

---

**An investigation into the genetic basis  
of autosomal recessive Osteogenesis  
Imperfecta (OI) III in a South African  
family of mixed ancestry**

---

Masters Thesis  
of  
**Susan Alicia Fernol**



A thesis submitted in fulfillment of the requirements for the  
degree Master of Science at the South African National  
Bioinformatics Institute in the Faculty of Natural Science,  
University of the Western Cape.

**Supervisor: Professor Alan Christoffels**  
**Co-Supervisor: Professor Manogari Chetty**

**Submission date: 2022**

# Keywords

Collagen

Genotype

Heterogeneity

Mixed ancestry heritage

Next-generation sequencing

Osteogenesis imperfecta

Phenotype

South African



UNIVERSITY *of the*  
WESTERN CAPE

# Abstract

**Background:** Osteogenesis Imperfecta (OI) is a rare skeletal dysplasia that is primarily characterized by bone fragility, recurrent fractures, and bone deformities. Over the years there has been an increase in the number of genes associated with OI. Currently there are twenty causative genes involved in OI spread across an autosomal dominant form, autosomal recessive form, and an X-linked form.

Among the different types of OI, the progressively deforming OI, has more than one causative OI gene associated with it, and both AD and AR mode of inheritance. A severe autosomal recessive form of OI type III has been studied in SA for more than 40 years. OI type3 has an estimated prevalence of 1 per 125 000 to 1 per 200 000 in the Black (referring to the various linguistic groups: Sesotho, Xhosa, Zulu, Tsonga, Venda) African population of SA. This high prevalence could be explained by the unaffected heterozygote having a biological advantage in the African environment and that the variant for OI type III in Africa occurred more than 2000 years ago in West and Central Africa, prior to migration to present day southern Africa.

**Method:** Whole genome sequencing (WGS) was used to understand the genetic basis of a progressively deforming, severe Osteogenesis Imperfecta in these twins of mixed ancestry. The WGS data was used to identify variants in all the known OI disease-causing genes and for the identification of novel variants in other parts of the genome.

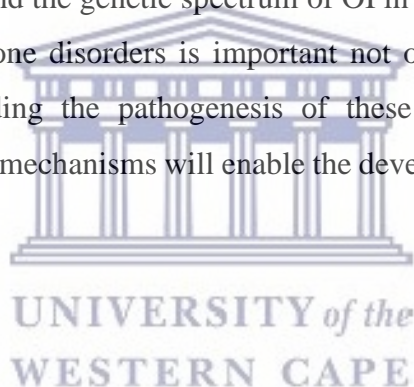
This study was approved by the Biomedical Research Ethics Committee (BMREC) of the University of the Western Cape. Saliva collection and DNA extraction was performed on four family members. The extracted DNA was submitted to Beijing Genomic Institute for sequencing on the BGISEQ 100PE with X30 coverage. Raw paired-end reads were subjected to quality trimming and adapter sequence removal conducted by BGI. Local FastQC analysis was performed on these reads. High quality reads were aligned to the GRCH38 of the human genome reference using BWA-MEM. The reads were then converted, sorted and PCR duplicates were marked. This was followed by GATK's quality score recalibration, Indel realignment, SNP and Indel discovery, and standard VSQR filtering. Finally, Variant annotation and variant filtering were performed. The filtering steps included mode of inheritance, SAHGP population filter, SNP and InDel separation, gene level annotations, disease/phenotype filter, protein-protein interaction, and SNP prediction. The study continued to look only at the SNPs in coding regions of the genome.

**Results:** The twins were between the ages of two and six when they had their first fractures, which occurred in their femurs. They have each had more than 100 fractures in multiple bones, these included their tibias, jaws, and ribs. Both twins had bowing of the long bones, vertebral fractures, and scoliosis. Because of the overlapping clinical manifestation of OI types I to IV with types V to XX, it becomes challenging to predict or diagnose the type of OI based solely on the phenotype.

This study did not identify any disease-causing variants in the *FKBP10* genes, like previous studies in the African population. This study reports on the identification of a promising *de novo* missense variant in the *COL1A2* gene (c.1478G>T, p.(Gly493Val)). According to ACMG guidelines this variant is classified as pathogenic based on various evidence and supporting criteria.

**Conclusion:** The results from this study indicates that the variant in the *COL1A2* could be reason for the phenotype observe and thus the cause of OI in this family.

The findings in this study expand the genetic spectrum of OI in mixed ancestry populations of South Africa. Studying rare bone disorders is important not only for arriving at a specific diagnosis but also understanding the pathogenesis of these conditions – only increased understanding of the molecular mechanisms will enable the development of targeted therapies.



# Declaration

The work in this thesis is based on research carried out at the South African National Bioinformatics Institute, Faculty of Science, University of the Western Cape.

I declare that '*An investigation into the genetic basis of Osteogenesis Imperfecta in a South African family of mixed ancestry*' is my own work.

I declare that it has not been submitted before for any degree or examination in any other university, and that all the sources I have used or quoted have been indicated and acknowledged as complete references.



Full Name: Susan Alicia Fernol

Date: August 2022.

Signed: ..........

# Acknowledgements

Ek wil graag van hierdie geleentheid gebruik maak, om ons Hemel se Vader te dank, vir sy genade en die krag wat hy my gegee het om hierdie graad te voltooi, en dat hy my nog elke dag deurdra.

Tweedens, will ek vir my Ma (Falica Gelderbloem) dankie se dat sy my by gestaan het deur alles. Dankie vir al Mammie opheringe oor die jare, ek weet dit was nie altyd maklik nie.

I would like to thank my supervisor, Professor Alan Christoffels, for giving me the opportunity to expand on the little knowledge I had in the bioinformatics field. This opportunity not only presented me with new skills and knowledge, it also allowed me to meet some really great people along the way. I thank you for all the support and guidance.

To Professor Manogari Chetty, my co-supervisor, thank you for having me as part of your Dental genetics group.

To my funding bodies, the South African DST/NRF research chairs initiative and the South African Research Chairs initiative, thank you for making this degree financially possible.

I would like to extend a thanks to our colleagues at the NSB Biobank who assisted with the DNA extraction and storing thereof.

I would also like to extend a hearty, thank you to the staff members at SANBI. You made me feel at home right from the start of this degree.

I would also like to thank my partner, Marshall Bryton Fisher. Thank you for all your support, all the late nights you sat up with me while I was working. Thank you for always believing in me and my goals.

Lastly, I would like to thank myself. Thank you for never giving up no matter how tough things got, through all the headaches, being stuck at certain points in the project, sleepless nights. Thank you for finding the excitement in the project and for enjoying every aspect this project had to offer. Thank you for being an inspiration and motivation to yourself.

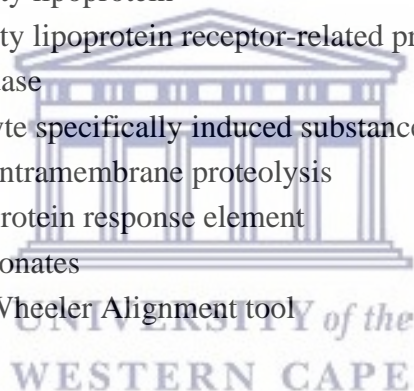
# Abbreviations

## General abbreviations

<b>Abbreviation</b>	<b>Definition</b>
OI	Osteogenesis imperfecta
DI	Dentinogenesis imperfecta
DNA	Deoxyribonucleic acid
SNP	Single nucleotide polymorphs, common genetic variant
Indel	Insertions/deletions in a gene or gene region
PCR	Polymerase chain reaction, increases amount of a defined DNA sequence
GWAS	Genome wide association studies
NGS	Next-generation sequencing
CADD	the Combined annotation dependent depletion
GWAVA	Genome wide annotation of variants
DNN	Deep neutral network
SVM	Support vector machine
nsSNV	Non-synonymous single nucleotide variants
SO	Sequence ontology
UWC	University of the Western Cape
SAM	Sequence alignment map
BAM	Binary alignment map
GATK	Genome Analysis Toolkit
BQSR	Base quality score recalibration
VEP	Variant Effect Predictor
FN	Fibronectin
ECM	Extracellular matrix
IRDIRC	International Rare Diseases Research Consortium
AD	Autosomal dominant
AR	Autosomal recessive
SA	South Africa
CMA	Cape Mixed Ancestry

(Continued)

<b>Abbreviation</b>	<b>Definition</b>
NGS	Next generation sequencing
WGS	Whole genome sequencing
WES	Whole exome sequencing
Gly	Glycine
NCPs	non-collagen proteins
ECM	extracellular matrix
BMD	bone mineral density
DXA	dual-energy absorptiometry
RNA	Ribonucleic acid
ER	Endoplasmic reticulum
PPIase	peptidyl-propyl cis/trans isomerase
SLRP	small leucine-rich protein
PEDF	pigment epithelium-derived factor
P4HB	propyl-4 hydroxylase-B
HSP47	Heat-shock protein 47
LDL protein	Low-density lipoprotein
LRP5	Low-density lipoprotein receptor-related protein 5
LOX	Lysyl oxidase
OASIS	old astrocyte specifically induced substance
RIP	regulated intramembrane proteolysis
UPRE	unfolded protein response element
BP	Bisphosphonates
BWA	Burrows-Wheeler Alignment tool





## Abbreviations for genes and proteins associated with Osteogenesis Imperfecta

<b>Abbreviation</b>	<b>Definition</b>
<i>BMP1</i>	Gene coding for bone morphogenetic protein-1 (BMP1), responsible for C-terminal cleavage of collagen (OI type XIII)
<i>COL1A1</i>	Gene coding for collagen type 1 alpha 1
<i>COL1A2</i>	Gene coding for collagen type 1 alpha 2
<i>CREB3L1</i>	Gene coding for OASIS, a regulator of type I procollagen (proposed OI type XVI)
<i>CRTAP</i>	Gene coding for cartilage-associated protein (CRTAP), involved in hydroxylation of certain collagen proline residues (OI type VII)
<i>FKBP10</i>	Gene coding for 65kDa FK-binding protein (FKBP65), a collagen chaperone (OI type XI)
<i>IFITM5</i>	Gene coding for interferon induced transmembrane protein 5, involved in bone formation (OI type V)
<i>LEPRE1</i>	Gene coding for prolyl 3-hydroxylase (P3H1), involved in hydroxylation of certain collagen proline residues (OI type VIII)
<i>PIIB</i>	Gene coding for cyclophilin B (CyPB), involved in C-terminal folding, trimer and helix formation and hydroxylation (OI type IX)
<i>SERPINF1</i>	Gene coding for pigment epithelium-derived factor, involved in bone mineralization (OI type VI)
<i>SERPINH1</i>	Gene coding for heat shock protein 47(HSP47), a collagen chaperone (OI type X)
<i>SPARC</i>	Gene coding for cysteine-rich acidic matrix-associated protein (SPARC), involved in extracellular matrix synthesis and promotion of changes to cell shape. Required for collagen in bone (OI type XVII)
<i>SP7</i>	Gene coding for osteoblast transcription factor, osterix (OI type XII)
<i>TMEM38B</i>	Gene coding for a channel involved in Ca <sup>2+</sup> release (OI type XIV)
<i>PIIB</i>	Gene coding for cyclophilin B (CyPB), involved in C-terminal folding, trimer and helix formation and hydroxylation (OI type IX)
<i>WNT1</i>	Gene coding for Wnt1, inducer of a major pathway in bone formation (OI type XV)

# 1 Chapter One: Introduction

## 1.1 Background

### 1.1.1 Rare Disease

A rare disease (RD) is a health condition affecting a small number of people in a population, compared to other diseases that are commonly identified in the population. The definition of RDs differ depending on where you are in the world, and there is currently no globally accepted definition<sup>1</sup>. The definition is based on a prevalence threshold. For instance, in the United States the definition of RDs is 1 per 200 000 of the population, whereas in Europe it is 1 per 2 000 of the population<sup>1,2</sup>. To date there is currently no official definition for RDs in South Africa (SA), as a result SA has adopted the Europe Union (EU) definition for RDs<sup>2</sup>. There are over 7000 RDs recorded and 72% are due to genetic factors, affecting at least 3.5-5.9% of populations worldwide<sup>1,2</sup>.

### 1.1.2 Osteogenesis Imperfecta

Osteogenesis Imperfecta (OI) is a genetically rare heterogeneous group of heritable disorders of the bones' connective tissue<sup>3-7</sup>. This condition is also known as "Fragilitas ossium", "Vrolijk disease" but is most often referred to as "brittle bone disease"<sup>8-12</sup>. OI is primarily characterized by bone fragility, short stature and recurrent fractures, as well as additional characteristics which may include hearing loss<sup>3-5,10,13,14</sup>, blue/grey sclerae, abnormal tooth development (dentinogenesis imperfecta), bone deformities (kyphoscoliosis, bowing of long bones), joint hypermobility<sup>11,12,15,16</sup>, hypotonia, craniofacial abnormalities, pulmonary function impairment<sup>11,15</sup> and cardiac valve abnormalities<sup>11,15</sup>

OI was originally classified into four types (types I-IV) based on clinical appearance and pattern of inheritance<sup>12,17,18</sup>. The original four types are in majority (more than 80%) of cases due to variations in the *COL1A1* and *COL1A2* genes<sup>13,15,18</sup>, transmitted in an autosomal dominant (AD) manner. The discovery of new subtypes lead to a number of revisions of the classification and nosology of OI<sup>5,18</sup>. These revisions resulted in expanding the original classification to include autosomal recessive (AR) inherited types (types VI-XVIII and XX) and X-linked type (type XIX). The AR form is due to variations found in 14 genes (*BMP1*, *SERPINF1*, *CRTAP*, *P3H1*, *PIIB*, *SERPINH1*, *FKBP10*, *TMEM38B*, *WNT1*, *CREB3L1*, *SPARC*, *TENTA5* (*FAM46A*), *MESD*, and *SP7*)<sup>10,18</sup> and two genes (*MBTPS2* and *PLS3*) in the X-linked form<sup>10,19,20</sup>.

The global estimated prevalence of OI ranges from 1/15 in 10 000 to 20 000 births<sup>21</sup>. Studies conducted in Europe and United States have an estimated prevalence of 0.3-0.7 per 10 000 births<sup>11</sup>. In the Black population of SA, Beighton and Versfeld, estimated the prevalence of OI type 1 to be 0.1 per 100 000 and for OI type 3 to be 0.6 per 100 000<sup>22</sup>. The high prevalence may be due to the unaffected heterozygote having a biological advantage in the African environment and the variant for this ‘OI type III’ in Africa occurred more than 2000 years ago in West and Central Africa prior to migration to present day southern Africa. This biological advantage may have allowed the population to adapt and reproduce and thrive in that specific environment. Over time the variant possibly became common in these environments, but with migration to southern Africa the population became different to the previous or ancestral population. As a result, the very same variant that may have been advantageous before has now become harmful in the southern African population. Moreover, the dominant trait was expressed in this population prior to migration, which could also mean because the variant had reduced penetrance probably resulting from a combination of genetic, environmental, and lifestyle factors.

According to Stephen *et al*<sup>4</sup>, a severe form of OI has been studied in South Africa for more than 40 years. This severe form of OI was presumed to be ‘AR OI type III’<sup>4</sup>. It was not until 2010, when Alanay *et al*<sup>23</sup>, identified a causative gene (*FKBP10*) closely resembling the various clinical and radiographic features of this severe form of ‘OI type III’.

This thesis notes that several authors switch between using ‘AR OI type III’<sup>4,24</sup>, ‘OI type III’<sup>25-27</sup> and/or ‘OI III’<sup>4,24</sup>, while other authors make use of ‘OI-3’ and/or ‘AR OI type 3’<sup>5,6</sup>. The authors were all making reference to the ‘progressively deforming OI’ based on their research patient’s clinical features. In other cases where ‘AR OI’ were used, it was with reference to a corresponding genotype<sup>28-32</sup>, indicating the genotype/molecular type of OI by the use of roman numerals. For the purpose of this thesis, ‘OI type 3’, with reference the ‘progressively deforming OI’, will be used.

### 1.1.3 OI type 3

OI type 3 has become the focus of medical and scientific interest. It is a genetically heterogenous form of OI that can be transmitted in either AD or AR pattern of inheritance<sup>5</sup>. In the South African population, OI type 3 is predominantly AR. The frequency for this severe form of OI is particularly high in the Black populations of southern Africa<sup>5,24</sup>. Due to the

clinical and radiological similarities in each of these African populations of southern Africa, a common variant (s) was proposed<sup>5</sup>.

In 2017, Vorster *et al*<sup>5</sup>, investigated the molecular basis of clinically diagnosed OI type 3 in the Black African population of SA. This study specifically looked at regions in the *FKBP10* gene, since disease-causing variants had previously been reported in this gene with a similar clinical description<sup>23,33</sup>. The study identified disease-causing variants in 45.1% (41/90) of the affected individuals. A frameshift variant (c.831dupC) was identified homozygous in 38.5% (35/91) of the individuals in the study cohort, which had previously been identified in two South African Venda individuals<sup>33</sup>. They also identified a novel frameshift variant (c.831delC) inherited in a compound heterozygous manner in four of the affected individuals. The frameshift variants both introduce premature termination of the amino acid sequence. Two novel nonsense variants (c.343C>T and c.1621C>T) that are predicted to truncate FKBP65 by 467 and 41 amino acids, respectively. Based on the haplotype analysis, the study suggest that the variant found in these populations is identical by descent, and the initial occurrence of the variant likely happened before the divergence of the linguistic groups in the study<sup>5</sup>. This study also suggests that because no variants were found for the remaining 50 affected individuals, with identical phenotypes, there may be more potential disease-causing variants in other regions of the *FKBP10* gene or even other genes, and that more research is required for clarification of the genetic heterogeneity observed in the *FKBP10* gene<sup>5</sup>.

In 2018, Chetty *et al*<sup>24</sup>, reported on the unusual phenotypic features of six individuals of Cape Mixed Ancestry (CMA) decent. The authors documented the severe DI and longevity, which were absent from majority of the African individuals with the progressively deforming OI. The affected persons were also negative for the variants in the *FKBP10* gene that was frequently identified in the Black African individuals<sup>5</sup>. Each of the CMA individuals has limited oral opening and all their teeth were discoloured, consistent with moderate to severe DI. Their sclerae were white and no hearing loss was observed. None of these individuals had received bisphosphonate therapy<sup>24</sup>.

The affected CMA twins are the focus of this thesis. They have unaffected, non-consanguineous parents, no history of OI in any of their parents' progenitors. The CMA twins were short in stature, and both had marked kyphoscoliosis (Figure 1.1, Panel A) and severe DI (Figure 1.1, Panel B & C)<sup>24</sup>. It must be emphasized that the CMA twins were *not managed by a clinician(s) or geneticist, no permission was given to work with or obtain their clinical*

information. As such the patients provided us with their clinical information and permission required and given was for the sequencing of their genomes.

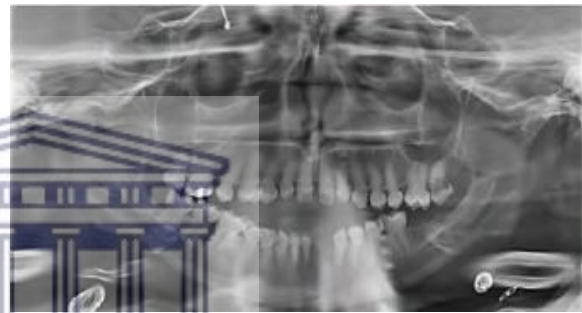
A.



B.



C.



**Figure 1.1:** Clinical presentation of OI type 3 and severe DI in twin sisters. Panel A: Affected twin sisters aged 28 (when picture was taken) and their unaffected mother; Panel B: An intraoral picture of the one twin; Panel C: A panorex of the second twin. Image obtained from Chetty et al.<sup>24</sup> (informed consent was given for the use of the images in thesis/publications)

#### 1.1.4 Rare genetic diseases and next-generation sequencing

Clinical manifestations that facilitate the recognition of the diseases have limitations in diagnostics of rare disease, since genetic disorders usually display wide phenotypic heterogeneity. Confirming the diagnosis at the molecular level is essential in modern practice for appropriate management of patient and patient's family<sup>34</sup>.

The advent of next-generation sequencing (NGS) radically changed the diagnostic workflow of rare genetic disease analysis by providing a rapid, powerful, and cost effective alternative to the traditional (Sanger sequencing) approach<sup>35-38</sup>. Processes that use to be chemically time-consuming, can be accomplished in a few weeks or even less with the NGS-based tools. These

NGS-based tools can point to the implication of a single gene (or a small number of genes) and help to establish a rapid diagnosis in a considerable percentage of cases<sup>35</sup>.

There are three major NGS-based tests used in the study of rare diseases. 1) Parallel sequencing of the coding sequences (exons) of related gene sets according to similar or overlapping phenotypes (gene panel). 2) Whole-exome sequencing (WES) where all known coding regions of the human genome are sequenced. 3) Whole-genome sequencing (WGS), which analyzes the entire human genome<sup>35</sup>. Strategies used to identify pathogenic genes include; 1) WES or WGS analysis of a group of patients with similar clinical features, and filtering variants located in a common gene of all or some members; and 2) Analysis of isolated patients with parents and/or informative family members and causative genes for variation filtering by different types of inheritance (autosomal dominant, recessive, sex-linked or *de novo*) to reduce the number of variants to a sufficiently low number to allow identification of the causal gene<sup>35</sup>.

NGS can simultaneously analyze from a few to hundreds of genes, and together with whole exome and genome sequencing can prove to be a significant advancement towards deciphering the genetic heterogeneity of rare diseases and enables the investigations of genes, that extends beyond the clinical hypothesis<sup>34</sup>. Exome sequencing success cases indicate that WGS of even a small number of individuals can identify causal variants<sup>34,39,40</sup>. Moreover, in recent years the use of WGS in genetic diagnosis have offered several advantages as oppose to targeted gene panels and WES (discussed in Chapter 2)<sup>41-43</sup>.

#### 1.1.5 WGS Bioinformatics Analysis

As the rate at which NGS data continues to propagate, data processing and analytics workflows must be updated to bridge the gap between big data and scientific discoveries.

Sequencing data analysis presents a fundamentally different challenge than analysing a targeted set of polymorphisms. The most fundamental difference is that in a sequencing study, all identified variants need to be evaluated, and many of these will not be included in the polymorphism databases. To process such data, an integrated computing and bioinformatics environment is required to manage, analyse and interpret the large amount of NGS data<sup>39,44,45</sup>.

NGS data analysis has three phases. Primary analysis generally defines the process by which device-specific sequence measurements are converted into a FASTQ file containing short read sequence data and sequence execution quality control metrics. In the secondary analysis, these sequence reads are matched against the human reference genome to detect discrepancies

between the patient sample and the reference. The most commonly used secondary analysis approach consists of five consecutive steps. These include initial read alignment<sup>46,47</sup>, duplicate read removal<sup>48,49</sup>, local realignment around known indels<sup>50</sup>, basic quality score recalibration, And variant detection<sup>50</sup>. The tertiary data analysis phase involves the discovery of clinically relevant variants<sup>51,52</sup>

### 1.1.6 Problem statement

For over 40 years, Osteogenesis imperfecta type 3 has become the focus of rare genetic research in southern Africa and Africa. Although rare, OI type 3 is common in the black African populations of Africa. OI affects individuals of different ethnic groups and of all age groups. Up to now, only individuals diagnosed with OI type 3 in the various Black African populations have been researched. Of the 20 disease-causing genes associated with OI, and more specifically the 17 genes associated with OI type 3, variants in the *FKBP10* gene has been assessed and identified as disease causing in the Black African populations. In addition to the *FKBP10* gene, a founder variant in the *P3H1* gene was identified in West Africa causing AR type VIII.

Although variants have been identified as disease-causing in some individuals, not all OI affected individuals shared these variants, but they were phenotypically similar. This could indicate that there may be other unknown genetic contributors which need to be elucidated, since it is possible for there to be multiple genes associated with a certain condition.

In addition, no study has investigated the genetic basis of OI in the CMA population of South Africa to date. Unlike the mild to no dentinogenesis imperfecta observed in the black African populations, the CMA group, and more specifically the twins, show severe presence of dentinogenesis imperfecta.

Considering the genetic diversity in the African populations and the difference in the phenotypic features presented by the CMA population group compared to that of the black, making use of the WGS approach could be beneficial due to the comprehensiveness of the test. This approach would not only allow for the identification of variants in all the known disease-causing genes but also allow for the identification of novel variants in other parts of the disease-causing candidate genes and variants in potentially new genes.

### 1.1.7 Aim and Objectives

#### Aim:

To understand the genetic basis of Osteogenesis Imperfecta in twins of Cape mixed ancestry decent

#### Objectives:

- a) To identify any unreported variants within the OI candidate genes.
- b) To identify potentially new variants in genes not previously associated with OI.
- c) To identify additional genes associated with biological pathways that map to published OI candidate genes.
- d) Prioritize the potentially disease-causing genes.

### 1.1.8 Chapter Outline

This thesis is composed of six different chapters, which are structured as follows:

**Chapter 1** provides context to OI in SA. It describes evidence for additional genetic basis of OI as seen in African populations. A case is made for the role of next generation sequencing to study the molecular basis of OI in a South African family. This chapter also highlights the purpose of the study aims and objectives.

**Chapter 2** is review of the current literature on OI and the role of NGS in identifying disease causing genes.

**Chapter 3** gives a description of the research methods that were followed in the study. It provides information on the participants and protocols for data collection. This chapter includes a discussion of the methods used to analyze the data and the ethics protocol that was followed.

**Chapter 4** describes the results for each filtering step and describes the identified SNPs and InDels.

**Chapter 5** provides an overview of the main findings. This chapter specifically discusses the results obtained as it pertains to OI and/or skeletal disorders. This chapter expands on the similarities between findings in this study and that of previous studies.

**Chapter 6** outlines the main findings, limitations, and implications of the findings. The chapter concludes with possible next steps.



## 2 Chapter Two: Literature Review

### 2.1 Introduction to bone formation

The process of bone formation is called osteogenesis or ossification. There are two main ways of bone formation, both of which involve the transformation of premature undifferentiated tissue into bone tissue<sup>53-55</sup>. The first process, called intramembranous ossification, is the direct conversion of undifferentiated tissue to bone, a process that occurs primarily in the bones of the skull and collarbone. The intramembranous ossification process results in the development of capillaries and osteoblasts, and the osteoblast eventually become osteocytes (bone cells). The second way bone is formed is through a process known as endochondral ossification, where cartilage tissue is formed from aggregated mesenchymal cells. Mesenchymal cells are transformed into cartilage cells (chondrocytes), which undergo five stages of differentiation until eventually all the cartilage is replaced by bone. The cartilage tissue serves as a template for skeletal components such as the vertebral column, pelvis, and the limbs<sup>54,55</sup>.

#### 2.1.1 Bone Composition

Skeletal tissue consists of a mineral called hydroxyapatite, organic matrix (type I collagen, non-collagen proteins, lipids), and water. These components contribute to both the mechanical and metabolic function of bone<sup>56-58</sup>. The surfaces of the bone is covered with osteoblasts and osteoclasts, which are responsible for constant bone turnover through simultaneous bone formation and resorption<sup>57</sup>. Osteocytes, the third major cellular component of bone, resides within bone tissue and communicate with adjacent osteocytes and osteoblast through channels called canaliculi<sup>57,59,60</sup>. Proteins in the extracellular matrix of bone can be divided as follows; (a) structural proteins (collagen and fibronectin) and (b) proteins with specialized functions, such as those that regulate collagen fibril diameter, and those that serve as signaling molecules, growth factors, and enzymes among other possible functions<sup>58,60,61</sup> (**Figure 2.1**).

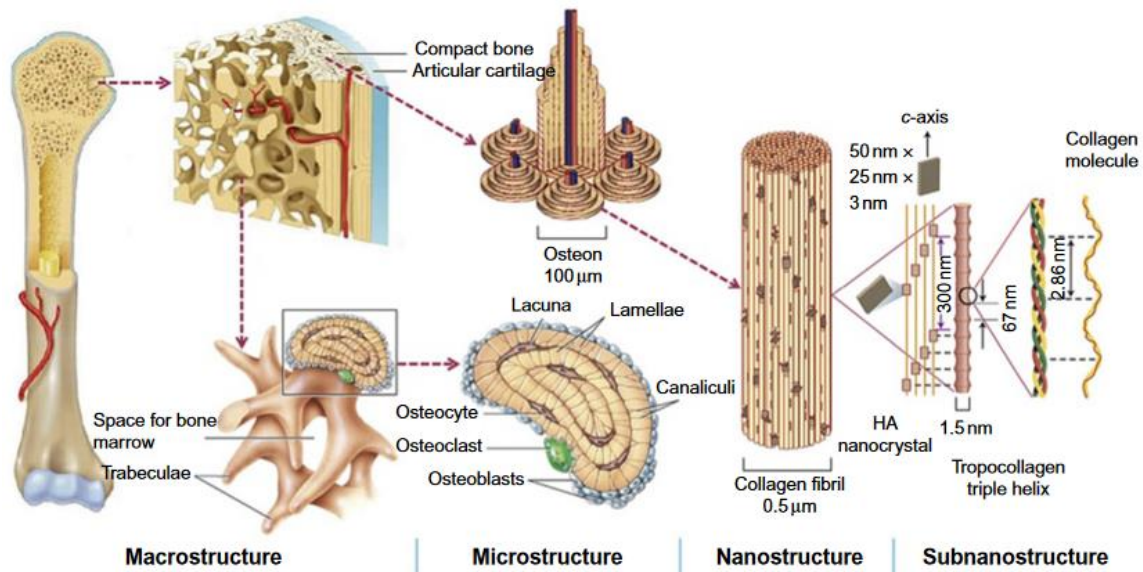


Figure 2.1: Hierarchical structure of bone from macro to sub-nano scales<sup>61</sup>.

### 2.1.1.1 Collagen

The basic building block of bone matrix is type I collagen, which comprises 80-90% of the organic matrix<sup>56-58,62</sup>. Type one collagen is present in skin, tendons, vasculature, as well as organs such as lungs, and heart. It forms the main component in the organic portion of the calcified tissue of bones and teeth. In hard tissue collagen is formed by osteoblasts in bone tissue and odontoblasts in dentin<sup>57,63</sup>.

Type I collagen is a triple helical molecule containing two identical alpha-1 chains (polypeptide chains) and a structurally similar, but genetically different alpha-2 chain. These chains are products of the genes, *COL1A1* and *COL1A2* respectively. Collagen alpha chains are characterized by glycine residue at every third position (Gly-X-Y repeating triplet, where X is usually a proline, and Y is often hydroxyproline) and undergo several post-translational modifications<sup>56,57,64,65</sup>.

The formation of collagen is a multistep process. One of three modification processes is the hydroxylation (introduction of a hydroxyl group -OH) of specific proline and lysine residues<sup>63,65</sup>. The prolyl 3-hydroxylation complex is a post-translational collagen modification system present in the endoplasmic reticulum (ER), and is composed of *CRTAP*, *P3H1*, and CyPB (encoded by the *PPIB* gene). This complex modifies a single proline residue (pro986) to 3-hydroxyproline on each alpha-1 chain of type I and type II collagen. Hydroxylation of proline and lysine residues occurs within the triple helical domain. Variants in the *CRTAP*,

*P3H1*, or *PPIB* strongly affect post-translational modifications of collagen, resulting in the complete absence of proline 3-hydroxylation and site-specific changes in collagen hydroxylation and glycosylation. Consequently, collagen folding is delayed and alterations in fibril assembly, cross-linking, and bone mineralization occurs<sup>63,66</sup>. Hydroxylysine molecules can form cross-links between collagen molecules in fibrils, and are sites for glycosylation and galactosylation. In essence, the collagen molecule goes through three post-translational processes to become procollagen. Procollagen molecules are transported into the Golgi apparatus and eventually secreted out of the cell<sup>63</sup>. The heat-shock protein 47 (HSP47), acting together with the FKBP10 protein (encoded by *FKBP10*), is important for the proper transport of type I collagen to the Golgi apparatus. *FKBP10* variants induce osteogenesis imperfecta with reduced collagen crosslinking, leading to collagen fiber deposition deficiencies and disturbances<sup>63</sup>.

On the outside of the Golgi apparatus the procollagen undergoes removal of the C- and N-propeptides by procollagen proteinases, respectively. A disintegrin and metalloproteinase with thrombospondin motifs (ADAMTS) enzymes 2, 3 and 14 are responsible for processing the N-terminal, whilst bone matrix protein-1 is liable for the C-terminal processing. The processed procollagen is termed tropocollagen. These two processes result in the formation of N-telopeptide and C-telopeptide, respectively<sup>63</sup>. Type I C-propeptide cleavage-site-variants disrupts extracellular collagen processes, resulting in osteogenesis imperfecta. Variants in *BMP1* results in decreased collagen maturation, hyperosteoridosis, and hypermineralization. Variants in *SERPINF1* (encoded by PEDF) exhibits a disordered bone matrix phenotype, a large amount of unmineralized osteoid, and an abnormal mineralization pattern. This is similar to the phenotype observed in OI due to variants at the type I C-propeptide cleavage site or in *BMP1*, suggesting PEDF may also play a role in procollagen processing<sup>61,63</sup>.

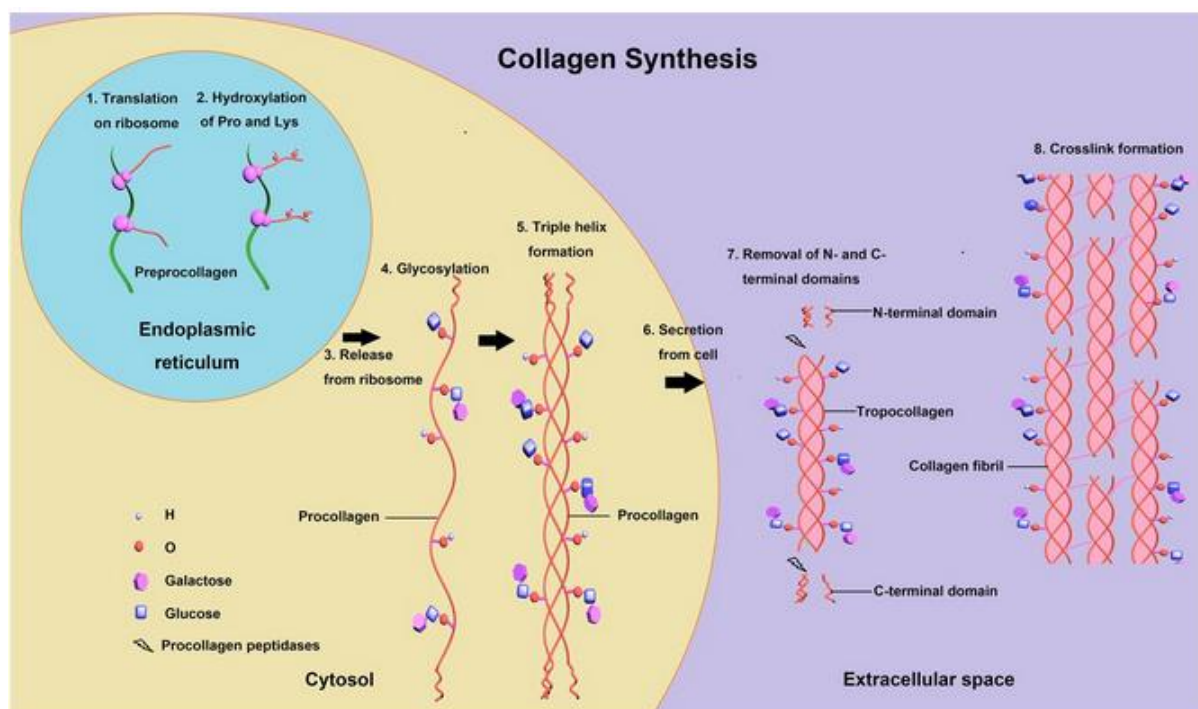


Figure 2.2: Illustration shows the synthesis of collagen<sup>1</sup>.

## 2.2 Osteogenesis Imperfecta

Osteogenesis Imperfecta is a genetic disease characterized by bone fragility and increase risk of fractures. The disease is typically characterized by bone fragility, low bone mass, recurrent fractures, progressive skeletal deformities, and growth deficiency<sup>14,67-69</sup>. In addition, there are several other characteristics that accompany OI and include the presence or absence of abnormal tooth growth, blue and/or grey sclerae, hearing loss, scoliosis, kyphoscoliosis, bowing of long bones, joint hypermobility, hypotonia, craniofacial abnormalities, pulmonary function impairment and cardiac valve abnormalities<sup>3,13,15,70-76</sup>.

### 2.2.1 OI Classifications and Phenotypic characteristics

#### 2.2.1.1 Sillence Classification

The Sillence classification divides OI into four categories (types I through IV), however the phenotype of OI varies greatly amongst them, ranging from mild symptoms with a normal life expectancy to perinatally fatal and deforming symptoms. Nonetheless, the clinical signs of several forms of OI overlap, making subclassification difficult, especially for children with no typical clinical symptoms in the early stages of the disease<sup>77-83</sup>.

Initially, abnormalities caused by *COL1A1* variants were assumed to be more severe. Later research, however, demonstrated that clinical variations in severity related to the damaged collagen helical position and types of amino acid substitution<sup>83</sup>.

### 2.2.1.1.1 Characteristics of OI types I to IV

Type I is the least severe form of OI, and individuals with this form of the disease generally present with multiple fractures, blue sclerae that persists into adulthood, brittle teeth, mild stunting and hearing loss<sup>79,84</sup>. As the least severe form, it is exceedingly rare to find fractures in new-borns. Individual with OI type I tend to have average expected height based on mid-parental height, long-bone, or spinal deformity. In type I individuals hearing loss increases with age<sup>78</sup> and in some affected individuals, specific manifestations of abnormal tooth dentin development which results in dentinogenesis imperfecta is observed<sup>84</sup>.

Compared to OI type I, type II is more severe. It can be identified radiographically by beaded ribs and "crumpled" thigh bones. Due to respiratory issues brought on by pulmonary hypoplasia, this perinatal variety typically results in death at birth or soon after, with rare reports of long-term survivals<sup>78</sup>. Severe bone deformation and fragility, poor mineralization, and numerous fractures that give the ribs their beaded look are some of the general signs. Specific manifestations are also described by Biggin and Munns<sup>78</sup>, including type II-A shortening and widening of long bones and wide ribs with fractures, type II-B shortening and widening of long bones with fractures and sparsely fractured ribs, and type II-C thin long bones with fractures and thin ribs<sup>78,79,84</sup>.

OI type III is progressively deforming and the most severe form in children that is not perinatally lethal. Specific signs include early-onset kyphoscoliosis and the presence of dentinogenesis imperfecta. General manifestations include moderate to severe bone deformities, blue or grey sclerae in infancy, triangular face, and potential hearing loss. Multiple in utero fractures that can be seen on a prenatal ultrasound scan are frequently sustained by patients with OI type III. Neonates are therefore born with numerous long-bone, rib, and limb abnormalities<sup>79,81,84</sup>.

Neonatal x-rays of a child with OI type III typically do not exhibit the crumpled long bones and beaded ribs of a child with OI type II, in spite of repeated fractures. However, it can be challenging to distinguish between these two kinds of OI because the child's underlying respiratory condition determines the clinical course. As previously indicated, disability will still occur despite all orthopedic interventions in children with OI type III due to the vast majority of them having short stature and being wheelchair reliant from an early age without a coordinated medical and orthopaedic treatment program<sup>78</sup>.

OI type IV is like type I, but with moderate to severe bone deformity of long bones and spinal

column, white sclerae, large head, easy bruising, moderate to severe stunting. More specific type IV-A individuals have normal teeth and type IV-B individuals have abnormal teeth<sup>79,81,84</sup>. OI type IV is characterized by bone fragility without the typical features of the type I phenotype (i.e., blue sclera and deafness). Fractures may present at any age and most of these patients have short stature. A small proportion of patients experience a severe, progressive lower limb deformity rather than recurrent fractures. Dentinogenesis imperfecta (genetic disorder of tooth development) is variable but when present is associated with a greater frequency of fractures.

#### 2.2.1.2 New Gene Discovery and Characteristics

In 2002, leading up to 2011, a novel variant was described that causes severe autosomal recessive OI<sup>80,81,85</sup>. This form clinically resembles the moderate to severe type IV or type III OI but has a characteristic “fish-scale” pattern of bone lamellation<sup>86,87</sup>. In addition to the fish-scale pattern observed, bone histo- morphometry shows osteomalacia, with “loser zones” and radiographic signs of bone fragility of the scapula, ribs, and long bone shafts<sup>1,2</sup> hyperosteoidosis, prolonged mineralization lag time and decreased mineral apposition rate. Individuals with OI type VI also has severe long-bone and spine fragility, white sclera, faint sclera or lack blue sclera, no dentinogenesis imperfecta, normal facial features, and they exhibit Wormian bones. Patients with OI type VI typically sustains their first fracture after the age of six months<sup>87</sup>.

OI type VII is a rare lethal autosomal recessive form caused by homozygous or compound heterozygous variants in the *CRTAP* gene. Variants in the *CRTAP* gene are said to be the first genetic cause of lethal recessive OI, associated with moderate to severe phenotype. OI type VII is characterized by early lower limb deformity, fractures at birth, bluish sclera, coxa vara and osteopenia. Rhizomelia (proximal limb disproportion) is a prominent clinical feature that distinguishes this form of OI from OI type III<sup>78,81,88</sup>.

Cabral et al,<sup>89</sup> described a form of severe or lethal autosomal recessive OI, called OI type VIII, which is characterized by severe growth deficiency, white sclera, severe growth impairment, extreme skeletal under mineralization, and bulbous metaphyses<sup>84,89</sup>. In 2007, variants in the *P3H1* gene were reported to be the cause of autosomal recessive OI type VIII. This was caused by the absence or severe deficiency of prolyl-hydroxylase activity<sup>81,89</sup>. Typically, biallelic *P3H1* variants are associated with a severe to lethal form of OI, characterized by rhizomelic limb shortening, white sclerae, severe skeletal under-mineralization, extreme growth deficiency and bulbous metaphyses<sup>90</sup>.

OI type IX is an autosomal recessive form corresponding to clinically severe types II or III of the Sillence classification<sup>91</sup>. It can be caused by homozygous and heterozygous variants in *PPIB* gene<sup>81,91,92</sup>. There are no reports of dentinogenesis imperfecta<sup>84</sup>. Some of the phenotypic manifestations due to variants in this gene includes fractures at birth, shortened long bones, and beaded ribs with small chest. Presence of mild rhizomelic shortening with tabulation and bowing of the long bones, hearing loss<sup>93</sup>.

OI type X is an autosomal recessive form caused by homozygous variant in the *SERPINH1* gene. It is characterized by bone deformities and multiple fractures, generalized osteopenia, dentinogenesis imperfecta, and blue sclera<sup>81,94</sup>. Additional characteristics may include thin ribs with healing fractures, platyspondyly, short limbs with bowed femora, joint laxity, and relative macrocephaly, respiratory distress, pyloric stenosis, renal stones, hypotonia, and respiratory failure<sup>94,95</sup>.

OI type XI is an autosomal recessive form caused by a homozygous variant in the *FKBP10* gene, related to chaperone defect<sup>96</sup>. Patients with type OI type XI have severe progressive deformation and may have joint contractures, and no dentinogenesis imperfecta<sup>97-99</sup>. *FKBP10* variants have been associated with significant pelvic abnormalities (protrusio acetabuli)<sup>78</sup>. Patients with variants in this gene are born with normal length and weight, early history of long bone fractures leading to progressive deformities of the limbs and are eventually wheelchair bound. Progressive kyphoscoliosis with flattening and wedging of the vertebral bodies is a distinctive feature of this recessive form of OI with the absence of dentinogenesis imperfecta and/or hearing loss<sup>92,98,100-102</sup>.

OI type XIII is caused by a homozygous variant in the *BMP1* gene<sup>103,104</sup>. The OI phenotype of individuals with BMP1 variants has been described as recurrent fractures, generalized bone deformity, osteopenia and wormian bones, bone fragility associated with an increase in areal bone mineral density (BMD) as measured by dual-energy absorptiometry (DXA), similar to variants that affect the C-propeptide domain of type I collagen<sup>105</sup>.

Shaheen et al,<sup>99</sup> described OI type XIV, as an autosomal recessive form caused by homozygous variants in the *TMEM38B* gene. Characterized by varying degrees of severity with multiple fractures and osteopenia, normal dentition, sclera, and hearing. Fractures occur prenatally or at approximately 6 years of age<sup>84,106</sup>.

OI type XV has been designated based on the identification of variants in *WNT1*<sup>107-109</sup>. Keupp et al,<sup>109</sup> reported that *WNT1* hypo-functional alleles result in phenotypes with low bone mass

in humans. They verified that variants in the recessive inherited gene lead to phenotypes of varying severity, ranging from mild to progressively deforming, which can occasionally lead to early infant death. They also detected families that had early osteoporosis with the autosomal dominant pattern of inheritance, with a heterozygous variant in *WNT1*<sup>84</sup>. OI type XV is characterized by early onset recurrent fractures of vertebrae and extremities<sup>110</sup>, right ptosis short stature<sup>72,108</sup>, bluish sclerae<sup>72,109</sup>, low bone density, bone deformity, severe vertebral compression, recurrent traumatic fractures of vertebrae, multiple long bone fractures<sup>107</sup>, low-bone-turn-over markers and a reduction of trabecular and cortical bone, severe early onset osteoporosis, low impact vertebral and peripheral, delayed development<sup>72,108,109,111</sup> Additional characteristics also include kyphoscoliosis, autism, intellectual disabilities, neurological abnormalities including seizures, absence of speech and inability to feed, brainstem and cerebellar hypoplasia<sup>112</sup>, quadriplegic<sup>111</sup>, hypotonia, disfigured skull, recurrent chest infection<sup>72,113–115</sup>.

OI type XVI is associated with variants in the *CREB3L1* gene. The phenotype includes multiple fractures of long bones, short tubular bones, multiple rib fractures with beaded appearance and narrow thorax. Soft calvaria, microretrognathia and short and bowed extremities due to fractures and flared metaphyseal regions. Thin/wavy ribs, mild platyspondyly, reduced skull mineralization, marked rhizomelic and mesomelic shortening are additional features<sup>116–118</sup>.

OI type XVII, caused by variants in the *SPARC* gene, is characterized by multiple vertebral fractures, kyphoscoliosis, long bone fractures, mild joint hyperlaxity, weak muscles of lower extremities, bowed humeri, generalized platyspondyly and thoracic kyphosis<sup>119</sup>.

The *MESD* gene was recently discovered and associated with causing OI type XX. There has only been one study thus far, documenting the clinical manifestations of this type of OI. These manifestations include prenatal fractures, bluish sclera in some individuals, no dentinogenesis imperfecta but presence of disorganized dentition/ clinical oligodontia, fractures in extremities, retarded gross motor function, vertebral/thoracic cage fractures, there was also evidence of intellectual disability in some<sup>120</sup>.

### **Classification**

In 2004, Rauch and Glorieux expanded the Sillence classification, due to initial quantitative deficiencies in the collagen type-I genes<sup>81,121</sup>. As a result, the Nosology and Classification of Genetic Skeletal disorders proposed a new classification, adding OI types V-VII, which takes into consideration clinical manifestations, radiological findings and molecular



alteration. OI types VI and VII were referred to as new findings and described as autosomal recessive types<sup>84,88,122</sup> and made up approximately 10% of OI cases.

In 2015, a new and what many consider simpler classification of OI was proposed by Forlino and Marini. This classification was based on functional pathways and considered each disease gene only once. These were divided into five groups; group 1 – primary defects in collagen synthesis and structure (types I-IV); group 2 – defects in bone mineralization (types V-VI); group 3 – defects in collagen modification (types VII-IX); group 4 – defects in collagen processing and crosslink (type X-XII); group 5 – defects in osteoblast differentiation and function (types XIII-XVIII)<sup>8,119,123–126</sup>.

There is a considerable overlap between the clinical presentation of OI types I to IV and OI types V to XX<sup>82</sup>. A recent review conducted by Chetty et al, documented the history and evolution of the nosology of OI from the four simple types described in 1979 to the heterogenous types described in 2019. The table below list these types according to the different types of OI with some removals and additions (**Table 2.1**).



Table 2.1: The updated classification of OI types and associated genes based on the 2019 Nosology<sup>18</sup>

Name of disorder	Mode of inheritance	OMIM number	Gene(s)	Molecular Diagnosis
<b>OI type 1</b>	AD	166200	<i>COL1A1, COL1A2</i>	OI type I
<b>OI type 2</b>	AD	166200	<i>COL1A1, COL1A2</i>	OI type I
	AR	610854	<i>CRTAP</i>	OI type VII
	AR	610915	<i>P3H1</i>	OI type VIII
	AR	259440	<i>PPIB</i>	OI type IX
<b>OI type 3</b>	AD	259240	<i>COL1A1, COL1A2</i>	OI type III
	AR	613982	<i>SERPINF1</i>	OI type VI
	AR	610682	<i>CRTAP</i>	OI type VII
	AR	610915	<i>P3H1</i>	OI type VIII
	AR	259440	<i>PPIB</i>	OI type IX
	AR	613848	<i>SERPINH1</i>	OI type X
	AR	610968	<i>FKBP10</i>	OI type XI
	AR	615066	<i>TMEM38B</i>	OI type XIII
	AR	112264	<i>BMP1</i>	OI type XIV
	AR/AD	615220	<i>WNT1</i>	OI type XV
	AR	616229	<i>CREB3L1</i>	OI type XVI
	AR	616507	<i>SPARC</i>	OI type XVII
	AR	617952	<i>TENTA5(FAM46A)</i>	OI type XVIII
AR	607783	<i>MESD</i>	OI type XX	
<b>OI type 4</b>	AD	166220	<i>COL1A1, COL1A2</i>	OI type IV
	AD	615220	<i>WNT1</i>	OI type XV
	AR	610854	<i>CRTAP</i>	OI type VII
	AR	259440	<i>PPIB</i>	OI type IX
	AR	610968	<i>FKBP10</i>	OI type XI
	AR	606633	<i>SP7</i>	OI type XII
	AR	610967	<i>IFITM4</i>	OI type V

## 2.3 Therapies and Treatment

Therapy is a type of care that tries to assist in resolving psychological or emotional problems. The three categories of medical treatment are, in theory, to cure a patient of a disease, treat the symptoms of a disease, and prevent the emergence of a disease.

### 2.3.1 Conventional Therapy

There are several therapies and treatments that individuals with OI undergo or are introduced to depending on the severity and type of OI. Treatments range from conventional management, which involves intensive physical rehabilitation, supplemented with orthopedic intervention<sup>127</sup>. The main objective of the physical rehabilitation in children with OI is to promote and sustain optimum functioning in their daily lives. The fractures, if recurrent, can result in significant

disability if optimal orthopaedic and medical management is not undertaken, however, even with orthopaedic and medical management there can still be significant disability<sup>78,81</sup>. This is achieved by a program combining early intervention, muscle strengthening and aerobic conditioning<sup>127</sup>. Motor milestones in individuals with OI may be delayed, for the most part because of muscle weakness. Isotonic strengthening exercises of the deltoids and biceps can be used to address muscle weakness in the upper appendage, and gluteal muscles and trunk extensors in the lower appendage. Strengthening of these muscle groups will ensure that individuals with OI, specifically children, are able to lift their limbs against gravity and move about independently<sup>128</sup>.

### 2.3.2 Pharmacological Therapy

Bisphosphonates (BP), synthetic compounds of inorganic pyrophosphates, are most used antiresorptive osteoporosis medication and have been proven effective in preventing osteoporosis and fragile fractures and increasing cortical thickness<sup>129</sup>. Bisphosphonates include alendronate, ibandronate, risedronate and zoledronic acid, pamidronate<sup>130</sup>. BPs can be administered orally or intravenously (IV), meaning “within a vein”. For oral BP administration alendronate or risedronate is suggested, and for patients with contradictions or intolerance to oral BPs, IV zoledronic acid and ibandronate is suggested because it has been shown to prevent fractures in clinical trials. Pamidronate is also administered via IV but is not recommended for osteoporosis for instance and has been superseded by zoledronic acid and in some cases IV ibandronate<sup>131</sup>.

Although BPs are a popular and effective anti-osteoporotic, its side effects prevent it from being used by everyone. As a result, it is only prescribed to people who already have pre-existing conditions like osteoporosis, metastatic bone disease, multiple myeloma, Paget's disease, polyostotic fibrous dysplasias, total joint arthroplasty, early stage avascular necrosis, osteogenesis imperfecta, and metastatic hypercalcemia<sup>132</sup>.

There have been concerns about the negative effects of BPs on fracture healing and bone remodelling, as it may prolong fracture healing due to bone resorption inhibition<sup>133,134</sup>. For instance, pamidronate is linked to a number of negative side effects, including osteonecrosis of the jaw, acute phase response, musculoskeletal discomfort, different ocular events, hypocalcemia, and the subsequent secondary hyperparathyroidism<sup>135</sup>. Improvements are evident when pamidronate is administered before and after surgical treatment, but it has also been observed that patients' conditions worsen when taken off the pamidronate treatment<sup>129</sup>.

The equivocal improvement in fractures in children is illuminated by data from bisphosphonate treatment of the *Brl* mouse<sup>136</sup>. The treatment increases bone volume and load to fracture of murine femora, but concomitantly decreases material strength and elastic modulus. Femurs become, ironically, more brittle after prolonged treatment, and bands of mineralized cartilage create matrix discontinuities that decrease bone quality. Prolonged treatment also alters osteoblast morphology. However, pamidronate treatment has not caused osteonecrosis of the jaw in any reported OI cases<sup>73,137–141</sup>.

The oral bisphosphonate risedronate has been administered to both children and adults with OI. A moderate improvement in fractures was reported in children during the first treatment year, but fracture incidence approached that of the placebo group during the second and third years of treatment. Adults treated with risedronate experienced an increase in bone density but not a decrease in fracture incidence<sup>142</sup>.

Bisphosphonates were reported to be marginally effective in OI type IV, caused by PDEF deficiency<sup>143</sup>. It was later postulated that because bisphosphonates bind to mineralized bone before they are ingested by osteoclasts, the increased amounts of unmineralized osteoid in OI type VI bone might disrupt bisphosphonate deposition<sup>144</sup>. Denosumab, which is not a BP, directly inhibits the receptor activator of nuclear factor kappa-B ligand (RANKL) pathway, was more effective than bisphosphonate in normalizing bone turnover for these patients in a short-term study. Denosumab also has the advantage of a much shorter half-life than pamidronate, 3-4 months versus 10 years. Bisphosphonates (BPs) appeared to have poor effects on the bone of OI type VI patients in comparison with other types of OI<sup>87</sup>.

The lack of published data about pamidronate treatment in SA patients, motivated Henderson *et al.*,<sup>3</sup> to assess subjective data regarding OI and pamidronate therapy in Black patients diagnosed with OI type III who were being treated at the Universitas Private Hospital in Bloemfontein. This study made use of researcher-administered questionnaires to assess the side effects experienced, impact on the quality of life, the patients' overall attitude towards the disorder and the value of their pamidronate treatment. Approximately 57.1% of patients tolerated the intravenous procedure. Some found the treatment problematic due to pain, discomfort, and fear. Pamidronate was shown to improve the physical and emotional wellbeing of 73.1% of patients in the study. The data in this study indicated that black South Africans affected by OI experienced similar symptoms and responses to the disorder as people elsewhere<sup>3</sup>.

## 2.4 Mendelian Inheritance

Diseases that are caused by variation within a single gene are referred to as Mendelian disorders or monogenic disorders. For many monogenic disorders, inheritance of a mutated copy or copies of a gene results in a characteristic phenotype, and inheritance of that phenotype follows a Mendelian segregation pattern<sup>145</sup>.

Inheritance patterns differ from genes on sex chromosomes (chromosomes X and Y) compared to genes located on autosomes, non-sex chromosomes (chromosome 1 to 22). Diseases caused by mutated genes located on the X chromosome can be inherited in either a dominant or recessive manner. Since males only have one X chromosome, any mutated gene on the X chromosome, dominant or recessive will result in disease<sup>146</sup>.

### 2.4.1 X-linked recessive inheritance

Males are more likely than females to have X-linked recessive diseases. If a father has an X-linked recessive disease, there is a 0% probability that he will convey the defective gene to his sons, because his sons will always get a Y chromosome from him (hemizygous father)<sup>147</sup>. Due to the fact that he only has one X chromosome, which contains the gene that is defective, he also has a 100% probability of transferring the gene to his daughters (carrier daughter). However, it's possible that the daughters won't develop the illness since they may get a copy of the healthy gene from their mother. If the father does not have the condition, then the daughter will be a carrier<sup>147</sup>.

A female will be afflicted by the disorder if she possesses two copies of the defective gene (homozygous female)<sup>147</sup>. She is referred to as a carrier (heterozygous female) if she possesses a working copy of the gene on one X chromosome and a copy of the defective gene on her other X chromosome. The disorder does not impair carriers, but they can still convey the defective gene to their offspring<sup>147</sup>.

A mother with an X-linked recessive disease will always pass the defective gene on to her sons, because she carries two copies of the defective gene. She will also will 100% certainty pass the defective gene on to her daughters (**Figure 2.3**). If the mother is a carrier and the father is healthy, each son has a 50% chance of being affected and each daughter has a 50% chance of inheriting the mother's genes<sup>147</sup>.

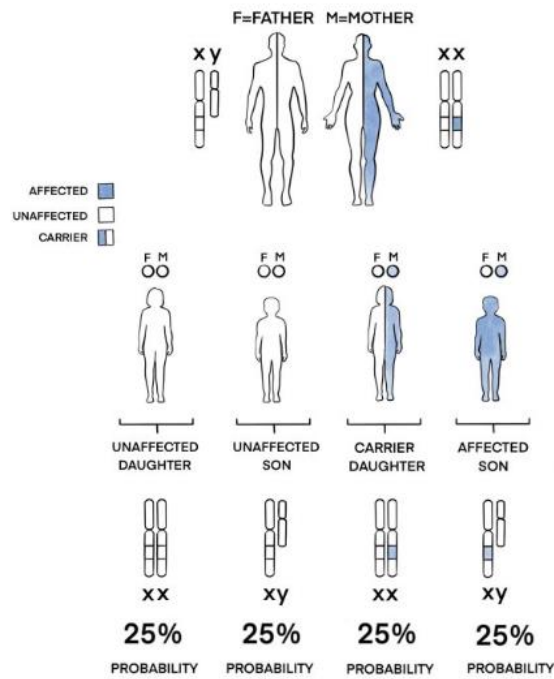


Figure 2.3 X-linked recessive inheritance.

## 2.4.2 X-linked dominant inheritance

X-linked dominant inheritance is less frequent than X-linked recessive inheritance. Because just one defective gene copy on the X chromosome is necessary to induce disease. As a result, both men and women can be afflicted (**Figure 2.4 and Figure 2.5**)<sup>147</sup>.

Because of unpredictable X-inactivation, females are typically less severely impacted than males. X-inactivation ensures that females, like males, have one functional copy of the X chromosome in each body cell, and so only one copy of the X-chromosome will be active<sup>147,148</sup>. Affected males will only transmit the affected gene to their daughters, not their sons<sup>147,148</sup>.

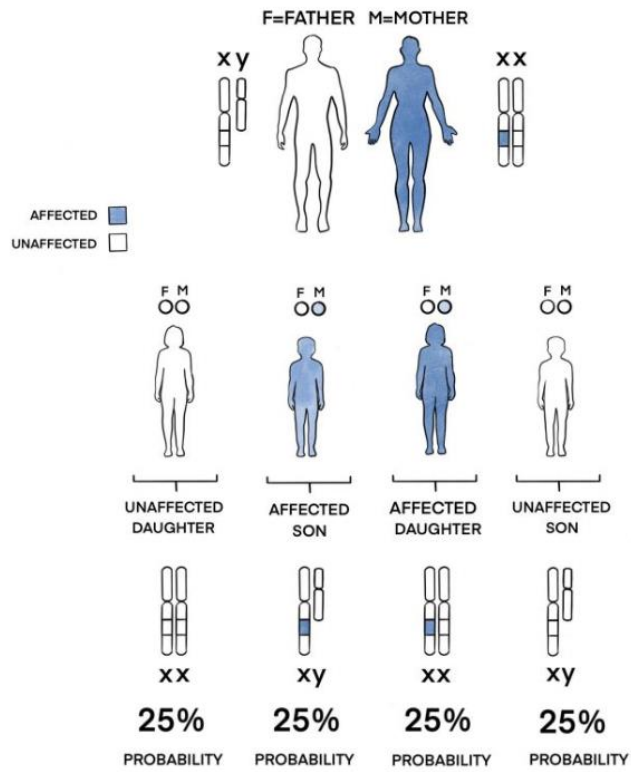


Figure 2.4: X-linked dominant inheritance with an affected mother.

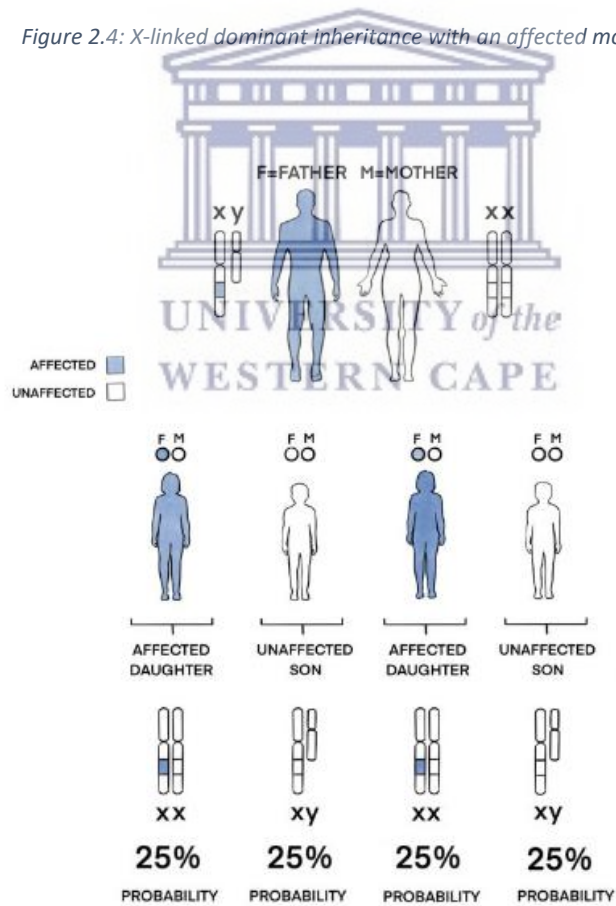


Figure 2.5: X-linked dominant inheritance with an affected father.

### 2.4.3 Autosomal Dominant Inheritance

In autosomal dominant inheritance the faulty gene is in one of the autosomes and only a single copy of the defective gene (from one parent) is enough to cause the disorder. The affected person has a 50% chance of passing the defected gene onto their offspring, affecting both genders and can be transferred by both genders and the disorder is seen in each generation of the family tree<sup>149–151</sup>. Sometimes, conditions inherited in this way have uneven expressivity and insufficient penetrance. Variable expressivity denotes a range in intensity from very mild to extremely severe, whereas incomplete penetrance denotes that while the defective gene is present in some individuals, the same individual may not have any symptoms at all (**Figure 2.6**).

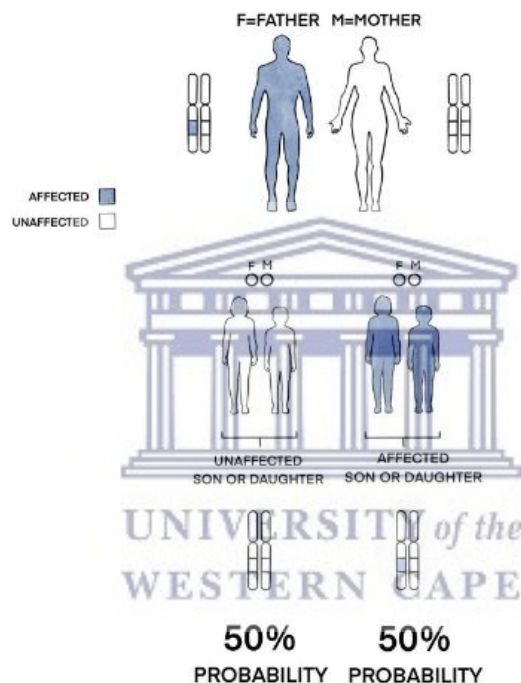


Figure 2.6: Autosomal dominant inheritance in an affected father.

### 2.4.4 Autosomal Recessive Inheritance

Two copies of the defective gene, one from each parent, are needed to induce disease in autosomal recessive diseases. Typically, carriers are unaware of their status until they have a kid who is affected or have had genetic testing done. Both parents would have to be carriers of the same gene variant for a child to be impacted. Affected individuals are generally born to unaffected (asymptomatic) and sometimes related (i.e., consanguineous) parents<sup>149</sup>. The likelihood of autosomal recessive genetic abnormalities is higher in children born from



consanguineous unions because blood relations are more likely to be silent carriers for the same recessive condition(s).

In the most likely scenario, where both parents carry the gene, there is a 25% risk that a child will be affected, a 25% chance that a child will be unaffected but not a carrier, and a 50% chance that a child will be an unaffected carrier<sup>152,153</sup>.

If one parent has the condition and the other is not a carrier, all of the offspring will be. A child has a 50% chance of being affected and a 50% chance of being a carrier if one parent has the condition and the other is a carrier. These risks are associated with each new pregnancy, and this process is random.

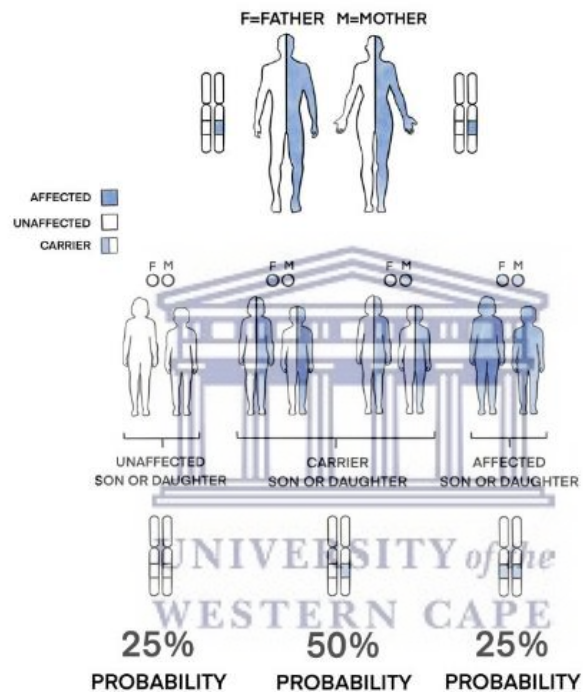


Figure 2.7: Autosomal recessive inheritance with father and mother carriers

## 2.5 De novo genetic alterations

A genetic variation that initially appears in one family member as a result of a variation in a parent's germ cell (egg or sperm) or a variation that develops in the fertilized egg during the early stages of embryogenesis<sup>154</sup>.

## 2.6 Complex diseases

Compared to monogenic diseases, where a single gene-variant is sufficient to cause a disease phenotype, complex diseases are caused by a combination of genetic, environment, and lifestyle factors. Contrary to single gene illnesses, complex diseases tend to run in families, although their inheritance patterns are less obvious. A person's likelihood of inheriting or passing on these illnesses is therefore difficult to predict<sup>155</sup>.

## 2.7 Penetrance of variants

Penetrance is defined as the percentage of individuals having a specific variant, who display a clinical phenotype of the related disease. Penetrance is often differentiated from variable expressivity, which describes the degree of heterogeneity in the clinical phenotype in people with a specific genotype<sup>145,156–159</sup>. Complete penetrance designates that all individuals who have the related disease-causing variant will develop clinical symptoms of the disease. Incomplete or reduced penetrance specifies that some individuals fail to express the trait, even though they carry the variant<sup>160</sup>. In recessive illnesses, incomplete penetrance is challenging to identify since being a carrier is enough to pass the disease on<sup>161</sup>. There may be an intermediate state of the variation with no clinical manifestation, which would account for incomplete penetrance<sup>161</sup>. Because the variant can vary from generation to generation, a normal person with a pre-variant may have numerous afflicted offspring in whom the variant is completely penetrant<sup>161</sup>. One specific example is Osteogenesis Imperfecta, in which the majority of people have dominant variations in one of the two genes that generate type I collagen. However, OI does not impact everyone who possesses *COL1A1* and *COL1A2* variations in the same way. The term "pseudo-incomplete penetrance" refers to a situation in which the clinical examination was unclear, or the symptoms had not yet materialized when the examination was conducted, and as a result, the reflection of non-penetrance was incorrect. Additionally, when the parents of many afflicted children with a dominant illness are healthy, germline mosaicism—which only manifests in the first generation—is noticed<sup>162</sup>.

Healthy individuals can have a huge number of possibly or mildly detrimental variants and imaginably tens of potentially severe disease variants without suffering any obvious ill effects<sup>163</sup>. The individual may be an asymptomatic carrier of a single recessive mutant variant or the variant is dominant, but the clinical phenotype might only be mild and lie within the range of normal healthy variation or become apparent only in later decades of life<sup>156</sup>.

## 2.8 Molecular Genetics of OI

Osteogenesis imperfecta is a single gene disorder (Mendelian), which follows an autosomal dominant mode of inheritance in approximately 90% of cases, associated with heterozygous variants in the *COL1A1* or *COL1A2* genes, encoding type I collagen<sup>72,101,164–167</sup>. However, the remainder of OI cases are due to recessively inherited variants in non-collagen genes and X-linked inherited variants. The sections below look at the molecular aspects of both dominantly inherited variants in collagen type I genes, as well as variants inherited in a recessive manner in non-collagen genes.

### 2.8.1 Dominantly inherited variants in collagen OI genes

Variants in the type I collagen genes (*COL1A1* and *COL1A2*) lead to quantitative and qualitative defects in collagen production<sup>102</sup>. Variants in the *COL1A1* gene that causes reduced synthesis of type I collagen via haploinsufficiency are referred to as ‘quantitative’ variants. These variants usually result from nonsense substitutions or frameshifts that cause premature codon termination and subsequent degradation of the transcript by the nonsense-mediated RNA degradation mechanism. These types of variants typically result in mild forms of OI (OI type I), where each cell produces less collagen but of normal quality and structure.

In contrast, missense variants in the alpha1(I) or alpha2(I) chain, especially glycine (Gly) substitutions in the Gly-X-Y repeat along the triple helical domain<sup>69,168</sup>, and splice site variants causing in-frame deletions or exon skipping, are usually referred to as ‘qualitative or structural’ variants and can cause either lethal, severe, or moderate forms of the disorder. These variants significantly impact the normal structure and assembly of the procollagen heterotrimer and may trigger a cascade of deleterious events at both the intracellular and extracellular level. At an intracellular level, this includes abnormal collagen post-translational modification, folding, trafficking and ER-stress. At an extracellular level this includes reduced collagen in the matrix, abnormal mineralization and altered matrix-to-cell signaling. Although thousands of different dominant variants in *COL1A1* and *COL1A2* have been described, the establishment of a genotype-phenotype correlation among the many different structural variants and explaining the spectrum of clinical severity ranging from moderate to lethal has been challenging. Some general rules have held true: (1) because the trimer assembly and the winding of the triple helical domain proceed in a zipper-like manner from the C-terminus to the N-terminus, variants closer to the C-terminus cause a more severe phenotype; (2) a greater proportion of the lethal variants affect the alpha1(I) chain compared to the alpha2(I); (3) one-third of all glycine substitutions

in the triple helical domain are lethal, especially when glycine is replaced by a charged or branched side chain amino acid; (4) lethal regions in the helical domain of  $\alpha 1(I)$ (2 lethal regions) and  $\alpha 2(I)$ (8 lethal regions) have identified and aligned with major ligand binding sites such as those for integrins, matrix metalloproteinases and various matrix molecules including proteoglycans<sup>169,170</sup>.

## 2.8.2 Recessively inherited variants in OI genes

### 2.8.2.1 Leprecan family members

The cartilage-associated protein (CRTAP), prolyl 3 hydroxylase 1 (P3H1) and peptidyl-propyl cis-trans isomerase B (PPIB) form part of a family of proteins called Leprecans. P3H1 forms a molecular complex with CRTAP and cyclophilin B encoded by PPIB, in a 1:1:1 ratio on the ER, which is responsible for one step in collagen post-translational modification, the prolyl 3-hydroxylation of proline residues, specifically  $\alpha 1(I)$ (Pro986)<sup>90,95,171-174</sup>. P3H1 provides the enzymatic activity of the complex. The absence or decrease of P3H1 or CRTAP results in a delay in triple helix formation with secondary prolonged exposure of the procollagen chains to other hydroxylation steps, leading to an over-modification of the chains<sup>90,172,175</sup>.

### 2.8.2.2 Peptidyl-prolyl cis-trans isomerases

*FKBP10* is located on chromosome 17 and encodes FKBP65, a member of FKBP-type peptidyl-propyl cis/trans isomerase (PPIase) family<sup>101,176</sup>. This protein is localized to the ER, and can form an ER chaperone complex with the Heat-shock protein 47 (HSP47) and the immunoglobulin heavy-chain-binding protein (BiP), which regulates the activity of LH2 encoded by *PLOD2*<sup>95,101,176</sup>. This protein is essential for the proper formation and balance between elastin and collagen molecules presented in the ER, with roles in collagen biosynthesis and folding<sup>101</sup>. Cross linking between collagen molecules creates durable collagen molecules which forms resistance against fractures, as such the role of FKBP65 is important in the hydroxylation of collagen prior to crosslinking. Therefore, any impaired FKBP65 can be deleterious in the establishment of durable collagen molecules<sup>101</sup>. At least three structurally distinct families of proteins have been linked by their ability to catalyse the bond preceding a proline residue between its *cis* and *trans* forms. The FKBP65 (FK506-binding) forms part of these distinct families.

### 2.8.2.3 Serpin family members

*SERPINF1* encodes pigment epithelium-derived factor (PEDF), a secreted glycoprotein of the Serpin superfamily, originally known for its neurotropic and antiangiogenic features<sup>80</sup>. Biochemically, patients with variants in this gene are characterized by a slight elevation in serum alkaline phosphatase and bone turn-over and by the absence of circulating levels of PEDF. At the histological level, iliac bone biopsies show a large amount of un-mineralized osteoid on their cancellous bone due to a mineralization defect and a ‘fish-scale’ pattern of bone deposition instead of normal bone lamellation<sup>87</sup>. The majority of people with type VI OI have variants in *SERPINF1* that cause loss of function and undetectable plasma levels of PEDF protein as a consequence of an early stop codon or in-frame deletions or insertions. It is a protein circulating in serum, and the pathogenetic mechanisms leading to forms of recessive OI are currently unclear. A lack of circulating PEDF was a specific hallmark of OI patients with *SERPINF1* variants because all OI patients with other candidate-gene variants, and carriers with one copy of *SERPINF1* variant, as well as healthy controls had normal detectable PEDF levels<sup>87</sup>. PEDF contains binding sites for type I collagen, which regulates osteoblastogenesis and osteoclast function through osteoprotegerin and sclerostin. In-vitro and in-vivo models had provided evidence that PEDF also inhibited osteoclasts differentiation through the OPG- RANK-RANKL pathway<sup>87</sup>. In the musculoskeletal system, type I collagen is bound by PEDF, which acts as an inhibitor of bone resorption by inhibiting osteoclast maturation via osteoprotegerin and receptor activator of nuclear factor- $\kappa$ B ligand. PEDF suppresses the expression of genes that inhibit mineralization, leading to enhanced osteoblastic differentiation and increased matrix mineralization<sup>80</sup>.

The HSP47 (encoded by *SERPINH1*) is a collagen-binding protein acting as an ER-chaperone that contributes to the proper assembly of the collagen triple helix<sup>94,171</sup>. Mechanistically, HSP47 regulates triple-helix stability via direct binding to specific arginine residues that lie at the interface between HSP47 and collagen. Loss of function variants in *SERPINH1* causes recessive OI due to aggregation and delayed secretion of procollagen molecules<sup>171</sup>. HSP47 may play a role in the monitoring function for proper triple helical structure assembly and, like FKBP65, acts downstream of the propyl 3-hydroxylation complex<sup>94,95</sup>.

### 2.8.2.4 The WNT family

The *WNT1* (Wingless-type MMTV integration site family, member 1) gene encodes for the secreted signaling protein WNT1. It is a member of a family of proteins that controls a variety

of cellular processes, including cell growth, differentiation, function, and death. Wnt signaling through the canonical Wnt/-catenin pathway is one of the pathways that is activated, and it leads to an increase in bone mass through a number of mechanisms, including stimulation of preosteoblast replication, induction of osteoblastogenesis, and inhibition of osteoblast and osteocyte apoptosis<sup>114,165</sup>. The canonical WNT pathway also entails attaching to the cell surface receptor LRP5 (lipoprotein-related receptor 5), which sets off a signaling cascade that causes beta catenin to go into the cell nucleus and activate target genes' transcription<sup>114,171</sup>. The beta-catenin signaling pathway is essential for skeletal development, homeostasis and remodeling<sup>114,171</sup>. Studies on humans and mice revealed that Wnt co-receptor low-density LRP5 loss or gain of function produced low or high bone mass phenotypes, respectively, highlighting the involvement of Wnt in bone development<sup>177</sup>. Loss of function variants in the WNT-ligand, have been associated with a spectrum of skeletal disorders<sup>114,171</sup>. Non-canonical signaling pathways, on the other hand, refers to all signaling pathways that are not mediated by  $\beta$ -catenin. Wnt5a and Wnt11 are ligands that activate the Wnt/Ca<sup>2+</sup> and Wnt/PCP pathways without causing an increase in intracellular  $\beta$ -catenin. Osteoblast differentiation is aided by Wnt5a's activation of non-canonical Wnt signaling. Osteoblast-specific Wnt5a knockout mice (Wnt5a-KO) produced via OSX-Cre transgenic mice had a phenotype with reduced bone production and resorption<sup>178</sup>.

### 2.8.2.5 The metalloproteases family

Bone morphogenetic proteins (BMPs) from part of the transforming growth factor  $\beta$  (TGF- $\beta$ ) superfamily which have been connected to a wide range of functions, such as cell proliferation, apoptosis, differentiation, and morphogenesis. Currently, roughly 20 distinct BMPs have been discovered and classified into subfamilies based on how similar their amino acid sequences are to one another. Because it is a metalloprotease that cleaves the C-terminus of procollagen I, II, and III and has the ability to induce cartilage formation in vivo, BMP1 does not belong to the TGF-superfamily<sup>179</sup>. BMP1 encodes a secreted procollagen C-peptinase that is closely related to the tolloid family of proteases and is functionally distinct from the other bone-inducing BMPs<sup>171</sup>. BMP1/Tolloid (TLD) acts as an astacin metalloprotease that plays an important role in the cleavage of the type I, II, III procollagen C-terminal propeptide and the amino-terminal propeptide from procollagen type V and XI<sup>76,169</sup>. In addition to procollagen, BMP1 has also been shown to exhibit protease activity on LOX and other extracellular matrix (ECM) proteins, which have a critical role in collagen cross-linking<sup>76,169,171</sup>.

Studies in BMP1/mTLD deficient patients with OI have demonstrated delayed cleavage of type I collagen C-propeptide and disorganization of type I or V collagen fibrils, as well as impaired processing of the small leucine-rich protein (SLRP) prodecorin. At the tissue level, bone of one OI patient with BMP1 variant was found to be hypermineralized compared to patients with OI caused by collagen I variants<sup>105</sup>. Functional studies have largely focused on the C-propeptide cleavage activity of BMP1 variants but BMP1/mTLD is also involved in processing of additional extracellular matrix components, in particular the processing of the SLRP prodecorin by impaired removal of the pro-domain. Decorin is known to influence both collagen assembly and regulate matrix mineralization<sup>105,169</sup>.

#### 2.8.2.6 The CREB/AFT family

The *CREB3L1* gene (cAMP responsive element binding protein 3 like 1) encodes the ER stress transducer ‘old astrocyte specifically induced substance’ (OASIS), a basic leucine zipper (bZIP) transcription factor which belongs to the cyclic adenosine monophosphate (AMP) responsive element binding protein/activating transcription factor (CREB/AFT) family<sup>116,171</sup>. OASIS is processed by regulated intramembrane proteolysis (RIP) in response to ER stress, and is highly expressed in osteoblasts<sup>116</sup>. *COL1A1* was identified as a target of OASIS, and Murakami *et al.*,<sup>180</sup> demonstrated with murine studies that OASIS activates the transcription of *COL1A1* through an unfolded protein response element (UPRE)-like sequence in the *COL1A1* promoter region, thereby revealing its critical role in bone formation<sup>116,171,180</sup>. Furthermore, the authors demonstrated that deficiency of OASIS affects transcription of several bone associated genes (*COL1A1*, *COL1A2*, *ALPL*, *IBSP* and *OPN*), reduces glycosaminoglycan levels in bone extracellular matrix and has negative effects on osteoblast<sup>116,171</sup>.

#### 2.8.2.7 The TRIC family

Two members of the T-complex protein Ring Complex (TRIC) family exist: TRIC-A (encoded by *TMEM38A*) and TRIC-B (encoded by *TMEM38B*)<sup>181</sup>. For this study, only the gene associated with OI will be discussed. Recessively inherited variants in the *TMEM38B* gene, which encodes the ubiquitously expressed endoplasmic reticulum protein trimeric intracellular cation channel (TRIC) type B, is expressed ubiquitously at low levels and functions to regulate intracellular calcium release. Similar to what is observed in mice, human fibroblasts from OI patients with TRIC-B variants showed decreased synthesis, secretion and deposition of type I collagen<sup>171,182,183</sup>. A study by Webb *et al.*,<sup>183</sup> demonstrated that the absence of TRIC-B disrupts

ER calcium flux kinetics is consistent with increased activation of the PERK/ATF4 pathways of ER stress<sup>183</sup>.





## 2.9 Sequencing

### 2.9.1 DNA sequencing

DNA sequencing is the act of determining the nucleotide composition and order of given DNA molecules<sup>184,185</sup>. The first revolution in the DNA sequencing field took place in the second half of the 1970s with methods published by Allan Maxam and Walter Gilbert<sup>186</sup> and Frederick Sanger and colleagues<sup>187</sup>. The Sanger method offered overall higher efficiency after a series of optimizations and dominated DNA sequencing for the past decades. New sequencing methods started to emerge and challenge the cost and supremacy of the Sanger dideoxy method<sup>184,185,188</sup>. These methods became known as next-generation sequencing.

### 2.9.2 Next Generation Sequencing

Next-generation sequencing (NGS), or Second-Generation Sequencing, employs massively parallel strategies that can produce large amounts of sequences from multiple samples at very high-throughput, and at a high degree of sequence coverage to allow for the loss of accuracy of individual reads when compared to Sanger sequencing. The time needed to generate the gigabase(Gb)-sized sequences by NGS was reduced from many years to only a few days or hours, with an accompanying massive price reduction<sup>187,189-193</sup>. Third-generation single-molecule sequencing technologies also emerged to reduce the price of sequencing and to simplify the preparatory procedures and sequencing methods<sup>193,194</sup>.

### 2.9.3 Next Generation Sequencing Application

Next-generation sequencing (NGS) technologies and applications, such as whole-genome sequencing (WGS) and whole-exome sequencing (WES), are increasingly used in the study of Mendelian, rare complex, and genetically heterogeneous disorders<sup>195</sup>.

In the field of human genetics, NGS is mainly used to identify putative disease-causing variants in Mendelian diseases and risk factors in complex diseases. Whole genome sequencing of affected individuals is now possible for the purpose of variant detection. However, sequencing a whole genome at high average coverage is still expensive. Thus, various targeted sequencing approaches are used for variant detection: (1) amplicon sequencing is used to identify variants in small regions, such as single genes, in large number of samples; (2) disease specific gene panels are used for the detection of variants in a group of known disease associated genes; (3) enrichment of all coding exons can be used for the identification of novel and known disease

causing variants and disease associated genes or if the list of candidate genes is too long<sup>196</sup>. This approach is called exome sequencing.

However, WES is unable to detect some classes of genetic variants, including regulatory domains such as those in the 5' and 3' untranslated regions, deeper intronic variants, and specific variant types e.g., triplet repeats, variants in the mitochondrial DNA. It is also limited in its ability to detect copy number variations (CNVs) and structural variants e.g., balanced or unbalanced translocations<sup>197</sup>. The use of whole genomic sequencing (WGS) reduces the before-mentioned problems with coverage because it does not require the exome capture step. It also allows for detection of all of the described classes of genetic variants except triplet repeats and balanced translocations. Difficulties with WGS include the greater cost and data-storage requirements, and the more complex bioinformatics analysis required. As these difficulties are addressed, it is anticipated that WGS will eventually replace WES in the analysis of MDs both internationally and locally.

## 2.10 Variant detection in NGS

The principles of sequencing analysis are essentially the same for both WES and WGS. The detection of Single Nucleotide Variants (SNVs) and small insertions and deletions (InDels) from raw NGS data can be separated into three parts. First, the sequencing output needs to be aligned to a human reference assembly to determine whether there are any differences in the aligned sections. Second, the potential effects of variants on the encoded protein need to be determined (e.g., non-synonymous or synonymous amino acid substitutions, splice site variant and Inel). Third, available databases, such as dbSNP, 1000 Genomes, and the Exome Sequencing Project are searched, and in many cases, a cohort of healthy control DNA samples may be genotyped to determine whether variants are novel or have been previously reported and the prevalence of the variant allele<sup>198</sup>. Basic concepts of these three tasks are introduced in the following paragraphs.

### 2.10.1 Alignment

'Alignment' is the process by which we discover how and where the read sequences are similar to the reference sequence. An 'alignment' is the result of from this process, specifically: an alignment is a way of 'lining up' some or all of the characters in a read with some characters from the reference in a way that reveals how they are similar. Massively parallel sequencing results in a large number of sequences, referred to as reads. Attached to each read are base

qualities for each nucleotide in the read, showing the predicted error rate for each base. For paired-end sequencing, the sequences come in pairs of two reads sequenced from opposite ends of a longer sequence of DNA, with a specific number of unknown sequences between them. Paired-end sequencing provides advantages to aid alignment. The known genomic distance between the pairs can be used to identify insertions and deletions relative to the reference genome<sup>199</sup>.

The genome of the sequenced individual must be constructed to make sense of these reads. Theoretically, this can be achieved solely based on the NGS reads without additional information, i.e., generating a de novo assembly. However, if a reference genome of the sequenced individual is available, NGS reads can be aligned to this reference genome. For human genomes, the most current and widely used reference sequences are GRCh37 (hg19) and GRCh38 (hg38).

One obvious issue when considering the choice of an alignment tool is the type of data that it was designed for or is suitable to align (DNA, RNA, miRNA, or bisulphite). Another dimension to consider is the sequencing platform that generated the high-throughput sequencing data. The read length supported by the mapper is a particularly important characteristic<sup>200</sup>. There are a great number of tools for alignment of sequences to the reference genome including the commonly used Burrows-Wheeler Alignment tool (BWA)<sup>201</sup> and Bowtie 2<sup>202,203</sup>.

These modern NGS aligners use string searching data structures to store the reference genome. For instance, BWA uses the Burrows-Wheeler Transformation (BWT) to efficiently store the genome in memory. For paired-end reads it first aligns both reads of a pair separately and then joins these alignments. If the reads of a pair could be mapped to different positions in the reference genome, positions where the two reads are close to each other are preferred. BWA is a software package for mapping low-divergent sequences against a large reference genome<sup>201</sup>, such as the human genome. It consists of three algorithms: BWA-backtrack, BWA-SW and BWA-MEM. The first algorithm is designed for Illumina sequence reads up to 100bp, while the other two are for longer sequences ranging from 70bp to 1Mbp. BWA-MEM and BWA-SW share similar features such as long-read support and split alignment, but BWA-MEM, which is the latest, is generally recommended for high quality queries as it is faster and more accurate. BWA-MEM also has better performance than BWA-backtrack for 70-100bp Illumina reads. The BWA-MEM algorithm performs local alignment and output SAM file<sup>201</sup>. The BWA alignment tool used in this study gives as output that is an unsorted alignment file in SAM

format, which is a text-based format for storing biological sequences that have been aligned to a reference sequence<sup>201</sup>. This is currently the most common file format for storing sequence data generated by NGS technologies.

Bowtie2 tend to be much faster and more memory efficient compared to BWA. However, BWA is the popular tool for genomic sequence data, whereas Bowtie2 is used for RNA-Seq projects. The most crucial difference between these mappers is that BWA performs gapped alignments and Bowtie2 does not.

### 2.10.2 SAM-TO-BAM

One of the key steps in any reads-to-variants workflow is post-alignment data processing to produce analysis-ready binary alignment/map (BAM) files. The output sequence alignment/map (SAM) formatted files are usually converted to BAM formatted files and these files are then sorted. The BAM formatted file is a binary version of the SAM file, which carries the same type of information but in a compressed format. The sorting allows all the alignments to be ordered either based on: (a) the order of the reference contigs listed in the original reference FASTA file; or (b) the coordinates within each contig. Sorting these files by coordinates is performed to avoid loading extra alignments into memory. The sorting and indexing of the BAM files aim to achieve fast retrieval of alignments overlapping a specific region without going through the entire alignment, making accessing data in these files much faster<sup>204,205</sup>. SAMTools is a tool that provides various utilities for manipulating alignments in the SAM format, including sorting, merging, indexing, and generating alignments in a per-position format. It takes SAM and/or BAM formatted files as input<sup>206</sup>.

Samples go through a polymerase chain reaction (PCR) amplification step that introduces GC bias, a major source of unwanted variations and errors in the sequencing coverage, which results in the sequencing of duplicate reads<sup>193</sup>. Since they share the same sequence and the same alignment position, they can lead to problems in variant detection. For example, during SNV calling, false-positive variants may arise as some alleles may be overrepresented due to amplification biases. To account for the PCR duplicates, reads are marked using the PICARD tool (version 2.20.1). The resulting duplicate reads are not informative and should not be counted as additional evidence for or against a putative variant. The duplicate marking process identifies these reads as such so that certain variant discovery tools, like GATK can ignore them. PICARD is a set of command line tools for manipulating high-throughput sequencing data and formats such as SAM/BAM/CRAM and VCF<sup>207</sup>.

### 2.10.3 Variant Calling

Variant calling is the process of identifying positions in the genome (or parts of the genome) of an individual that are different compared to the reference genome. The simplest way to call the genotype at a position is to create a list of all sequenced bases aligned to the position and calculate the proportion of bases that are different to the reference genome. This format is often referred to as a “pileup”. Two cut-offs for heterozygous variants such as 30% and 80% can then be used to call variants. However, this approach does not take properties such as base or mapping quality into account. Such an approach is error prone at low read depth.

Thus, more sophisticated variant callers have been developed for accomplishing this challenging task as alignment and sequencing artefacts complicate the process of variant calling. The most widely used state-of-the-art variant callers include, but are not limited to, Genome Analysis Toolkit (GATK)-haplotypeCaller<sup>208</sup> and SAMTools<sup>206</sup>.

Bayesian models are used by many modern variant callers, such as SAMtools mpileup<sup>206</sup> or GATK UnifiedGenotyper<sup>208</sup>. A simple Bayesian genotyper for SNVs was described by McKenna et al,<sup>208</sup> namely, GATK HaplotypeCaller. GATK HaplotypeCaller uses a different approach for variant calling. It first looks for regions that are potentially variable by searching for a significant number of mismatches in aligned reads. For every such region, it constructs a local *de novo* assembly. *De novo* assemblies are represented as graphs where different paths in the graph represents different haplotypes. GATK HaplotypeCaller identifies the most likely haplotype in each graph and if this haplotype contains a variant, i.e., is not representing the reference haplotype, it calls the variant<sup>208</sup>.

### 2.10.4 Variant Filtration and Annotation

SAMtools, VarFilter and GATK VariantFiltration are tools that apply filters on called variants. However, fixed thresholds may not be suitable for different datasets. For instance, the read depth at a position depends on the amount of total sequence. Thus, applying the same read depth threshold for samples with different amounts of sequence is not optimal. GATK VariantRecalibrator provides a more sophisticated method for filtering. It uses a set of known high confidence variants, as found in the HapMap project, and searches for these variants in the set of called variants. It then models the distribution of these variants relative to annotations such as read depth or mapping quality and clusters them. Scores are assigned to all variants based on their distance to the centre of these clusters. A variant is removed if that variant is too

far away from the centre of the cluster, i.e., its score is too low. The threshold for the score is based on the set of known variants: typically, the threshold is defined such that 99.9% of known variants in the dataset have a higher threshold and are therefore not filtered. The key assumption for this filter method is that known variants that occur at high frequency in a population are more likely to be true than novel variants that have not previously been seen<sup>208</sup>.

Variant annotation involves adding auxiliary metadata and knowledge to quality-filtered raw putative variant calls to enhance assessment of variant likely to impact function. Several computational variant annotation tools that produce integrated reports that can be used for further rule-based filtering have been developed. SnpEff<sup>209</sup> and Variant Effect Predictor (VEP)<sup>210</sup> are arguably the most widely used, with the latter gaining prominence. Both applications enable variant annotation based on Ensembl transcripts and also leverage the rich annotations in dbNSfp<sup>195</sup>, a database of curated annotations and functional effect predictions for all potential non-synonymous and splice-site single nucleotide variants in the human reference genome. SnpEff is an open-source tool that annotates variants and predicts their effect on genes by using an interval forest approach. SnpEff annotates variants based on their genomic locations such as intronic, untranslated region, upstream, downstream, splice site, or intergenic regions and predicts coding effects. SnpEff also generates extensive report files and is easily customizable. VEP, on the other hand, is an open-source, free-to-use toolset for analysis, annotation, and prioritization of genomic variants in coding and non-coding regions<sup>195</sup>.

The aim of all functional annotation tools is to annotate information of the variant effects/consequences, including but not limited to (i) listing which gene(s)/ transcript(s) are affected, (ii) determination of the effect of a variant on a protein sequence, (iii) correlation of the variant with known genomic annotations (e.g. coding sequence, intronic sequence, noncoding RNA and regulatory regions), and (iv) matching potential variants to known variants found in the different databases (e.g. dbSNP<sup>211</sup>, 1000 Genome Project<sup>212</sup>, ExAc<sup>213</sup>, gnomAD<sup>214</sup>, and ClinVar<sup>215</sup>).

## 2.10.5 Additional filtering procedures

### 2.10.5.1 Inheritance filtering

Working with an inheritance model is a useful way to filter out and reduce the number of variants that have to be analysed downstream<sup>195</sup>. Once a set of variants is selected based on correct segregation with the target group or individual of interest, further selection can be carried. There are a few flexible software frameworks for exploring genetic variation annotation as well as filter based on user-provided pedigree file and specified mode of inheritance. GEMINI (Genome MINing) is one of these software frameworks<sup>216</sup>. The GEMINI framework begins by loading a VCF file into a database, each variant is automatically annotated by comparing it to several genome annotations from sources such as ENCODE tracks, UCSC tracks, OMIM, dnSNP, KEGG, and HPRD<sup>216</sup>.

### 2.10.5.2 Knowledge-driven variant filtering

Even with strong experimental design and variant filtering protocols, WGS studies often produce large amounts of candidate variants with likely functional effects than what can be verified experimentally. While it is possible to rank variants based on predicted impact on a given protein, identification of the strongest candidate variant for a particular disease or phenotype is not always obvious. For this reason, assessing candidate genes possessing functional variants in the context of existing biochemical knowledge and their known biomolecular functions is an important step in producing a manageable set of variants for further validation or exploring. However, when doing a large study this may still leave a large number of variants remaining. Extant knowledge driven variant prioritization like clinical phenotype specific to the disease may help reduce the number of candidate variants even further.

There are a wide range of software and/or tools available to assist in the above assessment. Some of these software includes ToppGene, STRINGdb, GeneMania, Phenolyzer, Exome Walker, OMIM explorer, Exomiser, DisGeNet, Monarch Initiative, Mammalian Phenotype Ontology (MPO), REACTOME, and KEGG among many others<sup>195</sup>.

#### 2.10.5.2.1 Functional gene annotation

Gene Ontology (GO) is widely used in biological databases, annotation projects and computational analyses for annotating newly sequenced genomes and clinical applications. It is an important bioinformatics tool for genome-scale protein function annotation. It tries to

explain the roles of genes or proteins in eukaryotic cellular process through establishment of a controlled vocabulary. GO has two components: the ontologies themselves, which are the defined terms and the structured relationship between them (GO ontology); and the associations between gene products and the terms (GO annotations). GO provides both ontologies and annotations for three distinct areas of cell biology: molecular function (basic activities of a gene product at the molecular level, such as binding or catalysis), biological process (collection of molecular events or operation), and cellular components or locations (components of cells or extracellular)<sup>217-219</sup>.

Several tools have been developed to assist in explaining the role of genes or proteins in different species. These include but are not limited to Panther, GOnet, geneontology and the ToppGene suite. The ToppGene suite is a website that is free and open to all users and does not require a login to access. It is a one-stop portal for (i) gene list functional enrichment, (ii) candidate gene prioritization using either functional annotations or network analysis and (iii) identification and prioritization of novel disease candidate genes in the interactome. The ToppGene suite has an application call ToppFun, which detects functional enrichment of input gene list based on Transcriptome (gene expression), Proteome (protein domains and interactions), Regulome (TFBS and miRNA), Ontologies (GO, Pathway), Phenotype (human disease and mouse phenotype), Pharmacome (Drug-Gene associations) and Bibliome (literature co-citations). The application supported identifiers include NCBI Entrez gene IDs, approved human gene symbols, NCBI reference sequence accession numbers as a single gene list. The application then gives as output HTML files, tab-delimited downloadable text file and graphical charts<sup>220</sup>.

#### 2.10.5.2.2 Protein interactions

Proteins do not act alone but usually interact with one another to carry out a specific biological function. Proteins are often assembled into complexes that perform specific functions related to structure, metabolism, growth, and communication. Protein-Protein Interactions (PPIs) occur when two proteins physically bind together to form functional modules and pathways that carry out most cellular processes.

Much like genes encoding proteins that occur in the same pathway as a known disease-gene product may also cause the disease if mutated, so too may proteins that physically interact with known disease-gene products. The STRING database and associated tools are powerful resources for identifying interacting partners of candidate gene's product or to identify



interactions between the products of a set of genes that bear functional variants<sup>195,221,222</sup>. STRING not only returns physical protein-protein interactions but also functional interactions<sup>223</sup>. Besides being able to list the interacting proteins it also provides a scale or score indicating the confidence of interaction. Such scores are often calculated based on the number of evidence and the experimental techniques/prediction methods used/reported by researchers for identifying the interactions. STRING database among the few databases (e.g STRING, menthe, APHID and HuRI) allow the user to import multiple genes/proteins as query, where the output includes interactions among the query proteins<sup>224</sup>. According to Bajpai et al,<sup>224</sup> their comparative study showed that following nPRINT, STRING may be ideal for obtaining the maximum number of proteins interactions<sup>224</sup>.

The STRING database is one of the most established databases, provides, additional sources and methods, e.g., text mining, correlation studies, and biological experiments. There are several other resources that integrate PPI data from different databases, allowing for the construction, visualization, and analysis of protein interaction networks.

#### 2.10.5.2.3 Impact prediction filtering

There is a plethora of bioinformatic algorithms that combine various types of parameters from multiple sources to infer deleteriousness when detailed experimental evaluation of individual variants is unavailable<sup>198</sup>. However, many of these algorithms are typically restricted to SNVs falling within protein-coding regions of the genome, with a particular focus on non-synonymous SNVs (nsSNVs). Most candidate variants lie in non-coding sections of the genome, whose role in maintaining normal genome function is still not well understood. Most annotation methods can only annotate protein coding variants, excluding more than 98% of the human genome.

Recently two computational approaches – (GWAVA) and the Combined Annotation Dependent Depletion (CADD) – were published to predict the deleterious effect of variants genome-wide. These two methods utilize machine-learning models trained on potential pathogenic variants or nearly fixed or fixed human derived alleles to distinguish deleterious variants from neutral ones. CADD is used for integrating diverse genome annotations and scoring any possible human single nucleotide variant (SNV) or small insertion/deletion (InDel) event (both coding and non-coding). The basis of CADD is to contrast the annotations of fixed but not simulated variants. CADD therefore measures deleteriousness, a property that strongly correlates with both molecular functionality and pathogenicity<sup>225</sup>. Its predictions are based on

a logistic regression model that account evolutionary conservation, regulatory and transcript information, and protein-level scores. CADD trains a linear kernel support vector machine (SVM) to separate observed genetic variants from simulated genetic variants. CADD's SVM can only learn linear representations of the data, which limits its performance<sup>226</sup>.

To overcome CADDs' limitation, a deep neural network algorithm (DNN) was implemented, called deleterious annotation of genetic variants using neural networks (DANN). DANN uses the same feature set and training data as CADD to train a deep neural network (DNN). DNNs can capture non-linear relationships among features and are better suited than SVMs for problems with large number of samples and features<sup>226</sup>. A newer tool was developed namely FATHMM-MKL which is a machine-learning approach that integrates functional annotations from ENCODE with nucleotide-based sequence conservation measures. This tool outperforms the more recent GWAVA and CADD prediction algorithms<sup>227</sup>.

PredictSNP integrates the above four tools together with FunSeq2 as a consensus scoring procedure for the five best performing nucleotide-based tools. Funseq2 uses an empirical scoring system that integrates evolutionary constraints, epigenetic data and knowledge of transcription-binding motifs to assess the impact of variants<sup>228</sup>. The above six tools represent diverse predictive approaches leveraging different training datasets, machine learning models, and combinations of decision features.



### 3 Chapter Three: Research Methodology

In this chapter a detailed methodology is described. To briefly summarize, raw paired-end sequenced reads quality trimming and adapter sequence removal were carried out by BGI sequencing service. Subsequent steps were carried out as part to this thesis. Local FastQC analysis was performed on the raw sequenced reads. High-quality reads were then mapped to the human genome reference sequence (hg38) using BWA-mem. After marking PCR duplicates, GATK base quality scores recalibration, InDel realignment, SNP and InDel discovery, and standard VSQR filtering parameters across all samples were performed according to GATK's best practices recommendation. Identified variants were annotated using SnpEff and VEP. Pathogenicity of missense variants were predicted in silico using ensemble scores based on several component scores and predictions (PredictSNP, CADD, DANN, GWAVA, FUNSeq and FATHMM). Variant filtering was done using various standard filters including, mode of inheritance, variant location, knowledge-based information and the SAHGP control group.

#### 3.1 Study Participants

The four study participants consisted of non-consanguineous parents and twins who presented with a severe form of OI (**Figure 3.2**).

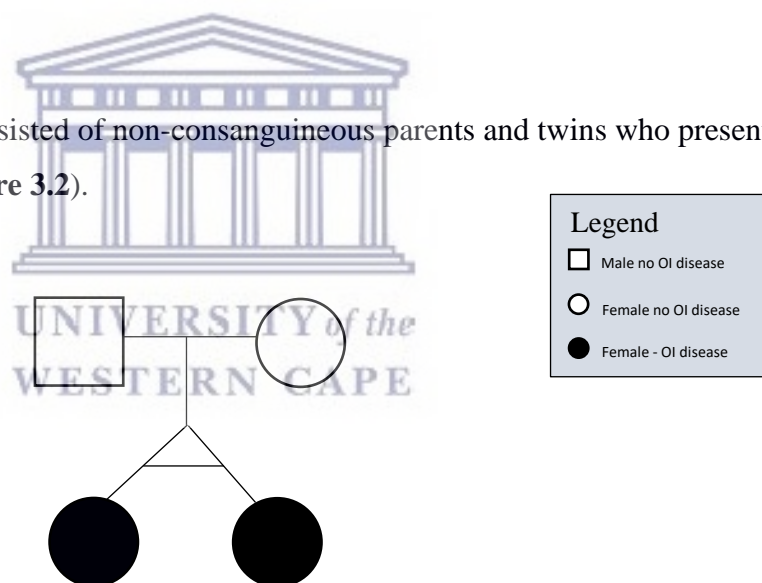


Figure 3.1: Pedigree the family in this study

The pedigree represents affected monozygotic (identical) twins and their unaffected non-consanguineous (unrelated) parents.

The family was chosen because they had severe form of OI and had not been previously genetically tested. The researcher worked in conjunction with the Department of Craniofacial Biology, University of the Western Cape (UWC), Tygerberg Dental Hospital, where the participants were recruited by Associate Professor Manogari Chetty.

### 3.2 Sample collection and DNA extraction

Written informed consent from all four participants were obtained (Appendix B, page 122), and saliva samples were collected. The Oragene DNA (OG-500) self-collection kit from DNA Genotek Inc (Ontario, Canada), was used for collection and storage of the saliva sample. The saliva collection kits were given to the four participants who took the kits home and then collected saliva as per the instruction accompanying the collection kits. The samples were then stored at room temperature at the Dental Genetics Department, from date of saliva collection (07 August 2018) until date of transfer to the Tygerberg Hospital Biobank based at the Haematology Department, Stellenbosch University (14 August 2018) where DNA extraction was taking place. The samples were kept at room temperature at the DNA extraction facility, until the day of extraction which was on the 10<sup>th</sup> of September 2018. This study made use of saliva samples as it was the less invasive method for getting DNA samples. A study by Garbier *et al*<sup>229</sup>, showed that extracting DNA from saliva can produce DNA of high quantities and good quality. The study by Garbier *et al*<sup>229</sup>, also indicate that compared to blood, saliva has the following advantages: allows for remote collection, can be frozen before DNA extraction, is painless and had a decreased risk of disease transmission and can be stored in tightly capped containers at room temperature for many years with the expectation that high molecular weight DNA will still remain present<sup>229,230</sup>.

The DNA was extracted using the Chemagen extraction kit according to manufacturers' instructions (PerkinElmer Inc, USA). This procedure was carried out according to the manufacturer's instructions. During the initial DNA extraction procedure, the DNA was treated with Proteinase K to remove all the proteins that may contaminate the samples. In addition to treating the DNA samples with Proteinase K, per the sequencing requirements, the DNA samples were also treated with two ribonucleases. The two ribonucleases allowed for double digestion of RNA because treatment with Ribonuclease A alone is not enough to degrade RNA into soluble-alcohol fragments. The quantification of the extracted DNA was done using the NanoDrop One according to manufacturers' instructions (Thermo Fisher Scientific Inc, UK).

### 3.3 DNA Sequencing

Prior to DNA sequencing, the DNA was stored at -20 degrees Celsius at the Tygerberg Hospital Biobank. The samples were then shipped to Hong Kong on dry ice to ensure a minimum freeze-thaw cycle. All library preparations and Illumina sequencing was performed at the Beijing Genomic Institute (BGI) in Hong Kong using the BGISEQ platform. Whole-genome

sequencing was undertaken at 30x coverage and 100bp paired-end reads. After sequencing, the raw reads were filtered using the FastQC tool (version 0.11.7) at the BGI sequencing facility. The FastQC filtering included removing adaptor sequences, contamination, and low-quality reads from raw reads.

### 3.4 Data pre-processing

The high-quality clean sequence reads received from BGI were in FASTQ format. The raw data received from BGI was stored on a one tera-byte hard drive, in four folders each containing 32 FASTQ files. These FASTQ files were merged using the bash command line function ‘cat’. The merging resulted in each of the four folders containing two FASTQ files (each sample file approximately 90 giga-bytes in size), one storing the forward read and the other storing the reverse read data.

To assess the quality of these sequence reads and to prevent any inconsistencies in the downstream analysis steps, these files were subjected to additional quality checking using the FastQC (version 0.11.7) tool.

#### 3.4.1 Alignment and mapping

The high-quality clean reads received from BGI were first indexed and then aligned to the human reference genome hg38 using the Burrow Wheeler Alignment tool (BWA - version 0.7.17). For this study, the BWA-MEM algorithm was used.

An additional index file was generated using SAMtools. The FASTA file index file was generated using the following SAMtools command<sup>47,231,232</sup>:

```
samtools faidx hg38.fa
```

This created a “ref.fa.fai” file, with one record per line for each of the contigs in the FASTA reference file.

Next, a sequence dictionary was generated by running the following Picard command<sup>232,233</sup>:

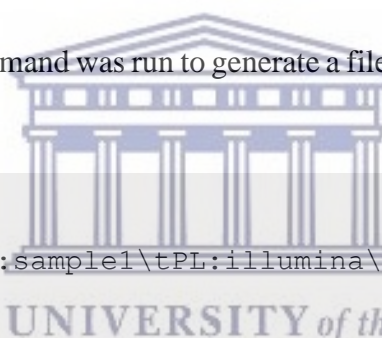
```
Java -jar CreateSequenceDictionary.jar \  
R= ~/Analysis_data/GATK/refs/hg38.fa \  
O = ref.dict
```

This created a file called “ref.dict” formatted like a SAM header, describing the content of the reference FATS file.

Note that the index (ref.fa.fai) and dictionary (ref.dict) files have the same prefix names and are required by the GATK software but not explicitly provided as input on the command line. Specifically, GATK assumes that these files are present in the “current” directory.

The final preparation before proceeding to the alignment and the rest of the analysis is to prepare the appropriate read group information. The read group information is where you enter the meta-data about the samples. This line is very important for all downstream analysis, since it will be the only meta-information that will be visible to analysis tools. There are multiple fields in the Read Group tag, but some of them are of critical importance, for example, the globally unique string identifier, name associated with the DNA sample in the file, platform used, DNA sequencing library identifier, and a platform unit identifier. This read group information is key for downstream GATK functionality as GATK will not work without a Read Group tag.

The following BWA-MEM command was run to generate a file containing the aligned reads<sup>232-234</sup>



```
bwa mem \  
-R '@RG\tID:group1\tSM:sample1\tPL:illumina\tLB:lib1\tPU:unit1' \  
-M -t 16\  
/tools/software/bcbio/genomes/Hsapiens/hg38/bwa/hg38.fa \  
combined_r1.fq.gz combined_r2.fq.gz > aligned1.sam
```

This created a SAM file called “aligned1.sam”. This file contains the reads aligned to the reference. The header sections contain contigs of aligned reference sequence, read groups (carrying platform, library, and sample information), and (optionally) data processing tools applied to the reads. The alignment section also includes information on the alignment of reads.

For the BWA-MEM part of the above command, the following link from the BWA manual was used (<http://bio-bwa.sourceforge.net/bwa.shtml>) and for the read group tags the link (<https://gatk.broadinstitute.org/hc/en-us/articles/360037226472-AddOrReplaceReadGroups-Picard->) was adapted. The human (hg38) reference genome as downloaded by bcbio-nextgen (<https://zenodo.org/record/5781867>), which refers to the github repo: (<https://github.com/bcbio/bcbio-nextgen>) was used.

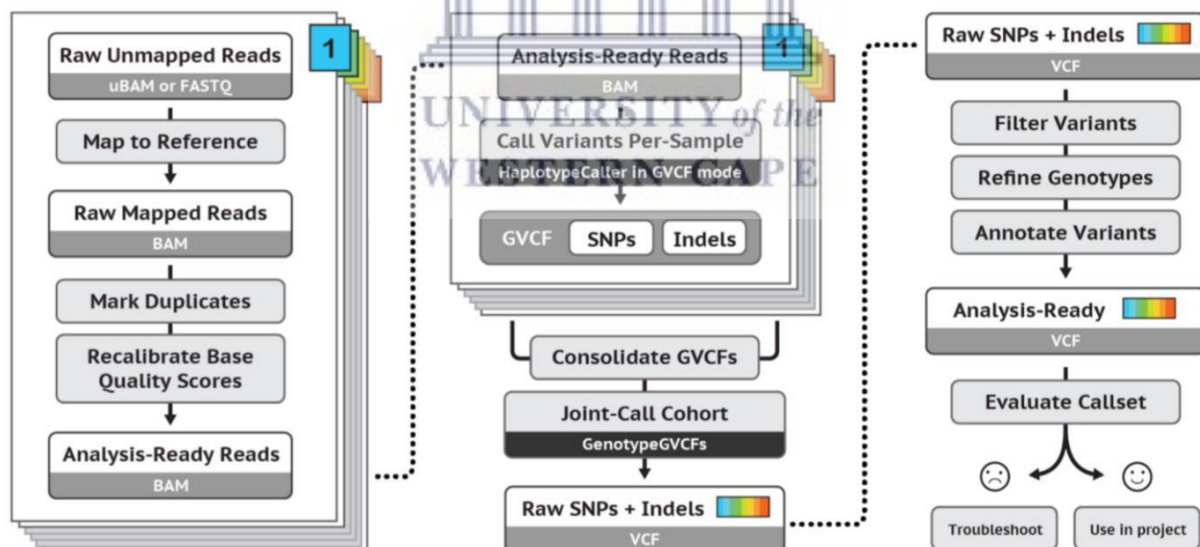
### 3.4.2 Marking duplicates, SAM-to-BAM conversion, and sorting

The output SAM formatted files (aligned1.sam) were then converted to BAM formatted files using the ‘view’ command<sup>233</sup>. Then sorting and indexing were done using the ‘sort’ and ‘index’ commands in SAMTools (version 1.9) (<http://www.htslib.org/doc/samtools.html>). To account for any PCR duplicates, these reads were marked using the Picard tool<sup>232,233</sup> (version 2.20.1). Script was adapted from the following link: <https://gatk.broadinstitute.org/hc/en-us/articles/360037872491--How-to-Fix-a-badly-formatted-BAM>.

```
java -jar /usr/local/share/picard-2.20.1-0/picard.jar
MarkDuplicates \
I=align1_sorted.bam \
O=marked_dup.bam \
M=marked_dup_metrics.txt
```

### 3.5 GATK’s best practice variant discovery

The raw scores produced by the sequencing machine are prone to technical errors, leading to over- or underestimated base quality scores. GATK’s best practice steps (**Figure 3.3**) allows for the correction of these base scores.



**Figure 3.2:** Workflow for NGS data analysis and GATK’s variant discovery used in this study.

Beginning at variant calling per sample to produce a file in GVCF format. Then perform joint genotyping, and finally applying VQSR filtering to produce the final multi- sample callset (<https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant-discovery-SNPs-Indels->).

### 3.5.1 Local realignment

The mapping algorithm (BWA) that is used in the initial step of aligning the data to the reference is prone to various types of artifacts. The realignment process identifies the most consistent placement of the reads with respect to the InDel to clean up these artifacts. It occurs in two steps: first the program identifies intervals that need to be realigned, then in the second step it determines the optimal consensus sequence and performs the actual realignment of reads. A target list of intervals to be realigned was used by running the following GATK command. Adapted from the GATK website like ([https://github.com/broadinstitute/gatk-docs/blob/master/gatk3-tutorials/\(howto\) Perform local realignment around InDels.md](https://github.com/broadinstitute/gatk-docs/blob/master/gatk3-tutorials/(howto) Perform local realignment around InDels.md)):

```
java -jar /usr/people/pvh/miniconda3/envs/mtbseq/opt/gatk-3.8/GenomeAnalysisTK.jar \  
-T RealignerTargetCreator \  
-R ~/Analysis_data/GATK/refs/hg38.fa \  
-known ~/Analysis_data/GATK/Mills_and_1000G_gold_standard.indels.hg38.vcf \  
-known ~/Analysis_data/GATK/Homo_sapiens_assembly38.known_indels.vcf \  
-I marked_dup.bam \  
-o targetlist.intervals
```

This created a file called “targetlist.intervals” containing the list of intervals that the program identified as needing realignment. The list of known InDel sites (Mills and 1000G gold standard InDels and Homo\_sapiens assembly38.known\_InDels) are used as targets for realignment (<https://gatk.broadinstitute.org/hc/en-us/articles/360035890811-Resource-bundle>). The realignment step was then performed using the ‘targetlist.intervals’.

The following GATK command was then used (adapted from [https://github.com/broadinstitute/gatk-docs/blob/master/gatk3-tutorials/\(howto\) Perform local realignment around InDels.md](https://github.com/broadinstitute/gatk-docs/blob/master/gatk3-tutorials/(howto) Perform local realignment around InDels.md))

This created a file called “InDelrealigned.bam” containing all the original reads, but with better local alignments in the regions that were realigned.



```

java -jar /usr/people/pvh/miniconda3/envs/mtbseq/opt/gatk-
3.8/GenomeAnalysisTK.jar \
-T IndelRealigner \
-R ~/Analysis_data/GATK/refs/hg38.fa \
-targetIntervals forIndelRealigner.intervals \
-known ~/Analysis_data/GATK/Mills_and_1000G_gold_standard.indels.hg38.vcf \
-known ~/Analysis_data/GATK/Homo_sapiens_assembly38.known_indels.vcf \
-l marked_dup.bam \
-maxReads 20000 \
-o InDelrealigned.bam

```

### 3.5.2 Base Quality Score Recalibration (BQSR)

This step assigns accurate quality scores to each sequenced base because the quality scores issued by the sequencer are sometimes inaccurate and biased. Base quality score recalibration (BQSR) is a machine learning approach that models these errors empirically and reads just the base quality scores accordingly. BQSR takes two complementary paths namely, (a) data processing and (b) plotting.

- (a) For the data processing path, the 'InDelrealigned1.bam' file was used to build a recalibration model. This was performed running the following GATK command<sup>232</sup> (both steps 'a' and 'b' was adapted from - [https://github.com/broadinstitute/gatk-docs/blob/master/gatk3-tutorials/\(howto\) Recalibrate base quality scores %3D run BQSR.md](https://github.com/broadinstitute/gatk-docs/blob/master/gatk3-tutorials/(howto) Recalibrate base quality scores %3D run BQSR.md)).

```

java -jar /usr/people/pvh/miniconda3/envs/mtbseq/opt/gatk-
3.8/GenomeAnalysisTK.jar \
-T BaseRecalibrator \
-R ~/Analysis_data/GATK/refs/hg38.fa \
--knownSites ~/Analysis_data/GATK/Mills_and_1000G_gold_standard.indels.hg38.vcf \
--knownSites ~/Analysis_data/GATK/Homo_sapiens_assembly38.known_indels.vcf \
--knownSites ~/Analysis_data/GATK/Homo_sapiens_assembly38.dbsnp138.vcf \
-l InDelrealigned.bam \
-o recal_data.table

```

The next step is applying numerical correction to each individual basecall based on the patterns identified in the first step (recorded in the recalibration table). This was executed using the following GATK command<sup>232</sup>:

```
java -jar /usr/people/pvh/miniconda3/envs/mtbseq/opt/gatk-3.8/GenomeAnalysisTK.jar \  
-T PrintReads \  
-R ~/Analysis_data/GATK/refs/hg38.fa \  
-I indelrealigned1.bam \  
-BQSR recal_data.table \  
-o recalibrated.bam
```

The above script created a 'recalibrated.bam' file, which has accurate base substitution, insertion, and deletion quality scores.

(b) The plot path evaluates what the data looks like after recalibration. The following GATK command was run:

```
java -jar /usr/people/pvh/miniconda3/envs/mtbseq/opt/gatk-3.8/GenomeAnalysisTK.jar \  
-T BaseRecalibrator \  
-R ~/Analysis_data/GATK/refs/hg38.fa \  
--knownSites ~/Analysis_data/GATK/Mills_and_1000G_gold_standard.indels.hg38.vcf \  
--knownSites ~/Analysis_data/GATK/Homo_sapiens_assembly38.known_indels.vcf \  
--knownSites ~/Analysis_data/GATK/Homo_sapiens_assembly38.dbsnp138.vcf \  
-I indelrealigned.bam \  
-BQSR recal_data.table \  
-o secpass_recal_data.table
```

To generate the plots based on the before/after recalibration tables, the following GATK command was run:

```
java -jar /usr/people/pvh/miniconda3/envs/mtbseq/opt/gatk-3.8/GenomeAnalysisTK.jar \  
-T AnalyzeCovariates \  
-R ~/Analysis_data/GATK/refs/hg38.fa \  
-before recal_data.table \  
-after secpass_recal_data.table \  
-csv BQSR.csv \  
-plots recalQC.pdf \  

```

This produced a pdf file with plots of before and after recalibration scores.

### 3.5.3 Variant Calling

For this study, only single nucleotide polymorphisms (SNPs) and insertion- deletions (InDels) was considered.

The variant calling was performed by running the following GATK command ([https://github.com/broadinstitute/gatk-docs/blob/master/gatk3-tutorials/\(howto\) Call variants with HaplotypeCaller.md](https://github.com/broadinstitute/gatk-docs/blob/master/gatk3-tutorials/(howto) Call variants with HaplotypeCaller.md)):

```
java -jar /usr/people/pvh/miniconda3/envs/mtbseq/opt/gatk-3.8/GenomeAnalysisTK.jar \  
-T HaplotypeCaller \  
-R ~/Analysis_data/GATK/refs/hg38.fa \  
-I recalibrated.bam \  
--variant_index_type LINEAR \  
--variant_index_parameter 128000 \  
--genotyping_mode DISCOVERY \  
--max_alternate_alleles 100 \  
-stand_call_conf 10 -ERC GVCF -o \  
-raw_variants.g.vcf
```

GATK's HaplotypeCaller was used to perform this task, which outputs a GVCF file containing raw variant calls. The HaplotypeCaller tool calls SNPs and InDels simultaneously via local re-

assembly of haplotypes in an active region, as opposed to GATK's UnifiedGenotyper.

### 3.5.3.1 Joint genotyping

The called variants were then merged into a single file that consisted of all four individuals. This step was accomplished using the GATK's 'joint genotyping' (<https://gatk.broadinstitute.org/hc/en-us/articles/360035890411-Calling-variants-on-cohorts-of-samples-using-the-HaplotypeCaller-in-GVCF-mode>), it takes output from the previous step and runs GenotypeGVCF on all the four genome files together to create raw SNP and InDel VCF files (raw\_variants.g.vcf).

```
java -jar /usr/people/pvh/miniconda3/envs/mtbseq/opt/gatk-3.8/GenomeAnalysisTK.jar \  
-T GenotypeGVCFs \  
-R ~/Analysis_data/GATK/refs/hg38.fa \  
-V raw_variants1.g.vcf \  
-V raw_variants2.g.vcf \  
-V raw_variants3.g.vcf \  
-V raw_variants4.g.vcf \  
--max_alternate_alleles 100 \  
-o jointcalls.g.vcf
```



UNIVERSITY of the  
WESTERN CAPE

### 3.5.4 Variant Quality Score Recalibration for SNPs

Following the joint genotyping step, raw SNP and InDels in the genomic Variant Call Format (GVCF) are obtained. These are then filtered through applying GATK's Variant Quality Score Recalibration (VQSR).

The aim is to assign a well-calibrated probability to each variant call to create accurate variant quality scores. In the first step of this two-step process, the program uses machine learning methods to assign a well-calibrated probability to each variant call in a raw call set. It then uses this variant quality score in the second step to filter the raw call set, producing a subset of calls with desired level of quality, fine-tuned to balance specificity and sensitivity.

To build the SNP recalibration model, the following GATK command was run (VQSR steps adapted from <https://gatk.broadinstitute.org/hc/en-us/articles/360035531112--How-to-Filter-variants-either-with-VQSR-or-by-hard-filtering> and <https://gatk.broadinstitute.org/hc/en->

[us/articles/360035531612-Variant-Quality-Score-Recalibration-VQSR- \):](https://www.ncbi.nlm.nih.gov/pmc/articles/360035531612-Variant-Quality-Score-Recalibration-VQSR-)

```
usr/people/pvh/miniconda3/envs/mtbseq/opt/gatk-3.8/GenomeAnalysisTK.jar \  
-T VariantRecalibrator \  
-R ~/Analysis_data/GATK/refs/hg38.fa \  
-input jointcalls.g.vcf \  
-resource:hapmap,known=false,training=true,truth=true,prior=15.0  
~/Analysis_data/GATK/hapmap_3.3.hg38.vcf \  
-resource:omni,known=false,training=true,truth=true,prior=12.0  
~/Analysis_data/GATK/1000G_omni2.5.hg38.vcf \  
-resource:1000G,known=false,training=true,truth=false,prior=10.0  
~/Analysis_data/GATK/1000G_phase1.snps.high_confidence.hg38.vcf \  
-resource:dbsnp,known=true,training=false,truth=false,prior=2.0  
~/Analysis_data/GATK/Homo_sapiens_assembly38.dbsnp138.vcf \  
-an DP -an QD -an FS -an SOR -an MQ -an MQRankSum -an  
ReadPosRankSum \  
-mode SNP \  
-tranche 100.0 -tranche 99.9 -tranche 99.0 -tranche 90.0 \  
-recalFile recalibrate_SNP.recal \  
-tranchesFile recalibrate_SNP1.tranches \  
-rscriptFile recalibrate_SNP_plots.R
```

This creates several files, the most important of these in the recalibration report, called 'recalibrate\_SNP.recal'. This file contains the recalibration data and is the file the program will use in the next step to generate a VCF file.

To apply the desired level of recalibration to the SNPs in the call set, the following GATK command was run:

```

java -jar /usr/people/pvh/miniconda3/envs/mtbseq/opt/gatk-
3.8/GenomeAnalysisTK.jar \
-T ApplyRecalibration \
-R ~/Analysis_data/GATK/refs/hg38.fa \
-input raw_variants.g.vcf \
-mode SNP \
--ts_filter_level 99.0 \
-recalFile recalibrate_SNP.recal \
-tranchesFile recalibrate_SNP.tranches \
-o recalibrated_snp_raw.g.vcf

```

This creates a VCF file called, ‘recalibrated\_snp\_raw.g.vcf’, which contains all the original variants from the original ‘raw\_variants.vcf’ file, but now the SNPs are annotated with their recalibrated quality scores (VQSLOD) and either PASS or FILTER depending on whether they are included in the selected tranche.

The output files from the VQSR, along with the raw\_SNP.vcf file serves as an input for GATK’s ApplyRecalibration tool. The output from this step was a recalibrated SNP file (‘recalibrated\_snp\_raw.g.vcf’).

#### 3.5.4.1 Variant Quality Score Recalibration for InDels

The GATK command used to build the InDel recalibration model was as follows:

```

java -jar /usr/people/pvh/miniconda3/envs/mtbseq/opt/gatk-
3.8/GenomeAnalysisTK.jar \
-T VariantRecalibrator \
-R ~/Analysis_data/GATK/refs/hg38.fa \
-input recalibrated_snp_raw_indels1.g.vcf \
-resource:mills,known=false,training=true,truth=true,prior=12.0
~/Analysis_data/GATK/Mills_and_1000G_gold_standard.indels.hg38.vcf \
-resource:dbsnp,known=true,training=false,truth=false,prior=2.0
~/Analysis_data/GATK/Homo_sapiens_assembly38.dbsnp138.vcf \
-an DP -an QD -an FS -an SOR -an MQ -an MQRankSum -an ReadPosRankSum \
-mode InDel\
-tranche 100.0 -tranche 99.9 -tranche 99.0 -tranche 90.0 \
--maxGaussians 4 \
-recalFile recalibrate_InDel1.recal \

```

To apply the desired level of recalibration to the InDels, the following GATK command was run:

```
java -jar /usr/people/pvh/miniconda3/envs/mtbseq/opt/gatk-3.8/GenomeAnalysisTK.jar \  
-T ApplyRecalibration \  
-R ~/Analysis_data/GATK/refs/hg38.fa \  
-input recalibrated_snp_raw_indels1.vcf \  
-mode InDel --ts_filter_level 99.0 \  
-recalFile recalibrate_InDel1.recal \  
-tranchesFile recalibrate_InDel1.tranches \  
-o recalibrated_variants1.g.vcf
```

The above script produced a final VCF file that contains both SNPs and InDels with recalibrated scores.

### 3.6 Decomposition and neutralization

In order to cater for multiallelic variants in the VCF file, the VT tools was used to decompose the variants so that variants with multiple alleles are expanded into distinct variant records; one record for each REF/ALT combination. The following VT command was run to perform this task:

```
vt decompose -s recalibrated_variants1.g.vcf -o decomp.g.vcf
```

The decomposed VCF files were neutralized so that variants are left aligned and represent the most parsimonious alleles. The following VT command was run for this purpose:

```
vt normalize -r ~/Analysis_data/GATK/refs/hg38.fa decomp.g.vcf -o norm.g.vcf
```

The decomposition and neutralization commands were adapted from <https://genome.sph.umich.edu/wiki/Vt#Decompose>.

### 3.7 Variant annotation

The recalibration files were annotated using the SnpEff annotation tool and GATKs' VariantAnnotator and Ensembl's Variant Effect Predictor (VEP). For each of the annotation procedures the following commands were used.

#### **SnpEff**

```
SnpEff -c ~/Analysis_data/annotation/SnpEff/SnpEff.config \  
-v GRCh38.92 \  
-o gatk \
```

#### VEP

```
vep -i norm.g.vcf \  
--cache --dir_cache /tools/databases/vep/vep102/ \  
-o ann_vep.g.vcf
```

### 3.8 Variant Filtering

#### 3.8.1 Mode of inheritance

GEMINI (v 0.30.2) was used to filter the variants based on inheritance. For autosomal dominant variants the commands labelled 'A' was used. For autosomal recessive (homozygous) variants 'B' was used. For compound heterozygous variants 'C' was used, and for de novo variants 'D' was used, all with default parameters.

- A. `gemini autosomal_dominant my_db.db`
- B. `gemini autosomal_recessive my_db.db`
- C. `gemini comp_het my_db.db`
- D. `gemini de_novo my_db.db`

UNIVERSITY of the  
WESTERN CAPE

The above commands extracted all (high, low, medium (moderate)) variants in the database created from the GVCF file (combined genome file of all participants in the study).

#### 3.9 Coding and non-coding separation

For each of the modes of inheritance the coding and non-coding variants were separated.

#### 3.10 SNPS and InDel separation

The results from 3.4.6.2 were then split into SNPs, insertions, and deletions. The python script in [Appendix G](#) was created and used to generate the SNP, insertions, and deletion files.



### 3.11 SAHGP population filter

'awk' command ([Appendix G, page 152](#)) was used to compare the remaining SNPs to data from the South African Human Genome Project (SAHGP). The SAHGP data set was made up of 23 healthy individuals' genome data that was grouped as follows: (1) seven individuals of mixed ancestry decent; (2) eight individuals of Sotho decent; (3) seven individuals of Xhosa decent; (4) one individual of Zulu decent<sup>235</sup>.

### 3.12 Gene level annotations and Pathway analysis

The ToppGene webserver (<https://www.toppgene.cchmc.org>) was used for GO enrichment, and REACTOME Pathway analyses. Specifically, the ToppFun module was selected from the main site and a total of 14 113 genes (**obtained from 3.4.6.2**) was used as input. The false discovery rate correlation cut-off parameter was set to p-value of 0.05 (p-value calculation method was based on probability density function). Of the 14 113 input genes, 3388 genes had duplicate names and/or the gene IDs were unknown (no synonyms could be identified). Finally, a total of 10 725 genes were used for the GO analysis.

A total of 7492 genes had at least one OI-candidate gene associated with GO enriched terms that were retained for further screening. GO-enriched terms that did not have an OI-candidate gene were ignored.

The 7492 genes were then used as input in REACTOME for pathway analysis. Pathways that had OI candidate genes featured in the gene list were retained for further analysis. A total of 4591 genes were left and used for further analysis.

### 3.13 Related disease and/or phenotype filter

The 4591 genes from Pathway analysis results was used as input in the DisGeNET database (v7.0) (<https://www.disgenet.org>) and/or Monarch Initiative Explorer database (<https://monarchinitiative.org>) to look for any gene-phenotype associations. Specific terms were used (based on the clinical phenotype of skeletal dysplasias) to reduce the number of gene-variants that could potentially play a role in the phenotype observed. At least one term per gene was enough to be included for further investigation. The terms used has been listed in the table below (**Table 3.1**). A total of 832 genes had skeletal dysplasias phenotype associations and were retained.

**Table 3.1:** List of terms used based on OI phenotype.

1. Fractures	2. Dentinogenesis imperfecta	3. Scoliosis
4. Long bone	5. Abnormality of dentin	6. Kyphoscoliosis
7. Craniofacial abnormalities	8. Short stature	9. Dwarfism
10. Osteopenia	11. Skeletal dysplasia	12. Hearing impairment
13. Abnormality of dentin enamel	14. Osteoporosis	15. Bowing of bones (short and long)
16. Blue sclerae	17. Abnormality of vertebral column	18. Abnormal form of the vertebral bodies
19. Rhizomelia	20. Tooth abnormalities	21. Kyphosis
22. Bone disease	23. Deformity of spine	24. Triangular face
25. Wormian bones		

### 3.14 Protein-Protein Interactions

The STRING database (v 11.0) was used to determine physical interactions of a total of 832 genes. The 832 genes were used as input in the query database, certain parameters were adjusted; ‘meaning of network edges’ were changes from ‘confidence’ (line thickness indicates the strength of data support) to ‘evidence’ (line colour indicates the type of interaction evidence). The network parameter was also set to exclude all nodes that were not linked. The rest of the settings were left unchanged. A total of 824 genes were linked in the STRING network.



### 3.15 Variant Pathogenicity

#### 3.15.1 SNP prediction

The prediction tool PredictSNP was used. This used as input a list of chromosome numbers, the start and end position, reference allele and alternate allele.

A total of 4295 variants (compound heterozygous), 9801 variants (homozygous) and 180 *de novo* variants were used as input in PredictSNP. Of the 4295 variants, 2 variants had reference allele that did not match with the selected genome assembly (GCHR38) and were not remappable to GCHR37 coordinates and were discarded by PredictSNP. The homozygous and *de novo* groups each had 147 and 5 variants that did not match and could not be remapped. The results obtained from PredictSNP was further filtered using bash awk commands, as follows:

(1) Predictions where all six tools predicted the variants to be ‘neutral’ were also discarded; (2) predictions where the variants were either predicted to be ‘neutral’ and/or had a ‘?’ for any of the prediction tools, were discarded. A total of 854 compound heterozygous variants, 3808 homozygous variants and 99 de novo variants remained.

### 3.16 Variant Prioritization

Variant prioritization aims to create a well-organized ranking of observed genetic variation. Due to large number of variants, it is important to identify, filter and prioritize those variants with association of the researchers’ target phenotype and to decrease the variants to a manageable number. **Figure 3.3** is a detailed breakdown of the variant prioritization that took place in this study.

### 3.17 Ethics Consideration

The researcher applied for ethical clearance to the University of the Western Cape Ethical Committee (BMREC). Upon approval of this application, data collection for this study began. In addition, ethical guidelines for informed consent, confidentiality and anonymity were adhered to.

The participants in this study were informed of the purpose of the study and were invited to participate in the research. The participants were assured anonymity and confidentiality with regards to publication of any specific details that could reveal their identity. The participants were also made aware of their ability to withdraw from the research study at any given time, date, and stage ([Appendix B, page 129](#)). Consent forms and information sheets were signed, upon agreement and understanding of the research and the role of each participant in this study.

## 4 Chapter Four: Results

### 4.1 Introduction

The purpose of this study was to identify genetic factors contributing to the phenotypic presentation of this skeletal dysplasia (OI type 3) presented by the two probands of mixed ancestry descent. Chapter four describes findings based on data collected from four family members; two unaffected non-consanguineous parents and twin siblings affected by OI. This chapter encompasses the demographic information, results from the NGS analysis: quality control analysis, mapping and/or alignment analysis results, the SNPs and InDel results, pathogenicity prediction, gene ontology, and protein-protein interaction results from a single family.

### 4.2 Participant demographics

The two female twins, aged 32 (when samples were obtained), presented with a clinically severe deforming OI. The parents were non-consanguineous and had no clinical representation of OI. This single family has no history of OI in the family.

The demographics of the study participants are summarized in **Table 4.1**. The birth weight of the probands were 1.5 kg and 1.6 kg with a height of 40 cm. The birth weight and height of the twins were low compared to the normal average weight of a healthy newborn baby, which is between 2.5 kg – 3 kg. There are factors like premature birth or mother's health condition that could contributed to the baby's low birth weight. In these situations, there are procedures to follow to get the baby back to normal weight. The average normal height of a healthy newborn baby is between 45.7 cm – 60 cm<sup>236</sup>. However, the weight and height of the twins did not normalize over the years, as their height seemed to have reached its peak at approximately 1 meter (height at the age of 32).

Proband-1 and Proband-2 were 2 weeks old and 6 weeks old, respectively, when they had their first fracture. These occurred in the femur. According to the probands' mother, they each suffered more than 100 fractures in all their bones and the fractures lessened over the years. The parents themselves had no history of fractures. The fractures in the twins w in the femurs, tibias, ulnar, jaws and ribs. The probands' showed no signs of muscle weakness and had 'relatively strong' muscles and had visible signs of triangular shaped faces. Both probands were born with bowed bones, the bowing was corrected with pins in their legs and arms. The bowing in their arms may have been escalated by their wheelchair usage. The twins made use

wheelchairs regularly but were not wheelchair dependent. They were able to put some weight on their legs, for instance proband-2 started walking independently at the age of seven but only for short distances and proband-1 uses crutches. Vertebral fractures, wormian bones and scoliosis curving to opposite sides were present in both twins. There is no history of intellectual disabilities. Up to now, both twins have had no history of bisphosphonate treatment. Both twins also have severe dentinogenesis imperfecta.

**Table 4.1:** Summary of study cohort demographics

Demographic Variable	Mother	Father	Proband-1	Proband-2
Age	N/A	N/A	32	32
Gender	F	M	F	F
Ethnic/Linguistic group	Mixed ancestry	Mixed ancestry	Mixed ancestry	Mixed ancestry
Dentinogenesis Imperfecta	Absent	Absent	Present	Present
Number of fractures	0	0	many	many
Wormian Bones	Absent	Absent	Present	Present
Bowing of bones (arms and legs)	Absent	Absent	Present	Present
Muscle weakness	Absent	Absent	Absent	Absent
Wheelchair dependent	Absent	Absent	Present	Present
Iscoliosis	Absent	Absent	Present	Present
Vertebral fractures	Absent	Absent	Present	Present
Mental retardation	Absent	Absent	Present	Present
Bisphosphonate treatment	N/A	N/A	None	None

N/A = not applicable; F = Female; M = Male; many = more than 100.

### 4.3 Quality control

The Beijing Genomic Institute (BGI) sequencing provider generated quality control data for the four sequenced genomes ([Appendix D](#)).

The local FastQC output includes, the ‘per base sequence quality’ that gives a general overview of the quality of the data ([Appendix E: Table E.1](#)). The overall base calls, as indicated by the yellow boxes located in the green area/background of the plots all fall within the ‘very good quality’ range of the quality scores. This summary was supported by other modules within the FastQC algorithm.

### 4.4 Read Alignment

The BWA–MEM tool was used to align the four human genome sequences to the hg38 reference genome (**Table 4.2**). On average 95.63% of the total reads mapped to the reference genome for all four genomes. No duplicate reads were observed for three of the four sequenced genomes. The genome sequence of the mother had a total of 30 445 785 duplicate reads, which

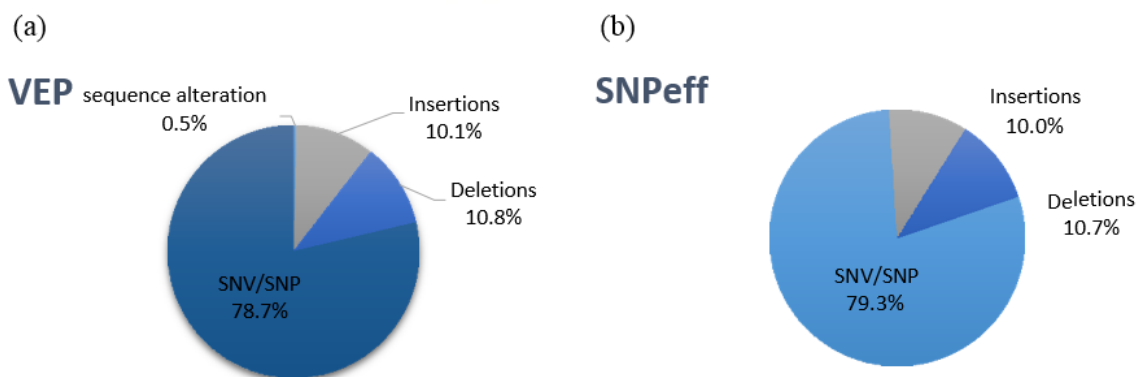
indicated 2.7% duplicate reads in the data (**Table 4.2**). This duplication may be due to the same DNA fragment being sequenced twice or during the sequencing procedure a copy of the same read was created and sequenced.

**Table 4.2:** The number of reads aligned and paired to the human reference genome GRCh38.

Flagstats groups	Mother	Father	Proband-2	Proband-1
<b>Total reads</b>	1111883311	1101743170	1108307613	1101980287
<b>Duplicates</b>	30445785	0	0	0
<b>Mapped</b>	1029992107 (92.63%)	1074823704 (97.56%)	1040212809 (93.86%)	1085156686 (98.47%)
<b>Paired in sequencing</b>	1103537098	1092571238	1100515358	1094067942
<b>Properly paired</b>	999213814 (90.55%)	1047463984 (95.87%)	1013850416 (92.13%)	1056785418 (96.59%)
<b>Singletons</b>	1526598 (0.14%)	1094996 (0.10%)	1516570 (0.14%)	737203 (0.07%)

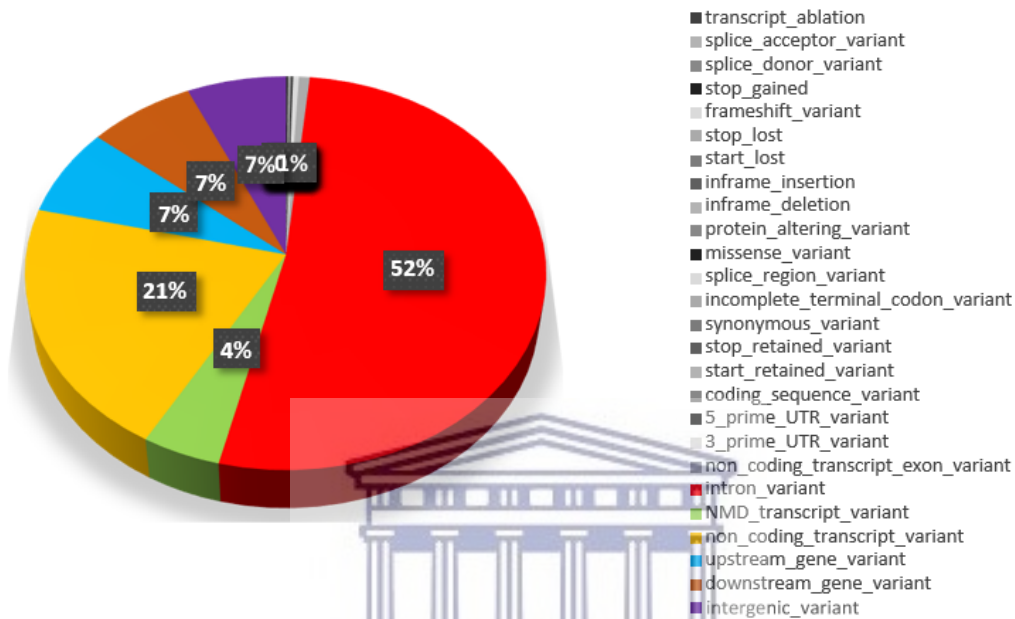
#### 4.5 Variant Discovery

More than eight million variants were identified for all participants by GATK. Variants were annotated by SnpEff (8 351 434) and VEP (8 242 224) as SNPs, deletions and insertions (**Figure 4.1**). For both annotation tools approximately 80% of the variants were described as variants, with insertions and deletions each accounting for approximately 10% of the rest of the variants identified



**Figure 4.1:** Number of variants identified by SNP annotation software. Variants annotated as SNPs, Insertions and deletions by (a) VEP and (b) SnpEff

The genomic location of all variants as annotated by VEP shows that 52% of variants are located in introns (Figure 4.3). At least half of the coding variants represent synonymous variants (see Appendix F, for the genomic location of all coding variants). The remainder variants have potential functional consequences for the corresponding gene. There is no evidence to suggest enrichment of these variants on specific chromosomes ([Appendix F: Figure F.1](#)).



**Figure 4.2:** Genomic location of all variants as annotated by VEP. Percentage of all variants grouped into their respective locations

## 4.6 Mode of inheritance

The use of inheritance patterns was the starting point to remove non-informative variants. There were no variants identified for the dominant mode of inheritance when using the GEMINI database tool ([Chapter 3, section 3.4.6.1](#)).

### 4.6.1 Homozygous Variants in both twins

Second, using the same GEMINI database tool, a total of 272 937 homozygous variants were identified when screening for autosomal recessive inherited variants (**Table 4.3**). This number consisted of high, low, and medium impacting variants across all genes found to segregate in an autosomal recessive manner. Seventy-three of these variants were high impacting, however none of these 73 variants were found in the OI-candidate genes. There were 959 variants found

with medium impact, of which the majority were missense variants with some ‘inframe insertion’ and ‘disruptive inframe deletion’ variants. None of these variants were located in OI-candidate genes. Majority of the homozygous variants (271 905) were annotated as low impacting. These variants were found in the following OI-candidate genes: *SERPINF1*, *BMP1*, *CREB3L1*, *TENT5A*, *COL1A2*, *PLOD2*, and *CRTAP*.

#### 4.6.2 Compound heterozygous variants in both twins

A total of 47 283 compound heterozygous variants were identified (**Table 4.3**). Compound heterozygous variants occur when the individual inherits one alternate allele from each of their parents, and these alleles are located at different loci on the same gene (see Chapter 2 section 2.3.1.2). Fourteen out of the 47 283 variants (0.03%) were high impacting and 4250 (8.9%) were of medium impact. The medium impacting variants were mostly missense variants, with the minority made up of ‘disruptive inframe insertion’, ‘disruptive inframe deletion’, ‘inframe insertion’ and ‘inframe deletion’. None of the high and medium impacting variants were located in OI-candidate genes. The low impacting variants included variants that were located in the *COL1A1* gene.

#### 4.6.3 *de novo* variants

GEMINI identified 35 625 *de novo* variants (**Table 4.3**). A total of 16 were high impact, 189 were medium impact and the remainder (35 420) were low impacting. No high impact variants were found for the OI-candidate genes. A missense variant among the medium impacting variants was found in the *COL1A2* gene and several variants were found in three OI-candidate genes (*TENT5A*, *BMP1*, *TMEM38B*) corresponding to low impacting variants.



**Table 4.3:** Summary of the variants grouped by mode of inheritance and variant impact.

Mode of inheritance	Impact of the variant as defined by the annotation algorithm	Variant annotation tools	
		SnEff (% of total variants)	VEP (% of total variants)
Autosomal Recessive - Homozygous	High	57 (0.03%)	73(0.03%)
	Medium	570 (0.33%)	959 (0.35%)
	Low	172 244 (99.64%)	271 905 (99.62%)
Autosomal Recessive – Compound Heterozygous	High	10 (0.02%)	14 (0.03%)
	Medium	3096 (5.63%)	4250 (8.99%)
	Low	51 920 (94.36%)	43 026 (90.98%)
<i>de novo</i>	High	15 (0.09%)	16 (0.05%)
	Medium	143 (0.83%)	189 (0.53%)
	Low	17 167 (99.09%)	35 420 (99.42%)

## 4.7 Variant split

### 4.7.1 Coding and non-coding

The homozygous, compound heterozygous and *de novo* variants identified were split into coding and non-coding variants. This resulted in 183 603 coding and 88 608 non-coding homozygous variants. A total of 18 488 coding and 17 137 non-coding *de novo* variants. All compound heterozygous variants were coding variants.

### 4.7.2 SNPs and InDels

The results for the coding variants can be viewed in **Table 4.4**. The 149 399 + 39 007 + 9797 corresponds to the total number of coding SNPs found in across all three modes of inheritance. The 11 508 + 3 154 + 3 684 corresponds to the total number of genes across the modes of inheritance. **Table 4.4** also give the total number of SNP, insertions, deletions, and genes found in each mode of inheritance.

**Table 4.4:** The number of variants found across the modes of inheritance separated into SNPs, insertions and deletions

Variant type	Homozygous (#genes)	Compound Heterozygous (#genes)	<i>de novo</i> (#genes)	Total variants per type (#genes)
SNPs	149 399 (11 243)	39 007 (3 154)	9 797 (3 684)	198 203 (18 081)
Insertions	15 731 (6 411)	4 083 (928)	4 298 (3 360)	24 112 (10 699)
Deletions	18 473 (6 964)	4 190 (1 043)	4 393 (3 411)	27 056 (11 418)
<b>Total variants per mode</b>	183 603 (24 618)	47 280 (5 125)	18 488 (10 455)	249 371 (40 198)

#### 4.8 SAHGP control

The 149 399 homozygous, 39 007 compound heterozygous and 9797 *de novo* coding SNPs that were compared to and found in the 23 healthy South African individuals' genomes were discarded, leaving a total of 146 397 homozygous SNPs, 38 739 compound heterozygous SNPs and 9 234 *de novo* variants. There were multiple duplicate variants found across the individuals in each population group and across the four populations (Table 4.5).

**Table 4.5:** The number of variants found in the four healthy populations in the SAHGP dataset

SAHGP population	# Of individuals (genome)	# Of homozygous SNPs	# Of compound heterozygous SNPs	# Of <i>de novo</i> SNPs
Mixed ancestry	7	2 255	182	463
Sotho	8	2 257	194	445
Xhosa	7	2 265	207	418
Zulu	1	1 121	77	201
<b>Total # SNP removed (excluding duplicates)</b>		3 002	350	563

#### 4.9 SNP: Gene ontology and Pathway enrichment

A total of 14 113 genes containing autosomal recessive inherited SNPs were used as input for GO enrichment analysis.

A total of 10 725 genes containing autosomal recessive inherited SNPs were used out of the 14 113 input genes in the GO enrichment analysis. GO terms that had at least one OI-candidate genes group under them were retained, resulting in a total number of 7492 genes remaining for

further analysis. Table 4.6 – Table 4.8 each show the molecular function, biological processes, and cellular components of GO terms, listed from most significant to least significant. Tables 4.6-4.8 also lists the p-values, and adjusted p-values using Bonferroni, FDR (Benjamini-Hochberg and Benjamini-Yekutieli). Table 4.6 represents the 10 (out of 28) most significant GO enrichment terms observed for molecular functions based on p-values. The ‘ion binding’ is the most significant enriched term for molecular function (p-value  $\leq 3.298 \times 10^{-17}$ ) and 2971 input genes. The GO enrichment term ‘multicellular organism development’ was the most significant term for the biological processes with a p-value of  $8.007 \times 10^{-38}$  and 1243 query genes (Table 4.7). The cellular component GO enrichment term ‘plasma membrane region’ was the most significant with a p-value of  $5.475 \times 10^{-48}$  and 1693 query genes (Table 4.8).

The most significant pathways for the query genes are listed in Table 4.10. This is an overrepresentation analysis: A statistical test that determines whether certain Reactome pathways are enriched in the submitted data. This test produces a probability score, which is corrected for false discovery rate (FDR) using the Benjamini-Hochberg method. A total of 4591 out of the 7492 genes in the sample were found in Reactome. A remainder of 2901 genes were not found neither mapped to any entity in Reactome. The pathways listed in Table 4.9 are the 20 (out of 359) more significant pathways featuring OI-candidate genes, the most significant out of the 20 being the “Collagen chain trimerization” pathway.

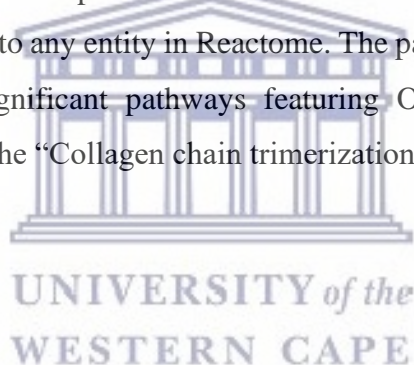


Table 4.6: Top 10 statistically enrichment GO Molecular function terms ordered by p-value ( $\leq 0.5$ ).

Category	ID	Name	P-value	q-value Bonferroni	q-value FDR B&H	q-value FDR B&Y
GO: Molecular Function	GO:0043167	ion binding	3.298E-17	1.26E-13	1.26E-13	1.112E-12
GO: Molecular Function	GO:0046872	metal ion binding	1.092E-09	0.000004173	2.455E-07	0.000002166
GO: Molecular Function	GO:0043169	cation binding	1.378E-09	0.000005264	2.924E-07	0.000002581
GO: Molecular Function	GO:0005509	calcium ion binding	1.518E-08	0.000058	0.000002572	0.0000227
GO: Molecular Function	GO:0016787	hydrolase activity	1.549E-08	0.00005916	0.000002572	0.0000227
GO: Molecular Function	GO:0140096	catalytic activity, acting on a protein	9.54E-08	0.0003644	0.00001301	0.0001149
GO: Molecular Function	GO:0005216	ion channel activity	1.666E-07	0.0006365	0.00002122	0.0001872
GO: Molecular Function	GO:0060090	molecular adaptor activity	1.758E-07	0.0006714	0.00002166	0.0001911
GO: Molecular Function	GO:0022857	transmembrane transporter activity	2.108E-07	0.0008053	0.0000244	0.0002154

Table 4.7: Top 10 statistically enrichment GO biological processes terms ordered by p-value ( $\leq 0.5$ )

Category	ID	Name	P-value	q-value Bonferroni	q-value FDR B&H	q-value FDR B&Y
GO: Biological Process	GO:0007275	multicellular organism development	8.007E-38	1.152E-33	1.152E-33	1.169E-32
GO: Biological Process	GO:0048856	anatomical structure development	7.965E-37	1.146E-32	5.728E-33	5.814E-32
GO: Biological Process	GO:0048731	system development	2.641E-36	3.798E-32	1.266E-32	1.285E-31
GO: Biological Process	GO:0032501	multicellular organismal process	3.616E-35	5.201E-31	1.3E-31	1.32E-30
GO: Biological Process	GO:0032502	developmental process	1.154E-33	1.66E-29	3.32E-30	3.371E-29
GO: Biological Process	GO:0007399	nervous system development	5.98E-33	8.601E-29	1.433E-29	1.455E-28
GO: Biological Process	GO:0048468	cell development	3.032E-29	4.36E-25	6.229E-26	6.323E-25
GO: Biological Process	GO:0051179	localization	2.856E-28	4.108E-24	5.135E-25	5.213E-24
GO: Biological Process	GO:0065008	regulation of biological quality	3.332E-28	4.792E-24	5.324E-25	5.405E-24
GO: Biological Process	GO:0023051	regulation of signaling	5.763E-28	8.289E-24	8.289E-25	8.414E-24

Table 4.8: Top 10 statistically enriched GO cellular terms ordered by p-value ( $\leq 0.5$ ).

Category	ID	Name	P-value	q-value Bonferroni	q-value FDR B&H	q-value FDR B&Y
GO: Cellular Component	GO:0098590	plasma membrane region	5.475E-48	9.756E-45	9.756E-45	7.866E-44
GO: Cellular Component	GO:0071944	cell periphery	3.589E-44	6.395E-41	3.198E-41	2.578E-40
GO: Cellular Component	GO:0042995	cell projection	1.655E-42	2.949E-39	9.829E-40	7.925E-39
GO: Cellular Component	GO:0120025	plasma membrane bounded cell projection	5.937E-42	1.058E-38	2.645E-39	2.132E-38
GO: Cellular Component	GO:0005886	plasma membrane	2.788E-39	4.968E-36	9.937E-37	8.012E-36
GO: Cellular Component	GO:0043005	neuron projection	3.148E-34	5.61E-31	9.35E-32	7.539E-31
GO: Cellular Component	GO:0030054	cell junction	1.003E-32	1.787E-29	2.553E-30	2.059E-29
GO: Cellular Component	GO:0045202	synapse	6.748E-28	1.203E-24	1.503E-25	1.212E-24
GO: Cellular Component	GO:0036477	somatodendritic compartment	1.021E-25	1.819E-22	2.021E-23	1.63E-22
GO: Cellular Component	GO:0030424	axon	7.895E-22	1.407E-18	1.172E-19	9.453E-19

Table 4.9: The 14 most significant pathways.

Pathway identifier	Pathway name	#Entities found	Entities pValue	Entities FDR
R-HSA-8948216	Collagen chain trimerization	33	0.008584594	0.999999675
R-HSA-2214320	Anchoring fibril formation	12	0.058994562	0.999999675
R-HSA-451326	Activation of kainate receptors upon glutamate binding	24	0.241038151	0.999999675
R-HSA-2022090	Assembly of collagen fibrils and other multimeric structures	42	0.251960639	0.999999675
R-HSA-2243919	Crosslinking of collagen fibrils	14	0.256473284	0.999999675
R-HSA-2672351	Stimuli-sensing channels	71	0.368873298	0.999999675
R-HSA-1442490	Collagen degradation	39	0.425320894	0.999999675
R-HSA-8957275	Post-translational protein phosphorylation	53	0.449589994	0.999999675
R-HSA-112308	Presynaptic depolarization and calcium channel opening	11	0.530654728	0.999999675
R-HSA-5576892	Phase 0 - rapid depolarisation	20	0.56645184	0.999999675

#### 4.10 SNP: Gene-phenotype association

With the use of 25 terms based on the phenotypic characteristics of OI, the 4591 genes from the GO enrichment step reduce to include only those genes sharing the phenotype from the list of 25 terms. This resulted in a total of 832 genes remaining. With the knowledge-based filtering step, this study identified 81 genes in 97 skeletal dysplasias. These genes are related to skeletal dysplasias, as listed in the newest and tenth version of the Nosology and Classification of genetic skeletal disorders<sup>230</sup>. Table F.2 shows the list of diseases with the genes well as the phenotype associated with OI.

#### 4.11 SNP: Protein-protein interaction

A total of 824 out of the 832 genes interacted in the protein network. The 824 genes were either directly or indirectly interacting with the OI-candidate genes. The OI-candidate genes included the SERPINH1, CRTAP, CREB3L1, COL1A1, COL1A2, and BMP1.

Of the 824 genes from the protein network, a total of 635 genes makes up 9801 homozygous SNPs, 304 genes make up 4295 compound heterozygous SNPs and 180 genes make up 316 SNPs.

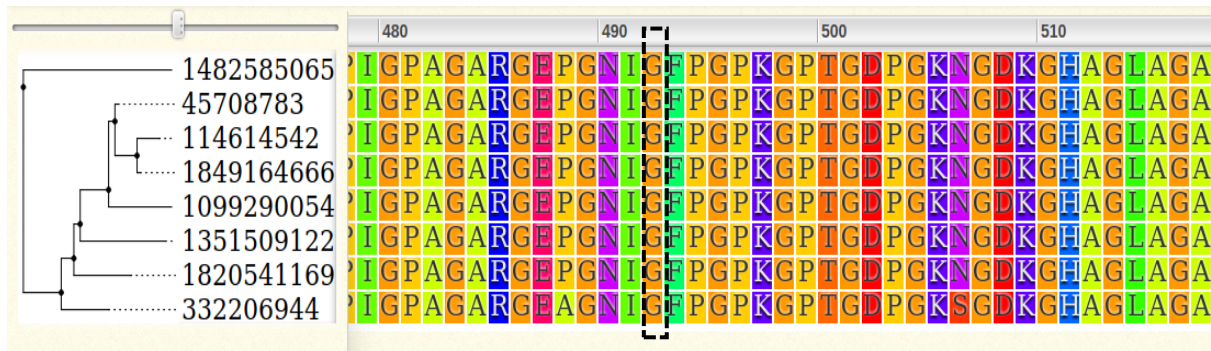
#### 4.12 SNP: Pathogenicity prediction

There was a total of 3808 homozygous variants where at least one out of the six prediction tools predicted the variant deleterious. This number was further reduced to include only those variants where at least three out of the six (50%) PredictSNP tools predicted the variants to be deleterious. This resulted in a total of 355 (178 genes) variants remaining. Table F.3 ([Appendix F](#)) shows the remaining variants. Majority of the homozygous variants were present in dbSNP but not in the ClinVar database and had no available publication. For those variants that had publications, the studies reported on the variants, but these studies had no relevant information to skeletal dysplasias.

A total number of 127 compound heterozygous variants remained after the SNP prediction, gene-phenotype and SAHGP control group filtering steps. The 127 variants are spread among 86 genes (Table F.4).

A total of five de novo variants remained, these consisted of *COL1A2*, *RASA2*, *OPAI*, *AK2*, and *MUC16* (Table F.5). The *COL1A2* and *AK2* genes each had a missense variant and the variants in the *OPAI* and *MUC16* genes were splice region variants. The *COL1A2* missense variant was

also found in a highly conserved location in the gene, shown across multiple closely related species (**Figure 4.3**).



**Figure 4.3:** Protein conservation of the COL1A2 gene across Homo sapiens and closely related species. Position 493 has a glycine that is conserved throughout the listed species. This indicates a highly conserved region in this gene.



## 5 Chapter Five: Discussion

### 5.1 Introduction

OI forms part of a list of more than 461 well characterized skeletal dysplasias classified primarily on their clinical, radiographic and/or molecular phenotype<sup>230</sup>

This study investigated the genetic cause of OI in a South African non-consanguineous family of Cape mixed ancestry through whole genome sequencing. The investigation included; (i) the identification of any unreported variants within the OI-candidate genes being to identify, (ii) identification of potentially new variants in genes not previously associated with OI, (iii) identification of genes associated with biological pathways that map to the published OI-candidate genes, and (iv) prioritizing the identified variants as potentially disease-causing variants in the identified genes.

OI forms part of a list of more than 461 well characterized skeletal dysplasias classified primarily on their clinical, radiographic and/or molecular phenotype<sup>230</sup>.

### 5.2 Participant demographics

The clinical features of OI present in the twins are on the severest end of the spectrum of clinical OI, which can be caused by variants in several genes. It is difficult to predict, based on phenotype alone, which gene would underly their OI. The clinical characteristics of OI observed due to variants in type I collagen genes, are also observed in the twins. For instance, dentinogenesis imperfecta, triangular face and wheelchair dependency is commonly observed in OI type III (see chapter 2, section 2.2.1.1). Additionally, there are characteristics common to the twins and OI type IV patients, these include severe bone deformity of long bones and spinal column, white sclerae, short stature and dentinogenesis imperfecta (see chapter 2, section 2.2.1.1). Given the clinical presentation of OI in the twins, it is difficult to make any direct association to a specific type of OI based solely on their phenotype. This is in line with the publication of Chetty *et al*<sup>18</sup>, which reports the considerable overlap between the Sillence classification of OI types I-IV and the expanded classifications of OI types V-XX.

### 5.3 Assessment of twenty OI candidate genes

This thesis reports on all identified variants based on the use of in silico tools within genes previously correlated with OI or OI related phenotype. We used a range of criteria to prioritize candidate SNPs. These criteria included inherited pattern (homozygous, compound heterozygous and *de novo*), location of variants (coding and non-coding), functional annotation



and pathway enrichment (gene ontology), protein-protein interactions (STRING database), pathogenicity prediction (benign or deleterious, variants with unknown significance (VUS)), gene-phenotype association and SAHGP control data (southern African healthy individuals).

Variants were identified in 11 OI-candidate genes (*COL1A1*, *COL1A2*, *BMP1*, *CREB3L1*, *SERPINH1*, *SERPINF1*, *TENT5A* and *CRTAP*) that form part of type 3 OI based on the latest classification<sup>237</sup>. **Table F.6** list the variants for the above-mentioned genes, the potential effect of the variant and whether the variant has been recorded or reported in databases such as ClinVar, dbSNP and LOVD. There were no pathogenic variants found in the *FKBP10* gene, which was previously shown to be disease-causing in many of the Black African populations of South Africa<sup>13</sup>. Many of these variants are considered likely benign or benign based on one or more of the following criteria: it occurs at a poorly conserved position in the protein, it is predicted to be benign by multiple in silico algorithms, and/ or has population frequency not consistent with disease. Based on in silico predictions, these variants are likely tolerated. Based on the ACGM guidelines, these variants were also categorized as benign or likely benign

Many variants identified in the OI-candidate genes were not present in the LOVD, ClinVar (NCBI), ClinVar Miner and the UCSC genome browser databases. This may be because these specific variants have not yet been identified and/ or reported. For those variants that were present in Clinvar and had publications, the reported variant had the same location and rsID but the alleles were different and/or the variant had no relevance to bone disease or the impact was not discussed in the articles. This does not come as a surprise, because these databases do not have mixed ancestry genomes. A *SERPINH1* variant was functionally annotated as low impacting but classified as deleterious by all six PredictSNP tool. A total of 78 uniquely identified variants in mixed ancestry patients were functionally annotated. Approximately six variants were classified as deleterious by the at least three of the six PredictSNP tool ([Appendix F, Table F.6](#)). The Algorithms underlying PredictSNP predicted the variants to be deleterious. Each prediction tool uses different algorithms for the predictions, these include utilizing machine-learning models trained on the potential to distinguish deleterious variants from neutral ones. CADD for instance, calculates its predictions based on a logistic regression model that account for evolutionary conservation, regulatory and transcript information, and protein-level scores (see Chapter 2, section 2.5.5.3.3). Therefore, the information beyond amino-acid composition (as seen in SnpEff) used in PredictSNP could provide a more refined analysis.

#### 5.4 Genes not associated with OI but associated with skeletal dysplasias

Variants were identified in 81 genes in 97 different disorders that are related to skeletal dysplasias, as listed in the newest and tenth version of the Nosology and Classification of genetic skeletal disorders<sup>230</sup> (see section 4.7.4). The phenotypic features in this list, is similar to the phenotype observed in the twins and other OI cases. These genes play key roles in the extracellular matrix organization, assembly of collagen fibrils, collagen biosynthesis and collagen formation. The processes and/or pathways are fundamental to the correct development of bones and skeletal system. Moreover, the genes of these variants form a physical and/or functional link with the OI-candidate genes as shown by the protein-protein interaction analysis. [Appendix F, Table F.5](#), lists the genes, disease, and associated phenotype that is similar to the clinical phenotype observed in OI cases. The variants for these genes were absent in the public databases and for those variants that were present the allele was different. There were also publications available for some of the variants, but they were irrelevant to bone disorders.

#### 5.5 Pathogenic or likely pathogenic variant

Variant that was annotated as pathogenic in collagen type 1 alpha 2 chain (*COL1A2*) gene. This gene has a key functional role in bone formation.

This study identified a de novo heterozygous missense mutation in the *COL1A2* (c.1478G>T, p. (493V)) gene, found on exon 25. The parents of the affected twins were both homozygous at position 1478 (G/G), however, the twins were both heterozygous (G/T). The glycine residue is highly conserved and there is a moderate physiochemical difference between glycine and valine. The variant in *COL1A2* may present a potential cause or influence for OI in this ethnic group. Sanger sequencing will need to be carried out to confirm the absence or presence of this variant in the parents and twins. This variant is a novel missense change affecting a residue that is known to be critical for normal protein structure, stability, and function.

As mentioned in the literature review ([Chapter 2, section 2.3.2.1](#), page 24-25), there are four standard rules that have to be applied as indication of severity based on certain substitutions in the collagen type I genes. The third rule is the substitution of glycine with a charged or branched side chain amino acids. The *de novo* missense variant identified in the *COL1A2* gene, could potentially be a variant of severe impact, as glycine is substituted for valine, which is a branched side chain amino acid. **Figure 4.3** indicates that this variant is located in

a highly conserved region as demonstrated by conservation of this region in eight of the selected primates.

According to the ACMG criteria the G493V is classified as a pathogenic variant. Multiple evidence based on ACMG guidelines are in support of the variant being pathogenic. The evidence code explanations can be found in [Appendix A, under ACMG guidelines](#).

The strength of evidence for code PM2 is strong because the position of the variant is strongly conserved (phyloP100way = 10 which is greater than 7.2). PhyloP100way scores are based on multiple alignments of 99 vertebrate genome sequences to the human genome. The greater the score, the more conserved the site. The variant was not found in gnomAD exomes, was also unable to check for gnomAD exomes coverage. The variant was also not found in gnomAD genomes, there was good gnomAD genome coverage = 31.2

For the evidence code PM5, the strength is strong, we have a novel missense amino acid change (NM\_000089.4(COL1A2):c.1478G>T (p.Gly493V) occurring at the same position as the pathogenic missense change (NM\_000089.4(COL1A2):c.1478G>A (p.Gly496Glu and NM\_000089.4(COL1A2):c.1477G>C(p.Gly493Arg)). The evidence strength is strong because two pathogenic alternative variants have been identified, Gly493Arg is classified as likely pathogenic and Gly493Glu is classified as pathogenic, both have been confirmed using ACMG.

The InterVar classifying system, which is a bioinformatics software tool for clinical interpretation of genetic variants by the ACMG/AMP 2015 guideline. The input to InterVar is an annotated file generated from ANNOVAR, while the output of InterVar is the classification of variants into 'benign', 'likely benign', 'uncertain significant', 'likely pathogenic', and 'pathogenic' together with detailed evidence codes. This tool classified *COL1A2*: c.1478G>T (p.Gly493Val) as likely pathogenic.

For the evidence code PM1, the variant is located in a mutational hotspot and/or critical and well-established functional domain without benign variation, and has a moderate strength. UniProt protein COL1A2\_HUMAN region of interest "Disordered" has 648 missense/in-frame variants (376 pathogenic variants, 248 uncertain variants and 24 benign variants), pathogenicity = 58.0%, which qualifies as pathogenic.

The strength of the evidence code PP2 was 'supporting', this code of evidence is defined as "missense variant in a gene that has a low rate of benign missense variation and in which

missense variants are a common mechanism of disease”. In this instance 434 out of 470 non-VUS missense variants in the *COL1A2* gene are pathogenic = 92.3% which is more than the threshold of 40.2%, and 528 out of 1197 clinically reported variants in the *COL1A2* are pathogenic = 44.1% which is more than threshold of 35.2%.

For the evidence code PP3 the strength is also supporting. This code considers multiple lines of computational evidence support a deleterious effect on the gene or gene product. For example, with the ACMG guidelines, pathogenic computational verdict was based on 12 pathogenic predictions from BayesDel\_addAF, DANN, DEOGEN2, EIGEN, FATHMM-MKL, LIST-S2, M-CAP, MVP, MutationAssessor, MutationTaster, PrimateAI and SIFT vs no benign predictions.



## 6 Chapter Six: Conclusion

### 6.1 Introduction

The development of sequencing technologies over the past few decades has revolutionized the methods used to find novel disease-causing genes. The purpose of this study was to identify possible disease-causing gene (known or novel) in a mixed ancestry population that presented with a severe form of OI. This study focused on a sample of the mixed ancestry population and not the entire population.

### 6.2 Summary of main points or findings

OI is a heterogeneous disease that can be caused through multiple different mechanisms. Moreover, the pathogenic mechanism in OI arises from gene variants that directly and indirectly affect the synthesis, structure, folding, secretion, and matrix organization of type I collagen. This study has identified a promising variant found in the *COL1A2* gene, which was located in a highly conserved region across multiple species, although the amino acid change was different to the two found in the databases; variant in this location has been reported as likely pathogenic according to the ACMG guidelines.

### 6.3 Limitations of the study

As with any type of research there exist opportunities for errors in the analysis. Because the analysis is not a completely automated process, there is a chance for human error in the input, examination and interpretation of the data and its relevance to the research question. The variant filtering strategy used followed closely with the suggested pipelines in literature, but with some deviations with regards to the application of certain filters. This in return, influenced the type of variants that were filtered out at each stage and lead to some variants making it to higher or lower levels of consideration than expected. The interpretation of the data was also limited by the current level of knowledge about the genes studies and the information available in databases such as dbSNP, OMIM, and ClinVar.

This case study also demonstrates the challenge of sequencing ethnic groups that have no genomic records in the public databases. As this may lead to inconclusive observations or conclusions being made about the findings, this can especially be true for the mixed ancestry populations of African and/or South African, since the exact genome information is not yet available for this ethnic group. Using existing databases also runs the risk of discarding

plausible candidate variants when filters are too stringent.

Despite being able to identify more variants, a major limitation of using WGS is the overwhelming yield of data, sorting through this amount of data can be very time consuming. Moreover, the vast majority of rare variants in the genome have no link to the patient's disorder. Therefore, understanding variation within the human genome, to differentiate normal variants from disease-causing variants, is very challenging.

Another concern is the large number of intergenic variants identified that have no information in the existing databases. Allocating functional impact to these variants was also challenging.

#### 6.4 Future Recommendations

There were a total number of 89 332 homozygous variants and 17 137 de novo non-coding variants that were not investigated in this study. In addition to these, 24 112 insertions and 27 056 deletions were identified but not investigated. These variants could be considered for future studies to look for any deleterious frame shifts, for example, due to large insertions or deletions occurring early in the sequence, which may alter the protein completely. These altered gene-products may be linked to OI-candidate genes and as such be contributing to the phenotype observed in the twins.

This is only a sample population and the variant identified in this study may not be the same as other OI affected individuals from the mixed ancestry population. Therefore, further studies could look at expanding this study to include more participants and include functional work as well.

#### 6.5 Conclusion

The accurate and timeous diagnosis of rare Mendelian disorders such as OI is of great importance, but traditional genetic methods have been laborious and often ineffectual. WGS and similar genomic technologies are promising both for both research and diagnostics. Its clinical utility in patient diagnosis and management is still being explored. Even when used in a diagnostic context it generates much information of research relevance especially in populations for which relatively little genomic information is available.

As researchers and clinicians in Africa become more familiar with the various techniques of WGS, it has become progressively more possible to implement clinical WGS. This also

open the possibility of offering WGS and similar genomic testing approaches (such as gene panels) for other clinical presentations, in the form of clinical testing or alternatively a niche research study, depending on factors such as the extent of functional genetic testing planned. In either case, there is a need for ethical handling of genomic data produced by WGS for a broader range of genetic conditions.

The goal of this research also includes creating a repository to capture all genotypic and phenotypic data of all the different ethnic groups in the African and South African populations with OI, then categorize this information, to make diagnosis and treatment of OI in these populations more precise. The COL1A2 variant is very promising with a CADD score of 25.30 and no rsID, in addition the ACGM criteria have supporting evidence that indicate the variant to be pathogenic. The variant identified in the *COL1A2* could support this being OI type 4-B.

The findings from this study will be added to existing databases as an initial step in creating this repository. As the knowledge around genetics underlying OI type 3 in African and/or South African populations grows, this will assist with future studies into the various presentations of OI and its genetics in the different ethnic groups of South Africa.



## Bibliography

1. Nguengang Wakap, S., Lambert, D. M., Olry, A., Rodwell, C., Gueydan, C., Lanneau, V., Murphy, D., Le Cam, Y. & Rath, A. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur. J. Hum. Genet.* **28**, 165–173 (2020).
2. Conradie, E. H., Malherbe, H., Hendriksz, C. J., Dercksen, M. & Vorster, B. C. An Overview of Benefits and Challenges of Rare Disease Biobanking in Africa, Focusing on South Africa. *Biopreservation Biobanking* (2021). doi:10.1089/bio.2020.0108
3. Henderson, B. D., Isaac, N., Mabele, O., Khiba, S., Nkayi, A. & Mokoena, T. Pamidronate treatment for osteogenesis imperfecta in black South Africans. *South Afr. Med. J. Suid-Afr. Tydskr. Vir Geneesk.* **106**, S47-49 (2016).
4. Stephen, L. X. G., Roberts, T., Van Hayden, E. & Chetty, M. Osteogenesis imperfecta type III in South Africa: Psychosocial challenges. *South Afr. Med. J. Suid-Afr. Tydskr. Vir Geneesk.* **106**, S90-93 (2016).
5. Vorster, A., Beighton, P., Chetty, M., Ganie, Y., Henderson, B., Honey, E., Maré, P., Thompson, D., Fieggen, K., Viljoen, D. & Ramesar, R. Osteogenesis imperfecta type 3 in South Africa: Causative mutations in FKBP10. *S. Afr. Med. J.* **107**, 457 (2017).
6. Sillence, D. Osteogenesis imperfecta in southern Africa: Peter Beighton's legacy. *S. Afr. Med. J.* **106**, 13 (2016).
7. Alharbi, S. A. A Systematic Overview of Osteogenesis Imperfecta. *Mol Biol* **5**, (2016).
8. Sá-Caputo, D. C., Dionello, C. da F., Frederico, É. H. F. F., Paineiras-Domingos, L. L., Sousa-Gonçalves, C. R., Morel, D. S., Moreira-Marconi, E., Unger, M. & Bernardo-Filho, M. WHOLE-BODY VIBRATION EXERCISE IMPROVES FUNCTIONAL PARAMETERS IN PATIENTS WITH OSTEOGENESIS IMPERFECTA: A SYSTEMATIC REVIEW WITH A SUITABLE APPROACH. *Afr. J. Tradit. Complement. Altern. Med. AJTCAM* **14**, 199–208 (2017).





9. Gupta, A., Kamal, G., Gupta, N. & Aggarwal, A. Combined Spinal–epidural Anesthesia With Dexmedetomidine-based Sedation for Multiple Corrective Osteotomies in a Child With Osteogenesis Imperfecta Type Iii: A Case Report. *Case Rep.* **9**, 60–63 (2017).
10. Tongkobpetch, S., Limpaphayom, N., Sangsin, A., Porntaveetus, T., Suphapeetiporn, K. & Shotelersuk, V. A novel de novo COL1A1 mutation in a Thai boy with osteogenesis imperfecta born to consanguineous parents. *Genet. Mol. Biol.* **40**, 763–767 (2017).
11. Marini, J. C., Forlino, A., Bächinger, H. P., Bishop, N. J., Byers, P. H., Paepe, A. D., Fassier, F., Fratzi-Zelman, N., Kozloff, K. M., Krakow, D., Montpetit, K. & Semler, O. Osteogenesis imperfecta. *Nat. Rev. Dis. Primer* **3**, 17052 (2017).
12. Mrosk, J., Bhavani, G. S., Shah, H., Hecht, J., Krüger, U., Shukla, A., Kornak, U. & Girisha, K. M. Diagnostic strategies and genotype-phenotype correlation in a large Indian cohort of osteogenesis imperfecta. *Bone* **110**, 368–377 (2018).
13. Umair, M., Alhaddad, B., Rafique, A., Jan, A., Haack, T. B., Graf, E., Ullah, A., Ahmad, F., Strom, T. M., Meitinger, T. & Ahmad, W. Exome sequencing reveals a novel homozygous splice site variant in the *WNT1* gene underlying osteogenesis imperfecta type 3. *Pediatr. Res.* **82**, 753–758 (2017).
14. Chetty, M., Roberts, T. S., Stephen, L. & Beighton, P. Craniofacial manifestations in osteogenesis imperfecta type III in South Africa. *BDJ Open* **3**, 17021 (2017).
15. Xu, X., Lv, F., Song, Y., Li, L., Asan, Wei, X., Zhao, X., Jiang, Y., Wang, O., Xing, X., Xia, W. & Li, M. Novel mutations in BMP1 induce a rare type of osteogenesis imperfecta. *Clin. Chim. Acta* **489**, 21–28 (2019).
16. Guillemyn, B., Kayserili, H., Demuyne, L., Sips, P., De Paepe, A., Syx, D., Coucke, P. J., Malfait, F. & Fransiska, S. A homozygous pathogenic missense variant broadens the phenotypic and mutational spectrum of CREB3L1-related osteogenesis imperfecta. *Hum. Mol. Genet.* (2019). doi:10.1093/hmg/ddz017
17. Van Dijk, F. S. & Sillence, D. O. Osteogenesis imperfecta: Clinic diagnosis, nomenclature and severity assessment. *AM J Med Genet A* **16A(6)**, 1470–1481 (2014).

18. Chetty, M., Roomaney, I. A. & Beighton, P. The evolution of the nosology of osteogenesis imperfecta. *Clin. Genet.* **99**, 42–52 (2021).
19. Lindert, U., Cabral, W. A., Ausavarat, S., Tongkobpetch, S., Ludin, K., Barnes, A. M., Yeetong, P., Weis, M., Krabichler, B., Srichomthong, C., Makareeva, E. N., Janecke, A. R., Leikin, S., Röthlisberger, B., Rohrbach, M., Kennerknecht, I., Eyre, D. R., Suphapeetiporn, K., Giunta, C., Marini, J. C. & Shotelersuk, V. MBTPS2 mutations cause defective regulated intramembrane proteolysis in X-linked osteogenesis imperfecta. *Nat. Commun.* **7**, 11920 (2016).
20. Hu, J., Li, L., Zheng, W., Zhao, D., Wang, O., Jiang, Y., Xing, X., Li, M. & Xia, W. A novel mutation in PLS3 causes extremely rare X-linked osteogenesis imperfecta. *Mol. Genet. Genomic Med.* **8**, e1525 (2020).
21. Forlino, A., Cabral, W. A., Barnes, A. M. & Marini, J. C. New perspectives on osteogenesis imperfecta. *Nat Rev Endocrinol* **7**, 540–57 (2011).
22. Beighton, P. & Versfeld, G. A. On the paradoxically high relative prevalence of osteogenesis imperfecta type III in the Black population of South Africa. *Clin. Genet.* **27**, 398–401 (1985).
23. Alanay, Y., Avaygan, H., Camacho, N., Utine, G. E., Boduroglu, K., Aktas, D., Alikasifoglu, M., Tuncbilek, E., Orhan, D. & Bakar, F. T. Mutations in the gene encoding the RER protein FKBP65 cause autosomal-recessive osteogenesis imperfecta. *Am J Hum Genet* **86(4)**, 551–9 (2010).
24. Chetty, M., Roberts, T., Stephen, L. X. G. & Beighton, P. Osteogenesis Imperfecta type III: A report on the unusual phenotypic features of six individuals of Cape mixed ancestry heritage. *South Afr. Dent. J.* **73**, 239–242 (2018).
25. Vij, N. & Belthur, M. V. Bony deformities in a child with untreated osteogenesis imperfecta type III. *JCIMCR* **3**, 2 (2022).
26. Reece, M. K. J., Quillin, K., Homewood, T. J. & Bunevich, J. Surgical Treatment of a Bilateral Mandibular Fracture in a Patient with Osteogenesis Imperfecta Type III. *Plast. Reconstr. Surg. Glob. Open* **9**, e3702 (2021).

27. Hüner, B., Handke-Vesely, A., Lato, K., Korzoum, A., Janni, W. & Reister, F. Mother and child with osteogenesis imperfecta type III. Pregnancy management, delivery, and outcome. *Case Rep. Perinat. Med.* **10**, (2021).
28. Xi, L., Lv, S., Zhang, H. & Zhang, Z.-L. Novel mutations in BMP1 result in a patient with autosomal recessive osteogenesis imperfecta. *Mol. Genet. Genomic Med.* **9**, e1676 (2021).
29. Ramzan, K., Alotaibi, M., Huma, R. & Afzal, S. Detection of a Recurrent TMEM38B Gene Deletion Associated with Recessive Osteogenesis Imperfecta. *Discoveries* **9**, e124
30. de Souza, L. T., Nunes, R. R., de Azevedo Magalhães, O. & Maria Félix, T. A new case of osteogenesis imperfecta type VIII and retinal detachment. *Am. J. Med. Genet. A.* **185**, 238–241 (2021).
31. Doyard, M., Bacrot, S., Huber, C., Di Rocco, M., Goldenberg, A., Aglan, M. S., Brunelle, P., Temtamy, S., Michot, C., Otaify, G. A., Haudry, C., Castanet, M., Leroux, J., Bonnefont, J.-P., Munnich, A., Baujat, G., Lapunzina, P., Monnot, S., Ruiz-Perez, V. L. & Cormier-Daire, V. FAM46A mutations are responsible for autosomal recessive osteogenesis imperfecta. *J. Med. Genet.* **55**, 278–284 (2018).
32. Moosa, S., Yamamoto, G. L., Garbes, L., Keupp, K., Beleza-Meireles, A., Moreno, C. A., Valadares, E. R., de Sousa, S. B., Maia, S., Saraiva, J., Honjo, R. S., Kim, G. A., Cabral de Menezes, H., Lausch, E., Lorini, P. V., Lamounier, A., Carniero, T. C. B., Giunta, C., Rohrbach, M., Janner, M., Semler, O., Beleggia, F., Li, Y., Yigit, G., Reintjes, N., Altmüller, J., Nürnberg, P., Cavalcanti, D. P., Zabel, B., Warman, M. L., Bertola, D. R., Wollnik, B. & Netzer, C. Autosomal-Recessive Mutations in MESD Cause Osteogenesis Imperfecta. *Am. J. Hum. Genet.* **105**, 836–843 (2019).
33. Kelley, B. P., Malfait, F., Bonafe, L., Baldrige, D., Homan, E. & Symoens, S. Mutations in FKBP10 cause recessive osteogenesis imperfecta and Bruck syndrome. *J Bone Min. Res* **26**, 666–72 (2011).
34. Vinkšelj, M., Witzl, K., Maver, A. & Peterlin, B. Improving diagnostics of rare genetic diseases with NGS approaches. *J. Community Genet.* **12**, 247–256 (2021).

35. Fernández-Marmiesse, A., Barbosa-Gouveia, S. & Couce, M. NGS Technologies as a Turning Point in Rare Disease Research , Diagnosis and Treatment. *Curr. Med. Chem.* **24**, (2017).
36. Schmidt, R. J., Macleay, A. & Le, L. P. VarGroup. *J. Mol. Diagn.* **21**, 384–389 (2019).
37. Adams, D. R. & Eng, C. M. Next-Generation Sequencing to Diagnose Suspected Genetic Disorders. *N. Engl. J. Med.* **379**, 1353–1362 (2018).
38. Liu, Z., Zhu, L., Roberts, R. & Tong, W. Toward Clinical Implementation of Next-Generation Sequencing-Based Genetic Testing in Rare Diseases: Where Are We? *Trends Genet. TIG* **35**, 852–867 (2019).
39. Cirulli, E. T. & Goldstein, D. B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat. Rev. Genet.* **11**, 415–425 (2010).
40. Bergant, G., Maver, A., Lovrecic, L., Čuturilo, G., Hodzic, A. & Peterlin, B. Comprehensive use of extended exome analysis improves diagnostic yield in rare disease: a retrospective survey in 1,059 cases. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **20**, 303–312 (2018).
41. Kumar, K. R., Davis, R. L., Tchan, M. C., Wali, G. M., Mahant, N., Ng, K., Kotschet, K., Siow, S.-F., Gu, J., Walls, Z., Kang, C., Wali, G., Levy, S., Phua, C. S., Yiannikas, C., Darveniza, P., Chang, F. C. F., Morales-Briceño, H., Rowe, D. B., Drew, A., Gayevskiy, V., Cowley, M. J., Minoche, A. E., Tisch, S., Hayes, M., Kummerfeld, S., Fung, V. S. C. & Sue, C. M. Whole genome sequencing for the genetic diagnosis of heterogenous dystonia phenotypes. *Parkinsonism Relat. Disord.* **69**, 111–118 (2019).
42. Kumar, K. R., Cowley, M. J. & Davis, R. L. Next-Generation Sequencing and Emerging Technologies. *Semin. Thromb. Hemost.* **45**, 661–673 (2019).
43. Lionel, A. C., Costain, G., Monfared, N., Walker, S., Reuter, M. S., Hosseini, S. M., Thiruvahindrapuram, B., Merico, D., Jobling, R., Nalpathamkalam, T., Pellecchia, G., Sung, W. W. L., Wang, Z., Bikangaga, P., Boelman, C., Carter, M. T., Cordeiro, D., Cytrynbaum, C., Dell, S. D., Dhir, P., Dowling, J. J., Heon, E., Hewson, S., Hiraki, L., Inbar-Feigenberg, M., Klatt, R., Kronick, J., Laxer, R. M., Licht, C., MacDonald, H., Mercimek-Andrews, S., Mendoza-Londono, R., Piscione,

- T., Schneider, R., Schulze, A., Silverman, E., Siriwardena, K., Snead, O. C., Sondheimer, N., Sutherland, J., Vincent, A., Wasserman, J. D., Weksberg, R., Shuman, C., Carew, C., Szego, M. J., Hayeems, R. Z., Basran, R., Stavropoulos, D. J., Ray, P. N., Bowdin, S., Meyn, M. S., Cohn, R. D., Scherer, S. W. & Marshall, C. R. Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. *Genet. Med.* **20**, 435–443 (2018).
44. Schmidt, B. & Hildebrandt, A. Next-generation sequencing: big data meets high performance computing. *Drug Discov. Today* **22**, 712–717 (2017).
45. Hartman, P., Beckman, K., Silverstein, K., Yohe, S., Schomaker, M., Henzler, C., Onsongo, G., Lam, H. C., Munro, S., Daniel, J., Billstein, B., Deshpande, A., Hauge, A., Mroz, P., Lee, W., Holle, J., Wiens, K., Karnuth, K., Kemmer, T., Leary, M., Michel, S., Pohlman, L., Thayanithy, V., Nelson, A., Bower, M. & Thyagarajan, B. Next generation sequencing for clinical diagnostics: Five year experience of an academic laboratory. *Mol. Genet. Metab. Rep.* **19**, 100464 (2019).
46. Alser, M., Rotman, J., Deshpande, D., Taraszka, K., Shi, H., Baykal, P. I., Yang, H. T., Xue, V., Knyazev, S., Singer, B. D., Balliu, B., Koslicki, D., Skums, P., Zelikovsky, A., Alkan, C., Mutlu, O. & Mangul, S. Technology dictates algorithms: recent developments in read alignment. *Genome Biol.* **22**, 249 (2021).
47. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
48. Chen, S., Zhou, Y., Chen, Y., Huang, T., Liao, W., Xu, Y., Li, Z. & Gu, J. Gencore: an efficient tool to generate consensus reads for error suppressing and duplicate removing of NGS data. *BMC Bioinformatics* **20**, 606 (2019).
49. Zhao, Q. A Study on Optimizing MarkDuplicate in Genome Sequencing Pipeline. in *Proc. 2018 5th Int. Conf. Bioinforma. Res. Appl.* 8–15 (Association for Computing Machinery, 2018).
- doi:10.1145/3309129.3309134

50. Van der Auwera, G., Carneiro, M., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K., Altshuler, D., Gabriel, S. & DePristo, M. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Curr. Protoc. Bioinforma.* **43**, 11.10.1-11.10.33 (2013).
51. Ugur Sezerman, O., Ulgen, E., Seymen, N. & Melis Durasi, I. in *Bioinforma. Tools Detect. Clin. Interpret. Genomic Var.* (eds. Samadikuchaksaraei, A. & Seifi, M.) (IntechOpen, 2019).  
doi:10.5772/intechopen.85524
52. Sefid Dashti, M. J. & Gamieldien, J. A practical guide to filtering and prioritizing genetic variants. *BioTechniques* **62**, 18–30 (2017).
53. Setiawati, R. & Rahardjo, P. in *Osteogenes. Bone Regen.* (ed. Yang, H.) (IntechOpen, 2019).  
doi:10.5772/intechopen.82452
54. Gilbert, S. F. Osteogenesis: The Development of Bones. *Dev. Biol.* **6th Ed.** (2000). at  
<<https://www.ncbi.nlm.nih.gov/books/NBK10056/>>
55. Biga, L. M., Dawson, S., Harwell, A., Hopkins, R., Kaufmann, J., LeMaster, M., Matern, P., Morrison-Graham, K., Quick, D. & Runyeon, J. 6.4 Bone Formation and Development. (2019). at  
<<https://open.oregonstate.edu/aandp/chapter/6-4-bone-formation-and-development/>>
56. Watts, N. B. Clinical Utility of Biochemical Markers of Bone Remodeling. *Clin. Chem.* **45**, 1359–1368 (1999).
57. Boskey, A. L. & Robey, P. G. in *Primer Metab. Bone Dis. Disord. Miner. Metab.* 84–92 (Wiley, 2018). doi:10.1002/9781119266594.ch11
58. Boskey, A. L. Bone composition: relationship to bone fragility and antiosteoporotic drug effects. *BoneKEy Rep.* **2**, (2013).
59. Markel, M. D. in *Equine Fract. Repair* 1–11 (John Wiley & Sons, Ltd, 2019).  
doi:10.1002/9781119108757.ch1
60. Meskinfam, M. in *Charact. Polym. Biomater.* 441–475 (Elsevier, 2017). doi:10.1016/B978-0-08-100737-2.00017-0

61. Taye, N., Karoulias, S. Z. & Hubmacher, D. The “other” 15–40%: The Role of Non-Collagenous Extracellular Matrix Proteins and Minor Collagens in Tendon. *J. Orthop. Res.* **38**, 23–35 (2020).
62. Saunders, J. T. & Schwarzbauer, J. E. Fibronectin matrix as a scaffold for procollagen proteinase binding and collagen processing. *Mol. Biol. Cell* **30**, 2218–2226 (2019).
63. Boraschi-Diaz, I., Wang, J., Mort, J. S. & Komarova, S. V. Collagen Type I as a Ligand for Receptor-Mediated Signaling. *Front. Phys.* **5**, (2017).
64. Holmes, D. F., Lu, Y., Starborg, T. & Kadler, K. E. in *Curr. Top. Dev. Biol.* **130**, 107–142 (Elsevier, 2018).
65. Sipilä, K. H., Drushinin, K., Rappu, P., Jokinen, J., Salminen, T. A., Salo, A. M., Käpylä, J., Myllyharju, J. & Heino, J. Proline hydroxylation in collagen supports integrin binding by two distinct mechanisms. *J. Biol. Chem.* **293**, 7645–7658 (2018).
66. Rappu, P., Salo, A. M., Myllyharju, J. & Heino, J. Role of prolyl hydroxylation in the molecular interactions of collagens. *Essays Biochem.* **63**, 325–335 (2019).
67. Vorster, A., Beighton, P., Chetty, M., Ganie, Y., Henderson, B., Honey, E., Maré, P., Thompson, D., Fieggen, K., Viljoen, D. & Ramesar, R. Osteogenesis imperfecta type 3 in South Africa: Causative mutations in FKBP10. *S. Afr. Med. J.* **107**, 457 (2017).
68. Stephen, L. X. G., Roberts, T., Van Hayden, E. & Chetty, M. Osteogenesis imperfecta type III in South Africa: Psychosocial challenges. *South Afr. Med. J. Suid-Afr. Tydskr. Vir Geneesk.* **106**, S90-93 (2016).
69. Lindert, U., Gnoli, M., Maioli, M., Bedeschi, M. F., Sangiorgi, L., Rohrbach, M. & Giunta, C. Insight into the Pathology of a COL1A1 Signal Peptide Heterozygous Mutation Leading to Severe Osteogenesis Imperfecta. *Calcif. Tissue Int.* **102**, 373–379 (2018).
70. Heike Hoyer-Kuhn, Laura Höbing, Julia Cassens, Eckhard Schoenau, & Oliver Semler. Children with severe Osteogenesis imperfecta and short stature present on average with normal IGF-I and IGFBP-3 levels. *J Pediatr Endocrinol Metab* **29(7)**, 813–818 (2016).

71. Tongkobpetch, S., Limpaphayom, N., Sangsin, A., Porntaveetus, T., Suphapeetiporn, K. & Shotelersuk, V. A novel de novo COL1A1 mutation in a Thai boy with osteogenesis imperfecta born to consanguineous parents. *Genet. Mol. Biol.* **40**, 763–767 (2017).
72. Ang, K., Sanchez Rangel, E., Yuan, Q., Wu, D., Carpenter, T. O. & Insogna, K. Skeletal disease in a father and daughter with a novel monoallelic WNT1 mutation. *Bone Rep.* **9**, 154–158 (2018).
73. Dwan, K., Phillipi, C. A., Steiner, R. D. & Basel, D. Bisphosphonate therapy for osteogenesis imperfecta. *Cochrane Database Syst. Rev.* **7:CD005088**, (2014).
74. Wang, J., Liu, Y., Song, L., Lv, F., Xu, X., San, A., Wang, J., Yang, H., Yang, Z., Jiang, Y., Wang, O., Xia, W., Xing, X. & Li, M. Novel Mutations in SERPINF1 Result in Rare Osteogenesis Imperfecta Type VI. *Calcif. Tissue Int.* **100**, 55–66 (2017).
75. Liu, Y., Asan, Ma, D., Lv, F., Xu, X., Wang, J., Xia, W., Jiang, Y., Wang, O., Xing, X., Yu, W., Wang, J., Sun, J., Song, L., Zhu, Y., Yang, H., Wang, J. & Li, M. Gene mutation spectrum and genotype-phenotype correlation in a cohort of Chinese osteogenesis imperfecta patients revealed by targeted next generation sequencing. *Osteoporos. Int.* **28**, 2985–2995 (2017).
76. Essawi, O., Symoens, S., Fannana, M., Darwish, M., Farraj, M., Willaert, A., Essawi, T., Callewaert, B., De Paepe, A., Malfait, F. & Coucke, P. J. Genetic analysis of osteogenesis imperfecta in the Palestinian population: molecular screening of 49 affected families. *Mol. Genet. Genomic Med.* **6**, 15–26 (2018).
77. Palomo, T., Vilaça, T. & Lazaretti-Castro, M. Osteogenesis imperfecta: diagnosis and treatment. *Curr. Opin. Endocrinol. Diabetes Obes.* **24**, 381–388 (2017).
78. Biggin, A. & Munns, C. F. Osteogenesis Imperfecta: Diagnosis and Treatment. *Springer Sci.* (2014). doi:10.1007/s11914-014-0225-0
79. Sillence, D. O., Senn, A. & Danks, D. M. Genetic heterogeneity in osteogenesis imperfecta. *J. Med. Genet.* **16**, 101–116 (1979).



80. Zhang, H., Xu, Y., Yue, H., Wang, C., Gu, J., He, J., Fu, W., Hu, W. & Zhang, Z. Novel mutations of the SERPINF1 and FKBP10 genes in Chinese families with autosomal recessive osteogenesis imperfecta. *Int. J. Mol. Med.* **41**, 3662–3670 (2018).
81. Alharbi, S. A. A Systematic Overview of Osteogenesis Imperfecta. *Mol Biol* **5**, (2016).
82. Chetty, M., Roomaney, I. A. & Beighton, P. The evolution of the nosology of osteogenesis imperfecta. *Clin. Genet.* **99**, 42–52 (2021).
83. Yang, L., Liu, B., Dong, X., Wu, J., Sun, C., Xi, L., Cheng, R., Wu, B., Wang, H., Tong, S., Wang, D. & Luo, F. Clinical severity prediction in children with osteogenesis imperfecta caused by COL1A1/2 defects. *Osteoporos. Int. J. Establ. Result Coop. Eur. Found. Osteoporos. Natl. Osteoporos. Found. USA* **33**, 1373–1384 (2022).
84. Valadares, E. R., Carneiro, T. B., Santos, P. M., Oliveira, A. C. & Zabel, B. What is new in genetics and osteogenesis imperfecta classification? *J Pediatr Rio J* **90(6)**, 536–541 (2014).
85. Balasubramanian, M., Parker, M. J., Dalton, A., Giunta, C., Lindert, U., Peres, L. C., Wagner, B. E., Arundel, P., Offiah, A. & Bishop, N. J. Genotype-phenotype study in type V osteogenesis imperfecta. *Clin. Dysmorphol.* **22**, 93–101 (2013).
86. Homan, E. P., Rauch, F. & Grafe, I. Mutations in SERPINF1 cause osteogenesis imperfecta type VI. *J Bone Min. Res* **26(12)**, 2798–2803 (2011).
87. Wang, J., Li, L., Zhang, Q., Liu, Y., Lv, F., Xu, X., Song, Y., Wang, O., Jiang, Y., Xia, W., Xing, X. & Li, M. Extremely low level of serum pigment epithelium-derived factor is a special biomarker of Chinese osteogenesis imperfecta patients with SERPINF1 mutations. *Clin. Chim. Acta* **478**, 216–221 (2018).
88. Barnes, A. M., Chang, W. & Morello, R. Deficiency of cartilage-associated protein in recessive lethal osteogenesis imperfecta. *N Engl J Med* **355 (26)**, 2757–2764 (2006).
89. Cabral, W. A., Chang, W., Barnes, A. M., Weis, M., Scott, M. A. & Leikin, S. Prolyl 3-hydroxylase 1 deficiency causes a recessive metabolic bone disorder resembling lethal/severe osteogenesis imperfecta. *Nat. Genet* **39**, 359–365 (2007).

90. Santana, A., Franzone, J. M., McGreal, C. M., Kruse, R. W. & Bober, M. B. A moderate form of osteogenesis imperfecta caused by compound heterozygous LEPRE1 mutations. *Bone Rep.* **9**, 132–135 (2018).
91. Van Dijk, F. S., Nesbitt, I. M., Zwikstra, E. H., Nikkels, P. G., Piersma, S. R. & Fratantoni, S. A. PPIB mutations cause severe osteogenesis imperfecta. *Am J Hum Genet* **85**, 521–7 (2009).
92. Moravej, H., Karamifar, H., Karamizadeh, Z., Amirhakimi, G., Atashi, S. & Nasirabadi, S. Bruck syndrome - a rare syndrome of bone fragility and joint contracture and novel homozygous FKBP10 mutation. *Endokrynol. Pol.* **66**, 170–174 (2015).
93. Rush, E., Caldwell, K., Kreikemeier, R., Lutz, R. & Esposito, P. Osteogenesis imperfecta caused by PPIB mutation with severe phenotype and congenital hearing loss. *J. Pediatr. Genet.* **03**, 029–034 (2015).
94. Christiansen, H. E., Schwarze, U., Pyott, S. M., AlSwaid, A., Al Balwi, M. & Alrasheed, S. Homozygosity for a missense mutation in SERPINH1, which encodes the collagen chaperone protein HSP47 results in severe recessive osteogenesis imperfecta. *Am J Hum Genet* **86**, 389–98 (2010).
95. Marshall, C., Lopez, J., Crookes, L., Pollitt, R. C. & Balasubramanian, M. A novel homozygous variant in SERPINH1 associated with a severe, lethal presentation of osteogenesis imperfecta with hydranencephaly. *Gene* **595**, 49–52 (2016).
96. Forlino, A., Cabral, W. A., Barnes, A. M. & Marini, J. C. New perspectives on osteogenesis imperfecta. *Nat Rev Endocrinol* **7**, 540–57 (2011).
97. Alanay, Y., Avaygan, H., Camacho, N., Utine, G. E., Boduroglu, K., Aktas, D., Alikasifoglu, M., Tuncbilek, E., Orhan, D. & Bakar, F. T. Mutations in the gene encoding the RER protein FKBP65 cause autosomal-recessive osteogenesis imperfecta. *Am J Hum Genet* **86(4)**, 551–9 (2010).
98. Kelley, B. P., Malfait, F., Bonafe, L., Baldrige, D., Homan, E. & Symoens, S. Mutations in FKBP10 cause recessive osteogenesis imperfecta and Bruck syndrome. *J Bone Min. Res* **26**, 666–72 (2011).

99. Shaheen, R., Al-Owain, M., Faqeih, E., Al-Hashmi, N., Awaji, A. & Al-Zayed, Z. Mutations in FKBP10 cause both Bruck syndrome and isolated osteogenesis imperfecta in humans. *Am J Med Genet* **155A**, 1448–1452 (2011).
100. Seyedhassani, S. M., Hashemi-Gorji, F., Yavari, M., Harazi, F. & Yassaee, V. R. Novel FKBP10 Mutation in a Patient with Osteogenesis Imperfecta Type XI. *Fetal Pediatr. Pathol.* **35**, 353–358 (2016).
101. Maghami, F., Tabei, S. M. B., Moravej, H., Dastsooz, H., Modarresi, F., Silawi, M. & Faghihi, M. A. Splicing defect in FKBP10 gene causes autosomal recessive osteogenesis imperfecta disease: a case report. *BMC Med. Genet.* **19**, 86 (2018).
102. Lietman, C. D., Lim, J., Grafe, I., Chen, Y., Ding, H., Bi, X., Ambrose, C. G., Fratzi-Zelman, N., Roschger, P., Klaushofer, K., Wagermaier, W., Schmidt, I., Fratzi, P., Rai, J., Weis, M., Eyre, D., Keene, D. R., Krakow, D. & Lee, B. H. Fkbp10 Deletion in Osteoblasts leads to Qualitative Defects in Bone. *J. Bone Miner. Res. Off. J. Am. Soc. Bone Miner. Res.* **32**, 1354–1367 (2017).
103. Asharani, P. V., Keupp, K., Semler, O., Wang, W., Li, Y. & Thiele, H. Attenuated BMP1 function compromises osteogenesis, leading to bone fragility in humans and zebrafish. *Am J Hum Genet* **90**, 661–674 (2012).
104. Martinez-Glez, V., Valencia, M., Caparros-Martin, J. A., Aglan, M., Temtamy, S. & Tenorio, J. Identification of a mutation causing deficient BMP1/mTLD proteolytic activity in autosomal recessive osteogenesis imperfecta. *Hum Mutat* **33**, 343-350. (2012).
105. Pollitt, R. C., Saraff, V., Dalton, A., Webb, E. A., Shaw, N. J., Sobey, G. J., Mughal, M. Z., Hobson, E., Ali, F., Bishop, N. J., Arundel, P., Högl, W. & Balasubramanian, M. Phenotypic variability in patients with osteogenesis imperfecta caused by BMP1 mutations. *Am. J. Med. Genet. A.* **170**, 3150–3156 (2016).
106. Shaheen, R., Alazami, A. M., Alshammari, M. J., Faqeih, E., Alhashmi, N. & Mousa, N. Study of autosomal recessive osteogenesis imperfecta in Arabia reveals a novel locus defined by TMEM38B mutation. *J Med Genet* **49**, 630-635. (2012).

107. Fahiminiya, S., Majewski, J., Mort, J. S., Moffatt, P., Glorieux, F. H. & Rauch, F. Mutations in WNT1 are a cause of osteogenesis imperfecta. *J Med Genet* **50**, 345–348 (2013).
108. Pyott, S. M., Tran, T. T., Leistriz, D. F., Pepin, M. G., Mendelsohn, N. J. & Temme, R. T. WNT1 Mutations in families affected by moderately severe and progressive recessive Osteogenesis Imperfecta. *Am J Hum Genet* **92**, 590–597 (2013).
109. Keupp, K., Beleggia, F., Kayserili, H., Barnes, A. M., Steiner, M. & Semler, O. Mutations in WNT1 cause different forms of bone fragility. *Am J Hum Genet* **92**, 565–574 (2013).
110. Won, J. Y., Jang, W. Y. & Lee, H. R. Novel missense loss-of-function mutations of WNT1 in an autosomal recessive osteogenesis imperfecta patient. *Eur J Med Genet* **60(8)**, 411–415 (2017).
111. Laine, C. M., Joeng, K. S. & Campeau, P. M. WNT1 mutations in early-onset osteoporosis and osteogenesis imperfecta. *N Engl J Med* **368 (19)**, 1809–1816 (2013).
112. Aldinger, K. A., Mendelsohn, N. J. & Chung, B. H. Variable brain phenotype primarily affects the brainstem and cerebellum in patients with osteogenesis im- perfecta caused by recessive WNT1 mutations. *J Med Genet* **53(6)**, 427–430 (2016).
113. Fageih, E., Shaheen, R. & Alkuraya, F. S. WNT1 mutation with recessive osteogenesis imperfecta and profound neurological phenotype. *J Med Genet* **50(7)**, 491–492 (2013).
114. Panigrahi, I., Didel, S., Kirpal, H., Bellampalli, R., Miyanath, S., Mullapudi, N. & Rao, S. Novel mutation in a family with WNT1 -related osteoporosis. *Eur. J. Med. Genet.* **61**, 369–371 (2018).
115. Stephen, J., Girisha, K. M., Dalal, A., Shukla, A., Shah, H., Srivastava, P., Kornak, U. & Phadke, S. R. Mutations in patients with osteogenesis imperfecta from consanguineous Indian families. *Eur. J. Med. Genet.* **58**, 21–27 (2015).
116. Guillemyn, B., Kayserili, H., Demuyneck, L., Sips, P., De Paepe, A., Syx, D., Coucke, P. J., Malfait, F. & Symoens, S. A homozygous pathogenic missense variant broadens the phenotypic and mutational spectrum of CREB3L1-related osteogenesis imperfecta. *Hum. Mol. Genet.* **28**, 1801–1809 (2019).

117. Cayami, F. K., Maugeri, A., Treurniet, S., Setijowati, E. D., Teunissen, B. P., Eekhoff, E. M. W., Pals, G., Faradz, S. M. & Micha, D. The first family with adult osteogenesis imperfecta caused by a novel homozygous mutation in CREB3L1. *Mol. Genet. Genomic Med.* **7**, e823 (2019).
118. Keller, R. B., Tran, T. T., Pyott, S. M., Pepin, M. G., Savarirayan, R., McGillivray, G., Nickerson, D. A., Bamshad, M. J. & Byers, P. H. Monoallelic and biallelic CREB3L1 variant causes mild and severe osteogenesis imperfecta, respectively. *Genet. Med.* **20**, 411–419 (2018).
119. Mendoza-Londono, R., Fahiminiya, S., Majewski, J., Care4Rare Canada Consortium, Tétreault, M., Nadaf, J., Kannu, P., Sochett, E., Howard, A., Stimec, J., Dupuis, L., Roschger, P., Klaushofer, K., Palomo, T., Ouellet, J., Al-Jallad, H., Mort, J. S., Moffatt, P., Boudko, S., Bächinger, H.-P. & Rauch, F. Recessive osteogenesis imperfecta caused by missense mutations in SPARC. *Am. J. Hum. Genet.* **96**, 979–985 (2015).
120. Moosa, S., Yamamoto, G. L., Garbes, L., Keupp, K., Beleza-Meireles, A., Moreno, C. A., Valadares, E. R., de Sousa, S. B., Maia, S., Saraiva, J., Honjo, R. S., Kim, C. A., Cabral de Menezes, H., Lausch, E., Lorini, P. V., Lamounier, A., Carniero, T. C. B., Giunta, C., Rohrbach, M., Janner, M., Semler, O., Beleggia, F., Li, Y., Yigit, G., Reintjes, N., Altmüller, J., Nürnberg, P., Cavalcanti, D. P., Zabel, B., Warman, M. L., Bertola, D. R., Wollnik, B. & Netzer, C. Autosomal-Recessive Mutations in MESD Cause Osteogenesis Imperfecta. *Am. J. Hum. Genet.* **105**, 836–843 (2019).
121. Starr, S. R., Roberts, T. T. & Fischer, P. R. Osteogenesis imperfecta: primary care. *Pediatr. Rev.* **31**, e54-64 (2010).
122. Rauch, F. & Glorieux, F. H. Osteogenesis imperfecta. *The Lancet* **363**, 1377–1385 (2004).
123. Forlino, A. & Marini, J. C. Osteogenesis imperfecta. *The Lancet* **387**, 1657–1671 (2016).
124. Garbes, L., Kim, K., Rieß, A., Hoyer-Kuhn, H., Beleggia, F., Bevot, A., Kim, M. J., Huh, Y. H., Kweon, H.-S., Savarirayan, R., Amor, D., Kakadia, P. M., Lindig, T., Kagan, K. O., Becker, J., Boyadjiev, S. A., Wollnik, B., Semler, O., Bohlander, S. K., Kim, J. & Netzer, C. Mutations in SEC24D, encoding a component of the COPII machinery, cause a syndromic form of osteogenesis imperfecta. *Am. J. Hum. Genet.* **96**, 432–439 (2015).

125. Lindert, U., Cabral, W. A., Ausavarat, S., Tongkobpetch, S., Ludin, K., Barnes, A. M., Yeetong, P., Weis, M., Krabichler, B., Srichomthong, C., Makareeva, E. N., Janecke, A. R., Leikin, S., Röthlisberger, B., Rohrbach, M., Kennerknecht, I., Eyre, D. R., Suphapeetiporn, K., Giunta, C., Marini, J. C. & Shotelersuk, V. MBTPS2 mutations cause defective regulated intramembrane proteolysis in X-linked osteogenesis imperfecta. *Nat. Commun.* **7**, 11920 (2016).
126. Rauch, F., Fahiminiya, S., Majewski, J., Carrot-Zhang, J., Boudko, S., Glorieux, F., Mort, J. S., Bächinger, H.-P. & Moffatt, P. Cole-Carpenter syndrome is caused by a heterozygous missense mutation in P4HB. *Am. J. Hum. Genet.* **96**, 425–431 (2015).
127. Marini, J. & Smith, S. M. Osteogenesis Imperfecta. (2015). at <http://www.ncbi.nlm.nih.gov/books/NBK279109/>
128. Marini, J. C. & Gerbert, N. L. Osteogenesis imperfecta. Rehabilitation and prospects for gene therapy. *Jama* **277**, 746–750 (1997).
129. Morello, R. & Esposito, P. W. Osteogenesis Imperfecta. *Osteogenesis* (2012). doi:10.5772/34775
130. General side effects of bisphosphonates and denosumab | Cancer Research UK. at <https://www.cancerresearchuk.org/about-cancer/cancer-in-general/treatment/bisphosphonates/side-effects/general>
131. Rosen, H. Bisphosphonate therapy for the treatment of osteoporosis - UpToDate. (2022). at <https://www.uptodate.com/contents/bisphosphonate-therapy-for-the-treatment-of-osteoporosis#H3422893804>
132. Gao, Y., Liu, X., Gu, Y., Song, D., Ding, M., Liao, L., Wang, J., Ni, J. & He, G. The Effect of Bisphosphonates on Fracture Healing Time and Changes in Bone Mass Density: A Meta-Analysis. *Front. Endocrinol.* **12**, (2021).
133. Lim, E. J., Kim, J.-T., Kim, C.-H., Kim, J. W., Chang, J. S. & Yoon, P. W. Effect of Preoperative Bisphosphonate Treatment on Fracture Healing after Internal Fixation Treatment of Intertrochanteric Femoral Fractures. *Hip Pelvis* **31**, 75–81 (2019).

134. Giusti, A. & Papapoulos, S. E. in *Primer Metab. Bone Dis. Disord. Miner. Metab.* 545–552 (Wiley, 2018). doi:10.1002/9781119266594.ch71
135. Ballard, T. & Chargui, S. in *StatPearls* (StatPearls Publishing, 2022). at <<http://www.ncbi.nlm.nih.gov/books/NBK551673/>>
136. Uveges, T. E., Collin-Osdoby, P., Cabral, W. A., Ledgard, F., Goldberg, L., Bergwitz, C., Forlino, A., Osdoby, P., Gronowicz, G. A. & Marini, J. C. Cellular mechanism of decreased bone in Brl mouse model of OI: imbalance of decreased osteoblast function and increased osteoclasts and their precursors. *J. Bone Miner. Res. Off. J. Am. Soc. Bone Miner. Res.* **23**, 1983–1994 (2008).
137. Li, J. S., Eisenstein, E. L., Grabowski, H. G., Reid, E. D., Mangum, B., Schulman, K. A., Goldsmith, J. V., Murphy, M. D., Califf, R. M. & Benjamin, D. K. J. Economic return of clinical trials performed under the pediatric exclusivity program. *Jama* **297**, 480–488 (2007).
138. Gatti, D., Antoniazzi, F., Prizzi, R., Braga, V., Rossini, M., Tato, L., Viapiana, O. & Adami, S. Intravenous neridronate in children with osteogenesis imperfecta: a randomized controlled study. *J. Bone Miner. Res. Off. J. Am. Soc. Bone Miner. Res.* **20**, 758–763 (2005).
139. Letocha, A. D., Cintas, H. L., Troendle, J. F., Reynolds, J. C., Cann, C. E., Chernoff, E. J., Hill, S. C., Gerber, L. H. & Marini, J. C. Controlled trial of pamidronate in children with types III and IV osteogenesis imperfecta confirms vertebral gains but not short-term functional improvement. *J. Bone Miner. Res. Off. J. Am. Soc. Bone Miner. Res.* **20**, 977–986 (2005).
140. Sackers, R., Kok, D., Engelbert, R., van Dongen, A., Jansen, M., Pruijs, H., Verbout, A., Schweitzer, D. & Uiterwaal, C. Skeletal effects and functional outcome with olpadronate in children with osteogenesis imperfecta: a 2-year randomised placebo-controlled study. *J. Bone Miner. Res. Off. J. Am. Soc. Bone Miner. Res.* **363**, 1427–1431 (2005).
141. Ward, L. M., Rauch, F., Whyte, M. P., D’Astous, J., Gates, P. E., Grogan, D., Lester, E. L., McCall, R. E., Pressly, T. A., Sanders, J. O., Smith, P. A., Steiner, R. D., Sullivan, E., Tyerman, G., Smith-Wright, D. L., Verbruggen, N., Heyden, N., Lombardi, A. & Glorieux, F. H. Alendronate for

- the treatment of pediatric osteogenesis imperfecta: a randomized placebo-controlled study. *J. Clin. Endocrinol. Metab.* **96**, 355–364 (2011).
142. Bradbury, L. A., Barlow, S., Geoghegan, F., Hannon, R. A., Stuckey, S. L., Wass, J. A., Russell, R. G., Brown, M. A. & Duncan, E. L. Risedronate in adults with osteogenesis imperfecta type I: increased bone mineral density and decreased bone turnover, but high fracture rate persists. *Oporosis Int. J. Establ. Result Coop. Eur. Found. Osteoporos. Natl. Osteoporos. Found. USA* **23**, 285–294 (2012).
143. Land, C., Rauch, F., Travers, R. & Glorieux, F. H. Osteogenesis imperfecta type VI in childhood and adolescence: effects of cyclical intravenous pamidronate treatment. *Bone* **40**, 638–644 (2007).
144. Semler, O., Netzer, C., Hoyer-Kuhn, H., Becker, J., Eysel, P. & Schoenau, E. First use of the RANKL antibody denosumab in osteogenesis imperfecta type VI. *J. Musculoskelet. Neuronal Interact.* **12**, 183–188 (2012).
145. Germain, D. P. & Jurca-Simina, I. E. in *Neurometab. Hered. Dis. Adults* (ed. Burlina, A. P.) 1–28 (Springer International Publishing, 2018). doi:10.1007/978-3-319-76148-0\_1
146. Boycott, K. M., Vanstone, M. R., Bulman, D. E. & MacKenzie, A. E. Rare disease genetics in the era of next-generation sequencing: discovery to translation. *Nat. Rev. Genet.* **14 (10)**, 681–690 (2013).
147. Basta, M. & Pandya, A. M. in *StatPearls* (StatPearls Publishing, 2022). at <http://www.ncbi.nlm.nih.gov/books/NBK557383/>
148. Migeon, B. R. X-linked diseases: susceptible females. *Genet. Med.* **22**, 1156–1174 (2020).
149. Newey, P. J., Whyte, M. P. & Thakker, R. V. in *Primer Metab. Bone Dis. Disord. Miner. Metab.* 341–350 (Wiley, 2018). doi:10.1002/9781119266594.ch42
150. Stauffer, S., Gardner, A., Duprez, W., Ungu, D. A. K. & Wismer, P. in *Labster Virtual Lab Exp. Basic Genet.* 1–11 (Springer Berlin Heidelberg, 2018). doi:10.1007/978-3-662-57999-2\_1



151. Bateson, W. & Mendel, G. Mendel's Principles of Heredity: A Defence with a Translation of Mendel's Original Papers on Hybridization the University Press. (1902).
152. Grant, M. Globins, Genes and Globinopathies. *Biol. Sci. Rev.* **9**, 2–5 (1997).
153. Read, A. Cystic fibrosis. Population genetics in action. *Biol Sci Rev* **4 (4)**, 18–20 (1992).
154. Pranckėnienė, L., Jakaitienė, A., Ambrozaitytė, L., Kavaliauskienė, I. & Kučinskas, V. Insights Into de novo Mutation Variation in Lithuanian Exome. *Front. Genet.* **9**, (2018).
155. Hunter, D. J. Gene-environment interactions in human diseases. *Nat. Rev. Genet.* **6**, 287–298 (2005).
156. Cooper, D. N., Krawczak, M., Polychronakos, C., Tyler-Smith, C. & Kehrer-Sawatzki, H. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Hum. Genet.* **132**, 1077–1130 (2013).
157. Lobo, I. Same genetic mutation, different genetic disease phenotype. *Nat. Educ.* **1(1)**, 64 (2008).
158. Shawky, R. M., Abd-Elkhalek, H. S., Gad, S., Mohammad, S. A. & Seifeldin, N. S. Cornelia-de Lange syndrome in an Egyptian infant with unusual bone deformities. *Egypt. J. Med. Hum. Genet.* **14(1)**, 109–112 (2013).
159. Shawky, R. M., Abd-Elkhalek, H. S. & Gad, S. Intrafamilial variability in Simpson–Golabi–Behmel syndrome with bilateral posterior ear lobule creases. *Egypt. J. Med. Hum. Genet.* **15(1)**, 87–90 (2014).
160. Gruber, C. & Bogunovic, D. Incomplete penetrance in primary immunodeficiency: a skeleton in the closet. *Hum. Genet.* **139**, 745–757 (2020).
161. Zlotogora, J. Penetrance and expressivity in the molecular age. *Genet. Med.* **5**, 347–352 (2003).
162. Hung, C. C., Lee, C. N., Chen, C. P., Lin, S. P., Chao, M. C., Chiou, S. S. & Su, Y. N. Low penetrance of retinoblastoma for p. V654L mutation of the RB1 gene. *BMC Med. Genet.* **12(1)**, 1 (2011).

163. Xue, Y., Chen, Y., Ayub, Q., Huang, N., Ball, E. V., Mort, M., Phillips, A. D., Shaw, K., Stenson, P. D., Cooper, D. N. & Tyler-Smith, C. Deleterious-and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. *Am. J. Hum. Genet.* **91(6)**, (2012).
164. Umair, M., Alhaddad, B., Rafique, A., Jan, A., Haack, T. B., Graf, E., Ullah, A., Ahmad, F., Strom, T. M., Meitinger, T. & Ahmad, W. Exome sequencing reveals a novel homozygous splice site variant in the WNT1 gene underlying osteogenesis imperfecta type 3. *Pediatr. Res.* **82**, 753–758 (2017).
165. Kausar, M., Siddiqi, S., Yaqoob, M., Mansoor, S., Makitie, O., Mir, A., Khor, C. C., Foo, J. N. & Anees, M. Novel mutation G324C in WNT1 mapped in a large Pakistani family with severe recessively inherited Osteogenesis Imperfecta. *J. Biomed. Sci.* **25**, (2018).
166. Maldonado, G., Ferro, C., Paredes, C. & Ríos, C. Use of denosumab in osteogenesis imperfecta: A case report. *Rev. Colomb. Reumatol. Engl. Ed.* **26**, 68–73 (2019).
167. Lafage-Proust, M.-H. & Courtois, I. The management of osteogenesis imperfecta in adults: state of the art. *Joint Bone Spine* (2019). doi:10.1016/j.jbspin.2019.02.001
168. Boskey, A. L. & Robey, P. G. in *Primer Metab. Bone Dis. Disord. Miner. Metab.* 84–92 (Wiley, 2018). doi:10.1002/9781119266594.ch11
169. Cho, S. Y., Asharani, P. V., Kim, O.-H., Iida, A., Miyake, N., Matsumoto, N., Nishimura, G., Ki, C.-S., Hong, G., Kim, S. J., Sohn, Y. B., Park, S. W., Lee, J., Kwun, Y., Carney, T. J., Huh, R., Ikegawa, S. & Jin, D.-K. Identification and In Vivo Functional Characterization of Novel Compound Heterozygous BMP1 Variants in Osteogenesis Imperfecta. *Hum. Mutat.* **36**, 191–195 (2015).
170. Morello, R. Osteogenesis imperfecta and therapeutics. *Matrix Biol.* **71–72**, 294–312 (2018).
171. Lim, J., Grafe, I., Alexander, S. & Lee, B. Genetic causes and mechanisms of Osteogenesis Imperfecta. *Bone* **102**, 40–49 (2017).
172. Fratzi-Zelman, N., Barnes, A. M., Weis, M., Carter, E., Hefferan, T. E., Perino, G., Chang, W., Smith, P. A., Roschger, P., Klaushofer, K., Glorieux, F. H., Eyre, D. R., Raggio, C., Rauch, F. &

- Marini, J. C. Non-Lethal Type VIII Osteogenesis Imperfecta Has Elevated Bone Matrix Mineralization. *J. Clin. Endocrinol. Metab.* **101**, 3516–3525 (2016).
173. Jiang, Y., Pan, J., Guo, D., Zhang, W., Xie, J., Fang, Z., Guo, C., Fang, Q., Jiang, W. & Guo, Y. Two novel mutations in the PPIB gene cause a rare pedigree of osteogenesis imperfecta type IX. *Clin. Chim. Acta Int. J. Clin. Chem.* **469**, 111–118 (2017).
174. Besio, R., Chow, C.-W., Tonelli, F., Marini, J. C. & Forlino, A. Bone biology: insights from osteogenesis imperfecta and related rare fragility syndromes. *FEBS J.* **286**, 3033–3056 (2019).
175. Huang, Y., Mei, L., Lv, W., Li, H., Zhang, R., Pan, Q., Tan, H., Guo, J., Luo, X., Chen, C., Liang, D. & Wu, L. Targeted exome sequencing identifies novel compound heterozygous mutations in P3H1 in a fetus with osteogenesis imperfecta type VIII. *Clin. Chim. Acta* **464**, 170–175 (2016).
176. Lv, F., Xu, X., Song, Y., Li, L., Asan, Wang, J., Yang, H., Wang, O., Jiang, Y., Xia, W., Xing, X. & Li, M. Novel Mutations in PLOD2 Cause Rare Bruck Syndrome. *Calcif. Tissue Int.* **102**, 296–309 (2018).
177. Jovanovic, M., Guterman-Ram, G. & Marini, J. C. Osteogenesis Imperfecta: Mechanisms and Signaling Pathways Connecting Classical and Rare OI Types. *Endocr. Rev.* **43**, 61–90 (2022).
178. Maeda, K., Kobayashi, Y., Koide, M., Uehara, S., Okamoto, M., Ishihara, A., Kayama, T., Saito, M. & Marumo, K. The Regulation of Bone Metabolism and Disorders by Wnt Signaling. *Int. J. Mol. Sci.* **20**, 5525 (2019).
179. Carreira, A. C. O., Zambuzzi, W. F., Rossi, M. C., Filho, R. A., Sogayar, M. C. & Granjeiro, J. M. in *Vitam. Horm.* **99**, 293–322 (Elsevier, 2015).
180. Murakami, T., Hino, S., Nishimura, R., Yoneda, T., Wanaka, A. & Imaizumi, K. Distinct mechanisms are responsible for osteopenia and growth retardation in OASIS-deficient mice. *Bone* **48**, 514–523 (2011).
181. Hu, Z. & Gulyaeva, O. TRIC-B: an under-explored druggable ion channel. *Nat. Rev. Drug Discov.* **18**, 657–657 (2019).

182. Lv, F., Xu, X., Wang, J., Liu, Y., Asan, Wang, J., Song, L., Song, Y., Jiang, Y., Wang, O., Xia, W., Xing, X. & Li, M. Two novel mutations in TMEM38B result in rare autosomal recessive osteogenesis imperfecta. *J. Hum. Genet.* **61**, 539–545 (2016).
183. Webb, E. A., Balasubramanian, M., Fratzl-Zelman, N., Cabral, W. A., Titheradge, H., Alsaedi, A., Saraff, V., Vogt, J., Cole, T., Stewart, S., Crabtree, N. J., Sargent, B. M., Gamsjaeger, S., Paschalis, E. P., Roschger, P., Klaushofer, K., Shaw, N. J., Marini, J. C. & Högl, W. Phenotypic Spectrum in Osteogenesis Imperfecta Due to Mutations in TMEM38B: Unraveling a Complex Cellular Defect. *J. Clin. Endocrinol. Metab.* **102**, 2019–2028 (2017).
184. Guzvic, M. The history of DNA sequencing. **32**, 301–312 (2013).
185. Franca, L., Carrilho, E. & Kist, T. A review of DNA sequencing techniques. *Q Rev Biophys* **35**, 169–200 (2002).
186. Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 560–564 (1977).
187. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5463–5467 (1977).
188. Barba, M., Czosnek, H. & Hadidi, A. Historic perspective, development and applications of next-generation sequencing in plant virology. *Viruses* **6**, 100–136 (2014).
189. Mardis, E. R. Next-Generation DNA Sequencing Methods. *Annu. Rev. Genomics Hum. Genet.* **9**, 387–402 (2008).
190. Mardis, E. R. A decade's perspective on DNA sequencing technology. *Nature* **470**, 198–203 (2011).
191. Metzker, M. L. Sequencing technologies — the next generation. *Nat. Rev. Genet.* **11**, 31–46 (2010).
192. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26**, 1135–1145 (2008).

193. Kulski, J. *Next Generation Sequencing: Advances, Applications and Challenges*. (BoD – Books on Demand, 2016).
194. Thompson, J. F. & Milos, P. M. The properties and applications of single-molecule DNA sequencing. *Genome Biol.* **12**, 217 (2011).
195. Sefid Dashti, M. J. & Gamielien, J. A practical guide to filtering and prioritizing genetic variants. *BioTechniques* **62**, 18–30 (2017).
196. Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. A. & Shendure, J. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* **12**, 745–755 (2011).
197. Rabbani, B., Tekin, M. & Mahdieh, N. The promise of whole-exome sequencing in medical genetics. *J. Hum. Genet.* **59**, 5–15 (2014).
198. Ohanian, M., Otway, R. & Fatkin, D. Heuristic Methods for Finding Pathogenic Variants in Gene Coding Sequences. *J. Am. Heart Assoc.* **1**, e002642 (2012).
199. Korb, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F., Kim, P. M., Palejev, D., Carriero, N. J., Du, L., Taillon, B. E., Chen, Z., Tanzer, A., Saunders, A. C. E., Chi, J., Yang, F., Carter, N. P., Hurles, M. E., Weissman, S. M., Harkins, T. T., Gerstein, M. B., Egholm, M. & Snyder, M. Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. *Science* **318**, 420–426 (2007).
200. Fonseca, N. A., Rung, J., Brazma, A. & Marioni, J. C. Tools for mapping high-throughput sequencing data. *EMBL Outstation Eur. Bioinforma. Inst. EBI Hinxton Camb. CB10 ISD* **28**, 3169–3177 (2012).
201. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
202. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

203. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
204. Sequence Alignment/Map Format Specifications. (2016). at <Available at <http://samtools.github.io/hts-specs/SAMv1.pdf>>
205. Center for Statistical Genetics. SAM/BAM. (2014). at <Available at:<http://genome.sph.umich.edu/wiki/SAM>.>
206. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
207. Picard Tool. *Broad Inst. GitHub Repos.* (2018). at <<http://broadinstitute.github.io/picard/>; Broad Institute>
208. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. & DePristo, M. A. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
209. Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X. & Ruden, D. M. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin)* **6**, 80–92 (2012).
210. McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P. & Cunningham, F. The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
211. Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M. & Sirotkin, K. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
212. Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korb, J. O., Lander, E. S., Lee, C., Lehrach, H., Mardis, E. R., Marth, G. T., McVean, G. A., Nickerson, D. A., Schmidt, J. P., Sherry, S. T., Wang, J.,

Wilson, R. K., Gibbs, R. A., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., Kovar, C., Lee, S., Muzny, D., Reid, J. G., Zhu, Y., Wang, J., Chang, Y., Feng, Q., Fang, X., Guo, X., Jian, M., Jiang, H., Jin, X., Lan, T., Li, G., Li, J., Li, Y., Liu, S., Liu, X., Lu, Y., Ma, X., Tang, M., Wang, B., Wang, G., Wu, H., Wu, R., Xu, X., Yin, Y., Zhang, D., Zhang, W., Zhao, J., Zhao, M., Zheng, X., Lander, E. S., Altshuler, D. M., Gabriel, S. B., Gupta, N., Gharani, N., Toji, L. H., Gerry, N. P., Resch, A. M., Flicek, P., Barker, J., Clarke, L., Gil, L., Hunt, S. E., Kelman, G., Kulesha, E., Leinonen, R., McLaren, W. M., Radhakrishnan, R., Roa, A., Smirnov, D., Smith, R. E., Streeter, I., Thormann, A., Toneva, I., Vaughan, B., Zheng-Bradley, X., Bentley, D. R., Grocock, R., Humphray, S., James, T., Kingsbury, Z., Lehrach, H., Sudbrak, R., Albrecht, M. W., Amstislavskiy, V. S., Borodina, T. A., Lienhard, M., Mertes, F., Sultan, M., Timmermann, B., Yaspo, M.-L., Mardis, E. R., Wilson, R. K., Fulton, L., Fulton, R., Sherry, S. T., Ananiev, V., Belaia, Z., Beloslyudtsev, D., Bouk, N., Chen, C., Church, D., Cohen, R., Cook, C., Garner, J., Hefferon, T., Kimelman, M., Liu, C., Lopez, J., Meric, P., O'Sullivan, C., Ostapchuk, Y., Phan, L., Ponomarov, S., Schneider, V., Shekhtman, E., Sirotkin, K., Slotta, D., Zhang, H., McVean, G. A., Durbin, R. M., Balasubramaniam, S., Burton, J., Danecek, P., Keane, T. M., Kolb-Kococinski, A., McCarthy, S., Stalker, J., Quail, M., Schmidt, J. P., Davies, C. J., Gollub, J., Webster, T., Wong, B., Zhan, Y., Auton, A., Campbell, C. L., Kong, Y., Marcketta, A., Gibbs, R. A., Yu, F., Antunes, L., Bainbridge, M., Muzny, D., Sabo, A., Huang, Z., Wang, J., Coin, L. J. M., Fang, L., Guo, X., Jin, X., Li, G., Li, Q., Li, Y., Li, Z., Lin, H., Liu, B., Luo, R., Shao, H., Xie, Y., Ye, C., Yu, C., Zhang, F., Zheng, H., Zhu, H., Alkan, C., Dal, E., Kahveci, F., Marth, G. T., Garrison, E. P., Kural, D., Lee, W.-P., Fung Leong, W., Stromberg, M., Ward, A. N., Wu, J., Zhang, M., Daly, M. J., DePristo, M. A., Handsaker, R. E., Altshuler, D. M., Banks, E., Bhatia, G., del Angel, G., Gabriel, S. B., Genovese, G., Gupta, N., Li, H., Kashin, S., Lander, E. S., McCarroll, S. A., Nemes, J. C., Poplin, R. E., Yoon, S. C., Lihm, J., Makarov, V., Clark, A. G., Gottipati, S., Keinan, A., Rodriguez-Flores, J. L., Korb, J. O., Rausch, T., Fritz, M. H., Stütz, A. M., Flicek, P., Beal, K., Clarke, L., Datta, A., Herrero, J., McLaren, W. M., Ritchie, G. R. S., Smith, R. E., Zerbino, D., Zheng-Bradley, X., Sabeti, P. C., Shlyakhter, I., Schaffner, S. F., Vitti, J., Cooper, D. N., Ball, E. V., Stenson, P. D., Bentley, D. R.,

Barnes, B., Bauer, M., Keira Cheetham, R., Cox, A., Eberle, M., Humphray, S., Kahn, S., Murray, L., Peden, J., Shaw, R., Kenny, E. E., Batzer, M. A., Konkel, M. K., Walker, J. A., MacArthur, D. G., Lek, M., Sudbrak, R., Amstislavskiy, V. S., Herwig, R., Mardis, E. R., Ding, L., Koboldt, D. C., Larson, D., Ye, K., Gravel, S., The 1000 Genomes Project Consortium, Corresponding authors, Steering committee, Production group, Baylor College of Medicine, BGI-Shenzhen, Broad Institute of MIT and Harvard, Coriell Institute for Medical Research, European Molecular Biology Laboratory, E. B. I., Illumina, Max Planck Institute for Molecular Genetics, McDonnell Genome Institute at Washington University, US National Institutes of Health, University of Oxford, Wellcome Trust Sanger Institute, Analysis group, Affymetrix, Albert Einstein College of Medicine, Bilkent University, Boston College, Cold Spring Harbor Laboratory, Cornell University, European Molecular Biology Laboratory, Harvard University, Human Gene Mutation Database, Icahn School of Medicine at Mount Sinai, Louisiana State University, Massachusetts General Hospital, McGill University & National Eye Institute, N. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

213. Karczewski, K. J., Weisburd, B., Thomas, B., Solomonson, M., Ruderfer, D. M., Kavanagh, D., Hamamsy, T., Lek, M., Samocha, K. E., Cummings, B. B., Birnbaum, D., Daly, M. J. & MacArthur, D. G. The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res.* **45**, D840–D845 (2017).
214. Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., Walters, R. K., Tashman, K., Farjoun, Y., Banks, E., Poterba, T., Wang, A., Seed, C., Whiffin, N., Chong, J. X., Samocha, K. E., Pierce-Hoffman, E., Zappala, Z., O’Donnell-Luria, A. H., Minikel, E. V., Weisburd, B., Lek, M., Ware, J. S., Vittal, C., Armean, I. M., Bergelson, L., Cibulskis, K., Connolly, K. M., Covarrubias, M., Donnelly, S., Ferriera, S., Gabriel, S., Gentry, J., Gupta, N., Jeandet, T., Kaplan, D., Llanwarne, C., Munshi, R., Novod, S., Petrillo, N., Roazen, D., Ruano-Rubio, V., Saltzman, A.,



- Schleicher, M., Soto, J., Tibbetts, K., Tolonen, C., Wade, G., Talkowski, M. E., Neale, B. M., Daly, M. J. & MacArthur, D. G. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
215. Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., McDaniel, K., Ovetsky, M., Riley, G., Zhou, G., Holmes, J. B., Kattman, B. L. & Maglott, D. R. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
216. Paila, U., Chapman, B. A., Kirchner, R. & Quinlan, A. R. GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations. *PLoS Comput. Biol.* **9**, e1003153 (2013).
217. Zhu, J., Zhao, Q., Katsevich, E. & Sabatti, C. Exploratory Gene Ontology Analysis with Interactive Visualization. *Sci. Rep.* **9**, 1–9 (2019).
218. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2019).
219. Pomaznoy, M., Ha, B. & Peters, B. GOnet: a tool for interactive Gene Ontology analysis. *BMC Bioinformatics* **19**, 470 (2018).
220. Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* **37**, W305–W311 (2009).
221. Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., Kuhn, M., Bork, P., Jensen, L. J. & von Mering, C. STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43**, D447–D452 (2015).
222. Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., Jensen, L. J. & Mering, C. von. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).

223. Andreopoulos, B. & Labudde, D. in *Expert Rev. Proteomics - EXPERT REV PROTEOMICS* **1**, (2013).
224. Bajpai, A. K., Davuluri, S., Tiwary, K., Narayanan, S., Oguru, S., Basavaraju, K., Dayalan, D., Thirumurugan, K. & Acharya, K. K. Systematic comparison of the protein-protein interaction databases from a user's perspective. *J. Biomed. Inform.* **103**, 103380 (2020).
225. Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M. & Shendure, J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
226. Quang, D., Chen, Y. & Xie, X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **31**, 761–763 (2015).
227. Shihab, H. A., Rogers, M. F., Gough, J., Mort, M., Cooper, D. N., Day, I. N. M., Gaunt, T. R. & Campbell, C. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* **31**, 1536–1543 (2015).
228. Bendl, J., Musil, M., Štourač, J., Zendulka, J., Damborský, J. & Brezovský, J. PredictSNP2: A Unified Platform for Accurately Evaluating SNP Effects by Exploiting the Different Characteristics of Variants in Distinct Genomic Regions. *PLoS Comput. Biol.* **12**, (2016).
229. GARBIERI, T. F., BROZOSKI, D. T., DIONÍSIO, T. J., SANTOS, G. F. & NEVES, L. T. das. Human DNA extraction from whole saliva that was fresh or stored for 3, 6 or 12 months using five different protocols. *J. Appl. Oral Sci.* **25**, 147–158 (2017).
230. Long-term storage of Oragene®/ saliva samples. DNA Genotek. PD-PR-012.
231. samtools(1) manual page. at <<http://www.htslib.org/doc/samtools.html>>
232. Bukowski, R., Sun, Q. & Wang, M. Variant calling: Part 1 [PowerPoint slides]. (2017).
233. Reeve, J. SNP calling pipeline for Pool-seq [PowerPoint slides]. (2018). at <[https://yeamanlab.weebly.com/uploads/5/7/9/5/57959825/snp\\_calling\\_pipeline.pdf](https://yeamanlab.weebly.com/uploads/5/7/9/5/57959825/snp_calling_pipeline.pdf)>
234. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv* **1303**, (2013).

235. Choudhury, A., Ramsay, M., Hazelhurst, S., Aron, S., Bardien, S., Botha, G., Chimusa, E. R., Christoffels, A., Gamielien, J., Sefid-Dashti, M. J., Joubert, F., Meintjes, A., Mulder, N., Ramesar, R., Rees, J., Scholtz, K., Sengupta, D., Soodyall, H., Venter, P., Warnich, L. & Pepper, M. S. Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. *Nat. Commun.* **8**, 2062 (2017).
236. Robertson, A., Malan, A., Greenfield, D., Mashao, L., Rhoda, N., Goga, A., Kerber, K. & Lawn, J. Newborn care chart: Routine care at birth and management of the sick and small newborn in hospitals [PowerPoint slides]. (2014). at <<http://www.kznhealth.gov.za/family/MCWH/KINC-May2014.pdf>>
237. Mortier, G. R., Cohn, D. H., Cormier-Daire, V., Hall, C., Krakow, D., Mundlos, S., Nishimura, G., Robertson, S., Sangiorgi, L., Savarirayan, R., Sillence, D., Superti-Furga, A., Unger, S. & Warman, M. L. Nosology and classification of genetic skeletal disorders: 2019 revision. *Am. J. Med. Genet. A.* **179**, 2393–2419 (2019).



## URL website references

Monarch Initiative:	<a href="https://monarchinitiative.org">https://monarchinitiative.org</a>
STRING database:	<a href="https://string-db.org">https://string-db.org</a>
UniProtKB:	<a href="https://www.uniprot.org">https://www.uniprot.org</a>
ClinVar:	<a href="https://www.ncbi.nlm.nih.gov/clinvar/">https://www.ncbi.nlm.nih.gov/clinvar/</a>
dbSNP:	<a href="https://ncbi.nlm.nih.gov/snp/">https://ncbi.nlm.nih.gov/snp/</a>
Blastp:	<a href="https://blast.ncbi.nlm.nih.gov/">https://blast.ncbi.nlm.nih.gov/</a>
WebPRANK:	<a href="https://www.ebi.ac.uk/goldman/webprank/">https://www.ebi.ac.uk/goldman/webprank/</a>
CADD:	<a href="https://cadd.gs.washington.edu/">https://cadd.gs.washington.edu/</a>
SIFT-InDel:	<a href="https://sift.bii.a-star.edu.sg/www/SIFT_InDels2.html">https://sift.bii.a-star.edu.sg/www/SIFT_InDels2.html</a>
DisGeNET:	<a href="https://www.disgenet.org">https://www.disgenet.org</a>
REACTOME:	<a href="https://www.reactome.org/">https://www.reactome.org/</a>
OMIM:	<a href="https://www.omimexplorer.com">https://www.omimexplorer.com</a>
ToppGene:	<a href="https://www.toppgene.cchmc.org">https://www.toppgene.cchmc.org</a>
LOVD:	<a href="https://oi.gene.le.ac.uk/">https://oi.gene.le.ac.uk/</a> <a href="https://databases.lovd.nl">https://databases.lovd.nl</a>



## GATK best practice variant discovery:

Local alignment:	<a href="https://github.com/broadinstitute/gatk-docs/blob/master/gatk3-tutorials/(howto)_Perform_local_realignment_around_InDels.md">https://github.com/broadinstitute/gatk-docs/blob/master/gatk3-tutorials/(howto)_Perform_local_realignment_around_InDels.md</a>
Base recalibration:	<a href="https://github.com/broadinstitute/gatk-docs/blob/master/gatk3-tutorials/(howto)_Recalibrate_base_quality_scores_%3D_run_BQSR.md">https://github.com/broadinstitute/gatk-docs/blob/master/gatk3-tutorials/(howto)_Recalibrate_base_quality_scores_%3D_run_BQSR.md</a>
Variant calling:	<a href="https://github.com/broadinstitute/gatk-docs/blob/master/gatk3-tutorials/(howto)_Call_variants_with_HaplotypeCaller.md">https://github.com/broadinstitute/gatk-docs/blob/master/gatk3-tutorials/(howto)_Call_variants_with_HaplotypeCaller.md</a>
Variant recalibration:	<a href="https://github.com/broadinstitute/gatk-docs/blob/master/gatk3-tutorials/(howto)_Recalibrate_variant_quality_scores_%3D_run_VQ_SR.md">https://github.com/broadinstitute/gatk-docs/blob/master/gatk3-tutorials/(howto)_Recalibrate_variant_quality_scores_%3D_run_VQ_SR.md</a>

## Appendices

### Appendix A: Glossary

**Mendelian disease:** A disease that is carried in families, in either a dominant or recessive manner, or that is typically controlled by variants of large effect in a single gene.

**Compound Heterozygote:** is the condition of having two or more heterogeneous recessive alleles at a particular locus that can cause genetic disease in a heterozygous state; that is, an organism is a compound heterozygote when it has two recessive alleles for the same gene, but with those two alleles being different from each other (for example, both alleles might be mutated but at different locations).

**Mapping:** assembling reads by aligning reads against a template, this is this reference sequence.

**Missense mutations:** Single DNA-base changes that occur in the coding regions of genes and alter the resulting encoded amino acid sequence.

**Nonsense mutations:** DNA-base changes that introduce termination codons in the coding sequences of genes, resulting in truncated proteins.

**Next-generation sequencing (Massively parallel):** DNA-sequencing methods that involve chemical assays other than the traditional Sanger dideoxy-chain-termination method. Next-generation sequencing methods produce much larger quantities of data at less expense, but the individual raw sequence reads that are generated from the individual amplified DNA- template sequences are shorter and have lower quality. / High-throughput sequencing nano- technology used to determine the base-pair sequence of DNA/RNA molecules at a much larger quantities.

**Annotation:** Computational process of attaching biological relevant information to genome sequence data.

**Coverage:** Also known as sequencing depth. Sequence coverage refers to the average number of reads per locus and differs from physical coverage, a term often used in genome assembly referring to cumulative length of reads or read pairs expressed as a multiple of genome size.

**GC content:** The content of guanine and cytosine bases in DNA/RNA sequence.

**InDel:** Insertion/deletion polymorphism.

**Paired-end sequencing:** Sequence information from two ends of a short DNA fragment,

usually a few hundred base pairs long.

**Read:** Short base-pair sequence inferred from the DNA/RNA template by sequencing.

**Resorption:** The organic process in which the substance of some differentiated structure that has been produced by the body undergoes lysis and assimilation

**Dentinogenesis Imperfecta:** is a disorder of tooth development.

**Lamellation:** an arrangement or structure in which there are thin layers, plates, or scales.

**Wormian bones:** A subset of the small intrasutural bones that lie between the cranial sutures formed by the bones of the skull vault, which are formed due to additional ossification centres in or near sutures.

**Histo-morphometry:** The quantitative study of the microscopic organization and structure of a tissue (as bone) especially by computer-assisted analysis of images formed by a microscope.

**Osteomalacia:** A bone disease in adults analogous to rickets in children, marked by bone demineralization caused by impaired metabolism or deficiency of vitamin D or phosphorus

**Hyperosteoidosis:** Excessive formation of osteoid, as seen in rickets and osteomalacia.

**Coxa vara:** a deformed hip joint in which the neck of the femur is bent downward.

**Rhizomelia:** Disproportion in the length of the most proximal segment of the limbs (upper arms and thighs)

**Protrusion acetabuli:** disease in which deformity occurs in acetabulum medial wall. This deformity leads to migration of femoral bone head into pelvic cavity.

**Microretrognathia:** A form of development hypoplasia of the mandible in which the mandible is mislocalised posteriorly.

**Mesomelia:** The condition of having abnormal short forearms and legs

**Consanguineous:** Of the same blood; related by birth; descended from the same parent or ancestor.

**Mineralization:** The process of mineralizing, or forming a mineral by combination of a metal with another element; also, the process of converting into a mineral, as a bone or a plant

**Catalyse:** To modify, especially to increase, the rate of (a chemical reaction) by catalysis.  
Toproduce fundamental change in; transform.

**Tropoelastin:** A water-soluble molecule that binds to form the roteinelastin.

**Osteoprotegerin:** A cytokine that can inhibit the production of osteoclasts.

**Osteoblastogenesis:** The production of osteoblasts

**Hypermineralization:** The state of being, or process of becoming hypermineralized.

**Decorin:** Any of a family of proteoglycans that "decorate" collagen fibres

**Glycosaminoglycan:** Any of a group of polysaccharides with high molecular weight that contain amino sugars and often form complexes with proteins

**Ubiquitously:** Being or seeming to be everywhere at the same time; omnipresent.

**Orthopedic:** relating to the prevention or cure of deformities of children or in general of the human body at any age

## ACMG Guidelines



UNIVERSITY of the  
WESTERN CAPE

### **PM2**

**Absent from controls (or at extremely low frequency if recessive) in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium. (Pathogenic, Moderate)**

We first established the gene's mode of inheritance.

The rule will trigger if the allele frequency is not found in GnomAD, with valid coverage, or:

- For dominant genes (including X-Linked and AD/AR) we check that the allele count is less than 5.
- For recessive genes (AR): the rule will trigger if the homozygous allele count is less than
  - 2. Alternatively we use the ACMG standard rule which fires if the allele frequency is less than 0.0001 (see ACMG Guidelines), however this is a more conservative threshold and our tests show it results in too many false negatives.

**In addition, the strength of rule PM2 is adjusted using the conservation score from PhyloP100Way:**

'supporting' if the variant does not alter the protein length and the position is not conserved (PhyloP < 1.4),

'strong' if the position is strongly conserved (PhyloP > 7.2)

### **PM5**

**Novel missense change at an amino acid residue where a different missense change determined to be pathogenic has been seen before. (Pathogenic, Moderate)**

This rule is a weaker version of PS1, it similarly only applies to missense variants, but considers all possible amino acid missense variants in the same codon. The rule will trigger if any pathogenic variants are identified in the clinical variants database. We then check whether they are independently confirmed pathogenic using the ACMG rules, and if not will reduce the rule strength to accordingly.

### **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. (Pathogenic, Moderate)**

This rule leverages the clinical variants database to evaluate how many missense/in-frame pathogenic variants are found in the region of the variant being classified:

Hot-Spot: using a region of 25 base-pairs on either side of the variant, the rule checks that there are at least 5 pathogenic variants (only using missense and inframe-indel variants), then weighs them by distance to compute a “proximity score”. The rule triggers with strength supporting, moderate or strong depending on the proximity and density of pathogenic and benign variants located within the hot-spot.

Protein Domains: if the variant is within a functional domain reported by UniProt, the rule tallies all the clinically reported missense/in-frame variants within the domain. It checks that the domain contains at least 2 pathogenic variants, and then triggers with strength supporting or strong based on the number of pathogenic, uncertain & benign variants reported within the domain.



The thresholds used by rule PM1 have been established through a careful calibration process and may change over time as further clinical evidence becomes available, or we refine the methodology.

**PP2**

Missense variant in a gene that has a low rate of benign missense variation and in which missense variants are a common mechanism of disease. (Pathogenic, Supporting)


**PP3**

Multiple lines of computational evidence support a deleterious effect on the gene or gene product (conservation, evolutionary, splicing impact, etc.) (Pathogenic, Supporting)



## Appendix B: Ethics Approval, information sheets and consent forms

### Ethics approval form



**OFFICE OF THE DIRECTOR: RESEARCH  
RESEARCH AND INNOVATION DIVISION**

Private Bag X17, Bellville 7535  
South Africa  
T: +27 21 959 4111/2948  
F: +27 21 959 3170  
E: [research-ethics@uwc.ac.za](mailto:research-ethics@uwc.ac.za)  
[www.uwc.ac.za](http://www.uwc.ac.za)

18 June 2018

Ms S Fernol, Prof A Christoffels and Prof M Chetty  
SANBI  
Faculty of Natural Science

**Ethics Reference Number:** BM18/4/11

**Project Title:** An investigation into the genetic basis of autosomal recessive osteogenesis imperfect III in a South African Family of mixed ancestry.


**Approval Period:** 14 June 2018 – 14 June 2019

I hereby certify that the Biomedical Science Research Ethics Committee of the University of the Western Cape approved the scientific methodology and ethics of the above mentioned research project.

Any amendments, extension or other modifications to the protocol must be submitted to the Ethics Committee for approval.

**Please remember to submit a progress report in good time for annual renewal.**

The Committee must be informed of any serious adverse event and/or termination of the study.



*Ms Patricia Josias  
Research Ethics Committee Officer  
University of the Western Cape*

**PROVISIONAL REC NUMBER -130416-050**

FROM HOPE TO ACTION THROUGH KNOWLEDGE

Information Sheets

**PARTICIPANTS DETAILS:**

Name:

\_\_\_\_\_

DOB: \_\_\_\_\_

Sex:

\_\_\_\_\_

Linguistic group: \_\_\_\_\_

Linguistic subgroup:

\_\_\_\_\_

Date: \_\_\_\_\_

Where seen: \_\_\_\_\_

File No:

\_\_\_\_\_

Patient address:

\_\_\_\_\_

\_\_\_\_\_

Contact details:

\_\_\_\_\_

Referred by: \_\_\_\_\_

Tel. No.

\_\_\_\_\_

Specific diagnosis:

\_\_\_\_\_

\_\_\_\_\_

**Brief clinical history:**

1. Pregnancy: \_\_\_\_\_

2. Birth: \_\_\_\_\_

3. Fractures: \_\_\_\_\_

4. Operations: \_\_\_\_\_



**Brief family history:**

5. Affected relatives:

---

6. Parent consanguinity:

---

**CLINICAL FEATURES:**

General condition:

---

Height: \_\_\_\_\_ Weight: \_\_\_\_\_ Head circumference:

---

Limb deformity:

---

Spinal malalignment:

---

Webbing:

---

Color of sclera::

1 \_\_\_\_\_ 2 \_\_\_\_\_ 3 \_\_\_\_\_ 4 \_\_\_\_\_ 5 \_\_\_\_\_

Other manifestations:

---

---



# RESEARCH PARTICIPANT INFORMATION SHEET:

**Study Title:** An investigation into the genetic basis of autosomal recessive Osteogenesis Imperfecta (OI) III in a South African family of mixed ancestry.

**Principal investigators name:** Prof Alan Christoffels

**Contact details of principal investigator email:** [alan@sanbi.ac.za](mailto:alan@sanbi.ac.za)

**Cell No:** 072 561 6518

**Co-supervisor:** Prof Manogari Chetty

**Thesis student Researcher:** Susan Alicia Fernol

You are invited to take part in a clinical research study.

Before you decide whether or not you wish to take part, you should read the information provided below. Take time to ask questions and do not feel under pressure to make a quick decision.

You should clearly understand the risks and benefits of taking part in this study. This process is known as informed consent.

You don't have to take part in this study. If you decide not to take part it won't affect your future medical or dental care.

You can change your mind about participating in the study at any time. Even if the study has begun, you can still decide to withdraw from the study. You don't have to give a reason. If you do withdraw, rest assured it won't affect the quality of treatment you get in the future.

---

## Why is the study being done?

This research study is being undertaken to document the molecular findings in an individual with autosomal recessive Osteogenesis Imperfecta type III in SA. This information will contribute to the understanding of the pathogenesis of the disorder and to facilitate the formulation of appropriate protocols for the management.

---

**Who is organizing and funding this study?**

This research project is being conducted by Susan Alicia Fernol together with Prof Christoffels and Prof Chetty as supervisor and co-supervisor respectively. This research study is for obtaining an MSc project in Bioinformatics from the University of the Western Cape, which will be funded by the DST/NRF South African research chair initiative (sarchi) grant.

---

**Why am I being asked to take part?**

You are being asked to take part because you may be a carrier of the recessive mutation/ have a confirmed diagnosis of Osteogenesis Imperfecta type III.

---

**How will the study be carried out?**

This study has a predominant molecular component in which saliva from the affected person and her immediate family will be collected using the Oragene saliva collection kit.

---

**What are the benefits?**

The findings of this study will contribute to the understanding of autosomal recessive Osteogenesis Imperfecta III in SA. Results will be made available to you and genetic counselling will be provided if necessary.

---

**What are the risks?**

There are no, if any risk at all to you as a research participant. This study necessitates no invasive procedures. There is no risk of physical, psychological, social or economic harm to the participant or his/her family during this study.

---

**Will it cost me anything to take part?**

There are no costs at all to you as a research participant. All expenses e.g., travel will be taken care of from research funds.

---

**Is the study confidential?**

- Written informed consent will be obtained from all participants on standardized forms which will be available in English, Afrikaans, Xhosa and if necessary any other indigenous language.
- All information will be stored in password protected computers. Written information will be stored in a locked office.
- All personal identifiers will be changed when the data are published.

- Photographs will only be used with the eyes hidden and with informed consent.

**If you have any further questions or need any further information now or at any time, please contact:** Dr Manogari Chetty

---



## RESEARCH PARTICIPANT INFORMATION SHEET: PARENT/GUARDIAN

**Study Title:** An investigation into the genetic basis of autosomal recessive Osteogenesis Imperfecta (OI) III in a South African family of mixed ancestry.

**Principal investigators name:** Prof Alan Christoffels

**Contact details of principal investigator email:** [alan@sanbi.ac.za](mailto:alan@sanbi.ac.za)

**Cell No:** 072 561 6518

**Co-supervisor:** Prof Manogari Chetty

**Thesis student Researcher:** Susan Alicia Fernol

Your child is invited to take part in a clinical research study.

Before you decide whether or not you wish for your child to take part, you should read the information provided below. Take time to ask questions and don't feel under pressure to make a quick decision.

You should clearly understand the risks and benefits of your child participating in this study. This process is known as informed consent.

Your child does not have to take part in this study. If you decide that you do not wish for him/her to take part it won't affect your child's medical or dental care.

You can change your mind about your child participating in the study at any time. Even if the study has begun, you can still decide to withdraw your child from the study. You don't have to give a reason. If you do withdraw your child from the study, rest assured it won't affect the quality of treatment your child will receive in the future.

---

### **Why is the study being done?**

This research study is being undertaken to document the molecular findings in an individual with autosomal recessive Osteogenesis Imperfecta type III in SA. This information will



contribute to the understanding of the pathogenesis of the disorder and to facilitate the formulation of appropriate protocols for the management.

---

**Who is organizing and funding this study?**

This research project is being conducted by Susan Alicia Fernol together with Prof Christoffels and Prof Chetty as supervisor and co-supervisor respectively. This research study is for obtaining a MSc project in Bioinformatics from the University of the Western Cape, which will be funded by the DST/NRF South African research chair initiative (sarchi) grant.

---

**Why is my child being asked to take part?**

Your child is being asked to take part because he/she may be a carrier of a recessive mutation/ has a confirmed diagnosis of Osteogenesis Imperfecta type III.

---

**How will the study be carried out?**

This study has a predominant molecular component in which saliva from the affected person and her immediate family will be collected using the Oragene saliva collection kit.

---

**What are the benefits?**

The findings of this study will contribute to the understanding of autosomal recessive Osteogenesis Imperfecta III in SA. Results will be made available to you and genetic counselling will be provided, if necessary. Any dental treatment that your child may require will be provided at no cost to you.

---

**What are the risks?**

There are no, if any risk at all to your child. This study necessitates no invasive procedures.

---

**Will it cost me anything to take part?**

There are no costs at all to the research participant. All expenses e.g., travel will be taken care of from research funds.

---

**Is the study confidential?**

- Written informed consent will be obtained from all participants on standardized forms which will be available in English, Afrikaans, Xhosa and if necessary any other indigenous language. In the case of minors, consent will be obtained from their parents and where possible, assent will be obtained from children.
- 

- All information will be stored in password protected computers. Written information will be stored in a locked office.
- All personal identifiers will be changed when the data are published.
- Photographs will only be used with the eyes hidden and with informed consent.

**If you have any further questions or need any further information now or at any time, please contact: Dr. Manogari Chetty**



## RESEARCH PARTICIPANT INFORMATION SHEET:

**Study Title:** An investigation into the genetic basis of autosomal recessive Osteogenesis Imperfecta (OI) III in a South African family of mixed ancestry.

**Principal investigators name:** Prof Alan Christoffels

**Contact details of principal investigator email:** [alan@sanbi.ac.za](mailto:alan@sanbi.ac.za)

**Cell No:** 072 561 6518

**Co-supervisor:** Prof Manogari Chetty

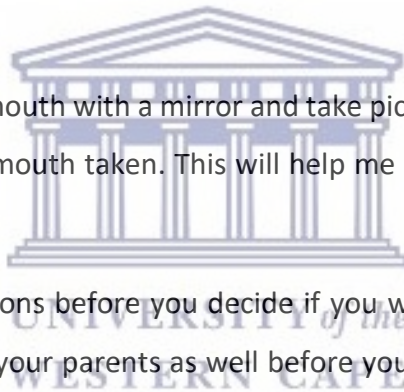
**Thesis student Researcher:** Susan Alicia Fernol You are invited to take part in a research study.

The results of this study will help identify the dental needs and treat other affected children just like you.

If you agree I will look in your mouth with a mirror and take pictures of your teeth. You might need to have an x ray of your mouth taken. This will help me treat you. You will experience no pain or discomfort.

You are welcome to ask questions before you decide if you would like me to examine you. You are welcome to chat with your parents as well before you agree. If you decide that you would not like to take part in this study, you just have to let me know. I will not ask you any questions and you will receive your medical and dental care as always.

If you agree for me to examine you, and you then change your mind, you just have to let me know and I will immediately stop.



Consent forms

## CONSENT FOR PARTICIPATION IN STUDY:

**STUDY TITLE: An investigation into the genetic basis of autosomal recessive Osteogenesis Imperfecta in a South African family of mixed ancestry.**

1. I (parent/guardian/participant) have read and understood the information sheet about this research project. The information has been fully explained to me and I have been able to ask questions, all of which have been answered to my satisfaction.

Yes..... No.....

2. I understand that (my children/ I) do not have to take part in this study and that I can have (him/her/myself) withdrawn from the study at any time. I understand that I don't have to give a reason for withdrawing (my children/myself) from the study and I understand that withdrawing (my children/myself) won't affect their future medical and dental care.

Yes..... No.....

3. I am aware of the potential risks of this research study to (my children/me).

Yes..... No.....

4. I give permission for researchers to look at (my children's/my) medical records to get information. I have been assured that the information about (my child/me) will be kept confidential.

Yes..... No.....

5. I (parent/guardian/participant) have been given a copy of the information sheet and this completed consent form.

Yes..... **No.....**

Patient Name (minor): .....

Parent Name: .....

Parent signature: .....

Child assent (7-17 years): .....

Guardian name: .....

Guardian signature: .....



**To be completed by the principal Investigator**

I, the undersigned, have taken the time to fully explain to the above patient the nature and purpose of this study in a way that they could understand. I have explained the risks involved as well as the possible benefits. I have invited them to ask questions on any aspect of the study that concerned them.

Principal Investigator Name: .....

Qualifications: .....

Signature: .....

Date: .....



## PATIENT CONSENT TO CLINICAL PHOTOGRAPHS AND PUBLICATION: TO WHOM IT MAY CONCERN

I, the undersigned ..... in my capacity  
as

(parent/guardian/participant) consent to photographs being taken of

.....as requested. I understand that these photographs will  
bestored appropriately, treated with the utmost confidentiality and be part of (my  
child's/my) dental records.

I hereby give consent for the images of (my child/myself) to be used ONLY for that I  
have indicated with a tick:

### **Record purposes and for (my child's/my) future management**

The photographic images will form part of the information collected for (my child's/my)  
care and treatment and will be kept confidential at all times.

### **Education and Training purposes**

The photographic images may be used for teaching purposes and viewed by health  
professionals outside of the UWC Faculty of Dentistry. The images may be used in talks,  
conference presentations, posters or on the Internet to help train other health  
professionals in the management of dental and oral diseases

### **Approved research purposes and publications**

This may involve the photographic images being used in medical or dental publications,  
journals, textbooks, conference material, e-publications and on the Internet. Images will  
be seen by health professionals and researchers who use the publications in their  
professional education. The images may be seen by the general public. Images will not  
be used with identifying information such as name, however, full confidentiality is not  
guaranteed.

**Other Purposes (please specify):** .....

- I understand that all efforts will be made to conceal (my child's/my) identity but that  
full confidentiality cannot be guaranteed

- I understand that my consent or refusal will in no way affect (my child's/my) dental care

Patient Name (print).....

Parent/Guardian if patient is under 18 years of age (print name): .....

Parent/Guardian Signature: .....

Date: .....

Child assent (7-17 years): .....

Principal Investigator print name: .....

Principal Investigator signature: .....

Date: .....

Requesting Clinician name (print).....		
Date: .....	Department: .....	Phone: .....
.....		
Patient Name (print): .....		
Views Required:		
Required for (tick):	Records _____	Teaching/ Lectures _____
Publication _____		Research _____
Images taken by: .....		Date: .....



UNIVERSITY of the  
WESTERN CAPE

## CONSENT FOR DNA ANALYSIS AND STORAGE

1. I, \_\_\_\_\_, request that an attempt be made using genetic material to assess that I / my child (delete where not applicable) might have inherited a disease causing mutation in the gene for  
.....

2. I understand that the genetic material for analysis is to be obtained from blood / saliva(delete where not applicable).

3. I request that no portion of the sample be stored for later use..... (tick if applicable)

I request that a portion of the sample be stored indefinitely, for 5 years, for 1 year, until the study is completed (delete where not applicable) for

1. Possible re-analysis
2. Analysis for the benefit of members of my immediate family
3. Research purposes, subject to the approval of the Research Ethics Committee, provided that all information will remain confidential
4. The result of the analysis will be made known to me, via my doctor(s), in accordance with the relevant protocol, if and when available.
5. If clinically relevant, I authorise that the results may be made known to family members.
6. I have been informed that:

- 
- the analysis procedure is specific to the genetic condition and cannot determine the complete genetic make-up of an individual.
  - the genetics laboratory is under an obligation to respect medical confidentiality
  - genetic analysis may not be informative for some families or family members
  - even under the best conditions, current technology of this type is not perfect and could lead to incorrect results
  - where biological material is used for research purposes, there may be no direct benefit to me



- 7. I understand that I may withdraw my consent for any aspect of the above at any time without this affecting my future medical care.
- 8. All of the above has been explained to me in a language that I understand and my questions answered by:

Doctor/Consultant: ..... Date:  
.....

Patient/Parent: ..... Date:  
.....

Laboratory Finding

---

---



## Appendix C: Variant calling annotation

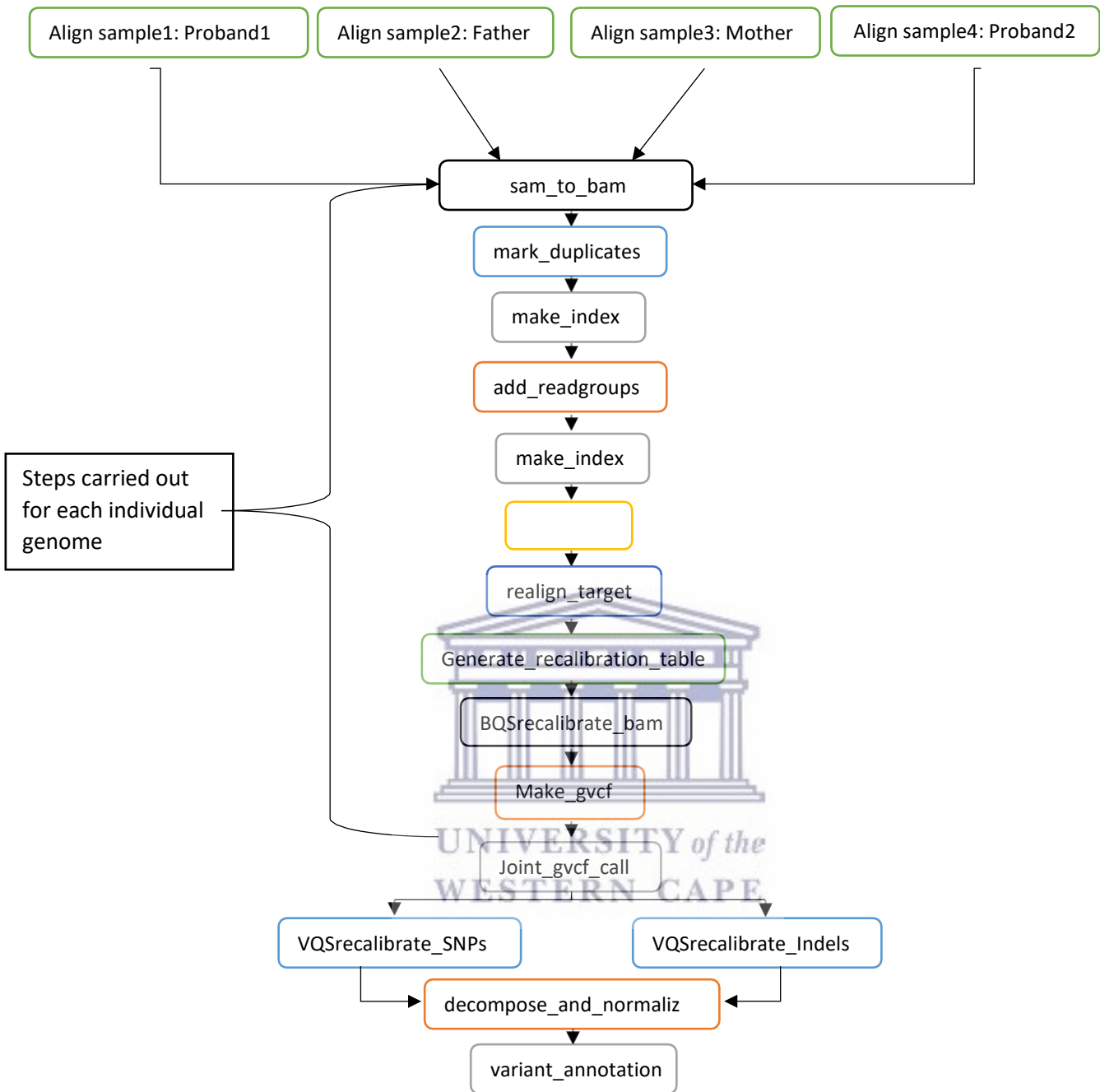


Figure C.1: Flowchart indicating each step taken from after the alignment to variant annotation.

## Appendix D: BGI quality control results

In the quality control step performed by BGI sequencers, the raw reads were put through quality assessment (**Figures D.1-D.2**). This assessment step included checking and removing adapter sequences, contamination, and low-quality reads from raw reads. Adapters are ligated to the 5' and 3' ends of each single DNA molecule during sequencing. These adapter sequences hold barcoding sequences, forward/reverse primers, and the binding sequences to immobilize the fragments to the flow cell and allow bridge amplification. Since adapter sequences are synthetic and are not seen in any genomic sequence, adapter contamination often leads to NGS alignment errors and an increased number of unaligned reads. Hence any adapter sequences need to be removed before mapping. In addition to adapter removal, trimming is performed to discard any low-quality reads, which generally occur at the 5' and 3' ends.

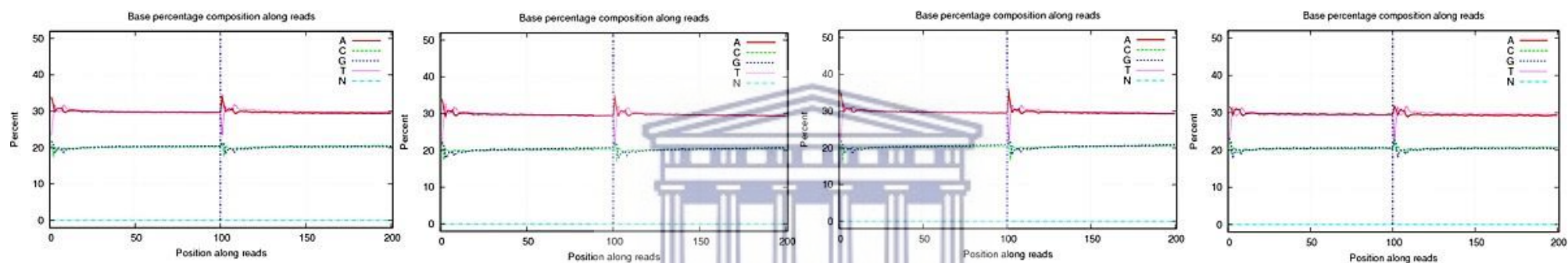


Figure D. 1: The distribution of qualities along reads in data filtering are shown for samples 1-4, from left to right

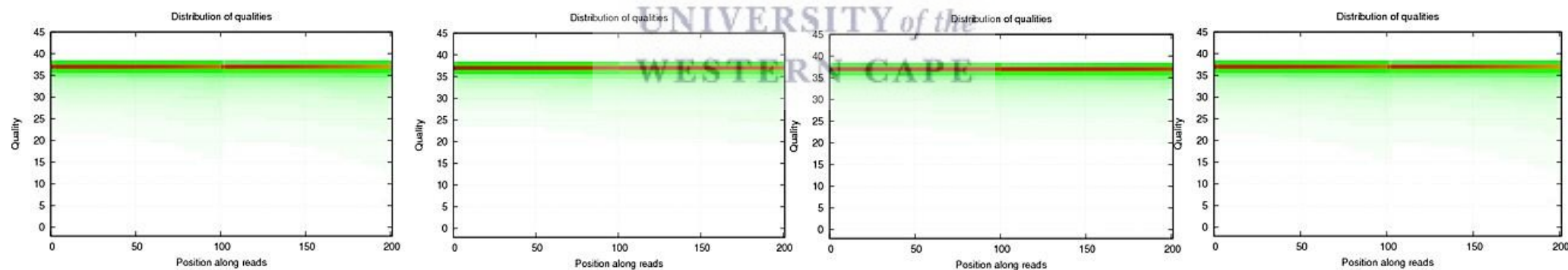


Figure D. 2: The distribution of base percentage along reads in data filtering are shown for samples 1 to sample 4, from left to right

## Appendix E: Local quality control

### 1. Per base sequence quality

The first module presents the per base sequence quality. This view shows an overview of the range of quality values across all bases at each position in the FASTQ file. For Each position a BoxWisker type plot is drawn (**Figure E.1**). The elements of the plot are as follows; the central red line, which in the below figures are shown at the top of the yellow box, is the median value. The yellow box represents the inter-quartile range (25 -75%). The upper and lower whiskers represent the 10 % and 90% points. The blue line represents the mean quality. The y-axis on the graph shows the quality scores. The higher the score the better the base call. The background of the graph divides the y-axis into very good quality (green), calls of reasonable quality (orange), and calls of poor quality (red). The quality of calls on most platforms will degrade as the run progresses, so it is common to see base calls falling into the orange area towards the end of a read.

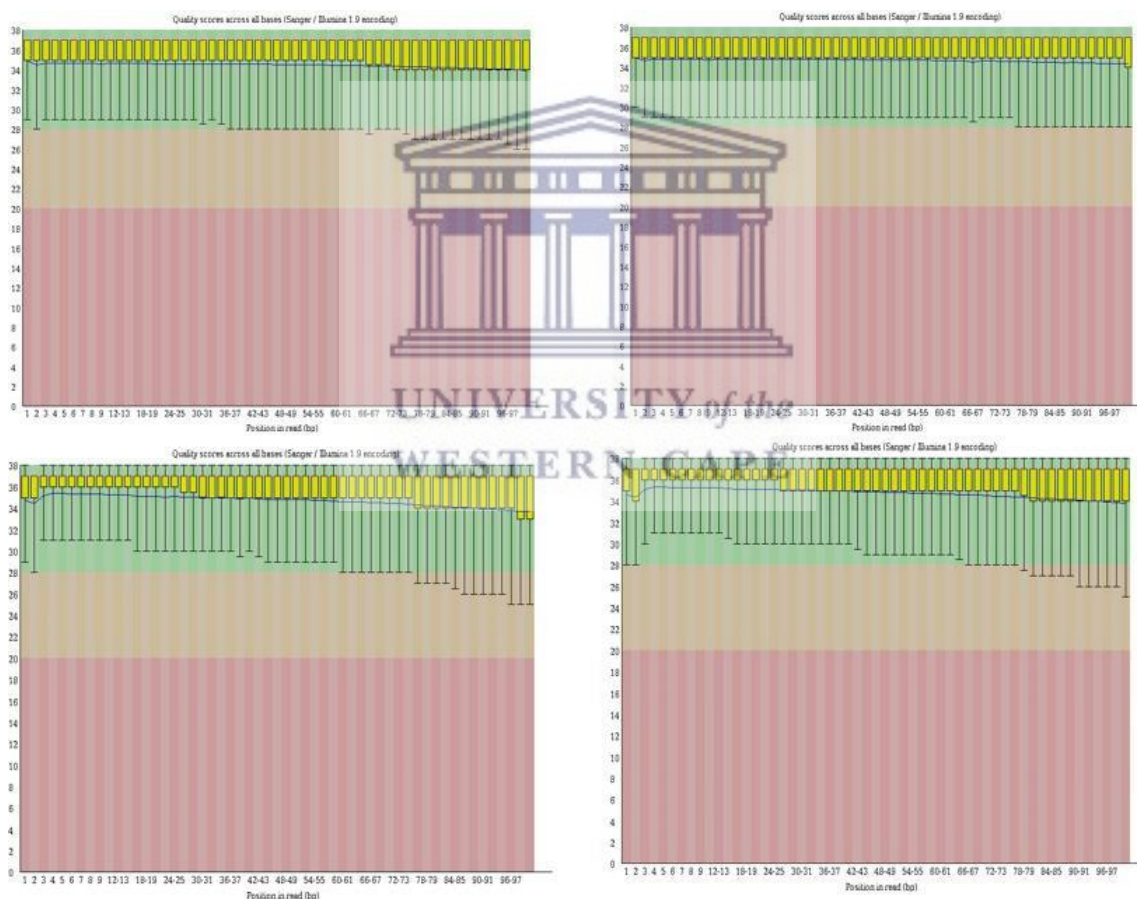
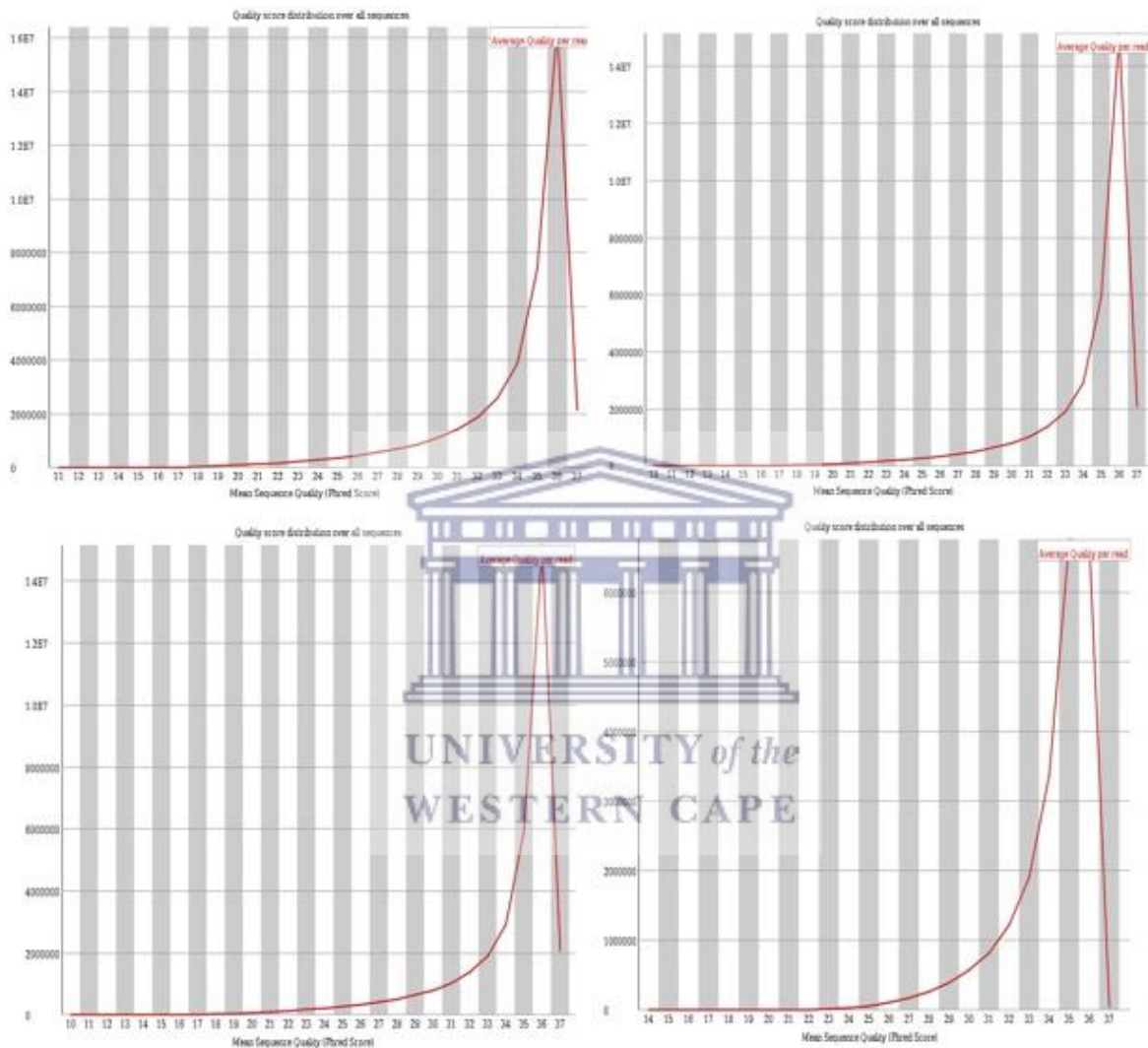


Figure E.1: The quality scores across all bases. From left to right samples 1- 4.

## 2. Per sequence quality score

The second module produced by the FastQC tool is the quality per sequence plot (**Figure E.2**). The per sequence quality score report allows you to see if a subset of the sequences have universally low-quality values. For each sequence, FastQC computed the mean quality (Phred) score across all bases of the sequence library, plotting out the distribution of the mean. In **Figure E.2**, the sequence displaces a very tight distribution with universally high quality.



**Figure E. 2:** Distribution of the quality of scores across all sequences in the library.

### 3. Per Base sequence content

The per base sequence content plots out the proportion of each base position in a file for which each of the four normal DNA bases has been called. These bases should be equally distributed in a diverse library. **Figure E.3** displays an even distribution of the four bases which are not subject to change at any position. The parallel lines across the plot display the distribution of the bases implying that the position looked at would not affect any base calls. Although there are spikes of over-representation of the bases in the start of the sequences, this is normal. A very bias position will have spikes reaching 50 to 80.



Figure E.3: The sequence base content across all bases.

#### 4. Per sequence G+C content

This module measures the GC content across the whole length of each sequence in a file and compares it to a modelled normal distribution of the GC content. In a normal random library you would expect to see roughly normal distribution of GC content where the central peak corresponds to the overall GC content of the underlying genome. A consistent GC content across all bases of the reads were observed (**Figure E.4**).

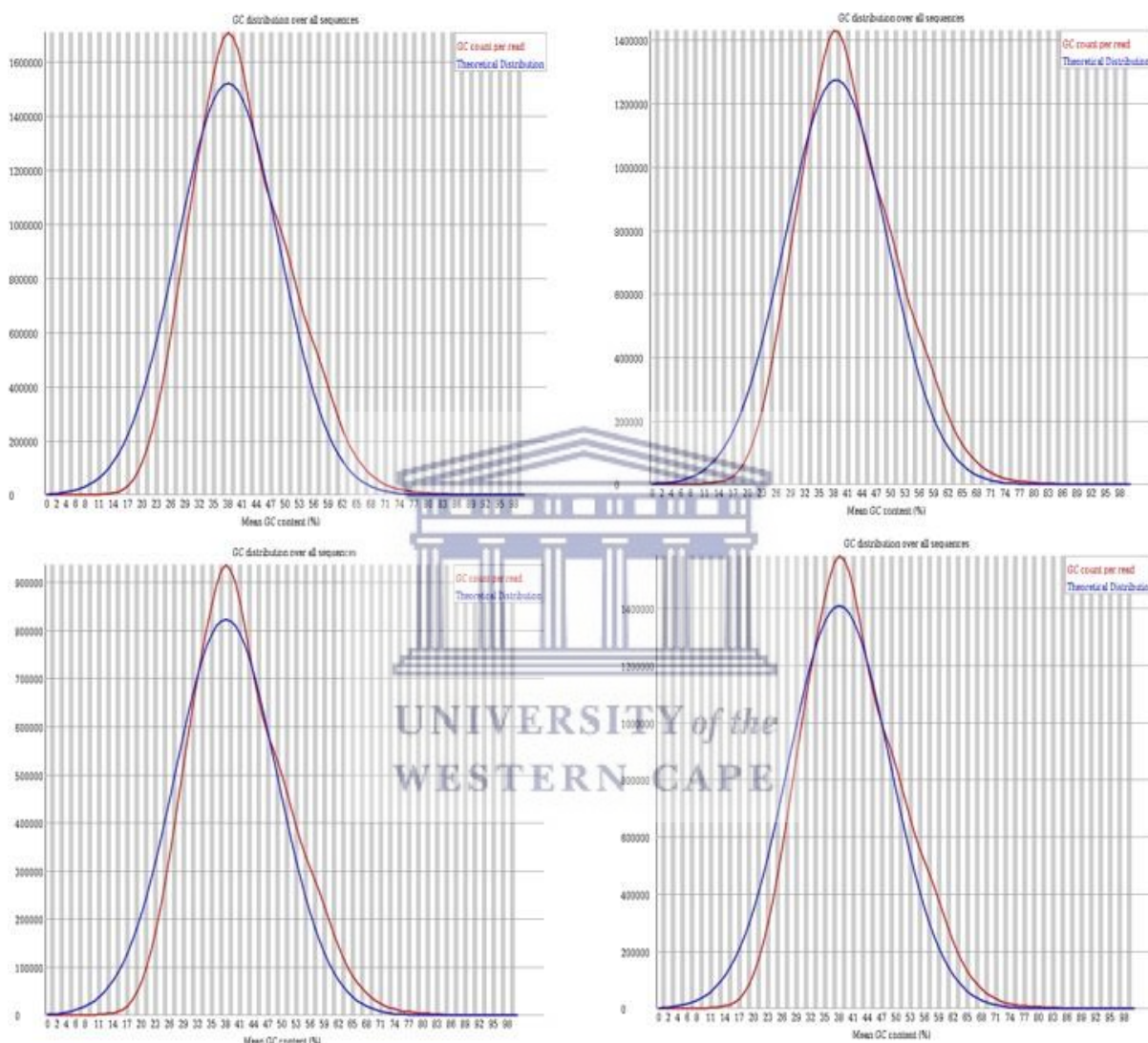


Figure E.4: G+C content for all bases per sample

## 5. Per base N content

If a sequencer is unable to make a base call with sufficient confidence, then it will normally substitute an N rather than a conventional base. This module plots out the percentage of base calls at each position for which an N was called. There were base N substitutions observed (Figure E.5)

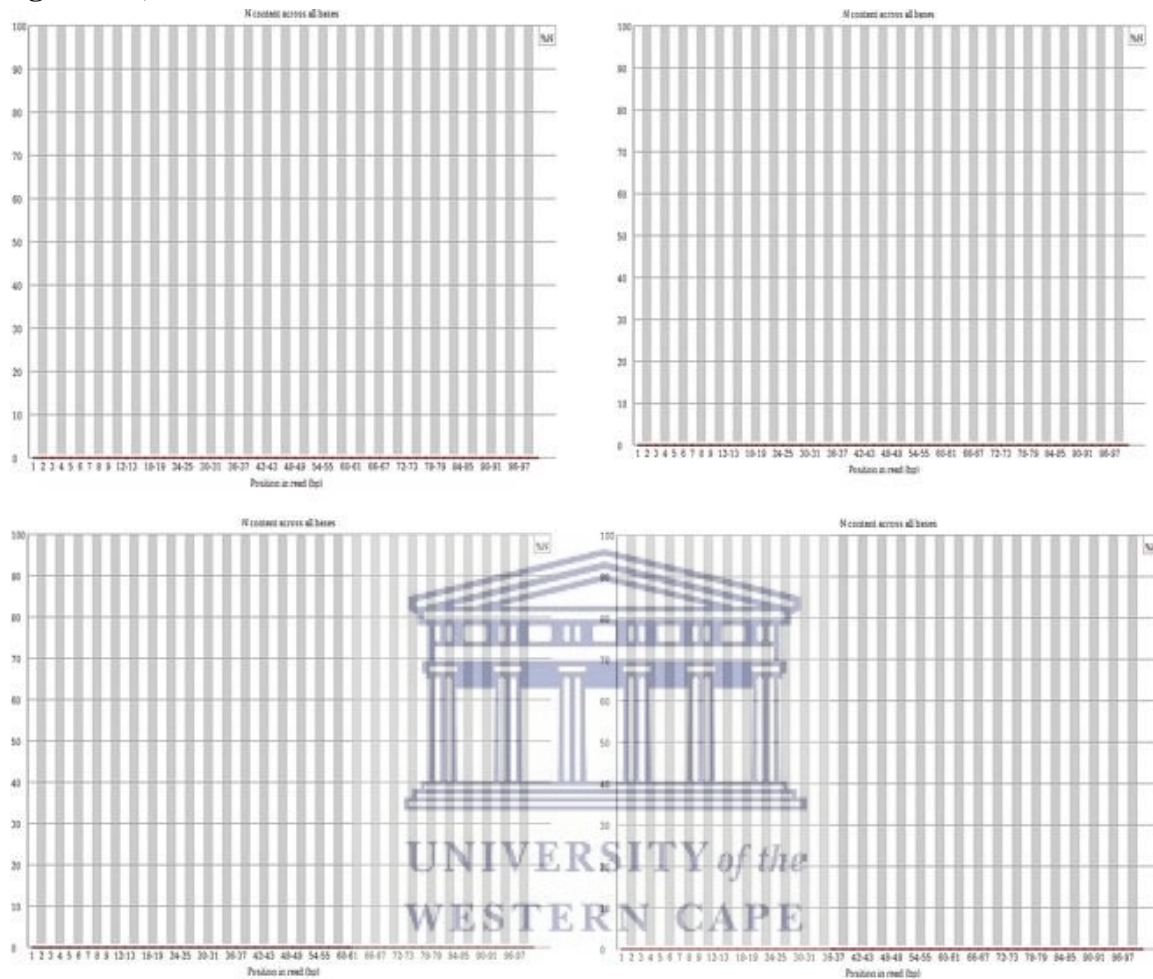


Figure E.6: Per base N content per sample.



## 6. Sequence length distribution

Some high throughput sequencers generate sequence fragments of uniform length, but others can contain reads of widely varying lengths. Even within uniform length libraries, some pipelines will trim sequences to remove poor quality base calls from the end. This module generates a graph (**Figure E.6**) showing the distribution of fragment sizes in the file which was analyzed. In many cases, as seen in this analysis, this will produce a simple graph showing a peak only at one size.

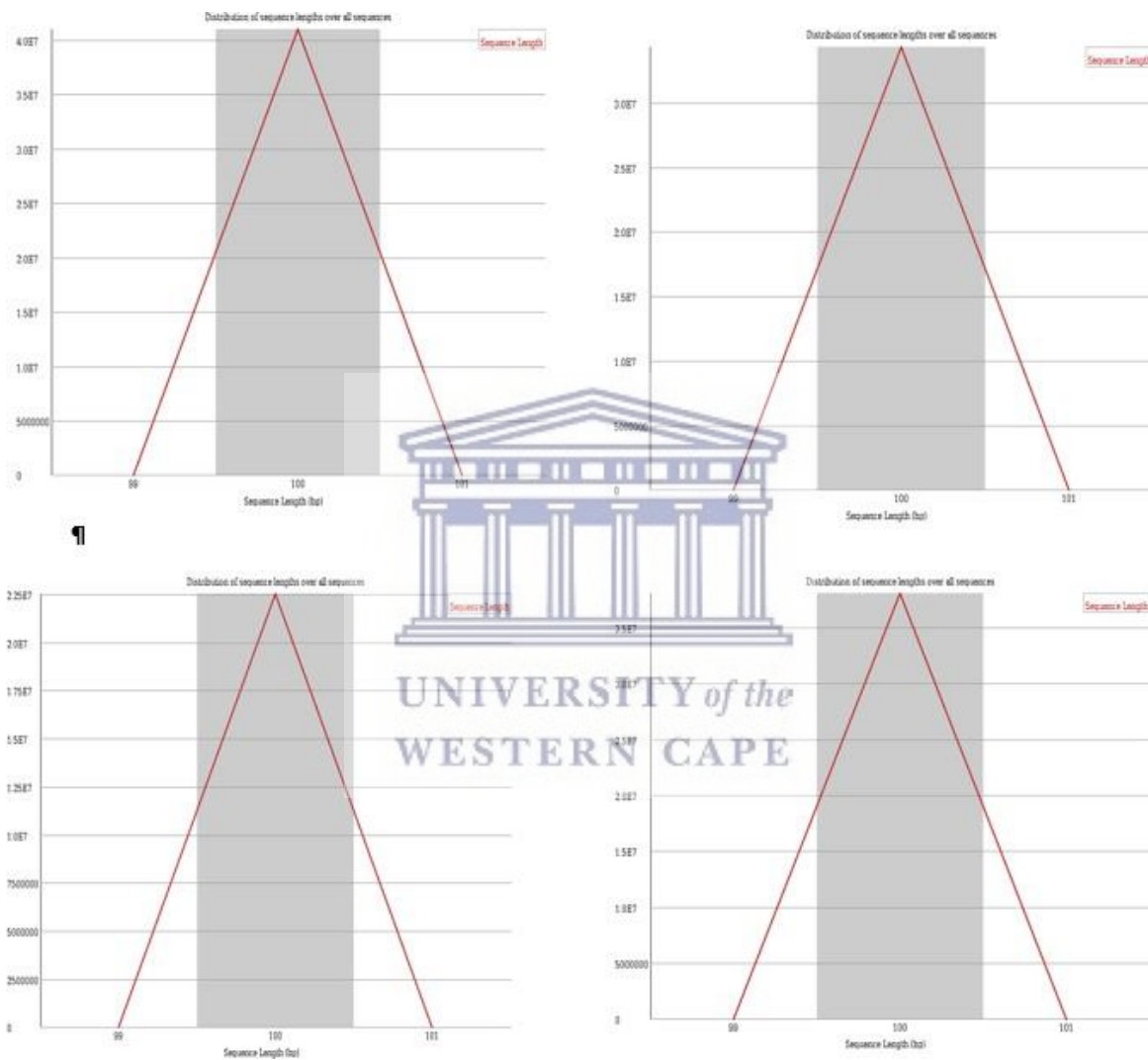


Figure E.7: Sequence length distribution.

## 7. Sequence duplication levels

In a diverse library most sequences will occur only once in the final set. A low level of duplication may indicate a very high level of coverage of the target sequence, but a high level of duplication is more likely to indicate some kind of enrichment bias (e.g., PCR over amplification.) This module counts the degree of duplication for every sequence in the set and creates a plot showing the relative number of sequences with different degrees of duplication. In **Figure E.7**, the plot rapidly drops to a sequence duplication level of two and then continues at zero throughout. For each of the samples analyzed, a certain percentage of the sequence remains if duplications are present. Sample 1 – 94.96 %, Sample 2 – 95.47%, Sample 3 – 96.53%, and Sample 4 – 95.99%.

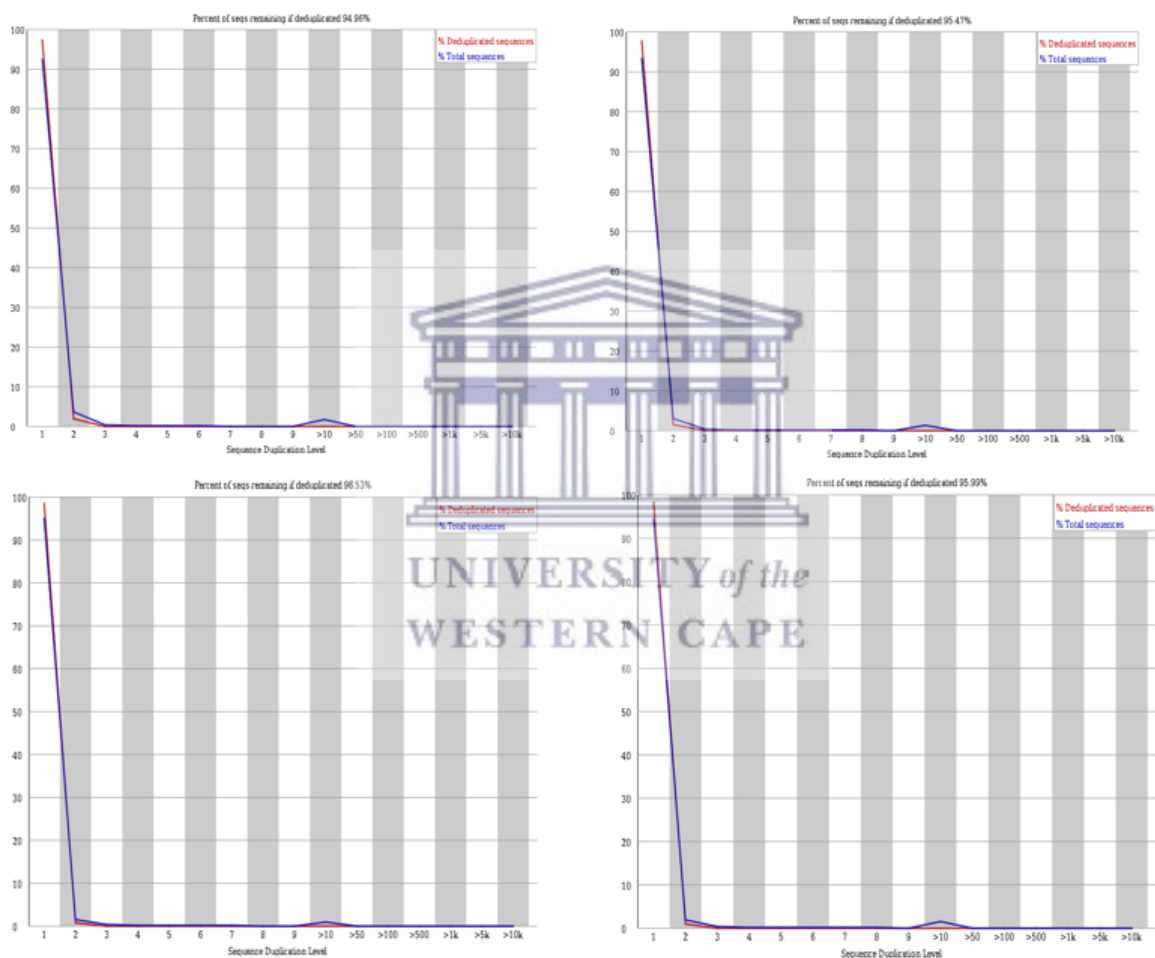


Figure E.8: Sequence duplication level.

## 8. Overrepresented sequences

A normal high-throughput library will contain a diverse set of sequences, with no individual sequence making up a tiny fraction of the whole. Finding that single sequence is very overrepresented in the set, it either means that it is highly biologically significant, or indicates that the library is contaminated, or not as diverse as one would expect. This module lists all the sequences which make up more than 0.1% of the total. For this module, no overrepresented sequences were found.

## 9. Adapter content

There were no samples found with any adapter contamination greater than 0.1% (**Figure E.8**).

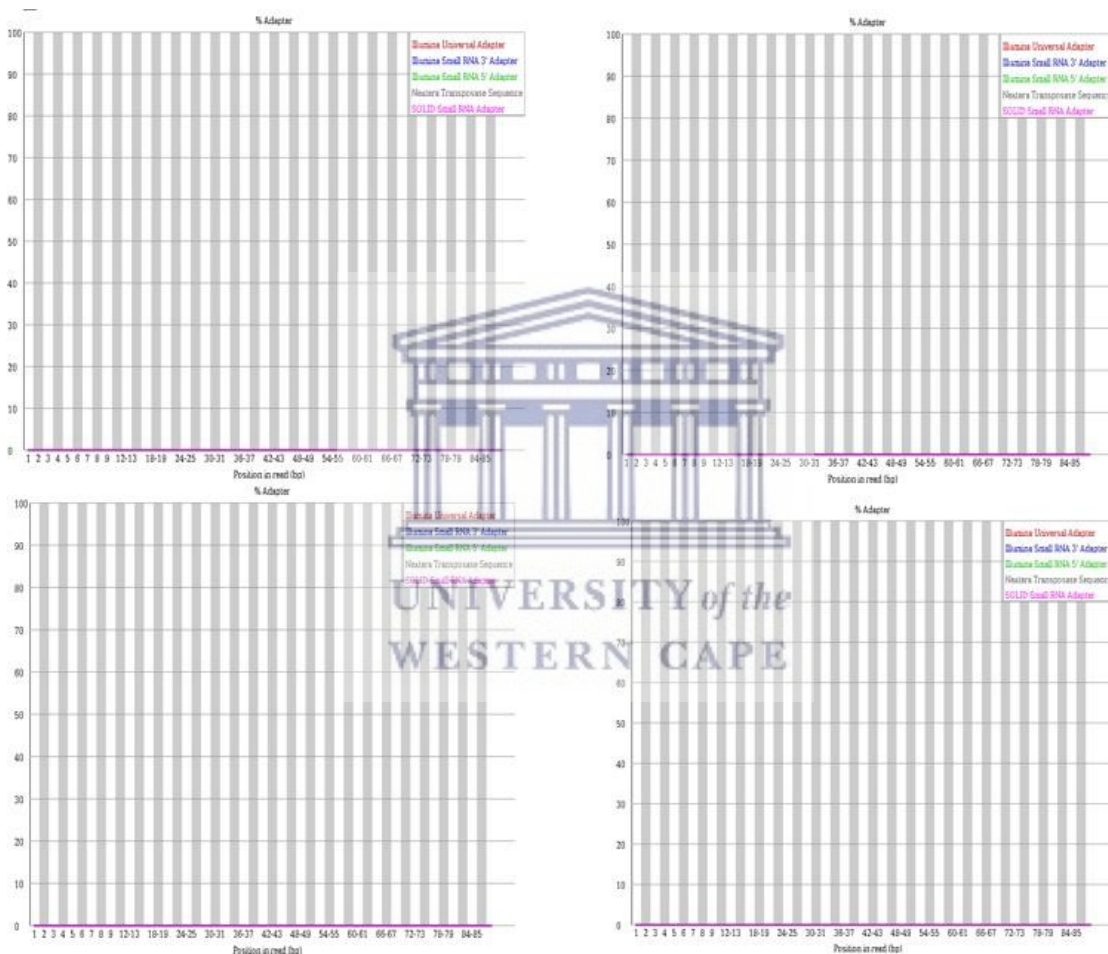


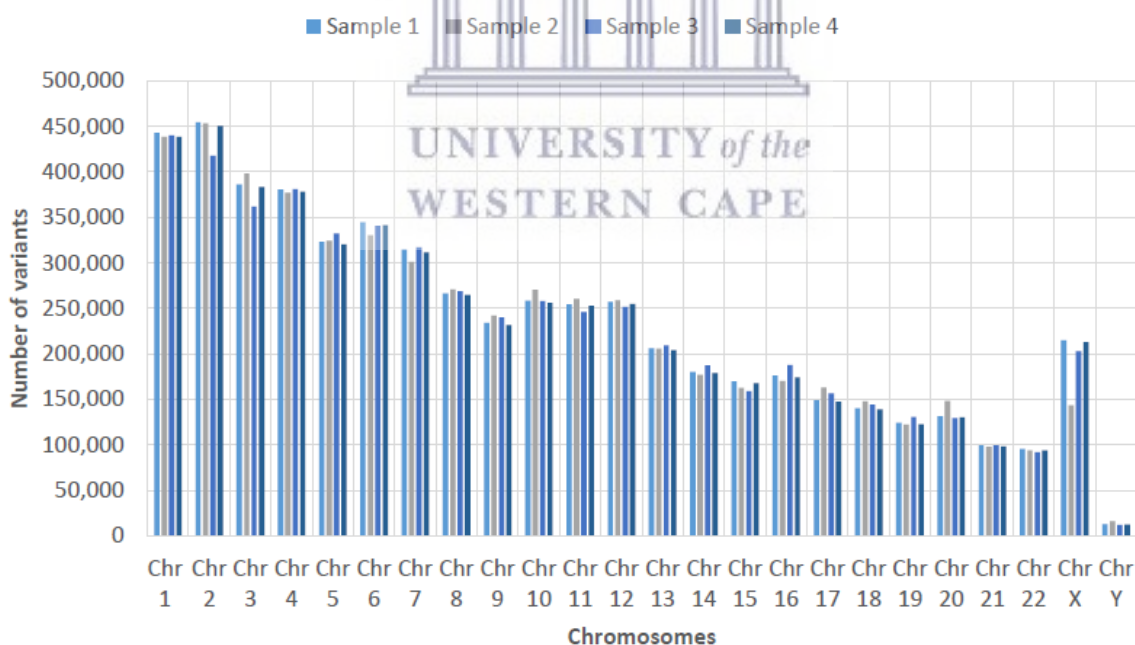
Figure E.9: Percentage of adapter sequence contamination.

## Appendix F: Variant Results

**Table F.1** presents the total number of variants across all coding sequences. The majority of the variants were synonymous variants making up 64 120 (51%) of all the coding variants. Missense variants make up the second largest (46%) part of the coding variants.

**Table F.1:** The number of variants for all coding sequences.

[Coding consequences]	Number of variants
stop_gained	605
frameshift_variant	1450
stop_lost	136
start_lost	139
inframe_insertion	832
inframe_deletion	1196
protein_altering_variant	13
missense_variant	57358
incomplete_terminal_codon_variant	12
start_retained_variant	6
synonymous_variant	64120
stop_retained_variant	91
coding_sequence_variant	107



**Figure F.1:** Number of variants found across all samples in all chromosomes.

**Table F.2:** Diseases and genes that share similar phenotype to that of OI.

<b>Gene</b>	<b>Phenotype</b>	<b>Disorder</b>
<i>ADAMTS17</i>	short stature	Weill–Marchesani syndrome homo
<i>ADAMTS2</i>	short stature, osteopenia, short, long bones	Geleophysic dysplasia
<i>AKT1</i>	kyphosis, scoliosis, short stature, kyphoscoliosis, skeletal dysplasia, abnormality of dental enamel, abnormal form of vertebral body	Proteus syndrome
<i>ANO5</i>	recurrent fractures, increase susceptibility to fractures, osteopenia, bowing of long bones, scoliosis	Gnathodiaphyseal dysplasia
<i>ARID1B</i>	short stature, kyphosis, scoliosis, abnormalities of dentition	Coffin–Siris syndrome
<i>ARSB</i>	Kyphosis	Mucopolysaccharidosis type 6
<i>B4GALT7</i>	short stature, osteopenia, blue sclerae, kyphoscoliosis, skeletal dysplasia, bowing of long bones, scoliosis, abnormality of primary teeth	Ehlers–Danlos syndrome, type spondylodysplastic
<i>CDH3</i>	short stature, abnormality of limb bone morphology, skeletal dysplasias	Ectrodactyly-ectodermal, dystrophy syndrome (EEM) dysplasia-macular
<i>CLCN7</i>	recurrent fractures, short stature, abnormality of the dentition, osteopetrosis, fractures of the long bones, bowing of the long bones, abnormality of dental morphology	Osteopetrosis, severe neonatal or infantile forms
		Osteopetrosis, intermediate form
		Osteopetrosis, late-onset form type 2 (OPTA2)
<i>COL11A1</i>	short stature, rhizomelia, femoral, ulna and radial bowing, abnormality of the dentition, abnormal form of the vertebral bodies	Stickler syndrome type 2
		Fibrochondrogenesis
<i>COL27A1</i>	short stature, scoliosis	Steel syndrome
<i>COL9A1</i>	short stature	Multiple epiphyseal dysplasia (MED)
		Stickler syndrome, recessive type
<i>CREBBP</i>	short stature, scoliosis	Rubinstein–Taybi syndrome
<i>CSGALNACT1</i>	short stature, scoliosis	CSGALNACT1 deficiency (joint dislocations and mild skeletal dysplasia)
<i>DNMT3A</i>	kyphoscoliosis, severe short stature, scoliosis	Tatton–Brown–Rahman syndrome
<i>DYNC2H1</i>	short stature, skeletal dysplasias, femoral bowing, scoliosis	Short rib–polydactyly syndrome (SRPS) type 1/3 (Saldino–Noonan/Verma–Naumoff)
		Asphyxiating thoracic dysplasia (ATD; Jeune)
<i>EIF2AK3</i>	short stature, kyphosis, recurrent fractures, osteoporosis, osteopenia, triangular face	SED with diabetes mellitus, Wolcott–Rallison type
<i>ENPP1</i>	short stature, tibial bowing, skeletal dysplasias, distal femoral bowing	Hypophosphatemic rickets, autosomal recessive, type 2 (ARHR2)
<i>EP300</i>	short stature	Rubinstein–Taybi syndrome
<i>EXT1</i>	short stature, abnormality of femur morphology, radial bowing abnormality of tibia morphology, abnormality of upper limbs, abnormality of dentition, scoliosis	Trichorhinophalangeal dysplasia type 2 (Langer–Giedion)
		Multiple cartilaginous exostoses (osteochondromas)

(continued)

<b>Gene</b>	<b>Phenotype</b>	<b>Disorder</b>
<i>EXTL3</i>	disproportionate short stature, abnormality of limb bone morphology, kyphosis, kyphoscoliosis, abnormality of the cervical spine,	SEMD with immune deficiency, EXTL3 type
<i>FAM20C</i>	short stature, bowing of long bones	Raine dysplasia (lethal and nonlethal forms)
<i>FGFR1</i>	short stature, scoliosis, skeletal dysplasias, osteopenia, abnormality of the dentition, abnormality of limbs, rhizomelia, abnormal form of the vertebral bodies, bowing of long bones, abnormality of dental morphology, recurrent fractures, osteoporosis	Osteoglophonic dysplasia
		Pfeiffer syndrome
		Hartsfield syndrome
<i>FGFR2</i>	short stature, abnormality of the cervical spine, femoral bowing, bowing of legs, abnormality of fibula morphology, abnormality of the lower limbs, recurrent fractures, osteopenia, ulnar bowing, scoliosis	Lacrimo-auriculo-dento-digital syndrome (LADD)
		Apert syndrome,
		Pfeiffer syndrome
		Craniosynostosis with cutis gyrate (Beare–Stevenson), Crouzon syndrome
<i>FIG4</i>	short stature	Yunis–Varon dysplasia
<i>FLNB</i>	scoliosis, radial and tibial, humerus bowing, rhizomelia, short stature	Atelosteogenesis type 3 (AO3),
		Larsen syndrome (dominant),
		Spondylocarpotarsal synostosis syndrome
<i>FNI</i>	short stature, recurrent fractures, kyphosis, scoliosis	SMD, Sutcliffe type or corner fractures type
<i>GLB1</i>	kyphosis, scoliosis, osteoporosis, short stature	Mucopolysaccharidosis type 4B,
		GM1 Gangliosidosis, several forms
<i>GLI3</i>	short stature, radial bowing, skeletal dysplasias	Preaxial polydactyly type 4 (PPD4)
		Greig cephalopolysyndactyly syndrome
		Pallister–Hall syndrome
<i>GNAS</i>	osteoporosis, kyphosis, recurrent fractures, osteopenia, bone fracture, scoliosis	Fibrous dysplasia, polyostotic form (McCune–Albright),
		Pseudohypoparathyroidism type IA
<i>GPC6</i>	short stature, rhizomelia,	Omodysplasia, recessive type
<i>HAAO</i>	short stature,	NAD deficiency syndrome
<i>HDAC4</i>	short stature	Brachydactyly–mental retardation syndrome
<i>HPGD</i>	osteoporosis, wormian bones, osteopenia, scoliosis	Hypertrophic osteoarthropathy
<i>HSPG2</i>	bowing of the long bones, short stature, kyphosis, kyphoscoliosis, skeletal dysplasia, wormian bones, osteoporosis, blue sclerae, scoliosis	Dyssegmental dysplasia, )
		Silverman–Handmaker and Rolland–Desbuquois types
		Schwartz–Jampel syndrome (myotonic chondrodystrophy)
<i>IFT43</i>	abnormality of dental enamel, Rhizomelia	SRPS unclassified

(continued)

<i>Genes</i>	<b>Phenotype</b>	<b>Disorder</b>
		Cranioectodermal dysplasia (Levin–Sensenbrenner) type 1, 2
<i>ILIRN</i>	osteopenia	Sterile multifocal osteomyelitis, periostitis, and pustulosis (CINCA/NOMID-like)
<i>LBR</i>	kyphosis, bowing of the long bones, abnormality of the dentition, lethal skeletal dysplasia, rhizomelia, mild short stature	Greenberg dysplasia
<i>LEMD3</i>	short stature, skeletal dysplasia, recurrent fractures, abnormality of the dentition	Osteopoikilosis Melorheostosis with osteopoikilosis
<i>LFNG</i>	short stature, kyphosis, scoliosis, abnormal form of the vertebral bodies	Spondylocostal dysostosis
<i>LRP5</i>	osteopenia, short stature, kyphoscoliosis, wormian bones, increase susceptibility to fractures, osteoporosis, abnormal lower limb bone morphology, abnormal form of the vertebral bodies, blue sclerae, crumpled long bones, abnormality of the vertebral column	Osteosclerosis Osteoporosis—AD form Osteoporosis-pseudoglioma syndrome
<i>LTBP3</i>	short stature, short long bones	Platyspondyly (brachyolmia) with amelogenesis imperfecta Geleophysic dysplasia Acromicric dysplasia
<i>OBSL1</i>	short stature, triangular face, kyphosis, abnormality of dental enamel, scoliosis	3-M syndrome
<i>OSTM1</i>	short stature, abnormality of skeletal system	Osteopetrosis, infantile form, with nervous system involvement (OPTB5)
<i>PCYT1A</i>	tibial bowing, femoral bowing, severe short stature, Rhizomelia, scoliosis	SMD with cone-rod dystrophy
<i>PDE4D</i>	short stature, abnormality of radius, abnormal form of vertebral bodies, scoliosis	Acrodysostosis
<i>PEX7</i>	kyphoscoliosis, rhizomelia	Rhizomelic CDP
<i>PLOD1</i>	kyphosis, thoracic scoliosis, osteoporosis, osteopenia, blue sclerae	Ehlers–Danlos syndrome, kyphoscoliotic type 1
<i>POLE</i>	short stature, osteopenia, scoliosis	IMAGE syndrome (intrauterine growth retardation, metaphyseal dysplasia, adrenal hypoplasia, and genital anomalies)
<i>POLR1D</i>	skeletal dysplasia, abnormality of dental enamel, abnormality of the vertebral column	Mandibulofacial dysostosis (Treacher Collins, Franceschetti–Klein)
<i>POR</i>	short stature, osteoporosis, femoral bowing	Antley–Bixler syndrome
<i>PRKARIA</i>	short stature, osteoporosis, scoliosis, kyphosis, osteopenia	Acrodysostosis
<i>PTDSS1</i>	short stature, kyphosis, abnormality of the dentition, abnormality of dental enamel, scoliosis	Lenz–Majewski hyperostotic dysplasia
<i>PTHLH</i>	short stature	Brachydactyly type E
<i>RBBP8</i>	short stature, scoliosis, abnormality of dental enamel	Microcephalic osteodysplastic primordial dwarfism (other types)
<i>ROR2</i>	short stature, kyphosis, blue sclerae, short long bones	Robinow syndrome, recessive type Brachydactyly type B
<i>RUNX2</i>	short stature, kyphosis, wormian bones, osteoporosis, abnormality of dental enamel, abnormality of dentition, multiple small vertebral fractures, scoliosis	Metaphyseal dysplasia with maxillary hypoplasia Cleidocranial dysplasia

(continued)

<b>Genes</b>	<b>Phenotype</b>	<b>Disorder</b>
<i>SGMS2</i>	femoral bowing, recurrent fractures, osteoporosis, severe short stature, osteopenia, scoliosis	Calvarial doughnut lesions with bone fragility
<i>SKI</i>	osteopenia, abnormal form of the vertebral bodies, scoliosis	Shprintzen–Goldberg syndrome
<i>SLCO2A1</i>	osteoporosis, scoliosis	Hypertrophic osteoarthropathy
<i>SMAD3</i>	scoliosis, osteoporosis	Loeys–Dietz syndrome (types 1–6)
<i>SMARCA4</i>	short stature, scoliosis, kyphosis	Coffin–Siris syndrome
<i>SMARCB1</i>	short stature, scoliosis, kyphosis, abnormality of the dentition	Coffin–Siris syndrome
<i>SRP54</i>	short stature, osteopenia abnormality of the skeletal system	Metaphyseal dysplasia with pancreatic insufficiency and cyclic neutropenia (Shwachman–Bodian–diamond syndrome, SBDS)
<i>SUMF1</i>	short stature	Multiple sulfatase deficiency
<i>TBCE</i>	short stature, abnormality of dental enamel, abnormality of dentition, scoliosis	Sanjad–Sakati syndrome
<i>TBXAS1</i>	abnormality of tibia morphology, abnormal form of the vertebral bodies, bowing of long bones	Hematodiaphyseal dysplasia Ghosal
<i>TGFB1</i>	scoliosis, abnormality of the ulna and radius, kyphosis, skeletal dysplasia, abnormality of the vertebral column	Diaphyseal dysplasia Camurati–Engelmann
<i>TGFBR2</i>	scoliosis, blue sclerae, osteoporosis	Loeys–Dietz syndrome (types 1–6)
<i>TP63</i>	short stature, abnormality of dental enamel, abnormality of dentition	Ankyloblepharon-ectodermal dysplasia-cleft palate (AEC), Ectrodactyly-ectodermal dysplasia cleft-palate syndrome type 3 (EEC3), Limb-mammary syndrome (including ADULT syndrome), Split-hand-foot malformation, isolated form, type 4 (SHFM4)
<i>TTC21B</i>	short stature, skeletal dysplasia, scoliosis	Asphyxiating thoracic dysplasia (ATD; Jeune)
<i>TWIST1</i>	short stature, scoliosis	Saethre–Chotzen syndrome
<i>VDR</i>	recurrent fractures, fibial bowing, femoral, bowing bowing of legs, fibular bowing, abnormality of the dentition, abnormal form of the vertebral bodies, scoliosis	Vitamin D-dependent rickets, type 2A
<i>WDR19</i>	short stature, skeletal dysplasia, osteoporosis, abnormality of dental enamel, rhizomelia	Asphyxiating thoracic dysplasia (ATD; Jeune), Cranioectodermal dysplasia (Levin–Sensenbrenner) type 1, 2
<i>WDR35</i>	short stature, rhizomelia, osteoporosis, abnormality of dental enamel, abnormality of the dentition, lethal skeletal dysplasia, short long bone	Chondroectodermal dysplasia (Ellis-van Creveld), SRPS type 5, SRPS unclassified, Cranioectodermal dysplasia (Levin–Sensenbrenner) type 1, 2
<i>WNT7A</i>	radial bowing, short stature, femoral bowing, bowing of long bones	Al-Awadi Raas-Rothschild limb-pelvis hypoplasia- aplasia, Fuhrmann syndrome
<i>XYLT1</i>	short stature, short long bones, blue sclerae, scoliosis	Desbuquois dysplasia type 2 (Baratela–Scott syndrome)



Table F.3: Remaining homozygous variants and their pathogenicity predictions using PredictSnP2

rsID	Gene	Chrom	Pos	REF	ALT	PredictSNP2		CADD		DANN		FATHMM-MKL		FunSeq2		GWAVA	
						score	conf	score	conf	score	conf	score	conf	score	conf	score	conf
rs7961282	ADCY6	chr12	48767613	G	T	1.00	0.97	11.85	0.69	0.84	0.62	0.91	0.88	0.94	0.68	0.66	0.84
rs76358362	ATG5	chr6	106293746	A	G	1.00	0.97	12.92	0.76	0.88	0.75	0.42	0.58	1.58	0.62	0.48	0.64
rs4718966	AUTS2	chr7	70575572	C	T	1.00	0.97	21.40	0.98	0.78	0.54	0.94	0.91	1.61	0.62	0.56	0.80
rs7108857	CBL	chr11	119206293	C	T	1.00	0.97	17.93	0.96	0.98	1.00	0.86	0.86	1.16	0.75	0.55	0.80
rs3003602	DNM1	chr9	128218785	C	T	1.00	0.97	10.28	0.67	0.90	0.82	0.85	0.82	3.17	0.67	0.46	0.64
rs6663178	EDEM3	chr1	184692508	G	A	1.00	0.97	13.85	0.83	0.92	0.89	0.91	0.88	1.09	0.76	0.58	0.80
rs8757	EIF4E	chr4	98880282	A	G	1.00	0.97	17.54	0.96	0.88	0.78	0.97	0.95	1.50	0.64	0.73	0.86
rs1050688	EIF4E	chr4	98880501	G	A	1.00	0.97	18.49	0.96	0.87	0.69	0.98	0.97	1.50	0.64	0.63	0.80
rs1359591	ENO1	chr1	8879251	G	A	1.00	0.97	9.32	0.67	0.94	0.96	0.42	0.58	3.65	0.67	0.50	0.64
rs6803	ERAL1	chr17	28860771	C	T	1.00	0.97	11.06	0.67	0.77	0.54	0.56	0.62	1.00	0.68	0.54	0.78
rs35694729	FSHR	chr2	48994625	T	C	1.00	0.97	20.50	0.98	0.84	0.62	0.98	0.94	1.01	0.68	0.49	0.64
rs2898820	FUT8	chr14	65412636	T	G	1.00	0.97	21.50	0.98	0.88	0.75	0.96	0.91	2.56	0.67	0.61	0.84
rs2881040	FYN	chr6	111793105	C	T	1.00	0.97	18.00	0.96	0.85	0.67	0.67	0.70	2.61	0.67	0.49	0.64
rs191998623	FYN	chr6	111873399	C	A	1.00	0.97	21.00	0.98	0.92	0.89	0.98	0.97	1.80	0.64	0.46	0.64
rs6960695	GLI3	chr7	42069062	C	A	1.00	0.97	21.40	0.98	0.81	0.54	0.98	0.94	2.40	0.67	0.50	0.64
rs7793034	GLI3	chr7	42075917	A	G	1.00	0.97	21.20	0.98	0.85	0.67	0.99	0.98	1.61	0.62	0.54	0.78
rs10760160	GSN	chr9	121241048	A	C	1.00	0.97	17.83	0.96	0.89	0.78	0.93	0.88	1.42	0.69	0.47	0.64
rs11801716	HIVEP3	chr1	41678228	C	T	1.00	0.97	20.50	0.98	0.88	0.75	0.92	0.88	1.75	0.64	0.48	0.64
rs997385	HIVEP3	chr1	41856381	G	T	1.00	0.97	18.03	0.96	0.81	0.54	0.97	0.92	0.91	0.68	0.50	0.64
rs7628439	IQCB1	chr3	121808672	A	T	1.00	0.97	9.77	0.67	0.83	0.56	0.51	0.64	0.91	0.68	0.50	0.64
rs78694778	ITGA7	chr12	55712148	A	C	1.00	0.97	14.85	0.86	0.87	0.69	0.41	0.58	0.98	0.68	0.60	0.82
rs4858770	KAT2B	chr3	20152931	C	T	1.00	0.97	12.77	0.76	0.79	0.54	0.93	0.91	1.09	0.76	0.55	0.80
rs11801871	KCNAB2	chr1	6066148	T	C	1.00	0.97	20.50	0.98	0.75	0.54	0.97	0.92	1.11	0.76	0.54	0.78
rs7964223	KRT8	chr12	52903536	G	A	1.00	0.97	21.80	0.98	0.93	0.96	0.94	0.91	2.10	0.67	0.46	0.64
rs8428	MAPK8	chr10	48436373	T	C	1.00	0.97	13.85	0.83	0.80	0.54	0.94	0.91	1.56	0.62	0.50	0.64
rs11598320	MAPK8	chr10	48439224	T	A	1.00	0.97	15.28	0.88	0.84	0.62	0.94	0.91	1.75	0.64	0.46	0.64

(continued)

rsID	Gene	Chrom	Pos	REF	ALT	PredictSNP2		CADD		DANN		FATHMM-MKL		FunSeq2		GWAVA	
						score	conf	score	conf	score	conf	score	conf	score	conf	score	conf
rs78516732	MECOM	chr3	169398068	T	C	1.00	0.97	12.70	0.76	0.78	0.54	0.98	0.94	1.09	0.76	0.54	0.78
rs753974	OBSL1	chr2	219576345	C	T	1.00	0.97	11.59	0.69	0.80	0.54	0.64	0.67	1.19	0.75	0.48	0.64
rs6877743	PDE4D	chr5	59158364	A	C	1.00	0.97	15.39	0.88	0.79	0.54	0.73	0.76	1.12	0.74	0.50	0.64
rs9875064	ROBO1	chr3	79599586	C	T	1.00	0.97	19.63	0.96	0.79	0.54	0.94	0.91	0.88	0.68	0.56	0.80
rs2268278	RUNX1	chr21	34812642	C	A	1.00	0.97	20.40	0.98	0.84	0.62	0.85	0.82	1.49	0.64	0.60	0.82
rs2284613	RUNX1	chr21	34845967	A	G	1.00	0.97	21.70	0.98	0.79	0.54	0.89	0.86	2.10	0.67	0.51	0.65
rs1555257	SEC23A	chr14	39038863	G	T	1.00	0.97	16.26	0.91	0.79	0.54	0.82	0.82	1.42	0.69	0.48	0.64
rs1743964	SGK1	chr6	134174974	G	A	1.00	0.97	11.58	0.69	0.89	0.78	0.97	0.95	2.27	0.67	0.48	0.64
rs4130912	SLC4A4	chr4	71457056	C	T	1.00	0.97	21.90	0.98	0.76	0.54	0.95	0.91	1.76	0.64	0.64	0.80
rs10516059	SLIT3	chr5	169114837	T	A	1.00	0.97	19.66	0.96	0.80	0.54	0.98	0.94	0.96	0.68	0.52	0.67
rs13144151	SMAD1	chr4	145482013	A	G	1.00	0.97	15.20	0.86	0.77	0.54	0.58	0.62	1.16	0.75	0.53	0.78
rs3755724	SYN2	chr3	12159406	C	T	1.00	0.97	19.44	0.96	0.89	0.78	0.83	0.82	2.41	0.67	0.50	0.64
rs7303658	SYT1	chr12	79404122	A	C	1.00	0.97	17.59	0.96	0.85	0.67	0.96	0.92	1.21	0.74	0.50	0.64
rs6781790	WDR6	chr3	49007334	C	T	1.00	0.97	14.40	0.85	0.87	0.69	0.88	0.86	2.14	0.67	0.57	0.76
rs3806557	WNT10A	chr2	218879152	G	A	1.00	0.97	22.30	0.98	0.94	0.98	0.98	0.94	1.61	0.62	0.55	0.80
rs7571600	WNT10A	chr2	218883665	C	A	1.00	0.97	12.95	0.76	0.88	0.75	0.81	0.82	1.37	0.64	0.59	0.84
rs11150090	WWOX	chr16	78587476	A	G	1.00	0.97	19.21	0.96	0.76	0.54	0.98	0.94	1.17	0.75	0.46	0.64
rs7460650	WWP1	chr8	86461856	A	G	1.00	0.97	19.11	0.96	0.87	0.69	0.77	0.80	1.31	0.68	0.51	0.65
rs72867801	ARHGEF3	chr3	56862732	T	C	1.00	0.97	19.30	0.96	0.76	0.54	0.90	0.88	1.06	0.72	0.44	0.53
rs77149297	ATG5	chr6	106305621	G	A	1.00	0.97	8.91	0.67	0.80	0.54	0.62	0.67	0.96	0.68	0.31	0.70
rs2454512	ATG7	chr3	11446032	T	C	1.00	0.97	13.68	0.77	0.83	0.56	0.91	0.88	1.80	0.64	0.40	0.53
rs11659758	BCL2	chr18	63213418	T	A	1.00	0.97	15.68	0.87	0.76	0.54	0.96	0.91	1.06	0.72	0.44	0.53
rs1013402	BDNF	chr11	27690834	A	G	1.00	0.97	15.80	0.89	0.92	0.89	0.93	0.91	0.92	0.68	0.39	0.56
rs11030119	BDNF	chr11	27706555	G	A	1.00	0.97	19.87	0.96	0.81	0.54	0.74	0.76	1.11	0.76	0.42	0.53
rs1463891	CASR	chr3	122251284	A	G	1.00	0.97	13.75	0.77	0.87	0.69	0.57	0.62	1.02	0.68	0.27	0.74
rs2298650	CBL	chr11	119284908	G	T	1.00	0.97	15.48	0.88	0.81	0.54	0.45	0.58	1.78	0.64	0.39	0.56
rs7196495	CDH1	chr16	68739957	C	T	1.00	0.97	17.00	0.95	0.81	0.54	0.93	0.91	2.47	0.67	0.40	0.53
rs7196661	CDH1	chr16	68740009	C	T	1.00	0.97	19.53	0.96	0.86	0.69	0.85	0.82	1.67	0.62	0.45	0.53

(continued)

rsID	Gene	Chrom	Pos	REF	ALT	PredictSNP2		CADD		DANN		FATHMM-MKL		FunSeq2		GWAVA	
						score	conf	score	conf	score	conf	score	conf	score	conf	score	conf
rs213954	CFTR	chr7	117580349	A	G	1.00	0.97	16.56	0.93	0.93	0.96	0.90	0.86	1.17	0.75	0.19	0.79
rs12505265	CLOCK	chr4	55545976	C	G	1.00	0.97	15.64	0.87	0.75	0.54	0.72	0.76	1.09	0.76	0.45	0.53
rs56173804	COL13A1	chr10	69802981	G	A	1.00	0.97	12.40	0.72	0.94	0.96	0.71	0.76	1.79	0.64	0.20	0.78
rs434564	CUX1	chr7	101948281	C	T	1.00	0.97	16.09	0.91	0.83	0.56	0.85	0.82	1.13	0.74	0.44	0.53
rs2207733	EDEM3	chr1	184718581	C	T	1.00	0.97	19.25	0.96	0.92	0.89	0.88	0.86	1.05	0.67	0.39	0.56
rs6604626	ESRRG	chr1	216517945	A	G	1.00	0.97	19.92	0.96	0.80	0.54	0.99	0.98	1.01	0.68	0.37	0.59
rs6673734	ESRRG	chr1	216518095	C	T	1.00	0.97	19.01	0.96	0.89	0.78	0.99	0.98	1.01	0.68	0.38	0.58
rs10746363	ESRRG	chr1	216518410	G	A	1.00	0.97	19.42	0.96	0.87	0.75	0.99	0.98	1.01	0.68	0.42	0.53
rs356675	FANCC	chr9	95184337	T	A	1.00	0.97	15.37	0.88	0.83	0.63	0.94	0.91	1.19	0.75	0.39	0.56
rs9832266	FGF12	chr3	192372835	A	T	1.00	0.97	18.86	0.96	0.88	0.78	0.97	0.95	1.07	0.72	0.39	0.56
rs12994034	FSHR	chr2	48999132	A	G	1.00	0.97	19.82	0.96	0.86	0.69	0.96	0.92	1.01	0.68	0.33	0.68
rs3788981	FSHR	chr2	49018422	A	C	1.00	0.97	18.15	0.96	0.83	0.56	0.92	0.88	1.01	0.68	0.43	0.53
rs2411351	FUT8	chr14	65704823	T	C	1.00	0.97	11.47	0.67	0.81	0.54	0.65	0.70	0.96	0.68	0.44	0.53
rs9384804	FYN	chr6	111755779	A	C	1.00	0.97	18.14	0.96	0.94	0.96	0.94	0.91	1.62	0.62	0.23	0.76
rs62413696	FYN	chr6	111797577	G	C	1.00	0.97	14.06	0.83	0.87	0.69	0.82	0.82	1.62	0.62	0.28	0.73
rs11896536	GLI2	chr2	120824981	G	T	1.00	0.97	11.13	0.67	0.89	0.78	0.53	0.64	0.92	0.68	0.21	0.78
rs2295583	GNAS	chr20	58903393	A	T	1.00	0.97	15.15	0.86	0.90	0.82	0.90	0.86	3.20	0.67	0.39	0.56
rs610412	KCNB1	chr20	49461665	A	C	1.00	0.97	15.93	0.89	0.77	0.54	0.71	0.76	2.60	0.67	0.34	0.65
rs2673409	KCNMA1	chr10	77633583	G	A	1.00	0.97	19.75	0.96	0.81	0.54	0.96	0.91	1.92	0.67	0.34	0.65
rs10930351	LRP2	chr2	169323396	A	G	1.00	0.97	11.39	0.67	0.88	0.78	0.77	0.80	1.53	0.62	0.19	0.79
rs12789028	LTBP3	chr11	65558683	G	A	1.00	0.97	12.43	0.72	0.94	0.96	0.44	0.58	1.17	0.75	0.44	0.53
rs2007924	NRXN3	chr14	79476506	A	G	1.00	0.97	16.86	0.95	0.91	0.89	0.85	0.82	0.96	0.68	0.33	0.68
rs2695217	PPP3CA	chr4	101160876	A	G	1.00	0.97	13.00	0.79	0.86	0.69	0.83	0.82	1.55	0.62	0.43	0.53
rs12023668	PRKAA2	chr1	56703752	A	G	1.00	0.97	10.18	0.67	0.87	0.69	0.58	0.62	0.91	0.68	0.30	0.68
rs6958	PRKAR1A	chr17	68532637	C	G	1.00	0.97	13.85	0.83	0.81	0.54	0.99	0.98	1.44	0.69	0.44	0.53
rs4687571	PRKCD	chr3	53170499	T	G	1.00	0.97	12.42	0.72	0.83	0.63	0.49	0.64	3.31	0.67	0.33	0.68
rs11102320	RAP1A	chr1	111626523	G	T	1.00	0.97	11.81	0.69	0.81	0.54	0.86	0.82	2.19	0.67	0.33	0.68
rs9945114	RBBP8	chr18	22884044	G	C	1.00	0.97	17.71	0.96	0.76	0.54	0.98	0.94	1.29	0.68	0.35	0.65

(continued)

rsID	Gene	Chrom	Pos	REF	ALT	PredictSNP2		CADD		DANN		FATHMM-MKL		FunSeq2		GWAVA	
						score	conf	score	conf	score	conf	score	conf	score	conf	score	conf
rs72739560	RORA	chr15	61146586	A	G	1.00	0.97	19.13	0.96	0.81	0.54	0.94	0.91	1.20	0.74	0.35	0.65
rs2834653	RUNX1	chr21	34858862	A	C	1.00	0.97	14.57	0.85	0.79	0.54	0.82	0.82	1.49	0.64	0.32	0.71
rs11088319	RUNX1	chr21	35431439	T	C	1.00	0.97	20.10	0.96	0.92	0.89	0.97	0.92	1.30	0.68	0.39	0.56
rs78990481	SLC4A4	chr4	71520506	A	T	1.00	0.97	9.11	0.67	0.78	0.54	0.73	0.76	0.98	0.68	0.35	0.65
rs4395527	SMAD1	chr4	145556278	T	A	1.00	0.97	9.20	0.67	0.76	0.54	0.59	0.67	1.59	0.62	0.19	0.79
rs11633026	SMAD3	chr15	67065420	C	T	1.00	0.97	10.49	0.67	0.94	0.96	0.49	0.64	2.59	0.67	0.44	0.53
rs5746091	SOD2	chr6	159693334	C	T	1.00	0.97	15.98	0.89	0.95	1.00	0.95	0.91	2.23	0.67	0.37	0.59
rs12418887	SOX6	chr11	16243165	A	G	1.00	0.97	17.00	0.95	0.90	0.82	0.79	0.80	1.12	0.74	0.39	0.56
rs1975299	SOX6	chr11	16299913	G	C	1.00	0.97	17.57	0.96	0.79	0.54	0.92	0.88	1.12	0.74	0.42	0.53
rs6738087	SPRED2	chr2	65365191	T	C	1.00	0.97	17.58	0.96	0.77	0.54	0.92	0.88	2.29	0.67	0.45	0.53
rs6770839	SUMF1	chr3	4157504	G	C	1.00	0.97	17.28	0.96	0.84	0.62	0.94	0.91	1.37	0.64	0.36	0.62
rs85440	TGIF2	chr20	36586948	C	T	1.00	0.97	17.68	0.96	0.95	0.98	0.70	0.70	1.13	0.74	0.29	0.70
rs58414032	VEGFA	chr6	43771786	G	C	1.00	0.97	18.50	0.96	0.89	0.78	0.91	0.88	1.60	0.62	0.31	0.70
rs8061900	WWOX	chr16	78222783	G	A	1.00	0.97	20.90	0.98	0.79	0.54	0.98	0.97	1.17	0.75	0.41	0.53
rs12446763	WWOX	chr16	78503200	A	G	1.00	0.97	21.80	0.98	0.90	0.82	0.94	0.91	2.16	0.67	0.42	0.53
rs2548843	WWOX	chr16	78648139	C	A	1.00	0.97	18.30	0.96	0.81	0.54	0.95	0.91	1.17	0.75	0.31	0.70
rs72804751	WWOX	chr16	78902483	G	A	1.00	0.97	18.90	0.96	0.89	0.78	0.97	0.95	1.36	0.66	0.22	0.75
rs12444827	ZFH3	chr16	73060594	T	A	1.00	0.97	19.30	0.96	0.76	0.54	0.90	0.88	1.06	0.72	0.44	0.53
rs5022069	ASIC2	chr17	33950920	T	A	1.00	0.97	17.13	0.95	0.92	0.89	0.96	0.91	0.00	0.80	0.58	0.80
rs8075397	BRIP1	chr17	61748761	C	T	1.00	0.97	13.48	0.77	0.78	0.54	0.57	0.62	0.46	0.68	0.48	0.64
rs75747180	CLDN11	chr3	170433436	G	A	1.00	0.97	9.83	0.67	0.83	0.63	0.83	0.82	0.31	0.83	0.46	0.64
rs11893842	INHA	chr2	219572251	A	G	1.00	0.97	13.38	0.79	0.84	0.62	0.74	0.76	0.49	0.62	0.48	0.64
rs41276509	KAT2B	chr3	20152763	A	G	1.00	0.97	10.55	0.67	0.81	0.54	0.91	0.88	0.46	0.68	0.57	0.76
rs2070592	PYY	chr17	43953963	C	T	1.00	0.97	17.49	0.96	0.88	0.78	0.95	0.91	0.35	0.81	0.82	0.86
rs7127006	SOX6	chr11	16357680	G	A	1.00	0.97	16.37	0.91	0.86	0.67	0.94	0.91	0.33	0.83	0.48	0.64
rs2285744	THSD7A	chr7	11469934	G	C	1.00	0.87	19.09	0.52	1.00	0.66	0.95	0.62	2.00	0.62	0.32	0.52
rs2614266	AHI1	chr6	135395394	A	T	0.33	0.91	14.40	0.85	0.93	0.89	0.11	0.85	1.11	0.76	0.56	0.80
rs11073890	ANPEP	chr15	89809380	A	G	0.33	0.91	11.82	0.69	0.88	0.78	0.16	0.85	1.92	0.67	0.48	0.64

(continued)

rsID	Gene	Chrom	Pos	REF	ALT	PredictSNP2		CADD		DANN		FATHMM-MKL		FunSeq2		GWAVA	
						score	conf	score	conf	score	conf	score	conf	score	conf	score	conf
rs4988351	BRIP1	chr17	61780448	C	G	0.33	0.91	10.20	0.67	0.89	0.78	0.22	0.75	1.42	0.69	0.59	0.84
rs2244346	LIMS2	chr2	127645125	A	T	0.33	0.91	10.20	0.67	0.88	0.75	0.16	0.85	1.10	0.76	0.46	0.64
rs28394191	LRPPRC	chr2	43945305	A	C	0.31	0.91	12.66	0.76	0.86	0.69	0.19	0.77	0.92	0.68	0.55	0.80
rs1391903	ATG7	chr3	11381797	C	G	0.52	0.91	19.18	0.96	0.72	0.62	0.86	0.82	1.99	0.67	0.51	0.65
rs9897928	BRIP1	chr17	61766669	A	T	0.39	0.91	8.82	0.67	0.68	0.65	0.57	0.62	1.27	0.68	0.54	0.78
rs2906649	CUX1	chr7	101933758	T	C	0.52	0.91	15.64	0.87	0.70	0.62	0.87	0.86	1.14	0.74	0.50	0.64
rs6750998	EIF2AK3	chr2	88583424	T	A	0.37	0.91	11.89	0.69	0.66	0.70	0.45	0.58	1.42	0.69	0.64	0.80
rs36061550	EIF4E	chr4	98930260	T	C	0.48	0.91	13.12	0.79	0.70	0.62	0.84	0.82	2.50	0.67	0.55	0.80
rs11806070	ESRRG	chr1	216500480	T	A	0.54	0.91	18.12	0.96	0.74	0.62	0.95	0.91	1.61	0.62	0.49	0.64
rs17032268	FANCD2	chr3	10026288	C	T	0.49	0.91	10.03	0.67	0.67	0.70	1.00	0.98	1.48	0.65	0.58	0.80
rs2883881	FOXO3	chr6	108576183	A	G	0.51	0.91	14.67	0.86	0.68	0.65	0.87	0.86	1.52	0.62	0.63	0.80
rs985909	FSHR	chr2	48978578	G	A	0.54	0.91	20.30	0.98	0.73	0.62	0.96	0.91	1.20	0.74	0.46	0.64
rs113091122	FYN	chr6	111676493	G	A	0.50	0.91	15.08	0.86	0.63	0.71	0.94	0.91	1.81	0.64	0.74	0.86
rs2182644	FYN	chr6	111817679	G	A	0.48	0.91	14.12	0.83	0.67	0.66	0.72	0.76	1.81	0.64	0.52	0.67
rs846272	GLI3	chr7	41996746	G	A	0.53	0.91	17.93	0.96	0.73	0.62	0.86	0.82	1.61	0.62	0.47	0.64
rs10951666	GLI3	chr7	42076289	C	T	0.56	0.91	21.80	0.98	0.68	0.65	0.99	0.98	1.61	0.62	0.55	0.80
rs17174756	GPC6	chr13	93929286	T	G	0.43	0.91	14.49	0.85	0.71	0.62	0.46	0.58	0.93	0.68	0.48	0.64
rs10760169	GSN	chr9	121295404	C	T	0.50	0.91	16.40	0.93	0.72	0.62	0.72	0.76	1.60	0.62	0.60	0.82
rs12069663	HIVEP3	chr1	41696581	T	C	0.30	0.91	11.60	0.69	0.40	0.80	0.89	0.86	0.91	0.68	0.53	0.78
rs1044153	KIF11	chr10	92655321	A	C	0.45	0.91	9.75	0.67	0.68	0.65	0.78	0.80	0.97	0.68	0.56	0.80
rs2715834	MAP2K6	chr17	69532581	C	G	0.56	0.91	17.82	0.96	0.67	0.66	0.98	0.94	1.11	0.76	0.56	0.80
rs9814561	MECOM	chr3	169400776	A	T	0.45	0.91	9.77	0.67	0.63	0.67	0.95	0.91	1.27	0.68	0.48	0.64
rs7521130	PAPPA2	chr1	176501844	G	A	0.51	0.91	15.85	0.89	0.70	0.62	0.83	0.82	1.00	0.68	0.62	0.82
rs675342	PCCA	chr13	100235386	C	T	0.37	0.91	11.37	0.67	0.59	0.75	0.67	0.70	1.48	0.65	0.57	0.76
rs1379805	RBBP8	chr18	22914589	T	C	0.40	0.91	9.12	0.67	0.72	0.62	0.59	0.67	1.11	0.76	0.51	0.65
rs331145	ROBO1	chr3	78868733	G	A	0.37	0.91	13.30	0.79	0.45	0.86	0.95	0.91	1.87	0.67	0.55	0.80
rs331199	ROBO1	chr3	78901732	A	G	0.53	0.91	17.99	0.96	0.71	0.62	0.94	0.91	1.07	0.72	0.48	0.64
rs6772919	ROBO1	chr3	78935416	C	T	0.31	0.91	11.87	0.69	0.37	0.80	0.92	0.88	0.88	0.68	0.53	0.78

(continued)

rsID	Gene	Chrom	Pos	REF	ALT	PredictSNP2		CADD		DANN		FATHMM-MKL		FunSeq2		GWAVA	
						score	conf	score	conf	score	conf	score	conf	score	conf	score	conf
rs9837046	ROBO1	chr3	79599513	A	G	0.54	0.91	15.24	0.88	0.67	0.66	0.95	0.91	0.88	0.68	0.54	0.78
rs2249650	RUNX1	chr21	34808689	A	G	0.53	0.91	20.80	0.98	0.73	0.62	0.87	0.86	1.49	0.64	0.51	0.65
rs2153277	RUNX2	chr6	45558952	C	T	0.43	0.91	16.04	0.89	0.52	0.85	0.93	0.88	2.26	0.67	0.49	0.64
rs1634605	SEMA3A	chr7	84410449	C	A	0.43	0.91	16.36	0.91	0.52	0.85	0.92	0.88	0.96	0.68	0.55	0.80
rs7933830	SOX6	chr11	16355573	C	T	0.41	0.91	14.65	0.86	0.50	0.88	0.95	0.91	0.94	0.68	0.50	0.64
rs34866095	SOX6	chr11	16355810	A	G	0.44	0.91	8.98	0.67	0.72	0.62	0.81	0.80	0.94	0.68	0.59	0.84
rs11023928	SOX6	chr11	16361573	A	T	0.46	0.91	12.02	0.69	0.72	0.62	0.89	0.86	1.12	0.74	0.50	0.64
rs9909600	VPS53	chr17	673921	T	C	0.55	0.91	17.82	0.96	0.67	0.66	0.98	0.97	1.17	0.75	0.61	0.84
rs59756710	WIF1	chr12	65093836	A	G	0.55	0.91	18.99	0.96	0.68	0.65	0.98	0.97	1.40	0.69	0.48	0.64
rs1074963	WWOX	chr16	78136263	C	G	0.49	0.91	15.72	0.87	0.61	0.72	0.86	0.82	1.17	0.75	0.53	0.78
rs9935941	WWOX	chr16	78138277	A	G	0.41	0.91	13.31	0.79	0.55	0.82	0.91	0.88	1.36	0.66	0.48	0.64
rs2548861	WWOX	chr16	78624496	T	G	0.52	0.91	19.18	0.96	0.72	0.62	0.86	0.82	1.99	0.67	0.51	0.65
rs1672718	ZBTB16	chr11	114080464	C	G	0.39	0.91	8.82	0.67	0.68	0.65	0.57	0.62	1.27	0.68	0.54	0.78
rs9923980	ZFH3	chr16	72911221	C	A	0.52	0.91	15.64	0.87	0.70	0.62	0.87	0.86	1.14	0.74	0.50	0.64
rs59590642	ZFH3	chr16	73092389	C	T	0.37	0.91	11.89	0.69	0.66	0.70	0.45	0.58	1.42	0.69	0.64	0.80
rs506388	ACTA1	chr1	229435314	T	G	0.53	0.91	16.82	0.95	0.69	0.62	0.92	0.88	1.36	0.66	0.80	0.86
rs4428403	ALDH7A1	chr5	126527433	G	A	0.52	0.91	15.22	0.88	0.67	0.66	0.87	0.86	1.70	0.62	0.54	0.78
rs11716312	ARHGEF3	chr3	56897542	A	C	0.50	0.91	16.55	0.93	0.58	0.75	0.97	0.95	1.30	0.68	0.55	0.80
rs9898582	ASIC2	chr17	33894162	G	T	0.52	0.91	15.98	0.89	0.71	0.62	0.90	0.86	1.42	0.69	0.67	0.84
rs35350386	ATP8A2	chr13	25560035	T	C	1.00	0.97	15.20	0.86	0.83	0.56	0.51	0.64	0.19	0.83	0.37	0.59
rs2419339	ATP8A2	chr13	25790769	A	T	1.00	0.97	15.88	0.89	0.77	0.54	0.91	0.88	0.57	0.55	0.30	0.68
rs3019605	CPT1A	chr11	68817878	G	A	1.00	0.97	9.66	0.67	0.89	0.78	0.54	0.62	0.08	0.80	0.20	0.78
rs9552216	CRYL1	chr13	20500208	T	C	1.00	0.97	17.34	0.96	0.76	0.54	0.95	0.91	0.00	0.80	0.41	0.53
rs10167567	DNMT3A	chr2	25261089	T	G	1.00	0.97	11.56	0.69	0.81	0.54	0.78	0.80	0.00	0.80	0.19	0.79
rs7547965	DOCK7	chr1	62460700	A	G	1.00	0.97	9.17	0.67	0.80	0.54	0.42	0.58	0.00	0.80	0.29	0.70
rs17123705	DOCK7	chr1	62533219	T	C	1.00	0.97	11.21	0.67	0.76	0.54	0.84	0.82	0.27	0.82	0.38	0.58
rs59505425	DOCK7	chr1	62571608	T	C	1.00	0.97	15.73	0.87	0.96	1.00	0.61	0.67	0.29	0.82	0.24	0.78
rs147752820	DOCK7	chr1	62645898	A	C	1.00	0.97	15.56	0.87	0.95	0.98	0.74	0.76	0.46	0.68	0.31	0.70

(continued)

rsID	Gene	Chrom	Pos	REF	ALT	PredictSNP2		CADD		DANN		FATHMM-MKL		FunSeq2		GWAVA	
						score	conf	score	conf	score	conf	score	conf	score	conf	score	conf
rs4699687	EIF4E	chr4	98868376	G	C	1.00	0.97	13.43	0.77	0.94	0.96	0.56	0.62	0.54	0.59	0.27	0.74
rs17511498	EIF4E	chr4	98934090	A	G	1.00	0.97	15.07	0.86	0.95	1.00	0.72	0.76	0.56	0.59	0.41	0.53
rs609324	FANCC	chr9	95146389	G	A	1.00	0.97	14.71	0.86	0.97	1.00	0.73	0.76	0.54	0.59	0.20	0.78
rs6798854	FGF12	chr3	192336452	G	A	1.00	0.97	15.69	0.87	0.95	1.00	0.81	0.80	0.54	0.59	0.41	0.53
rs1103996	FIG4	chr6	109696392	A	G	1.00	0.97	16.38	0.91	0.90	0.78	0.59	0.67	0.19	0.83	0.32	0.71
rs9386839	FIG4	chr6	109725419	C	A	1.00	0.97	12.89	0.76	0.87	0.75	0.89	0.86	0.00	0.80	0.33	0.68
rs2300440	FSHR	chr2	48967418	T	A	1.00	0.97	14.77	0.86	0.92	0.89	0.67	0.70	0.56	0.55	0.17	0.79
rs7464569	GDAP1	chr8	74474293	T	C	1.00	0.97	11.60	0.69	0.77	0.54	0.58	0.62	0.45	0.68	0.27	0.74
rs3828024	GREM2	chr1	240530814	G	A	1.00	0.97	11.39	0.67	0.81	0.54	0.47	0.64	0.15	0.83	0.39	0.56
rs61992487	HSP90AA1	chr14	102079333	G	C	1.00	0.97	14.81	0.86	0.94	0.98	0.71	0.76	0.15	0.83	0.34	0.65
rs4676750	IQCB1	chr3	121777704	A	C	1.00	0.97	9.95	0.67	0.89	0.78	0.42	0.58	0.39	0.82	0.30	0.68
rs10512998	ITGA1	chr5	52809029	A	T	1.00	0.97	11.61	0.69	0.90	0.82	0.41	0.58	0.15	0.83	0.25	0.75
rs56729913	KAT2B	chr3	20143361	T	A	1.00	0.97	14.10	0.83	0.81	0.54	0.94	0.91	0.27	0.82	0.22	0.75
rs11187105	KIF11	chr10	92624670	C	T	1.00	0.97	11.11	0.67	0.85	0.67	0.83	0.82	0.19	0.83	0.28	0.73
rs11187117	KIF11	chr10	92652401	A	C	1.00	0.97	13.65	0.77	0.92	0.89	0.46	0.58	0.11	0.83	0.22	0.75
rs11588109	MFN2	chr1	11977186	A	G	1.00	0.97	12.04	0.69	0.91	0.89	0.73	0.76	0.42	0.77	0.24	0.78
rs35735	NR1H4	chr12	100556593	G	T	1.00	0.97	16.05	0.89	0.95	0.98	0.70	0.70	0.46	0.68	0.20	0.78
rs221494	NRXN3	chr14	79623343	A	C	1.00	0.97	14.17	0.83	0.93	0.96	0.58	0.62	0.35	0.81	0.36	0.62
rs9486793	OSTM1	chr6	108068101	A	G	1.00	0.97	13.42	0.77	0.91	0.89	0.63	0.67	0.35	0.81	0.34	0.65
rs60631199	PCYT1A	chr3	196234383	C	T	1.00	0.97	13.47	0.77	0.94	0.96	0.63	0.67	0.00	0.80	0.14	0.79
rs55998224	PDE4D	chr5	59175833	T	C	1.00	0.97	13.91	0.83	0.91	0.89	0.74	0.76	0.26	0.82	0.30	0.68
rs186464895	PROKR2	chr20	5298416	G	A	1.00	0.97	17.13	0.95	0.87	0.75	0.96	0.91	0.15	0.83	0.27	0.74
rs113224513	RAB31	chr18	9708630	C	T	1.00	0.97	13.08	0.79	0.86	0.69	0.82	0.82	0.17	0.83	0.27	0.74
rs9961942	RBBP8	chr18	22808992	T	C	1.00	0.97	12.11	0.72	0.86	0.67	0.65	0.70	0.33	0.81	0.43	0.53
rs13241054	RNF216	chr7	5776079	T	C	1.00	0.97	13.01	0.79	0.95	1.00	0.53	0.62	0.31	0.83	0.26	0.73
rs7615834	ROBO1	chr3	79624110	G	A	1.00	0.97	12.47	0.72	0.97	1.00	0.55	0.62	0.19	0.83	0.20	0.78
rs12079656	SDCCAG8	chr1	243289817	G	T	1.00	0.97	11.57	0.69	0.93	0.96	0.54	0.62	0.27	0.82	0.05	0.79
rs7187490	SH2B1	chr16	28854672	A	G	1.00	0.97	11.89	0.69	0.97	1.00	0.42	0.58	0.48	0.62	0.32	0.71

(continued)

rsID	Gene	Chrom	Pos	REF	ALT	PredictSNP2		CADD		DANN		FATHMM-MKL		FunSeq2		GWAVA	
						score	conf	score	conf	score	conf	score	conf	score	conf	score	conf
rs224591	SLC11A2	chr12	51006844	C	A	1.00	0.97	11.76	0.69	0.92	0.89	0.52	0.64	0.31	0.83	0.16	0.79
rs12643326	SLC4A4	chr4	71261433	T	C	1.00	0.97	12.99	0.76	0.96	1.00	0.54	0.62	0.26	0.82	0.33	0.68
rs28801405	SLC4A4	chr4	71503942	C	T	1.00	0.97	14.16	0.83	0.87	0.69	0.86	0.82	0.31	0.83	0.35	0.65
rs16932966	SOX6	chr11	16361336	G	A	1.00	0.97	14.64	0.86	0.91	0.89	0.64	0.67	0.28	0.82	0.19	0.79
rs211456	SYNGAP1	chr6	33421604	G	T	1.00	0.97	13.14	0.79	0.93	0.96	0.47	0.64	0.15	0.83	0.43	0.53
rs10740307	TET1	chr10	68661944	C	T	1.00	0.97	18.91	0.96	0.85	0.67	0.95	0.91	0.35	0.81	0.27	0.74
rs10861170	TXNRD1	chr12	104222179	A	G	1.00	0.97	13.91	0.83	0.91	0.89	0.57	0.62	0.35	0.81	0.42	0.53
rs6741654	UNC80	chr2	209951216	A	G	1.00	0.97	11.61	0.69	0.77	0.54	0.80	0.80	0.31	0.83	0.37	0.59
rs7213115	VPS53	chr17	621086	G	A	1.00	0.97	12.23	0.72	0.79	0.54	0.45	0.58	0.43	0.77	0.29	0.70
rs12325758	VPS53	chr17	656868	G	A	1.00	0.97	12.05	0.69	0.81	0.54	0.91	0.88	0.00	0.80	0.16	0.79
rs77240075	WNT7A	chr3	13881117	C	G	1.00	0.97	17.12	0.95	0.95	1.00	0.87	0.86	0.58	0.55	0.39	0.56
rs62039373	WWOX	chr16	78626812	A	C	1.00	0.97	14.18	0.83	0.96	1.00	0.68	0.70	0.15	0.83	0.28	0.73
rs1424162	WWOX	chr16	78651283	C	T	1.00	0.97	9.22	0.67	0.85	0.62	0.42	0.58	0.36	0.81	0.36	0.62
rs7820234	WWP1	chr8	86338721	A	C	1.00	0.97	9.25	0.67	0.77	0.54	0.90	0.88	0.55	0.59	0.36	0.62
rs4782026	XYLT1	chr16	17160447	A	T	1.00	0.97	13.18	0.79	0.81	0.54	0.47	0.64	0.19	0.83	0.26	0.73
rs61613546	ZFHX3	chr16	73591833	A	T	1.00	0.97	11.56	0.69	0.88	0.75	0.78	0.80	0.57	0.55	0.22	0.75
rs66694826	ZFPM2	chr8	105643938	C	T	1.00	0.97	9.07	0.67	0.86	0.69	0.88	0.86	0.55	0.59	0.34	0.65
rs10953305	ACHE	chr7	100894947	G	A	0.86	0.91	14.33	0.85	0.95	0.98	0.04	0.95	1.10	0.76	0.45	0.53
rs7755506	ATG5	chr6	106316486	T	C	0.83	0.91	11.24	0.67	0.89	0.78	0.19	0.75	1.77	0.64	0.34	0.65
rs1383596	BCL2	chr18	63241629	C	A	0.32	0.91	14.64	0.86	0.80	0.54	0.23	0.75	2.05	0.67	0.38	0.58
rs3743674	CDH1	chr16	68737469	C	T	0.33	0.91	10.95	0.67	0.92	0.89	0.15	0.91	1.85	0.67	0.13	0.79
rs1599926	COPB2	chr3	139394307	T	A	0.84	0.91	14.04	0.83	0.85	0.67	0.19	0.77	1.92	0.67	0.36	0.62
rs6994556	FGFR1	chr8	38467098	G	A	0.34	0.91	11.34	0.67	0.90	0.82	0.16	0.85	2.05	0.67	0.41	0.53
rs9556314	GPC6	chr13	93659745	T	G	0.38	0.91	13.83	0.83	0.91	0.89	0.16	0.85	0.93	0.68	0.25	0.75
rs12153725	ITGA1	chr5	52812197	T	C	0.40	0.91	13.13	0.79	0.92	0.89	0.17	0.85	1.04	0.67	0.37	0.59
rs11227217	LTBP3	chr11	65539931	C	T	0.39	0.91	10.39	0.67	0.95	1.00	0.23	0.75	1.17	0.75	0.38	0.58
rs555183	MGLL	chr3	127721299	A	G	0.33	0.91	14.08	0.83	0.88	0.78	0.14	0.91	1.37	0.64	0.26	0.73
rs517237	PHGDH	chr1	119742365	T	C	0.38	0.91	12.83	0.76	0.91	0.89	0.17	0.85	1.76	0.64	0.32	0.71

(continued)



rsID	Gene	Chrom	Pos	REF	ALT	PredictSNP2		CADD		DANN		FATHMM-MKL		FunSeq2		GWAVA	
						score	conf	score	conf	score	conf	score	conf	score	conf	score	conf
rs2952272	PRKAR1A	chr17	68528424	T	G	0.41	0.91	14.57	0.85	0.91	0.89	0.21	0.75	1.44	0.69	0.27	0.74
rs500184	RAP1A	chr1	111620924	C	A	0.37	0.91	15.99	0.89	0.86	0.69	0.18	0.85	1.56	0.62	0.19	0.79
rs6762889	SLCO2A1	chr3	134029323	C	T	0.39	0.91	10.72	0.67	0.95	0.98	0.17	0.85	1.05	0.67	0.23	0.76
rs17749633	SUMF1	chr3	3962743	T	C	0.34	0.91	14.80	0.86	0.84	0.62	0.18	0.82	0.93	0.68	0.31	0.70
rs381847	SYNGAP1	chr6	33433347	C	G	0.34	0.91	13.10	0.79	0.88	0.75	0.19	0.75	1.23	0.74	0.20	0.78
rs4555139	XYLT1	chr16	17470092	G	A	0.37	0.91	9.37	0.67	0.92	0.89	0.19	0.77	0.99	0.68	0.20	0.78
rs1075855	ZFH3	chr16	73064192	G	C	0.35	0.91	14.59	0.85	0.86	0.69	0.21	0.75	1.61	0.62	0.36	0.62
rs985136	ABCC8	chr11	17476247	C	G	0.43	0.91	13.44	0.77	0.70	0.62	0.60	0.67	1.56	0.62	0.38	0.58
rs580739	ADRA1A	chr8	26856190	C	G	0.42	0.91	13.62	0.77	0.74	0.62	0.49	0.64	0.94	0.68	0.38	0.58
rs4737009	ANK1	chr8	41772887	G	A	0.41	0.91	11.58	0.69	0.69	0.62	0.69	0.70	1.54	0.62	0.25	0.75
rs9384521	ARID1B	chr6	156985765	A	G	0.41	0.91	12.44	0.72	0.54	0.82	0.98	0.94	1.24	0.74	0.39	0.56
rs7763624	ARID1B	chr6	157097826	G	A	0.48	0.91	16.67	0.93	0.58	0.79	0.94	0.91	2.05	0.67	0.42	0.53
rs9898220	ASIC2	chr17	34066996	A	G	0.43	0.91	10.23	0.67	0.64	0.67	0.77	0.80	1.61	0.62	0.31	0.70
rs200966269	BMP4	chr14	53953937	A	C	0.51	0.91	15.06	0.86	0.69	0.62	0.87	0.86	2.36	0.67	0.40	0.53
rs7774274	CD109	chr6	73704887	G	T	0.44	0.91	11.96	0.69	0.66	0.70	0.85	0.82	1.27	0.68	0.37	0.59
rs12090767	CHRM3	chr1	239851082	G	A	0.50	0.91	13.67	0.77	0.72	0.62	0.94	0.91	0.98	0.68	0.41	0.53
rs11160668	DYNC1H1	chr14	102046749	G	A	0.39	0.91	9.11	0.67	0.58	0.79	0.83	0.82	1.61	0.62	0.32	0.71
rs7433757	EPHA3	chr3	89168179	G	A	0.55	0.91	21.40	0.98	0.69	0.62	0.99	0.98	1.39	0.64	0.40	0.53
rs35001652	EPHB2	chr1	22756174	G	A	0.53	0.91	14.93	0.86	0.72	0.62	0.95	0.91	2.15	0.67	0.27	0.74
rs1025654	ESRRG	chr1	216907989	C	G	0.42	0.91	11.72	0.69	0.56	0.81	0.97	0.95	1.01	0.68	0.38	0.58
rs6444632	FGF12	chr3	192337832	A	G	0.47	0.91	11.27	0.67	0.73	0.62	0.94	0.91	0.88	0.68	0.24	0.78
rs76226586	FGF12	chr3	192488737	A	C	0.48	0.91	12.21	0.72	0.64	0.67	0.96	0.92	0.88	0.68	0.35	0.65
rs2300439	FSHR	chr2	48967563	A	G	0.43	0.91	9.11	0.67	0.75	0.62	0.74	0.76	1.01	0.68	0.23	0.76
rs2148709	FYN	chr6	111800025	G	A	0.38	0.91	12.62	0.76	0.55	0.82	0.85	0.82	1.62	0.62	0.43	0.53
rs12055398	FYN	chr6	111812327	T	C	0.46	0.91	16.50	0.93	0.72	0.62	0.47	0.64	1.81	0.64	0.38	0.58
rs10168559	GLI2	chr2	120842295	C	T	0.51	0.91	18.29	0.96	0.72	0.62	0.82	0.82	1.54	0.62	0.28	0.73
rs2141173	GLI3	chr7	41994802	G	A	0.51	0.91	17.23	0.96	0.67	0.66	0.75	0.76	1.61	0.62	0.38	0.58
rs10513365	GSN	chr9	121258647	C	T	0.54	0.91	17.59	0.96	0.75	0.62	0.97	0.92	1.23	0.74	0.41	0.53

(continued)

rsID	Gene	Chrom	Pos	REF	ALT	PredictSNP2		CADD		DANN		FATHMM-MKL		FunSeq2		GWAVA	
						score	conf	score	conf	score	conf	score	conf	score	conf	score	conf
rs11175942	HMGA2	chr12	65845861	G	C	0.39	0.91	9.09	0.67	0.68	0.66	0.49	0.64	1.86	0.67	0.45	0.53
rs12313942	HMGA2	chr12	65848548	G	A	0.49	0.91	13.13	0.79	0.66	0.70	0.89	0.86	1.05	0.67	0.31	0.70
rs7707371	ITGA1	chr5	52811584	C	T	0.47	0.91	16.18	0.91	0.56	0.81	0.95	0.91	2.02	0.67	0.45	0.53
rs7730842	ITGA1	chr5	52811602	T	C	0.41	0.91	11.29	0.67	0.60	0.72	0.89	0.86	1.41	0.69	0.36	0.62
rs183253987	LEMD3	chr12	65198470	T	C	0.45	0.91	11.52	0.67	0.63	0.67	0.95	0.91	1.49	0.64	0.37	0.59
rs12089591	LHX4	chr1	180255070	C	G	0.47	0.91	12.76	0.76	0.74	0.62	0.81	0.80	1.76	0.64	0.31	0.70
rs11172113	LRP1	chr12	57133500	T	C	0.44	0.91	14.98	0.86	0.61	0.68	0.62	0.67	0.97	0.68	0.43	0.53
rs9919504	MAPK8	chr10	48410736	A	C	0.47	0.91	10.69	0.67	0.68	0.66	0.87	0.86	1.12	0.74	0.39	0.56
rs2543577	NRXN3	chr14	79606151	T	C	0.53	0.91	16.87	0.95	0.65	0.65	0.96	0.91	0.96	0.68	0.35	0.65
rs1501641	PDE4D	chr5	59892278	A	G	0.44	0.91	14.86	0.86	0.55	0.82	0.93	0.91	1.73	0.64	0.37	0.59
rs7073334	PPA1	chr10	70232762	T	C	0.45	0.91	14.42	0.85	0.68	0.66	0.49	0.64	1.97	0.67	0.25	0.75
rs56886418	PRDM1	chr6	106090852	A	G	0.52	0.91	15.52	0.87	0.65	0.65	0.89	0.86	2.55	0.67	0.45	0.53
rs7547486	PRKAA2	chr1	56650158	G	A	0.37	0.91	11.23	0.67	0.53	0.85	0.97	0.92	1.53	0.62	0.29	0.70
rs12073253	PRKAA2	chr1	56650829	T	G	0.43	0.91	13.46	0.77	0.70	0.62	0.63	0.67	1.53	0.62	0.24	0.78
rs12964830	RBBP8	chr18	22938970	G	T	0.51	0.91	12.70	0.76	0.68	0.65	0.95	0.91	1.29	0.68	0.30	0.68
rs58423115	ROBO1	chr3	78722702	T	G	0.39	0.91	9.12	0.67	0.68	0.65	0.54	0.62	0.88	0.68	0.37	0.59
rs58792018	ROBO1	chr3	78798407	G	A	0.33	0.91	12.18	0.72	0.47	0.86	0.85	0.82	1.07	0.72	0.40	0.53
rs162428	ROBO1	chr3	78888587	A	G	0.36	0.91	11.32	0.67	0.49	0.88	0.95	0.91	0.88	0.68	0.44	0.53
rs9809381	ROBO1	chr3	79671812	C	T	0.40	0.91	15.47	0.88	0.38	0.80	0.97	0.92	0.88	0.68	0.45	0.53
rs72748757	RORA	chr15	60619280	A	G	0.36	0.91	13.61	0.77	0.45	0.86	0.90	0.88	1.20	0.74	0.18	0.79
rs1351546	RORA	chr15	60935507	C	T	0.50	0.91	19.00	0.96	0.57	0.79	0.98	0.94	1.01	0.68	0.33	0.68
rs72739558	RORA	chr15	61146565	G	A	0.53	0.91	19.29	0.96	0.61	0.72	0.98	0.97	1.20	0.74	0.30	0.68
rs4817695	RUNX1	chr21	34847020	A	G	0.41	0.91	11.58	0.69	0.61	0.72	0.73	0.76	1.48	0.65	0.30	0.68
rs1040329	RUNX1	chr21	35382946	C	G	0.38	0.91	13.08	0.79	0.44	0.85	0.98	0.94	1.49	0.64	0.45	0.53
rs7531548	SDCCAG8	chr1	243386220	T	G	0.38	0.91	11.63	0.69	0.55	0.82	0.90	0.86	0.93	0.68	0.35	0.65
rs17242890	SEMA3A	chr7	84193529	T	A	0.51	0.91	14.77	0.86	0.60	0.72	0.97	0.95	1.76	0.64	0.44	0.53
rs10861941	SYT1	chr12	79366608	A	G	0.40	0.91	11.75	0.69	0.56	0.81	0.91	0.88	1.56	0.62	0.24	0.78
rs61927383	SYT1	chr12	79417553	G	A	0.41	0.91	10.14	0.67	0.63	0.71	0.77	0.80	1.21	0.74	0.34	0.65

(continued)

rsID	Gene	Chrom	Pos	REF	ALT	PredictSNP2		CADD		DANN		FATHMM-MKL		FunSeq2		GWAVA	
						score	conf	score	conf	score	conf	score	conf	score	conf	score	conf
rs1399773	TP63	chr3	189868241	T	G	0.48	0.91	12.25	0.72	0.61	0.72	0.96	0.92	1.95	0.67	0.34	0.65
rs6789961	TP63	chr3	189869302	A	G	0.44	0.91	10.23	0.67	0.61	0.68	0.87	0.86	1.96	0.67	0.35	0.65
rs461251	VPS53	chr17	715922	A	G	0.46	0.91	12.33	0.72	0.69	0.62	0.78	0.80	1.16	0.75	0.35	0.65
rs1446530	WIF1	chr12	65085724	C	A	0.37	0.91	12.61	0.76	0.53	0.85	0.90	0.86	1.40	0.69	0.42	0.53
rs8057015	WVOX	chr16	78126565	C	A	0.52	0.91	16.23	0.91	0.66	0.70	0.86	0.86	1.17	0.75	0.39	0.56
rs2550599	WVOX	chr16	78620854	C	G	0.30	0.91	9.74	0.67	0.40	0.80	0.87	0.86	2.16	0.67	0.36	0.62
rs2550619	WVOX	chr16	78638597	C	G	0.53	0.91	17.87	0.96	0.72	0.62	0.92	0.88	2.16	0.67	0.30	0.68
rs2881483	WVOX	chr16	78920778	A	G	0.46	0.91	18.84	0.96	0.55	0.82	0.96	0.92	1.36	0.66	0.31	0.70
rs8052534	WVOX	chr16	78929352	G	C	0.53	0.91	16.88	0.95	0.67	0.66	0.95	0.91	1.36	0.66	0.33	0.68
rs1468577	BRIP1	chr17	61793256	A	G	0.35	0.91	7.65	0.76	0.90	0.82	0.62	0.67	1.09	0.76	0.45	0.53
rs1550223	RORA	chr15	61145563	A	C	0.33	0.91	10.98	0.67	0.90	0.82	0.16	0.85	0.39	0.82	0.46	0.64
rs7651006	TRAK1	chr3	42106807	A	G	0.33	0.91	13.22	0.79	0.85	0.67	0.18	0.85	0.52	0.62	0.50	0.64
rs17123688	DOCK7	chr1	62510677	T	C	0.46	0.91	3.81	0.67	0.65	0.65	0.92	0.88	0.54	0.59	0.58	0.80
rs331182	ROBO1	chr3	78888626	T	C	0.47	0.91	15.54	0.87	0.57	0.79	0.86	0.82	0.26	0.82	0.48	0.64
rs17563	BMP4	chr14	53950804	A	G	-0.14	0.67	22.00	0.52	0.90	0.84	0.94	0.59	3.00	0.61	0.25	0.51
rs3766597	AGL	chr1	99877228	C	T	0.36	0.91	14.39	0.85	0.86	0.69	0.17	0.89	0.26	0.82	0.17	0.79
rs12695921	AGTR1	chr3	148740058	T	C	0.38	0.91	13.89	0.83	0.90	0.82	0.16	0.85	0.50	0.62	0.13	0.79
rs4870500	ARID1B	chr6	156960846	A	G	0.32	0.91	13.23	0.79	0.89	0.78	0.15	0.91	0.43	0.77	0.37	0.59
rs7461612	BMP1	chr8	22187714	T	A	0.34	0.91	10.78	0.67	0.90	0.82	0.21	0.75	0.56	0.55	0.38	0.58
rs3729496	CHAT	chr10	49613145	G	T	0.31	0.91	8.90	0.67	0.88	0.78	0.16	0.85	0.49	0.62	0.44	0.53
rs10998973	COL13A1	chr10	69802681	G	A	0.06	0.93	21.30	0.79	0.97	0.87	0.32	0.73	0.00	0.93	0.34	0.56
rs10124024	COL27A1	chr9	114226442	C	A	0.35	0.91	14.94	0.86	0.85	0.62	0.20	0.75	0.19	0.83	0.25	0.75
rs12185464	CYB5A	chr18	74296255	C	T	0.32	0.91	9.31	0.67	0.88	0.78	0.23	0.75	0.00	0.80	0.19	0.79
rs13271325	EXTL3	chr8	28616074	T	C	0.31	0.91	10.68	0.67	0.91	0.89	0.15	0.91	0.24	0.83	0.30	0.68
rs188422682	FCGR2B	chr1	161676104	T	C	0.33	0.91	9.40	0.67	0.90	0.82	0.16	0.85	0.39	0.82	0.44	0.53
rs67957534	FUT8	chr14	65657867	T	G	0.32	0.91	9.12	0.67	0.88	0.78	0.23	0.75	0.15	0.83	0.26	0.73
rs7411138	GREM2	chr1	240553161	C	A	0.35	0.91	14.20	0.83	0.86	0.69	0.16	0.85	0.08	0.80	0.32	0.71
rs55927556	HIVEP3	chr1	41600820	A	G	0.31	0.91	8.70	0.67	0.92	0.89	0.13	0.89	0.11	0.83	0.35	0.65

(continued)

rsID	Gene	Chrom	Pos	REF	ALT	PredictSNP2		CADD		DANN		FATHMM-MKL		FunSeq2		GWAVA	
						score	conf	score	conf	score	conf	score	conf	score	conf	score	conf
rs12304905	HMGA2	chr12	65829525	T	G	0.34	0.91	13.97	0.83	0.86	0.69	0.18	0.85	0.24	0.83	0.45	0.53
rs75744075	HSD11B1	chr1	209688220	T	G	0.31	0.91	12.19	0.72	0.87	0.69	0.20	0.75	0.19	0.83	0.28	0.73
rs28707318	INPP4B	chr4	142362820	C	T	0.41	0.91	12.61	0.76	0.94	0.98	0.18	0.82	0.00	0.80	0.28	0.73
rs72741269	PLXNA2	chr1	208248349	C	A	0.41	0.91	16.37	0.91	0.90	0.82	0.16	0.85	0.50	0.62	0.29	0.70
rs9951973	RBBP8	chr18	22974353	C	G	0.35	0.91	13.71	0.77	0.87	0.75	0.22	0.75	0.52	0.62	0.21	0.78
rs6771681	ROBO1	chr3	79621484	T	A	0.33	0.91	9.63	0.67	0.89	0.78	0.19	0.75	0.26	0.82	0.23	0.76
rs922782	RORA	chr15	60777984	G	T	0.32	0.91	14.18	0.83	0.76	0.54	0.21	0.75	0.39	0.82	0.28	0.73
rs10020604	SGMS2	chr4	107868042	A	G	0.32	0.91	12.24	0.72	0.93	0.89	0.05	0.95	0.00	0.80	0.38	0.58
rs12493289	SLC25A20	chr3	48867728	G	A	0.33	0.91	13.26	0.79	0.93	0.96	0.10	0.87	0.31	0.83	0.34	0.65
rs13127327	SLC4A4	chr4	71474802	A	G	0.34	0.91	12.50	0.72	0.89	0.78	0.16	0.85	0.17	0.83	0.32	0.71
rs62022768	THSD4	chr15	71555960	C	T	0.35	0.91	11.31	0.67	0.91	0.89	0.18	0.85	0.00	0.80	0.24	0.78
rs1029465	THSD7A	chr7	11472487	T	G	0.32	0.91	14.09	0.83	0.81	0.54	0.18	0.85	0.11	0.83	0.26	0.73
rs12447267	WVOX	chr16	78626873	A	G	0.35	0.91	11.69	0.69	0.91	0.82	0.16	0.85	0.55	0.59	0.19	0.79
rs6009961	ACR	chr22	50744057	A	G	0.46	0.91	13.92	0.83	0.75	0.62	0.65	0.67	0.11	0.83	0.40	0.53
rs34863395	ADCY6	chr12	48772984	C	T	0.52	0.91	16.61	0.93	0.64	0.67	0.94	0.91	0.31	0.83	0.28	0.73
rs8080422	ASIC2	chr17	33539314	T	G	0.48	0.91	14.76	0.86	0.59	0.75	0.94	0.91	0.00	0.80	0.36	0.62
rs34314793	ASIC2	chr17	33938117	C	G	0.50	0.91	13.28	0.79	0.74	0.62	0.93	0.91	0.19	0.83	0.39	0.56
rs4816342	BACH1	chr21	29223048	C	T	0.47	0.91	9.98	0.67	0.65	0.65	0.96	0.92	0.43	0.77	0.27	0.74
rs12937232	BRIP1	chr17	61742809	C	A	0.42	0.91	10.44	0.69	0.68	0.65	0.65	0.70	0.46	0.68	0.34	0.65
rs2518872	CFTR	chr7	117574319	G	C	0.46	0.91	13.18	0.79	0.74	0.62	0.70	0.70	0.57	0.55	0.37	0.59
rs3019606	CPT1A	chr11	68817888	A	G	0.41	0.91	10.87	0.67	0.68	0.66	0.64	0.67	0.27	0.82	0.41	0.53
rs425768	CUX1	chr7	101967234	T	C	0.45	0.91	9.93	0.67	0.67	0.66	0.76	0.76	0.15	0.83	0.25	0.75
rs16954731	DCC	chr18	52346271	C	G	0.46	0.91	11.31	0.67	0.61	0.68	0.95	0.91	0.36	0.81	0.28	0.73
rs2835783	DYRK1A	chr21	37530219	T	A	0.37	0.91	11.26	0.67	0.73	0.62	0.46	0.58	0.00	0.80	0.27	0.74
rs67999841	HIVEP3	chr1	41599751	T	A	0.40	0.91	11.55	0.67	0.70	0.62	0.63	0.67	0.11	0.83	0.22	0.75
rs889295	ITGA1	chr5	52814876	A	G	0.35	0.91	9.71	0.67	0.64	0.67	0.42	0.58	0.52	0.62	0.18	0.79
rs7644014	KALRN	chr3	124374600	C	A	0.37	0.91	9.01	0.67	0.75	0.62	0.44	0.58	0.39	0.82	0.22	0.75
rs7998989	PCCA	chr13	100159063	C	G	0.46	0.91	11.15	0.67	0.71	0.62	0.86	0.86	0.06	0.80	0.17	0.79

(continued)

rsID	Gene	Chrom	Pos	REF	ALT	PredictSNP2		CADD		DANN		FATHMM-MKL		FunSeq2		GWAVA	
						score	conf	score	conf	score	conf	score	conf	score	conf	score	conf
rs4724730	RNF216	chr7	5786200	C	T	0.44	0.91	11.61	0.69	0.71	0.62	0.81	0.80	0.00	0.80	0.12	0.79
rs329825	ROBO1	chr3	78913726	T	C	0.54	0.91	16.92	0.95	0.69	0.65	0.94	0.91	0.45	0.68	0.22	0.75
rs4748311	RSU1	chr10	16694477	C	A	0.40	0.91	13.96	0.83	0.51	0.85	0.94	0.91	0.38	0.82	0.36	0.62
rs10926978	SDCCAG8	chr1	243268352	T	C	0.42	0.91	10.25	0.67	0.65	0.65	0.74	0.76	0.49	0.62	0.45	0.53
rs1228937	SEMA3A	chr7	84492220	G	C	0.51	0.91	15.07	0.86	0.72	0.62	0.88	0.86	0.33	0.81	0.44	0.53
rs10475467	SEMA5A	chr5	9415605	G	A	0.45	0.91	12.29	0.72	0.60	0.72	0.89	0.86	0.19	0.83	0.41	0.53
rs4756029	SLC35C1	chr11	45815243	T	C	0.42	0.91	9.55	0.67	0.66	0.70	0.71	0.76	0.00	0.80	0.11	0.79
rs10893501	ST3GAL4	chr11	126381138	A	G	0.33	0.91	11.20	0.67	0.59	0.75	0.55	0.62	0.44	0.77	0.21	0.78
rs7653848	TP63	chr3	189868063	C	T	0.42	0.91	9.87	0.67	0.61	0.72	0.85	0.82	0.35	0.81	0.34	0.65
rs6531721	UGDH	chr4	39495771	C	A	0.34	0.91	10.63	0.67	0.64	0.67	0.44	0.58	0.19	0.83	0.38	0.58
rs2729615	ZFH3	chr16	73713642	T	A	0.54	0.91	16.98	0.95	0.72	0.62	0.97	0.95	0.19	0.83	0.41	0.53
rs2957454	ZFPM2	chr8	105356118	T	G	0.46	0.91	10.85	0.67	0.74	0.62	0.90	0.88	0.19	0.83	0.30	0.68
rs2165646	ZFPM2	chr8	105642045	G	A	0.47	0.91	15.26	0.88	0.59	0.75	0.89	0.86	0.19	0.83	0.42	0.53
rs8085984	DOK6	chr18	69608073	A	G	0.33	0.91	8.42	0.76	0.84	0.63	0.83	0.82	0.13	0.83	0.25	0.75
rs7851602	GNE	chr9	36235624	A	C	0.33	0.91	7.21	0.76	0.90	0.82	0.46	0.58	0.50	0.62	0.19	0.79
rs60055270	CACNA1B	chr9	138115389	C	T	0.18	0.73	9.26	0.67	0.79	0.54	0.07	0.93	0.52	0.62	0.52	0.67
rs1859143	COL25A1	chr4	109302164	C	T	0.28	0.73	8.78	0.67	0.92	0.89	0.08	0.89	0.29	0.82	0.53	0.78
rs1043313	GNE	chr9	36214974	G	A	0.17	0.73	8.79	0.67	0.76	0.54	0.08	0.91	0.31	0.83	0.46	0.64
rs1304037	IL1A	chr2	112774659	T	C	0.28	0.73	10.85	0.67	0.85	0.62	0.19	0.75	0.41	0.77	0.47	0.64
rs161058	PEX3	chr6	143485139	A	G	0.22	0.73	11.98	0.69	0.77	0.54	0.14	0.91	0.21	0.83	0.47	0.64
rs331162	ROBO1	chr3	78858426	T	A	0.22	0.73	10.54	0.67	0.88	0.75	0.11	0.85	0.45	0.68	0.49	0.64
rs200886825	ATG5	chr6	106246396	A	T	-0.46	0.77	2.22	0.83	0.77	0.54	0.19	0.77	0.96	0.68	0.52	0.67
rs11642413	CDH1	chr16	68756491	G	A	-0.47	0.77	0.91	0.86	0.81	0.54	0.21	0.75	1.66	0.62	0.49	0.64
rs6770844	CLDN11	chr3	170433024	T	C	-0.43	0.77	7.44	0.76	0.83	0.56	0.22	0.75	0.93	0.68	0.53	0.78
rs58691769	FYN	chr6	111837865	C	T	-0.47	0.77	7.13	0.76	0.83	0.56	0.09	0.89	1.99	0.67	0.46	0.64
rs7776191	FYN	chr6	111862031	T	G	-0.48	0.77	6.17	0.76	0.79	0.54	0.10	0.85	1.18	0.75	0.50	0.64
rs2076316	GPLD1	chr6	24489513	A	C	-0.45	0.77	6.31	0.76	0.81	0.54	0.13	0.89	1.02	0.68	0.73	0.86
rs2371494	IL23A	chr12	56334216	G	A	-0.39	0.74	5.51	0.82	0.86	0.69	0.22	0.75	1.77	0.64	0.48	0.64

(continued)

rsID	Gene	Chrom	Pos	REF	ALT	PredictSNP2		CADD		DANN		FATHMM-MKL		FunSeq2		GWAVA	
						score	conf	score	conf	score	conf	score	conf	score	conf	score	conf
rs2720006	KCNMA1	chr10	77637916	G	A	-0.25	0.73	8.07	0.76	0.91	0.89	0.18	0.82	1.12	0.74	0.51	0.65
rs6581616	LEMD3	chr12	65172173	A	G	-0.45	0.77	6.68	0.76	0.78	0.54	0.15	0.91	1.48	0.65	0.52	0.67
rs8019656	SLC7A7	chr14	22815939	T	C	-0.46	0.77	8.28	0.76	0.82	0.56	0.08	0.89	1.94	0.67	0.62	0.82

rsID= reference SNP ID assigned by dbSNP; Chrom= chromosome; Pos= start coordinate of the variant; REF= reference allele; ALT= alternative allele; conf= confidence

Table F.4: Remaining compound heterozygous variants and their pathogenicity predictions using PredictSnP2.

rsID	Gene	Chrom	Pos	REF	ALT	PredictSNP2		CADD		DANN		FATHMM-MKL		FunSeq2		GWAVA	
						score	conf	score	conf	score	conf	score	conf	score	conf	score	conf
rs56164415	BDNF	chr11	27700188	G	A	1,00	0,97	22,70	0,98	0,95	1,00	0,93	0,91	1,11	0,76	0,83	0,86
rs116336234	BRAF	chr7	140726391	T	A	1,00	0,97	14,66	0,86	0,87	0,69	0,89	0,86	1,57	0,62	0,54	0,78
rs3829814	BRAF	chr7	140726457	A	G	1,00	0,97	14,49	0,85	0,91	0,89	0,93	0,91	1,57	0,62	0,65	0,82
rs117460857	CACNA1A	chr19	13599970	G	A	1,00	0,97	9,73	0,67	0,77	0,54	0,96	0,92	1,54	0,62	0,82	0,86
rs41272781	COL2A1	chr12	47973112	G	A	1,00	0,97	14,13	0,83	0,86	0,69	0,97	0,92	1,40	0,69	0,53	0,78
,	COL2A1	chr12	47973325	A	G	1,00	0,97	17,01	0,95	0,77	0,54	0,97	0,94	1,22	0,74	0,62	0,82
rs11551376	CYFIP2	chr5	157266547	G	A	1,00	0,97	15,64	0,87	0,92	0,89	0,90	0,86	2,20	0,67	0,57	0,76
rs417309	DGCR8	chr22	20111021	G	A	1,00	0,97	14,84	0,86	0,79	0,54	0,94	0,91	2,37	0,67	0,50	0,64
rs3738493	ESRRG	chr1	216505524	G	A	1,00	0,97	12,66	0,76	0,78	0,54	0,49	0,64	1,01	0,68	0,68	0,86
rs36070315	FANCD2	chr3	10063832	A	C	1,00	0,82	23,80	0,59	0,99	0,64	0,97	0,69	3,00	0,62	0,60	0,62
,	FANCD2	chr3	10074664	G	A	1,00	0,82	24,50	0,59	0,99	0,64	0,88	0,65	3,00	0,62	0,50	0,54
rs41258305	FGFR2	chr10	121597968	T	C	1,00	0,97	19,34	0,96	0,90	0,82	0,89	0,86	1,72	0,62	0,59	0,84
rs191030169	FUT8	chr14	65411847	C	T	1,00	0,97	12,11	0,72	0,88	0,78	0,75	0,76	1,92	0,67	0,48	0,64
rs11542051	GABPA	chr21	25769498	G	A	1,00	0,97	14,67	0,86	0,90	0,82	0,89	0,86	1,62	0,62	0,68	0,86
rs67171462	IDH1	chr2	208266043	G	C	1,00	0,97	13,55	0,77	0,78	0,54	0,96	0,92	2,84	0,67	0,47	0,64
,	KCNAB2	chr1	6046066	G	A	1,00	0,97	17,94	0,96	0,93	0,96	0,95	0,91	1,91	0,67	0,84	0,86
rs113446944	KRAS	chr12	25208045	T	C	1,00	0,97	21,50	0,98	0,90	0,78	0,94	0,91	1,75	0,64	0,50	0,64

(continued)

rsID	Gene	Chrom	Pos	REF	ALT	PredictSNP2		CADD		DANN		FATHMM-MKL		FunSeq2		GWAVA	
						score	conf	score	conf	score	conf	score	conf	score	conf	score	conf
rs113980111	KRT8	chr12	52897215	T	C	1,00	0,97	17,00	0,95	0,88	0,75	0,96	0,92	1,49	0,64	0,53	0,78
rs116829253	LRP1	chr12	57149010	A	G	1,00	0,97	19,41	0,96	0,95	1,00	0,92	0,88	3,00	0,67	0,54	0,78
rs150795346	LRP1	chr12	57185191	G	A	1,00	0,87	24,40	0,60	1,00	0,70	0,95	0,60	4,00	0,62	0,68	0,51
rs13058	MAPK1	chr22	21763213	A	C	1,00	0,97	12,34	0,72	0,81	0,54	0,93	0,91	2,42	0,67	0,59	0,84
rs11627567	MARK3	chr14	103385438	C	T	1,00	0,97	13,67	0,77	0,91	0,89	0,46	0,58	3,31	0,67	0,57	0,76
,	PAX5	chr9	37034204	G	A	1,00	0,97	22,70	0,98	0,95	1,00	0,97	0,92	3,37	0,67	0,67	0,84
rs13328226	PDE4D	chr5	59276089	G	A	1,00	0,97	20,90	0,98	0,93	0,96	0,94	0,91	1,12	0,74	0,49	0,64
rs12549853	PLEC	chr8	143946468	G	A	1,00	0,97	17,84	0,96	0,91	0,89	0,85	0,82	1,64	0,62	0,59	0,84
rs147713970	PLEC	chr8	143953830	G	A	1,00	0,97	21,00	0,98	0,94	0,98	0,84	0,82	2,62	0,67	0,64	0,80
rs3796407	PPARGC1A	chr4	23883088	G	A	1,00	0,97	21,70	0,98	0,91	0,89	0,98	0,97	1,77	0,64	0,63	0,80
rs181234898	PTEN	chr10	87965825	C	T	1,00	0,97	16,61	0,93	0,94	0,98	0,98	0,97	1,09	0,76	0,66	0,84
rs138309082	PTEN	chr10	87966260	T	C	1,00	0,97	9,14	0,67	0,81	0,54	0,58	0,62	1,09	0,76	0,75	0,86
,	PTEN	chr10	87967417	G	A	1,00	0,97	12,87	0,76	0,77	0,54	0,96	0,92	1,06	0,72	0,47	0,64
rs147568240	RCN1	chr11	32091763	G	A	1,00	0,97	15,69	0,87	0,93	0,89	0,81	0,80	2,42	0,67	0,55	0,80
rs112212583	SEC23A	chr14	39103185	C	A	1,00	0,97	15,21	0,86	0,87	0,75	0,95	0,91	3,21	0,67	0,66	0,84
,	SERPINH1	chr11	75572087	C	G	1,00	0,97	10,76	0,67	0,83	0,63	0,49	0,64	1,17	0,75	0,47	0,64
,	SLC4A4	chr4	71568446	C	G	1,00	0,97	13,79	0,77	0,86	0,69	0,98	0,97	0,98	0,68	0,55	0,80
rs17122776	SLC7A7	chr14	22819744	A	G	1,00	0,97	16,74	0,93	0,79	0,54	0,79	0,80	3,54	0,67	0,67	0,84
rs2236135	SLC7A8	chr14	23126512	A	G	1,00	0,97	13,85	0,83	0,87	0,69	0,76	0,80	1,39	0,64	0,56	0,80
rs7032831	SMARCA2	chr9	2160692	A	G	1,00	0,97	19,10	0,96	0,88	0,75	0,99	0,98	1,75	0,64	0,48	0,64
rs1059310	SOS1	chr2	38985496	T	C	1,00	0,97	13,81	0,77	0,92	0,89	0,98	0,97	1,41	0,69	0,54	0,78
rs17416291	SPTBN1	chr2	54660091	C	T	1,00	0,97	16,97	0,95	0,80	0,54	0,57	0,62	1,30	0,68	0,62	0,82
rs117911387	STXBP1	chr9	127684557	G	A	1,00	0,97	19,77	0,96	0,76	0,54	0,99	0,98	1,19	0,74	0,54	0,78
rs142128221	TET2	chr4	105233902	A	C	1,00	0,97	13,84	0,83	0,97	1,00	0,90	0,88	3,00	0,67	0,64	0,80
rs113779084	THSD7A	chr7	11832161	G	A	1,00	0,97	15,45	0,88	0,95	0,98	0,91	0,88	1,53	0,62	0,74	0,86
,	TTN	chr2	178526801	T	C	1,00	0,97	20,10	0,96	0,88	0,78	0,99	0,98	1,35	0,66	0,54	0,78
rs2562830	TTN	chr2	178720666	A	G	1,00	0,97	19,82	0,96	0,83	0,56	0,98	0,94	1,35	0,66	0,64	0,80

(continued)

rsID	Gene	Chrom	Pos	REF	ALT	PredictSNP2		CADD		DANN		FATHMM-MKL		FunSeq2		GWAVA	
						score	conf	score	conf	score	conf	score	conf	score	conf	score	conf
rs8179187	UBE3A	chr15	25407179	T	G	1,00	0,97	18,72	0,96	0,93	0,96	0,96	0,91	1,32	0,68	0,64	0,80
,	VAC14	chr16	70781941	C	A	1,00	0,87	24,20	0,60	0,99	0,60	0,99	0,81	4,00	0,62	0,64	0,51
rs11789583	VLDLR	chr9	2644954	C	T	1,00	0,97	21,80	0,98	0,86	0,69	0,98	0,94	0,98	0,68	0,51	0,65
rs3421	VLDLR	chr9	2654048	G	A	1,00	0,97	13,23	0,79	0,87	0,69	0,96	0,92	1,16	0,75	0,64	0,80
rs12143648	WNT4	chr1	22143813	C	T	1,00	0,97	22,40	0,98	0,94	0,96	0,88	0,86	1,23	0,74	0,48	0,64
rs1047898	WNT5A	chr3	55465753	T	C	1,00	0,97	15,09	0,86	0,85	0,67	0,98	0,94	1,05	0,67	0,48	0,64
rs143794132	YWHAG	chr7	76328721	A	C	1,00	0,97	12,80	0,76	0,76	0,54	0,82	0,82	1,18	0,75	0,51	0,65
rs113497421	ZFH3	chr16	72950840	C	T	1,00	0,87	20,40	0,52	1,00	0,67	0,98	0,79	4,00	0,62	0,62	0,51
rs1139403	ACTG1	chr17	81512766	G	A	1,00	0,97	10,04	0,67	0,86	0,67	0,87	0,86	3,76	0,67	0,39	0,56
rs17819603	ARSB	chr5	78883246	C	A	1,00	0,97	13,86	0,83	0,87	0,69	0,43	0,58	1,23	0,74	0,11	0,79
rs3087519	CIITA	chr16	10929351	T	C	1,00	0,97	10,84	0,67	0,76	0,54	0,91	0,88	1,60	0,62	0,45	0,53
rs3753841	COL11A1	chr1	102914362	G	A	1,00	0,82	26,50	0,68	1,00	0,71	1,00	0,78	3,00	0,62	0,34	0,55
rs112371022	FOSB	chr19	45468266	G	C	1,00	0,97	12,60	0,76	0,85	0,62	0,87	0,86	1,09	0,76	0,42	0,53
rs3738346	GJB4	chr1	34761865	A	C	1,00	0,87	20,80	0,52	1,00	0,68	0,99	0,82	3,00	0,61	0,17	0,55
,	GMPPA	chr2	219506633	C	T	1,00	0,97	19,94	0,96	0,93	0,96	0,90	0,88	1,27	0,68	0,44	0,53
rs5742714	IGF1	chr12	102396074	C	G	1,00	0,97	13,32	0,79	0,78	0,54	0,98	0,94	1,27	0,68	0,43	0,53
rs1800191	LRP1	chr12	57196975	C	T	1,00	0,97	13,74	0,77	0,87	0,69	0,46	0,58	1,40	0,69	0,18	0,79
rs148782031	MITF	chr3	69967570	T	C	1,00	0,97	12,47	0,72	0,77	0,54	0,98	0,94	1,28	0,68	0,42	0,53
rs456551	PEX26	chr22	18089340	T	A	1,00	0,97	11,48	0,67	0,79	0,54	0,61	0,67	1,57	0,62	0,44	0,53
rs1065837	PLEC	chr8	143916167	G	A	1,00	0,97	11,44	0,67	0,89	0,78	0,91	0,88	1,64	0,62	0,43	0,53
rs6704	SERPINH1	chr11	75572608	C	A	1,00	0,97	15,01	0,86	0,93	0,96	0,64	0,67	1,17	0,75	0,45	0,53
rs2305689	VAC14	chr16	70762612	G	A	1,00	0,97	13,22	0,79	0,91	0,89	0,89	0,86	0,89	0,68	0,29	0,70
rs17055392	WNT5A	chr3	55466672	A	G	1,00	0,97	13,86	0,83	0,89	0,78	0,98	0,94	1,24	0,74	0,43	0,53
rs11858	ARID1B	chr6	157209024	T	A	1,00	0,97	12,06	0,69	0,77	0,54	0,69	0,70	0,43	0,77	0,54	0,78
,	DOCK5	chr8	25184786	A	C	1,00	0,97	10,77	0,67	0,88	0,78	0,46	0,58	0,33	0,81	0,63	0,80
rs1055259	GSTM3	chr1	109734239	T	C	1,00	0,97	14,16	0,83	0,92	0,89	0,57	0,62	0,23	0,83	0,47	0,64
rs139529706	ROBO1	chr3	79018735	A	G	1,00	0,97	17,39	0,96	0,86	0,69	0,55	0,62	0,45	0,68	0,62	0,82

(continued)



rsID	Gene	Chrom	Pos	REF	ALT	PredictSNP2		CADD		DANN		FATHMM-MKL		FunSeq2		GWAVA	
						score	conf	score	conf	score	conf	score	conf	score	conf	score	conf
,	ST3GAL4	chr11	126355773	C	G	1,00	0,97	16,18	0,91	0,92	0,89	0,58	0,62	0,25	0,83	0,63	0,80
rs111468751	DPP4	chr2	162074074	C	T	0,41	0,91	12,93	0,76	0,96	1,00	0,21	0,75	2,27	0,67	0,54	0,78
rs35782247	FANCD2	chr3	10048005	T	G	0,48	0,59	24,10	0,59	1,00	0,68	0,79	0,65	3,00	0,62	0,48	0,53
rs9847297	FGF12	chr3	192408733	C	T	0,32	0,91	14,65	0,86	0,83	0,63	0,16	0,85	1,43	0,69	0,48	0,64
rs140704996	GABPA	chr21	25735457	A	C	0,31	0,91	13,40	0,79	0,84	0,62	0,22	0,75	1,80	0,64	0,53	0,78
,	IGF1R	chr15	98649444	C	T	0,34	0,91	9,02	0,67	0,91	0,89	0,23	0,75	1,00	0,68	0,68	0,86
rs113762479	LIFR	chr5	38556583	G	C	0,32	0,91	13,14	0,79	0,92	0,89	0,10	0,87	1,01	0,68	0,53	0,78
rs112771035	ST3GAL4	chr11	126355981	C	G	0,36	0,91	13,32	0,79	0,88	0,75	0,20	0,75	1,23	0,74	0,52	0,67
rs188478638	TMEM216	chr11	61392546	G	A	0,34	0,91	9,55	0,67	0,90	0,82	0,20	0,75	1,54	0,62	0,60	0,82
rs41276924	ANPEP	chr15	89806729	A	G	0,50	0,91	12,26	0,72	0,68	0,65	0,95	0,91	2,56	0,67	0,71	0,86
,	BANF1	chr11	66004062	C	T	0,44	0,91	9,86	0,67	0,67	0,70	0,85	0,82	1,27	0,68	0,55	0,80
rs71651665	GABPA	chr21	25771524	C	T	0,47	0,91	10,06	0,67	0,64	0,67	0,97	0,92	1,81	0,64	0,47	0,64
rs7217	ITGB1	chr10	32900729	C	T	0,36	0,91	10,91	0,67	0,51	0,85	0,93	0,91	0,92	0,68	0,50	0,64
rs2116830	KCNMA1	chr10	76886778	G	T	0,33	0,91	11,55	0,67	0,51	0,85	0,81	0,80	0,94	0,68	0,64	0,80
rs113413307	NPHP4	chr1	5905690	G	C	0,13	0,82	24,60	0,63	0,99	0,60	0,98	0,74	3,00	0,61	0,58	0,51
rs3731749	TTN	chr2	178541464	C	T	0,13	0,82	22,30	0,52	0,99	0,62	0,98	0,80	3,00	0,61	0,28	0,51
rs55801134	TTN	chr2	178575436	C	G	0,48	0,59	21,20	0,59	0,95	0,65	0,99	0,76	3,00	0,62	0,46	0,51
rs16971436	ZFH3	chr16	72958864	T	G	0,45	0,59	22,90	0,58	0,93	0,69	0,95	0,63	4,00	0,67	0,66	0,65
rs77984596	PCLO	chr7	82756284	G	T	1,00	0,97	13,18	0,79	0,77	0,54	0,96	0,92	0,29	0,82	0,31	0,70
rs12160860	PEX26	chr22	18093175	A	G	1,00	0,97	11,03	0,67	0,78	0,54	0,83	0,82	0,15	0,83	0,31	0,70
rs5992997	PEX26	chr22	18093623	T	G	1,00	0,97	13,46	0,77	0,94	0,96	0,89	0,86	0,15	0,83	0,28	0,73
rs467924	PEX26	chr22	18101128	C	A	1,00	0,97	10,61	0,67	0,89	0,78	0,81	0,80	0,33	0,81	0,36	0,62
rs112981680	SEMA5A	chr5	9042861	C	A	1,00	0,97	16,83	0,95	0,83	0,56	0,97	0,92	0,00	0,80	0,42	0,53
rs8176799	TPH1	chr11	18019036	T	C	1,00	0,97	12,91	0,76	0,88	0,78	0,93	0,91	0,13	0,83	0,26	0,73
rs6145	VLDLR	chr9	2651962	G	T	1,00	0,97	10,96	0,67	0,81	0,54	0,63	0,67	0,36	0,81	0,45	0,53
rs11549222	ACTG1	chr17	81511519	G	A	0,52	0,93	16,79	0,58	0,95	0,87	0,26	0,73	3,00	1,00	0,20	0,60
rs9273455	HLA-DQB1	chr6	32660053	C	T	0,32	0,91	9,09	0,67	0,88	0,78	0,19	0,77	1,20	0,74	0,37	0,59

(continued)

rsID	Gene	Chrom	Pos	REF	ALT	PredictSNP2		CADD		DANN		FATHMM-MKL		FunSeq2		GWAVA	
						score	conf	score	conf	score	conf	score	conf	score	conf	score	conf
rs1054693	JMJD1C	chr10	63465690	C	T	0,41	0,91	12,21	0,72	0,97	1,00	0,17	0,85	1,99	0,67	0,45	0,53
rs2269682	SLC11A2	chr12	51024453	G	A	0,32	0,91	9,48	0,67	0,88	0,78	0,21	0,75	0,95	0,68	0,30	0,68
rs12552433	SMARCA2	chr9	2043406	C	T	0,31	0,91	13,19	0,79	0,83	0,56	0,21	0,75	3,35	0,67	0,41	0,53
rs111572677	ESR2	chr14	64294244	C	A	0,41	0,91	11,76	0,69	0,65	0,65	0,65	0,70	2,76	0,67	0,37	0,59
rs6988932	EXT1	chr8	117798373	G	A	0,45	0,91	10,51	0,67	0,62	0,71	0,90	0,88	0,98	0,68	0,43	0,53
rs1041403	NRIP1	chr21	14963194	C	T	0,50	0,91	14,46	0,85	0,70	0,62	0,88	0,86	1,50	0,64	0,40	0,53
rs199720608	PLEC	chr8	143921332	T	C	0,45	0,59	23,10	0,58	0,96	0,63	0,88	0,65	3,00	0,62	0,25	0,60
rs7002152	PLEC	chr8	143925888	T	C	0,32	0,91	9,72	0,67	0,53	0,85	0,75	0,76	1,02	0,68	0,25	0,75
rs112895108	OBSCN	chr1	228208104	G	A	0,36	0,91	12,30	0,72	0,91	0,89	0,18	0,82	0,37	0,82	0,54	0,78
rs113542146	OBSCN	chr1	228208120	C	T	0,38	0,91	10,50	0,67	0,93	0,96	0,22	0,75	0,37	0,82	0,53	0,78
rs559428	SLC35A3	chr1	100026912	G	A	0,48	0,91	16,65	0,93	0,57	0,79	0,93	0,88	0,22	0,83	0,47	0,64
rs10800841	SYT2	chr1	202596308	T	C	0,32	0,91	7,69	0,76	0,76	0,54	0,78	0,80	0,39	0,82	0,58	0,80
rs55742743	TTN	chr2	178537015	C	T	0,07	0,63	20,40	0,52	0,97	0,69	0,98	0,80	3,00	0,61	0,64	0,51
rs16866406	TTN	chr2	178592420	G	A	-0,08	0,65	19,16	0,52	0,92	0,83	0,97	0,68	3,00	0,61	0,59	0,51
rs2288569	TTN	chr2	178593864	C	T	0,02	0,63	22,20	0,51	0,95	0,81	1,00	0,83	3,00	0,61	0,57	0,50
rs1047483	ITGA1	chr5	52953782	A	T	0,33	0,91	13,84	0,83	0,85	0,67	0,23	0,75	0,42	0,77	0,26	0,73
rs7933887	ST3GAL4	chr11	126405869	C	A	0,35	0,91	11,54	0,67	0,91	0,89	0,17	0,89	0,25	0,83	0,36	0,62
rs74713675	TPH1	chr11	18018225	T	C	0,34	0,91	9,79	0,67	0,90	0,78	0,18	0,85	0,13	0,83	0,29	0,70
rs28969414	CYP2B6	chr19	41017171	T	A	0,47	0,91	9,75	0,67	0,68	0,65	0,88	0,86	0,00	0,80	0,29	0,70
rs62357176	ITGA1	chr5	52958655	T	C	0,38	0,91	9,71	0,67	0,71	0,62	0,49	0,64	0,42	0,77	0,30	0,68
rs3183950	PAPPA2	chr1	176844641	A	T	0,30	0,91	4,74	0,80	0,85	0,62	0,85	0,82	0,19	0,83	0,32	0,71
rs61730222	TG	chr8	133096215	G	C	1,00	0,87	18,65	0,56	1,00	0,66	0,90	0,56	2,00	0,62	0,22	0,52
rs6748924	WDR35	chr2	19913460	A	G	0,36	0,91	6,39	0,76	0,76	0,54	0,95	0,91	0,31	0,83	0,44	0,53
rs187458	CYFIP2	chr5	157266552	C	G	-0,43	0,77	7,74	0,76	0,78	0,54	0,16	0,85	1,57	0,62	0,60	0,82
rs112836371	KRAS	chr12	25204997	A	C	-0,43	0,77	5,36	0,80	0,83	0,56	0,18	0,82	0,94	0,68	0,49	0,64
rs2289124	NOX4	chr11	89491309	G	A	-0,22	0,73	8,47	0,76	0,94	0,96	0,15	0,85	1,05	0,67	0,58	0,80
rs188029913	ROBO1	chr3	79018954	C	A	-0,43	0,77	7,81	0,76	0,79	0,54	0,20	0,75	1,25	0,74	0,54	0,78
rs561149	SLC35A3	chr1	100034860	T	C	-0,42	0,74	8,10	0,76	0,85	0,62	0,15	0,91	1,60	0,62	0,52	0,67
rs565708	SLC35A3	chr1	100035359	C	A	-0,45	0,77	8,15	0,76	0,79	0,54	0,14	0,91	0,98	0,68	0,46	0,64

(continued)

Table F.5: Ten remaining de novo variants and their pathogenicity predictions using PredictSnP2.

rsID	Gene	Chrom	Pos	REF	ALT	PredictSNP2		CADD		DANN		FATHMM-MKL		FunSeq2		GWAVA	
						score	conf	score	conf	score	conf	score	conf	score	conf	score	conf
rs112880266	DOCK5	chr8	25184822	G	A	1,00	0,97	15,04	0,86	0,94	0,98	0,95	0,91	0,96	0,68	0,63	0,80
.	GNA13	chr17	65013650	C	A	1,00	0,97	14,32	0,85	0,79	0,54	0,96	0,91	1,43	0,69	0,70	0,86
rs58826872	TRIP12	chr2	229921921	G	C	1,00	0,97	16,56	0,93	0,89	0,78	1,00	0,98	1,49	0,64	0,66	0,84
rs113484714	CTDP1	chr18	79692148	G	A	1,00	0,97	11,42	0,67	0,92	0,89	0,93	0,91	0,93	0,68	0,22	0,75
.	COL1A2	chr7	94412657	G	T	0,13	0,82	25,30	0,71	0,99	0,60	0,98	0,74	4,00	0,62	0,62	0,51
.	STS	chrX	7204634	C	G	1,00	0,97	11,81	0,69	0,92	0,89	0,43	0,58	0,00	0,80	0,13	0,79
rs13272960	KAT6A	chr8	41930730	T	A	1,00	0,97	9,62	0,67	0,77	0,54	0,62	0,67	0,00	0,80	0,35	0,65
rs73617839	IL11	chr19	55369857	C	T	0,31	0,91	9,29	0,67	0,94	0,98	0,12	0,85	1,63	0,62	0,24	0,78
.	TCF7L2	chr10	112964811	G	A	0,41	0,91	11,34	0,67	0,58	0,75	0,93	0,88	0,55	0,59	0,29	0,70
.	PCYT1A	chr3	196250096	T	A	0,30	0,91	8,84	0,67	0,39	0,80	0,89	0,86	0,33	0,81	0,21	0,78



Table F.6: Osteogenesis imperfecta variant pathogenicity predictions and citation information.

Position	REF	ALT	Gene	impact	PredictSNP2		CADD		DANN		FATHMM-MKL		FunSeq2		GWAVA		Databases			
					score	conf	score	conf	score	conf	score	conf	score	conf	score	conf	ClinVar	dbSNP	Cited	LOVD
1761352	C	T	SERPINF1	LOW	-1.00	0.88	2.62	0.81	0.56	0.81	0.03	0.95	0.00	0.80	0.38	0.58	X	✓	X	X
1768196	C	T	SERPINF1	LOW	-1.00	0.88	2.21	0.83	0.35	0.80	0.06	0.95	0.08	0.80	0.20	0.78	X	✓	X	X
1769315	A	G	SERPINF1	LOW	1.00	0.97	9.98	0.67	0.89	0.78	0.24	0.48	0.08	0.80	0.10	0.79	X	✓	X	X
1769333	C	T	SERPINF1	LOW	-0.42	0.74	1.34	0.86	0.87	0.75	0.05	0.95	0.08	0.80	0.14	0.79	X	✓	X	X
1776714	C	T	SERPINF1	LOW	-0.51	0.93	21.20	0.79	0.86	0.90	0.40	0.72	0.00	0.93	0.18	0.60	benign	✓	X	X
22180196	A	G	BMP1	LOW	-1.00	0.88	3.27	0.80	0.71	0.62	0.13	0.87	0.37	0.82	0.17	0.79	benign	✓	X	X
22184638	G	T	BMP1	LOW	-1.00	0.88	1.40	0.86	0.37	0.80	0.08	0.89	0.56	0.59	0.27	0.74	X	✓	X	X
22185268	G	A	BMP1	LOW	-1.00	0.88	1.80	0.86	0.43	0.82	0.03	0.95	0.37	0.82	0.33	0.68	X	✓	X	X
22187607	C	T	BMP1	LOW	-0.40	0.74	5.45	0.82	0.88	0.75	0.12	0.85	0.56	0.59	0.26	0.73	X	✓	X	X
22187714	T	A	BMP1	MED	0.34	0.91	10.78	0.67	0.90	0.82	0.21	0.75	0.56	0.55	0.38	0.58	X	✓	X	X
22188910	C	T	BMP1	LOW	-0.39	0.74	7.54	0.76	0.88	0.75	0.07	0.93	3.42	0.67	0.42	0.53	X	✓	✓	X
22189593	G	A	BMP1	LOW	-1.00	0.88	4.57	0.76	0.57	0.79	0.03	0.95	0.37	0.82	0.27	0.74	X	✓	X	X
22192267	A	G	BMP1	LOW	-1.00	0.88	3.23	0.80	0.72	0.62	0.16	0.85	0.56	0.59	0.13	0.79	benign	✓	✓	X
22192927	A	C	BMP1	LOW	-1.00	0.88	0.87	0.86	0.44	0.85	0.13	0.87	0.56	0.59	0.25	0.75	X	✓	X	X
22193733	A	G	BMP1	LOW	-1.00	0.88	0.16	0.86	0.20	0.80	0.02	0.95	0.37	0.82	0.35	0.65	benign	✓	X	X
22195610	T	C	BMP1	LOW	-1.00	0.88	1.52	0.86	0.33	0.80	0.10	0.85	0.37	0.82	0.31	0.70	benign	✓	✓	X
22195791	T	C	BMP1	LOW	-1.00	0.88	0.34	0.86	0.24	0.80	0.03	0.95	0.37	0.82	0.28	0.73	benign	✓	X	X
22196684	T	C	BMP1	LOW	-1.00	0.96	15.37	0.73	0.73	0.93	0.21	0.77	0.00	0.93	0.56	0.69	benign	✓	X	X
33112754	T	A	CRTAP	LOW	-1.00	0.88	4.02	0.79	0.61	0.68	0.03	0.95	0.00	0.80	0.39	0.56	X	✓	X	X
33117420	G	A	CRTAP	LOW	-1.00	0.88	1.81	0.88	0.40	0.80	0.04	0.95	0.15	0.83	0.06	0.79	X	✓	X	X
33119661	A	C	CRTAP	LOW	-0.50	0.77	2.20	0.83	0.75	0.54	0.11	0.85	0.33	0.81	0.21	0.78	X	✓	X	X
33119673	G	T	CRTAP	LOW	-1.00	0.88	2.72	0.79	0.49	0.88	0.04	0.95	0.92	0.68	0.23	0.76	X	✓	X	X
33119863	G	A	CRTAP	LOW	-1.00	0.88	1.40	0.86	0.58	0.75	0.05	0.95	0.15	0.83	0.15	0.79	X	✓	X	X
33119912	T	C	CRTAP	LOW	0.27	0.73	11.72	0.69	0.84	0.63	0.19	0.75	0.77	0.45	0.22	0.75	X	✓	X	X
33119942	G	A	CRTAP	LOW	-1.00	0.88	4.74	0.80	0.50	0.88	0.09	0.87	0.15	0.83	0.10	0.79	X	✓	X	X
33120237	A	G	CRTAP	LOW	-1.00	0.88	3.62	0.81	0.59	0.75	0.13	0.89	0.15	0.83	0.15	0.79	X	✓	X	X
33120253	A	G	CRTAP	LOW	-1.00	0.88	4.33	0.76	0.71	0.62	0.07	0.93	0.15	0.83	0.15	0.79	X	✓	X	X

(continued)

Chrom	REF	ALT	Gene	Impact	PredictSNP2		CADD		DANN		FATHMM-MKL		FunSeq2		GWAVA		Databases			
					score	conf	score	conf	score	conf	score	conf	score	conf	score	conf	ClinVar	dbSNP	Cited	LOVD
33120406	C	T	CRTAP	LOW	-0.64	0.95	15.84	0.58	0.46	0.97	0.08	0.90	0.00	0.93	0.20	0.60	benign	✓	✓	✓
33120757	A	C	CRTAP	LOW	-1.00	0.88	1.81	0.88	0.72	0.62	0.17	0.89	0.15	0.83	0.15	0.79	X	✓	X	X
33120871	T	A	CRTAP	LOW	-1.00	0.88	2.08	0.83	0.41	0.80	0.03	0.95	0.15	0.83	0.11	0.79	X	✓	X	X
33121040	A	C	CRTAP	LOW	-0.45	0.77	8.19	0.76	0.76	0.54	0.14	0.91	0.15	0.83	0.16	0.79	X	✓	X	X
33121130	T	C	CRTAP	LOW	-1.00	0.88	1.71	0.86	0.31	0.80	0.03	0.95	0.15	0.83	0.12	0.79	X	✓	X	X
33121256	A	G	CRTAP	LOW	-1.00	0.88	3.36	0.82	0.42	0.82	0.06	0.95	0.33	0.81	0.24	0.78	X	✓	X	X
33122679	G	C	CRTAP	LOW	-1.00	0.88	0.49	0.86	0.13	0.80	0.03	0.95	0.15	0.83	0.17	0.79	X	✓	X	X
33123488	C	T	CRTAP	LOW	-1.00	0.88	4.58	0.80	0.66	0.70	0.05	0.95	0.15	0.83	0.14	0.79	X	✓	X	X
33123944	T	A	CRTAP	LOW	-1.00	0.88	3.69	0.81	0.36	0.80	0.03	0.95	0.15	0.83	0.18	0.79	X	✓	X	X
33124940	T	C	CRTAP	LOW	-1.00	0.88	0.71	0.86	0.39	0.80	0.02	0.95	0.15	0.83	0.35	0.65	X	✓	X	X
33126309	G	A	CRTAP	LOW	-1.00	0.88	1.34	0.86	0.64	0.67	0.05	0.95	0.15	0.83	0.33	0.68	X	✓	X	X
33126361	T	C	CRTAP	LOW	-1.00	0.88	2.45	0.82	0.44	0.85	0.05	0.95	0.15	0.83	0.30	0.68	X	✓	X	X
33139705	T	G	CRTAP	LOW	-1.00	0.88	2.04	0.83	0.24	0.80	0.06	0.95	0.33	0.81	0.24	0.78	X	✓	X	X
46299128	C	A	CREB3L1	LOW	-0.42	0.74	4.15	0.79	0.66	0.70	0.35	0.48	1.83	0.64	0.45	0.53	X	✓	X	X
46301658	G	A	CREB3L1	LOW	-1.00	0.88	0.19	0.86	0.52	0.85	0.04	0.95	0.33	0.81	0.31	0.70	X	✓	X	X
46302720	C	T	CREB3L1	LOW	-1.00	0.88	1.62	0.86	0.44	0.85	0.05	0.95	0.33	0.81	0.33	0.68	X	✓	X	X
46305552	C	T	CREB3L1	LOW	-0.47	0.77	0.35	0.86	0.83	0.63	0.15	0.91	0.52	0.62	0.44	0.53	X	✓	X	X
46306017	G	A	CREB3L1	LOW	-1.00	0.88	0.27	0.86	0.74	0.62	0.04	0.95	0.33	0.81	0.28	0.73	X	✓	X	X
50184005	G	A	COL1A1	LOW	-1.00	0.88	6.87	0.76	0.63	0.71	0.17	0.89	0.82	0.45	0.50	0.64	X	✓	✓	X
50184032	A	G	COL1A1	LOW	-1.00	0.88	2.50	0.81	0.58	0.79	0.22	0.75	1.44	0.69	0.46	0.64	X	✓	X	X
50184491	A	G	COL1A1	LOW	1.00	0.97	9.06	0.67	0.83	0.63	0.27	0.48	1.78	0.64	0.40	0.53	benign, VUS	✓	✓	X
50184758	G	A	COL1A1	LOW	1.00	0.97	12.82	0.76	0.90	0.82	0.84	0.82	0.82	0.45	0.51	0.65	benign	✓	✓	X
50199874	C	A	COL1A1	LOW	-0.55	0.71	21.70	0.57	0.94	0.67	0.84	0.65	0.00	0.81	0.36	0.48	benign	✓	✓	✓
75563105	G	T	SERPINH1	LOW	-0.42	0.74	4.78	0.80	0.74	0.62	0.36	0.48	1.16	0.75	0.22	0.75	X	✓	✓	X
75563585	T	C	SERPINH1	LOW	-0.45	0.77	6.12	0.76	0.77	0.54	0.14	0.89	1.16	0.75	0.26	0.73	X	✓	X	X
75566712	C	G	SERPINH1	LOW	-0.50	0.93	11.18	0.88	0.65	0.96	0.91	0.77	0.00	0.93	0.39	0.52	benign	✓	X	✓
75568801	C	T	SERPINH1	LOW	-1.00	0.96	12.78	0.77	0.85	0.90	0.04	0.96	0.00	0.93	0.17	0.60	benign	✓	X	✓
75571837	G	A	SERPINH1	LOW	-0.57	0.93	11.63	0.83	0.86	0.90	0.80	0.57	0.00	0.93	0.22	0.59	benign	✓	X	✓

(continued)

Chrom	REF	ALT	Gene	Impact	PredictSNP2		CADD		DANN		FATHMM-MKL		FunSeq2		GWAVA		Databases			
					score	conf	score	conf	score	conf	score	conf	score	conf	score	conf	ClinVar	dbSNP	Cited	LOVD
75572087	C	G	SERPINH1	LOW	1.00	0.97	10.76	0.67	0.83	0.63	0.49	0.64	1.17	0.75	0.47	0.64	X	X	X	X
75572608	C	A	SERPINH1	LOW	1.00	0.97	15.01	0.86	0.93	0.96	0.64	0.67	1.17	0.75	0.45	0.53	benign	✓	X	X
81496028	G	A	TENT5A	LOW	-1.00	0.88	2.32	0.82	0.42	0.82	0.08	0.91	0.39	0.82	0.14	0.79	X	✓	X	X
81497987	C	T	TENT5A	LOW	-1.00	0.88	2.86	0.79	0.62	0.68	0.09	0.89	0.39	0.82	0.25	0.75	X	✓	X	X
81500692	C	A	TENT5A	LOW	-0.39	0.74	7.89	0.76	0.85	0.67	0.16	0.85	0.39	0.82	0.28	0.73	X	✓	X	X
81500960	G	C	TENT5A	LOW	-1.00	0.88	2.70	0.79	0.51	0.85	0.11	0.85	0.39	0.82	0.28	0.73	X	✓	X	X
81503605	G	T	TENT5A	LOW	-1.00	0.88	0.09	0.86	0.21	0.80	0.04	0.95	0.39	0.82	0.22	0.75	X	✓	X	X
81506683	G	A	TENT5A	LOW	-1.00	0.88	0.19	0.86	0.23	0.80	0.04	0.95	0.39	0.82	0.22	0.75	X	✓	X	X
81509077	G	T	TENT5A	LOW	-1.00	0.88	0.06	0.86	0.39	0.80	0.03	0.95	0.39	0.82	0.22	0.75	X	✓	X	X
81515679	G	T	TENT5A	LOW	-1.00	0.88	2.66	0.81	0.74	0.62	0.15	0.91	1.37	0.64	0.27	0.74	X	✓	X	X
81529262	T	A	TENT5A	LOW	-1.00	0.88	5.43	0.82	0.74	0.62	0.16	0.85	0.39	0.82	0.24	0.78	X	✓	X	X
81542439	G	A	TENT5A	LOW	-1.00	0.88	0.15	0.86	0.26	0.80	0.02	0.95	0.39	0.82	0.18	0.79	X	✓	X	X
81577968	G	A	TENT5A	LOW	-1.00	0.88	5.48	0.82	0.56	0.81	0.10	0.84	0.39	0.82	0.29	0.70	X	✓	X	X
81588567	A	T	TENT5A	LOW	-1.00	0.88	0.87	0.86	0.39	0.80	0.12	0.85	0.39	0.82	0.37	0.59	X	✓	X	X
81636889	A	G	TENT5A	LOW	-1.00	0.88	1.34	0.86	0.42	0.82	0.09	0.87	0.57	0.55	0.16	0.79	X	✓	X	X
81639572	C	T	TENT5A	LOW	-1.00	0.88	0.13	0.86	0.55	0.82	0.01	0.95	0.39	0.82	0.28	0.73	X	✓	X	X
81669858	C	T	TENT5A	LOW	-1.00	0.88	3.23	0.80	0.71	0.62	0.11	0.85	0.39	0.82	0.28	0.73	X	✓	X	X
81683814	C	A	TENT5A	LOW	-0.34	0.74	9.80	0.67	0.75	0.62	0.14	0.91	1.38	0.64	0.24	0.78	X	✓	X	X
94401587	T	C	COL1A2	LOW	-0.60	0.93	1.95	0.95	0.48	0.97	0.83	0.57	0.00	0.93	0.33	0.56	benign	✓	✓	✓
94412657	G	T	COL1A2	MED	0.13	0.82	25.30	0.71	0.99	0.60	0.98	0.74	4.00	0.62	0.62	0.51	X	X	X	X
94429083	T	C	COL1A2	LOW	-1.00	0.88	1.03	0.86	0.47	0.86	0.15	0.85	0.53	0.62	0.24	0.78	benign	✓	✓	X
105779069	G	A	TMEM38B	LOW	-1.00	0.88	2.08	0.83	0.68	0.65	0.04	0.95	0.00	0.80	0.42	0.53	X	✓	X	X
105779481	G	T	TMEM38B	LOW	-1.00	0.88	0.95	0.86	0.68	0.66	0.04	0.95	0.00	0.80	0.40	0.53	X	✓	X	X

## Appendix G: Technical documentation

### 1. Python scripts

#### I. OI candidate gene mutation extraction

```
import sys, os
```

```
vcf_files = open("/home/user/Desktop/MSc_SAF_2018/Data_files/new_results/working_results/comp/  
snpeff_ann_variants2_2.vcf", "r")
```

This line opens a file for view whatever content is inside.

```
OI_candidate_gene = open("/home/user/Desktop/MSc_SAF_2018/Data_files/new_results/  
Start_a_fresh(29_October_2019)/OI_candidate_genes/OI_candidate_genes_S2_2.txt", "w")
```

```
can_genes = []
```

This line creates an empty list

This line create and opens a file for writing new information to.

```
for line in vcf_files:  
    if '##' not in line:  
        e1 = line.split('\t')[0:5]  
        e1_1 = line.replace(';', '\t').replace('|', '\t').split('\t')[-12:]  
        can_genes.append(e1+e1_1)  
vcf_files.close()
```

```
for ext_lines in can_genes:
```

```
    if ext_lines[9] == 'MESD' or ext_lines[10] == 'MESD':
```

```
        print(ext_lines, sep='\t', file=OI_candidate_gene)
```

```
    elif ext_lines[9] == 'CRTAP' or ext_lines[10] == 'CRTAP':
```

```
        print(ext_lines, sep='\t', file=OI_candidate_gene)
```

```
    elif ext_lines[9] == 'P3H1' or ext_lines[10] == 'P3H1':
```

```
        print(ext_lines, sep='\t', file=OI_candidate_gene)
```

```
    elif ext_lines[9] == 'PPIB' or ext_lines[10] == 'PPIB':
```

```
        print(ext_lines, sep='\t', file=OI_candidate_gene)
```

```
    elif ext_lines[9] == 'SERPINH1' or ext_lines[10] == 'SERPINH1':
```

```
        print(ext_lines, sep='\t', file=OI_candidate_gene)
```

```
    elif ext_lines[9] == 'FKBP10' or ext_lines[10] == 'FKBP10':
```

```
        print(ext_lines, sep='\t', file=OI_candidate_gene)
```

```
    elif ext_lines[9] == 'PLOD2' or ext_lines[10] == 'PLOD2':
```

```
        print(ext_lines, sep='\t', file=OI_candidate_gene)
```

```
    elif ext_lines[9] == 'BMP1' or ext_lines[10] == 'BMP1':
```

```
        print(ext_lines, sep='\t', file=OI_candidate_gene)
```

```
    elif ext_lines[9] == 'SP7' or ext_lines[10] == 'SP7':
```

```
        print(ext_lines, sep='\t', file=OI_candidate_gene)
```

```
    elif ext_lines[9] == 'TMEM38B' or ext_lines[10] == 'TMEM38B':
```

```
        print(ext_lines, sep='\t', file=OI_candidate_gene)
```

```
    elif ext_lines[9] == 'WNT1' or ext_lines[10] == 'WNT1':
```

```
        print(ext_lines, sep='\t', file=OI_candidate_gene)
```

```
    elif ext_lines[9] == 'CREB3L1' or ext_lines[10] == 'CREB3L1':
```

```
        print(ext_lines, sep='\t', file=OI_candidate_gene)
```

```
    elif ext_lines[9] == 'SPARC' or ext_lines[10] == 'SPARC':
```

```
        print(ext_lines, sep='\t', file=OI_candidate_gene)
```

```
    elif ext_lines[9] == 'FAM46A' or ext_lines[10] == 'FAM46A':
```

```
        print(ext_lines, sep='\t', file=OI_candidate_gene)
```

1<sup>st</sup>: the code goes through each line in the file called vcf\_files.

2<sup>nd</sup>: if the a line does not have a "##" in it

3<sup>rd</sup>: it will take that line and using a tab delimiter split it into columns. Only retaining the first 5 columns.

4<sup>th</sup>: if the line has a ';' in it, replace that with a tab, also if the line has '|' sign replace thatwith a tab as well and just retain the last 12 columns fromthe back of the line for every line in the file.

5<sup>th</sup>: save all these lines from the file into the empty 'can\_gene' list.

1<sup>st</sup>: the code goes through eachline in the new populated list.

It then checks to see in each of the rows, whether columns 9 and 10 has the word in the single quotes in it.

If so, it then saves these rows into the empty 'OI\_candidate\_gene' file

## II. SNPs and InDel separation

These lines open file for reading existing information and writing information to new empty files, indicated by the circled 'r' and 'w'

```
import sys, os
```

```
mutations = open("/home/user/Desktop/MSc_SAF_2018/Data_files/new_results/  
Start_a_fresh(29_October_2019)/OI_candidate_genes/all_match_dup_removed.txt", "r")
```

```
point_mutation = open('/home/user/Desktop/MSc_SAF_2018/Data_files/new_results/  
Start_a_fresh(29_October_2019)/All_moderate/point_mutations.txt', 'w')  
insertion = open('/home/user/Desktop/MSc_SAF_2018/Data_files/new_results/  
Start_a_fresh(29_October_2019)/OI_candidate_genes/insertions.txt', 'w')  
deletion = open('/home/user/Desktop/MSc_SAF_2018/Data_files/new_results/  
Start_a_fresh(29_October_2019)/All_moderate/deletions.txt', 'w')
```

```
for line in mutations:  
    e1 = line.split(',')  
    if len(str(e1[2])) == 1 and len(str(e1[3])) == 1:  
        print(e1, sep='\t', file=point_mutation)  
    elif len(str(e1[2])) > 1 and len(str(e1[3])) == 1:  
        print(e1, sep='\t', file=deletion)  
    elif len(str(e1[2])) == 1 and len(str(e1[3])) > 1:  
        print(e1, sep='\t', file=insertion)
```

1<sup>st</sup>: the code reads each row of information in the 'mutations' file.

2<sup>nd</sup>: it then separates the columns by a comma

3<sup>rd</sup>: If the length of a 'word' or 'character' in the 2<sup>nd</sup> column is 1 and the length of the 'character' in the 3<sup>rd</sup> column is also 1. Then save this in the point mutation file.

The rest of the lines does the same as the 3<sup>rd</sup>. for the following conditions.

If the length of the 'character' in the 2 is more than 1 and that of the 3<sup>rd</sup> is, then save this into the deletion file.

The opposite of the above condition is true for saving variants into the insertion file.

UNIVERSITY of the  
WESTERN CAPE



## 2. AWK commands

### i. SAHGP Filter

The command below goes into each file and checks if the first two columns of the first file and the first two columns of the second file is the same. If the columns match the matching files are printed to a third file (**match.sh**).

```
awk 'NR=FNR {a[$1FS$2]=$0;next} a[$1FS$2] {print a[$1FS$2],"->" , $0}' inputfile1.txt  
inputfile2.txt > outputfile.txt
```

### ii. Phenotype filter

The command below opens the file with the list of phenotypes. The command then opens a second file containing the genes, disease and phenotype information for that specific disease and gene. Then the command checks to see if any of the words(phenotypes) in the first file matches that of the second. If the command finds a match, it then prints out these matches to a third

```
awk -F "\t" 'NR=FNR {a[$1]++;next} {if (a[$3]) print a[$1], $0; else print "not found",  
$0;}' inputfile1.txt inputfile2.txt > outputfile.txt
```

