi: 10.1016/j.jmoldx.2016.07.006.

Musio, A. *et al.* (2006) 'X-linked Cornelia de Lange syndrome owing to SMC1L1 mutations', *Nature Genetics*, 38(5), pp. 528–530. doi: 10.1038/ng1779.

Newkirk, D. A. *et al.* (2017) 'The effect of Nipped-B-like (Nipbl) haploinsufficiency on genome-wide cohesin binding and target gene expression: modeling Cornelia de Lange syndrome', *Clinical Epigenetics*. Clinical Epigenetics, 9(1), pp. 1–20. doi: 10.1186/s13148-017-0391-x.

Ngcungcu, T. *et al.* (2017) 'Duplicated Enhancer Region Increases Expression of CTSB and Segregates with Keratolytic Winter Erythema in South African and Norwegian Families', *American Journal of Human Genetics*. ElsevierCompany., 100(5), pp. 737–750. doi: 10.1016/j.ajhg.2017.03.012.

O'Rawe, J. A. *et al.* (2015) 'TAF1 Variants Are Associated with Dysmorphic Features, Intellectual Disability, and Neurological Manifestations', *American Journal of Human Genetics*, 97(6), pp. 922–932. doi: 10.1016/j.ajhg.2015.11.005.

Oliver, C. *et al.* (2010) 'Cornelia de Lange syndrome: Extending the physical and psychological phenotype', *American Journal of Medical Genetics, Part A*, 152A(5), pp. 1127–1135. doi: 10.1002/ajmg.a.33363.

Olley, G. *et al.* (2018) 'BRD4 interacts with NIPBL and BRD4 is mutated in a Cornelia de Lange–like syndrome', *Nature Genetics*, 50(3), pp. 329–332. doi: 10.1038/s41588-018-0042-y.

Parenti, I. *et al.* (2017) 'Mutations in chromatin regulators functionally link Cornelia de Lange syndrome and clinically overlapping phenotypes', *Human Genetics*. Springer Berlin Heidelberg, 136(3), pp. 307–320. doi: 10.1007/s00439-017-1758-y.

Pezic, D., Weeks, S. L. and Hadjur, S. (2017) 'More to cohesin than meets the eye: complex diversity for fine-tuning of function', *Current Opinion in Genetics and Development*. Elsevier Ltd, 43, pp. 93–100. doi: 10.1016/j.gde.2017.01.004.

Ptacek, L. J. *et al.* (1963) 'The Cornelia de Lange syndrome', *The Journal of Pediatrics*, 63(5), pp. 1000–1020.

Quang, D., Chen, Y. and Xie, X. (2015) 'DANN: A deep learning approach for annotating the pathogenicity of genetic variants', *Bioinformatics*, 31(5), pp. 761–763. doi: 10.1093/bioinformatics/btu703.

Remeseiro, S. *et al.* (2012) ‘A unique role of cohesin-SA1 in gene regulation and development’, *EMBO Journal*, 31(9), pp. 2090–2102. doi: 10.1038/emboj.2012.60.

Rentzsch, P. *et al.* (2019) ‘CADD: predicting the deleteriousness of variants throughout the human genome’, *Nucleic Acids Research*. Oxford University Press, 47(D1), pp. D886–D894. doi: 10.1093/nar/gky1016.

Richards, S. *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(5), pp. 405–423. doi: 10.1038/gim.2015.30.

Rollins, R. A., Morcillo, P. and Dorsett, D. (1999) ‘Nipped-B, a Drosophila homologue of chromosomal adherins, participates in activation by remote enhancers in the cut and Ultrabithorax genes’, *Genetics*, 152(2), pp. 577–593.

Sanger, F. *et al.* (1977) ‘Nucleotide sequence of bacteriophage φX174 DNA’, *Nature*, 265, pp. 687–695.

Santen, G. W. E. *et al.* (2012) ‘Mutations in SWI/SNF chromatin remodeling complex gene ARID1B cause Coffin-Siris syndrome’, *Nature Genetics*. Nature Publishing Group, 44(4), pp. 379–380. doi: 10.1038/ng.2217.

Schwarz, J. M. *et al.* (2014) ‘MutationTaster2: mutation prediction for the deep-sequencing age’, *Nature Methods*, 11(4), pp. 361–362. doi: 10.1038/nmeth.2890.

Shihab, H. A. *et al.* (2012) ‘Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models’, *Human Mutation*, 34(1), pp. 57–65. doi: 10.1002/humu.22225.

Sim, N. L. *et al.* (2012) ‘SIFT web server: Predicting effects of amino acid substitutions on proteins’, *Nucleic Acids Research*, 40(W1), pp. 452–457. doi: 10.1093/nar/gks539.

Sjögren, C. and Nasmyth, K. (2001) ‘Sister chromatid cohesion is required for postreplicative double-strand break repair in Saccharomyces cerevisiae’, *Current Biology*, 11(12), pp. 991–995. doi: 10.1016/S0960-9822(01)00271-8.

Taylor, A. *et al.* (2018) ‘Consensus for genes to be included on cancer panel tests offered by UK genetics services: guidelines of the UK Cancer Genetics Group’, *Journal of Medical Genetics*, 55, pp. 372–377. doi: 10.1136/jmedgenet-2017-105188.

Tonkin, E. T. *et al.* (2004) ‘NIPBL, encoding a homolog of fungal Scc2-type sister chromatid cohesion proteins and fly Nipped-B, is mutated in Cornelia de Lange syndrome’, *Nature Genetics*, 36(6), pp. 636–641. doi: 10.1038/ng1363.

Tsurusaki, Y. *et al.* (2012) ‘Mutations affecting components of the SWI/SNF complex cause Coffin-Siris syndrome’, *Nature Genetics*. Nature Publishing Group, 44(4), pp. 376–378. doi: 10.1038/ng.2219.

Untergasser, A. *et al.* (2012) ‘Primer3-new capabilities and interfaces’, *Nucleic Acids Research*, 40(15), pp. 1–12. doi: 10.1093/nar/gks596.

Vega, H. *et al.* (2005) ‘Roberts syndrome is caused by mutations in ESCO2, a human homolog of yeast ECO1 that is essential for the establishment of sister chromatid cohesion’, *Nature Genetics*, 37(5), pp. 468–470. doi: 10.1038/ng1548.

Vega, H. *et al.* (2010) ‘Phenotypic variability in 49 cases of ESCO2 mutations, including novel missense and codon deletion in the acetyltransferase domain, correlates with ESCO2 expression and establishes the clinical criteria for Roberts syndrome’, *Journal of Medical Genetics*, 47, pp. 30–37. doi: 10.1136/jmg.2009.068395.

Vrolick, W. (1849) *Tabulae ad illustrandam embryogenesin hominis et mammalium tam naturalem quasm abnormem, Amsterdam: Londonck.*

Watrin, E. *et al.* (2006) ‘Human Scc4 Is Required for Cohesin Binding to Chromatin, Sister-Chromatid Cohesion, and Mitotic Progression’, *Current Biology*, 16(9), pp. 863–874. doi: 10.1016/j.cub.2006.03.049.

Wildeman, M. *et al.* (2009) ‘Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker’, *Human Mutation*, 29(1), pp. 6–13. doi: 10.1002/humu.

Yohe, S. *et al.* (2015) ‘Clinical validation of targeted next-generation sequencing for inherited disorders’, *Archives of Pathology and Laboratory Medicine*, 139(2), pp. 204–210. doi: 10.5858/arpa.2013-0625-OA.

Yuan, B. *et al.* (2015) ‘Global transcriptional disturbances underlie Cornelia de Lange syndrome and related phenotypes’, *Journal of Clinical Investigation*, 125(2), pp. 636–651. doi: 10.1172/jci77435.

Yuan, B. *et al.* (2018) ‘Clinical exome sequencing reveals locus heterogeneity and phenotypic variability of cohesinopathies’, *Genetics in Medicine*. Springer US, 0(0), pp. 1–13. doi: 10.1038/s41436-018-0085-6.

Zerbino, D. R. *et al.* (2018) ‘Ensembl 2018’, *Nucleic Acids Research*, 46(D1), pp. D754–D761. doi: 10.1093/nar/gkx1098.

Zhang, A., Yeung, P. L. and Li, C. W. (2004) ‘Identification of a novel family of ankyrin repeats containing cofactors for p160 nuclear receptor coactivators.’, *J Biol Chem*, 279, pp. 33799–33805.

Zhang, B. *et al.* (2007) ‘Mice lacking sister chromatid cohesion protein PDS5B exhibit developmental abnormalities reminiscent of Cornelia de Lange syndrome’, *Development*, 134(17), pp. 3191–3201. doi: 10.1242/dev.005884.

Zhang, B. *et al.* (2009) ‘Dosage Effects of Cohesin Regulatory Factor PDS5 on Mammalian Development: Implications for Cohesinopathies’, *PLoS ONE*, 4(5), p. e5232. doi: 10.1371/journal.pone.0005232.