Genetic Characterisation of selected probands/kindreds with congenital heart disease

Giulia Aquilina

[redacted]

Submitted to the Faculty of Medicine and Surgery at the University of Malta in fulfilment of the requirements for the degree of Master of Science in BioChemistry.

Supervisor: Prof Nikolai Paul Pace M.D.(Melit.),Ph.D.(Melit)

Co-supervisor: Prof Jean Calleja Aguis M.D.(Melit.),Ph.D.(Lond.),F.R.C.O.G.(Lond.),F.R.C.P.I.(Dublin),M.Sc.(Clinical Embryology)(Leeds)

ACKNOWLEDGEMENTS

My first thanks are to my supervisors, Prof Nikolai P. Pace M.D. Ph.D and Prof. Jean Calleja Aguis M.D. Ph.D.,F.R.C.O.G.,F.R.C.P.I.,M.Sc. for their guidance, expertise, and invaluable feedback throughout the course of this research.

I would like to express my deepest gratitude to my family for their support throughout this journey. Their encouragement, understanding, and belief in me have provided me with the strength and motivation to persevere through challenges and pursue my academic goals.

I am also grateful to my partner for their endless love, patience, and encouragement. Their unwavering support, understanding, and belief in my abilities.

Last but not least, I would like to thank all those who have contributed as research participants. Your trust in this research project will forever be cherished and I am truly grateful for your contributions.

ABSTRACT

Congenital Heart Disease (CHD) is a rare multifactorial disease affecting around 1% of the population. CHD has a complex aetiology involving genetic, environmental, and maternal factors. The genetic factors driving CHD are also numerous, with syndromic, chromosomal, monogenic and polygenic factors implicated in its development. Within the past few decades, a minority of patients diagnosed with this disorder were known to survive until adulthood, hence research in this field is ever growing. This dissertation entitled "Genetic Characterisation of Selected Probands/Kindreds with Congenital Heart Disease" presents findings from genomic analysis of an adult proband with complex cyanotic CHD complicated by pulmonary hypertension and Eisenmenger syndrome. Trio whole exome sequencing was performed on the proband and unaffected parents followed by bioinformatic analysis to identify deleterious variants segregating using different disease models (dominant with reduced penetrance, recessive, de-novo) that could possibly explain the observed phenotype. Variant filtering and prioritisation was performed with reference to a large gene panel (n = 635 loci) derived from the literature having possible association with CHD. A *de novo* pathogenic missense variant in *BMPR2* p.Arg491Trp rs137852746 was identified in the proband that was absent from an ethnically matched reference cohort. Molecular modelling was applied to evaluate physiochemical properties of this missense variant located in the kinase domain. This molecular modelling provided evidence for a deleterious effect of this variant on protein stability. Our findings identify a monogenic driver for pulmonary hypertension in the proband that holds significant implications for clinical practice. This locus has been previously associated with pulmonary hypertension and CHD. *BMPR2* has been associated with microvascular and aortic endothelial cell development along with vascular development, whilst also being the primary gene associated to the development of PAH in humans.

Key Words:

Congenital Heart Disease, Pulmonary Arterial Hypertension, Genetic research, Mendelian genetic model, *BMPR2*.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ACMG	American College of Medical Genetics
AD	Autosomal dominant
AI	Artificial intelligence
aiVCE	Artificial intelligence-based variant classification engine
AMP	Association for Molecular Pathology
APVR	Anomalous pulmonary venous return
AR	Autosomal recessive
AS	Aortic stenosis
ASD	Atrial septal defects
AV	Atrioventricular
AVC	Atrioventricular canal
AVSD	Atrioventricular septal defects
B	Benign
BAV	Bicuspid aortic valve
BAV	Bicuspid aortic valve
BMP	Bone morphogenic protein
BMPR	Bone morphogenic protein receptor
BMPR2	Bone morphogenic protein receptor 2
CADD	Combined annotated dependent depletion
CAP	College of American pathologists
CHD	Congenital heart disease
CNV	Copy number variants
CoA	Coarctation of the Aorta
COPD	Chronic obstructive pulmonary disease
CTD	Conotruncal defects
CTEPH	Chronic thromboembolic pulmonary hypertension
Del	Deleterious
DILV	Double inlet left ventricle
DNA	Deoxyribonucleic acid
DORV	Double outlet right ventricle
ES	Eisenmenger syndrome
EtBr	Ethidium bromide
ExAC	Exome aggregation consortium
FHF	First heart field
FISH	Fluorescence in situ hybridisation
FN	False negative
FP	False positive
FPR	False positive rate
GDF	Growth and differentiation factors
GWAS	Genome-Wide association studies
HGMD	Human gene mutation database
HHT	Hereditary haemorrhagic telangiectasia
HLHS	Hypoplastic Left heart syndrome
HMM	Hidden Markov model
HOA	Hypertrophic osteoarthropathy

HRHS	Hypoplastic right heart syndrome
HTX	Heterotaxy
ILD	Interstitial lung disease
LB	Likely benign
LDR	Low density ladder
LOF	Loss of function
LP	Likely pathogenic
LV	Left ventricle
LVO	Left ventricular obstruction
MAF	Minor allele frequency
MESP1	Mesoderm posterior 1
NGS	Next generation sequencing
P	Pathogenic
PAH	Pulmonary arterial hypertension
PAH	Phenylalanine hydroxylase
PAPVR	Partial anomalous pulmonary venous return
PCR	Polymerase chain reaction
PDA	Patent ductus arteriosus
PDB	Protein data bank
PH	Pulmonary hypertension
PKU	Phenylketonuria
PPH	Primary pulmonary hypertension
PTM	Post-translational modification
PVD	Pulmonary vascular disease
PVH	Pulmonary venous hypertension
PVS	Pulmonary vein stenosis
RF	Random forest
RV	Right ventricle
SHF	Second heart field
SIFT	Sorting intolerant from tolerant
SIT	Situs inversus totalis
SMAD	Suppressor of Mothers Against Decapentaplegic
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SV	Single variant
SVM	Support vector machine
TF	Transcription factor
TGA	Transposition o the great arteries
TGF-β	Transforming Growth Factor β
TN	True negative
TOF	Tetralogy of Fallot
TP	True positive
TPR	True positive rate
UTR	Untranslated region
VAF	Variant allele frequency
VCF	Variant call file

VEGF	Vascular endothelial growth factor
VSD	Ventricular septal defects
VUS	Variant of uncertain significance
WES	Whole exome sequencing

Chapter 1 – Literature Review – Congenital Heart Disease

1.0 Overview on CHD

Congenital heart disease (CHD) can be defined as the presence of structural abnormalities within the heart or great vessels occurring before birth and/or, being present at birth. CHD results via the perturbations of the normal cardiac developmental program. Throughout the years, it has been referred to as one of the most common congenital anomalies found in new-borns. Nowadays this disease encompasses approximately 1% of all livebirths. (Dolbec & Mick, 2011; Van der Linde et al., 2011; Rohit & Rajan, 2020; Sun et al., 2015; Williams et al., 2019). Within the past few decades, a minority of patients diagnosed with this disorder were known to survive until adulthood. Other than the drastic increase in long-term survival of such patients, recent studies have shown a decrease in reoperation. A dramatic improvement in long-term survival as well as decrease in reoperation on the majority of CHD cases may be attributed towards the progression of surgical treatment and diagnosis available (Bouma & Mulder, 2017; Erikssen et al., 2015).

The optimal situation for diagnosis would be in neonates, and the typical scenario for neonatal diagnosis tends to present with a combination of cyanosis, congestive heart failure, and shock. The type of physical signs exhibited are indicative of right or left-sided obstructions, right-sided lesions tend to present with cyanosis due to the lack of oxygen towards the lungs, whilst left-sided lesions tend to present with inadequate system perfusion and overall shock (Dolbec & Mick, 2011). However, the optimal situation may not always be the case in a clinical setting, and diagnosis and or re-assessment of adult patients may need to occur. The clinical examination plays an imperative role in the assessment and follow-up of unoperated as well as palliated and repaired CHD cases. Clinical examination includes both physical clinician examination such as inspection of any thoracic scaring and peripheral upper and lower limp pulses, as well as machinery examination including but not limited to echocardiography and electrocardiography (Sun et al., 2015; Graziani & Delogu, 2016).

Historically, the categorisation of CHD has been based on a combination of anatomical and physiological phenotypes, including but not limited to conotruncal defects

(CTD) which affect the ventricular septum and hence any outflow tract, left ventricular obstruction (LVO) causing defects, abnormal left to right relationship (heterotaxia/HTX), and abnormalities within the mitral and tricuspid valves which may influence the inflow (Gelb, 2015). Due to the variety of heart defects which encompass congenital heart disease, the symptoms presented may vary greatly. The generally presented symptoms include extreme tiredness, clubbed fingernails, rapid heartbeat, shortness of breath, poor feeding, and cyanosis, amongst others (Sun et al., 2015). Although the development of this disease occurs shortly after birth, the development of symptoms generally occurs in early childhood and teenage years, having further complications arising in adulthood. Examples of these adulthood complications include hypertension, endocarditis, pulmonary hypertension, and increased infection in the respiratory tract (Sun et al., 2015). The disorders classified under this disease come in a wide spectrum having different degrees of severity. Increasing severity tends to be related to a lesser prevalence in population, being exemplified by severe lesions including hypoplastic left heart syndrome and truncus arteriosus. An approximation of around a third of CHD patients fall under the severe classification and tend to require surgical intervention within the first year of life (Zaidi & Brueckner, 2017). Whilst other lesions are less rare such as those involving ventricular or atrial septal defects. (Hoffman & Kaplan, 2002). Irrespective of severity however, CHD is still the leading cause of mortality due to birth defects in the United States (Go et al., 2013). Nowadays however, childhood survival rate has become more likely for varying severities, including serious severity lesions such as hypoplastic left heart syndrome (Bouma & Mulder, 2017).

As of 2010, the population of adults diagnosed with congenital heart disease was estimated at around 1.2 million in Europe and 1 million in the United States of America. As described by the CONCOR registry, the mortality of two thirds of adult patients diagnosed with CHD may be attributed towards cardiac causes. The majority of CHD mortality cases (26%) may be attributed towards chronic heart failure, the remaining percentages include non-cardiovascular deaths (23%), malignancy (9%), and pneumonia (4%) (Verheugt et al., 2010). A common challenge faced by the CHD population is all that encompasses ageing. Lifelong surveillance is required for such patients; thus, it is imperative that transition from paediatric to adult cardiology is seamless so as to prevent any loss of follow-up which might occur (Bouma & Mulder, 2017).

The revolutionary procedure in 1945 allowed for an almost universally lethal condition to progress into a more approachable and treatable one through surgical and catheter-based medical interventions. The surgical intervention first published by Helen Taussig, Vivien Thomas and Alfred Blalock which described the successful treatment for 'blue' babies in the 1940's. The procedure includes the introduction of a systemic to pulmonary artery shunt. (Blalock & Taussig, 1984). The development of this surgical treatment has allowed 80% of the modern era patients undergoing CHD surgery in developed countries to have a 10-year survival including those having complex CHD.

Table 1.0.1 The incidence of specific lesions classified as CHD along with their estimated percentage. (Dolbec & Mick, 2011).The table shows the specific lesion and the relative estimated percentage of the lesion identified within CHD.

Lesion	Estimated percentage of CHD
Ebstein anomaly	1%
Hypoplastic left heart syndrome	1%
Interrupted aortic arch	1%
Pulmonary stenosis	1%
Tricuspid atresia	1%
Atrial septal defect	5%
Critical aortic stenosis	5%
Transposition of the great vessels	5%
Coarctation of the aorta	10%
Patent ductus arteriosus	10%
Tetralogy of Fallot	10%
Ventricular septal defect	20%

The above **Table 1.0.1** shows the major lesions which may be present upon the diagnosis of CHD along with their respective percentage prevalence in ascending order.

The classification of the various cardiac defects encompassing CHD can be clinically quite challenging. Amongst many classifications, a paper published by Williams et al., in 2019 suggested that CHD classification should occurs according to four key features namely; (1) The nature of the structural heart defect, hence encompassing the type, defect complexity and pattern of malformation (Botto et al., 2007), (2) the resulting blood flow pattern arising from said deformity, (3) observed familial recurrence risk (Ellesøe et al., 2018), and (4) shared susceptibility genes (Houyel et al., 2011). Phenotypes are also typically classified into major categories according to the lesion namely: Right-sided, Left-sided, Conotruncal, Laterality defects, and isolated

septal defects. **Table 1.0.2** below shows these phenotypes major subclassing as well as examples of lesions falling within the categories.

Other than the above-mentioned classification, others have also been mentioned in literature, including specific classification for adults released by Connelly et al., 1998 & Warnes et al., 2001. In comparison to the above-mentioned, these classifications are stratified according to severity on the basis of care required. Literature also classifies cases of CHD according to severity, having different lesions listed under the subcategories of severe, moderate and mild. One paper by Hoffman & Kaplan, in 2002 describes the lesions classified under these sub-categories, for instance severe CHD is described as the majority of patients presenting as severely ill at a new-born stage or early infancy whilst also includes all patients with cyanotic heart disease as well as acyanotic lesions. Moderate CHD encompasses patients requiring expert care but are less intensive than the previously described, examples of lesions under this subcategory include mild/moderate atrio-septal or aortic incompetence, complex forms of Ventricular Septal Defects (VSD), amongst others. Mild CHD on the other hand is described as being the most numerous group having patients classified as asymptomatic, may not present with significant murmurs and tend to undergo spontaneous early resolution of the lesions present. Examples of lesions falling within this subcategory include small VSD, Bicuspid aortic valve (BAV) without aortic stenosis (AS), amongst others.

Table 1.0.2. The major subdivisions of phenotypic categories encompassing CHD split into 5 along with lesions falling in those categories as described by Williams et al., in 2019.

	Right Sided	**Left Sided**	**Conotruncal**	**Laterality Defects**	**Isolated Septal Detects**
Lesion	Hypoplastic Right Heart Syndrome	Bicuspid Aortic Valve	Tetralogy Of Fallot	Heterotaxy	Atrial Septal Defects
	Ebstein Anomaly	Aortic Stenosis	Pulmonary Atresia	Atrioventricular Septal Defects	Ventricular Septal Defects
	Pulmonary Artery Atresia	Coarctation Of the Aorta	Truncus Arteriosus	Anomalous Pulmonary Venous Return	
		Hypoplastic Left Heart Syndrome	Double Outlet Right Ventricle	Transposition of the Great Arteries	
				Malproposed Vessels	
				Dextrocardia	
				Situs Inversus Totalis	

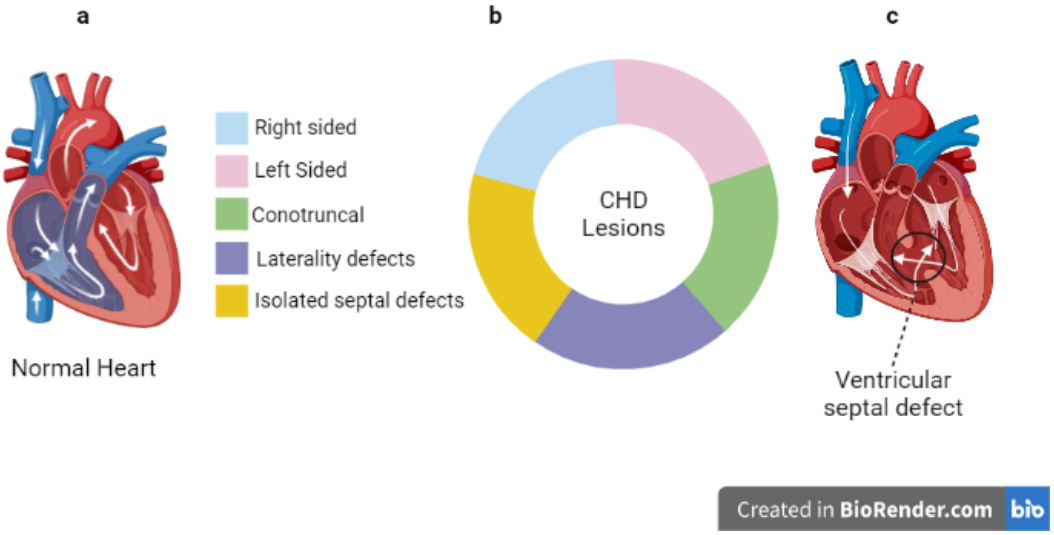


Figure1.0.1. Graphical representation of congenital heart disease (CHD) and the range of heart and great vessel malformations. (a) The typical heart and the typical pattern of blood flow. (b) The major subclassifications of lesions classified under CHD s described by Williams et al., in 2019. The proportions of the pie chart are not representative of the percentage population. (c) One of the most common congenital heart defects known as ventricular septal defect (VSD) having a malformation between in the septum separating the two ventricles thus causing a backflow of oxygenated and deoxygenated blood. Created with BioRender.com

1.0.1 Double Outlet Right Ventricle (DORV)

DORV incorporates a variety of heart malformations whereby the great vessels are misaligned with respect to their corresponding ventricles (Hutson & Kirby, 2009). Normal physiology describes the pulmonary trunk arising from the right ventricle (RV), whilst the aorta arises from the left ventricle (LV), DORV is characterised by the arousal of both the pulmonary artery and the aorta from the right atrium either entirely or predominantly (Hutson & Kirby, 2009; Yim et al., 2018). Due to both outflow vessels originating predominantly from the RV, this defect is present in conjunction with VSDs (Hutson & Kirby, 2009). The classification of DORV in terms of congenital heart disease/defects has 3 facets made up of the morphology, connections and relationship (Yim et al., 2018). The term DORV therefore although defines the connection at the atrioventricular (AV) junction, the relationship and morphology are not accounted for. This subcategory includes a large variety of the above 3 facets and thus is highly heterogenous. This congenital defect falls within the classification of conotruncal lesions being majorly present alongside ventricular septal defects as well as other septal and valvular defects, thus contributing a towards 20% of CHD cases (Dolbec & Mick, 2011; Hoffman & Kaplan, 2002). The classification of DORV is based on the position of the

VSD being either sub-aortic, sub-pulmonary, doubly-committed or non-committed (Anderson et al., 2001; Ebadi et al., 2017; Hutson & Kirby, 2009; Mahle et al., 2008). In the case of the VSD being situated below the aorta and the great arteries situated adjacent to each other, the DORV classifies as subaortic (Hutson & Kirby, 2009). DORV with sub-pulmonary VSD is classified via the presence of a VSD below the pulmonary artery (Hutson & Kirby, 2009). Doubly-committed VSD in DORV refers to the VSD being present below both of the outflow vessels, thus having two VSDs (Hutson & Kirby, 2009). Whilst the VSD described in noncommitted VSD is present remotely and not dedicated to either vessel (Hutson & Kirby, 2009). Nowadays, other than via neonatal echocardiography, in-utero echocardiography allows accurate diagnosis and allows adequate parental counselling in terms of neonatal prognosis and treatment plan, as well as pre-planned delivery (Gedikbasi et al., 2008).

Although surgical intervention is available for most types of DORV, it is not as clear cut whether all types of DORV should undergo such surgery, and which subtypes classify for which types of surgery. Various research has been performed to narrow the surgical treatment plan for the various DORV subtypes having an overall increase in survival rate post-surgically (Bradley et al., 2007). Throughout the years, research efforts have been made to identify the best surgical strategic plans for such patients, having previously thought that single ventricular repair provided a higher survival rate in comparison to biventricular repair (Bradley et al., 2007). However nowadays the margin between these two surgeries is decreasing further (Oladunjoye et al., 2019).

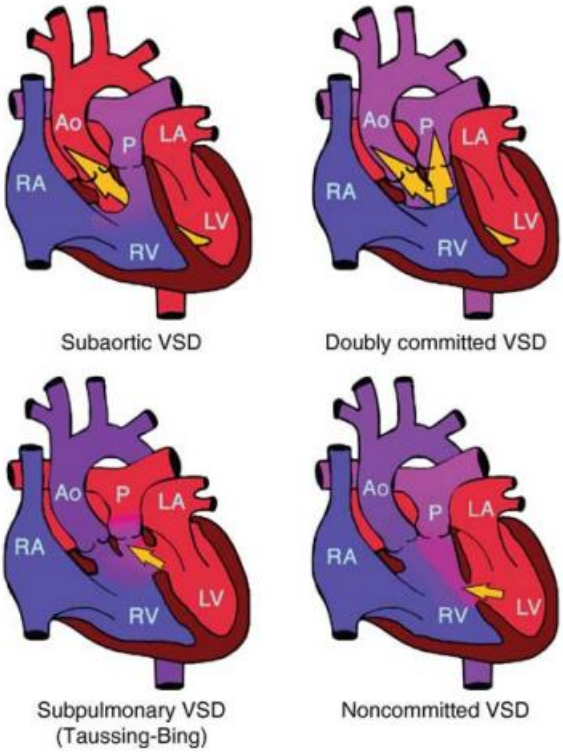


Figure 1.0.1.1. The 3 subtypes of Ventricular septal defects (VSDs) found alongside DORV classification namely, subaortic VSD, Doubly committed VSD, Subpulmonary VSD and noncommitted VSD (Hutson & Kirby, 2009).

1.1 Aetiology

Although being heavily researched throughout the years, the aetiology of CHD is largely unknown. No more than approximately 15% of all worldwide CHD cases have a currently identified cause (Bouma & Mulder, 2017). This limitation in underlying aetiology could also be linked to the worldwide birth prevalence currently known. Although CHD is estimated to have a prevalence of approximately 1%, this percentage only includes the cases accounted for. Variations in referrals, access to care as well as referral bias might allow several cases to go undiagnosed, and hence this percentage population prevalence to be less accurate (van der Bom et al., 2011). Another factor which might influence the current percentage of known diagnosed CHD cases is the referral to a centre whereby the case would be added to the worldwide population. This could occur when a physician encounters a defect which they consider manageable, and thus would not refer the patient to a centre, and hence not being included in the worldwide population (Hoffman & Kaplan, 2002). Additionally, some physically subtle lesions, such as atrial septal defects, may not be detected until later life stages, whilst other severe lesions may result in neonatal death without cardiologic or autopsy (Abu-Harb et al., 1994; Kuehl et al., 1999; Rostad & Sørland, 1981; Seldon et al., 1962).

The little evidence that has been compiled with relation to the aetiology of this disorder has nevertheless allowed identification of multifactorial causes. Although relatively low, 2-10% of non-syndromic CHD cases may be attributed towards environmental factors. Some of these include risk factors such as maternal diabetes and phenylketonuria, which have proven to have an increased risk for foetal development of CHD. Other risk factors associated to gestational predisposition include maternal obesity, exposure to alcohol, febrile illness, early onset pre-eclampsia, rubella infection as well as exposure to teratogens/drugs including thalidomide and retinoic acid. (Jenkins et al., 2007; Hedermann et al., 2021; Nora, 1968; Zhang et al., 2021). Another important factor known to contribute towards the development of CHD can be seen from a genetics point of view. The development of technology such as Next-Generation sequencing (NGS), has allowed for the clarification of part of the aetiology of CHD, the genetic aspect. As of 2011, around 40 genes were identified to be implicated in the development of CHD (van der Bom et al., 2011), nowadays this list is evermore increasing. A recent study published

in 2019 estimated over 400 genes which may be associated to the pathogenesis of CHD (Williams et al., 2019).

Karyotyping approaches have allowed the identification of numerous malformation syndromes, including congenital heart disease, which may be attributed to the presence of chromosomal aneuploidy. Different cases of aneuploidy pertain to different prevalence of CHD, an example of which being Down's syndrome, having a CHD prevalence of approximately 45% (Vis et al., 2009). Overall, 8-10% of CHD cases may be accounted for by chromosomal aneuploidies including, but not limited to trisomy 13, 18, 21, Turner syndrome, Klinefelter syndrome and DiGeorge syndrome.

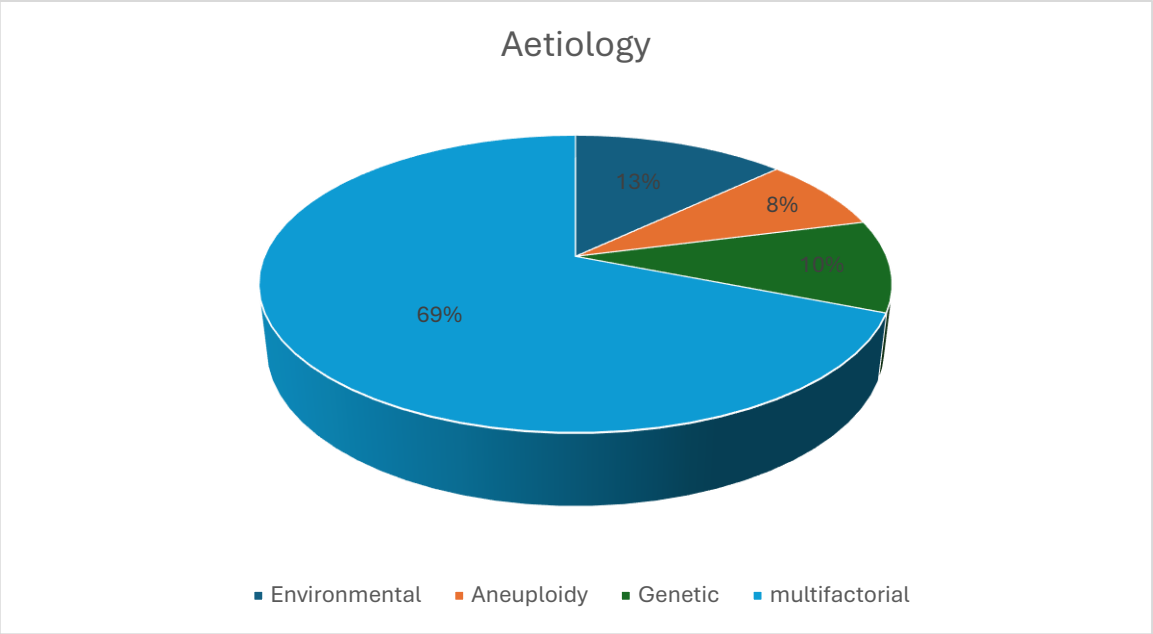


Figure 1.1.1. A diagrammatic overview of the different aetiologies contributing towards the development of congenital heart disease (CHD) namely environmental, Aneuploidy, Genetic and multifactorial having percentages of 13%, 8%, 10% and 69% respectively. (Jenkins et al., 2007; Hedermann et al., 2021; Nora, 1968, van der Bom et al., 2011, Vis et al., 2009, Zhang et al., 2021)

The remaining percentage of diagnosed CHD aetiology are considered to be multifactorial, suggesting the interaction between the environmental and genetic aetiologies (van der Bom et al., 2011). The issue with the multifactorial explanation for CHD aetiology is the lack of reports available. Few reports have been made on the interaction and accumulated effects of various factors on the prevalence of CHD (Joziasse et al., 2009; Nora, 1968; Smith et al., 2009).

Throughout the years, efforts have been made to compile evidence towards modifiable risk factors. However, the goal of achieving said risk factors in an unbiased and precise manner has proven challenging. (Botto et al., 2007). Two related challenges typically found in aetiological studies are those of case classification and risk estimation. This challenge can be viewed upon from different aspects, one being that of phenotypic classification. If classification, and thus analysis, had to be performed based solely upon phenotype, the product would be that of several groups having too low a percentage of cases to acquire a risk factor estimation (Botto et al., 2007; Clark, 1996, 2001).

Table 1.1.1. The percentage probability of congenital heart disease (CHD) occurrence according to the familial phenotypic representation (van der Bom et al., 2011)

Probability of occurrence in overall CHD	Type of family member affected
1-6%	Unaffected parents, one affected sibling.
3-10%	Unaffected parents, Two affected siblings.
2-20%	Affected Mother
1-5%	Affected Father.

1.2 Epidemiology

As previously mentioned, CHD is the most frequently described congenital disorder in new-borns and contributes to the leading cause of infant deaths arising from birth defects (Tennant et al., 2010). The prevalence in livebirths is estimated at around 8 per 1000 cases having a range from 3 to 10. This vast range arises due to several difficulties found in birth prevalence studies including variations in referral, access to care as well as referral bias in terms of active echocardiography screening (Bouma & Mulder, 2017). The individual lesions making part of the classification of CHD have their own prevalence, isolated septal defects are accounted for as the most common lesion in CHD, Ventricular Septal Defects (VSD) have an estimated prevalence of 3570 per million livebirths, and Atrial Septal Defects at 941 per million births (Diab et al., 2021). Prevalence can also be segregated according to severity, having moderate and severe defects at a prevalence of 1.5 per 1000 livebirths each. Prevalence may be further influenced by the increased detection via prenatal ultrasounds, this accounts for up to one third of defects as well as 57-85% of severe lesions to be detected before the end of gestation (Bouma & Mulder, 2017). Epidemiological studies performed in 2015 by Cowan & Ware, suggests that 20-30% of CHD cases may be associated to an

environmental or genetic trigger. 3-5% of CHD cases may be attributed towards single-gene disorders, 8-10% to aneuploidies and 3-25% to pathogenic copy number variants (CNV). One of the largest genetic studies via next generation sequencing (NGS) on CHD performed by Jin et al., in 2017 suggests that 8% of CHD cases may be attributed towards *de novo* autosomal dominant genes whilst 2% may be attributed towards inherited autosomal recessive variants.

There is relatively little evidence behind any temporal or geographic variation in general incidence of CHD to suggest any environmental triggers. However, slight differences in CHD types can be witnessed in different populations such as, an increased incidence report of LVO lesions in Caucasian children, in comparison to the increased incidence report of RV obstruction amongst Chinese children, thus suggesting contributions of population-specific genetics (Jacobs et al., 2000; van der Linde et al., 2011).

A review performed in 2011 by van der Linde et al., identified a substantial increase in CHD birth prevalence between the years 1930 and after 1995, having previously had a prevalence of 0.6 per 1000 live births, to 9.1 per 1000 live births after 1995. In terms of birth prevalence with relation to geographical incidence, significant difference was accounted for by this review. Out of the continents, Asia was ranked as having the highest total reported birth prevalence of CHD, whilst Europe had the second highest (van der Linde et al., 2011).

The survival of patients diagnosed with CHD has kept increasing over the years. A major Norwegian study showed an increase of survival of patients up until the age of 16 diagnosed with CHD from 62% in 1971 to 87% in 2011 (Erikssen et al., 2015). As of 2011, the median age of patients with severe congenital heart disease increased from 11 years of age in 1985 to 17 years of age in 2000. Whilst in the past two decades, mortality amongst CHD patients has declined between 50-70%, depending on the particular defect. Upon comparison between age groups, the largest reduction in mortality was identified in the patients aged 1-4 years followed by 5-14 years of age.(van der Bom et al., 2011). Epidemiological changes including a reduction in reoperation required a well as an increase (from 12% to 34%) in operations for patients with simple defects, contribute towards the sustained survival and hence increase within the adult CHD population (Bouma & Mulder, 2017).

As previously mentioned, the CONCOR registry is a Dutch national database which compiled over 11,400 patients with CHD over the age of 18 between the year 2001 and 2009. Data on mortality was gathered by this register, which identified that sex differences were accounted for and present in adult men diagnosed with type 2 atrial septal defect. These men had a worse overall survival than the general CHD men population. Whilst an in-hospital 20-day mortality for CHD identified a higher correlation with younger men than women of the reproductive age.

1.3 Pathophysiology of CHD- Embryological cardiac development.

For an in depth understanding of the development of CHD, the molecular as well as anatomical understanding of heart structural development is necessary. A review by Kloesel et al., in 2016 described a brief background on the structural development of the heart in 9 summative steps:

1. The formation of the three germ layers; Endoderm, Mesoderm and Ectoderm. This occurs during a process known as gastrulation.
2. Establishment of the first heart field (FHF) and Second heart field (SHF). Giving rise to segments of the linear heart tube and the arterial and venous pole cells respectively.
3. Heart tube formation due to the embryonic folding at the craniocaudal and lateral axis, allowing the formation of the endocardial tubes.

The endocardial tubes consist of myocardial cells which form the myocardium and the endothelial cells forming the endocardium. these two cell types are separated by an extracellular matrix, and the fusion of these cell types (occurring due to folding) allow the formation of the heart tube. The epicardium is formed at a later stage via the migration of proepicardial precursor cells. In the early stages after uterine implantation, two cell layers are formed, known as the epiblast and the hypoblast. A primitive streak forms at the caudal region of the epiblast and extends cranially.The first step concerns the formation of the three germ layers which occurs during a process known as gastrulation.

4. Cardiac looping, convergence, and wedging.

   Ranging from day 23 to 28, the bending of the endocardial tube allows the configuration of a cardiac loop. This process is driven by the elongation of the endocardial tube via migration of the precursor cells. Convergence defines the midline

alignment of the outflow tract and the AV canal. Wedging describes the process resulting in the creation of the pulmonary and systemic trunks.

5. Septa formation.

   The developmental process towards the formation of the septa is a long process, not being entirely complete till after birth. One of the first septa to begin development is referred to as the septum primum, which initiates the division of the left and right atrium, leaving a hole at 2 distinct sites, namely the ostium primum and the ostium secundum. The second septa to be developed at a later stage, approximately day 33, is known as the foramen ovale. These two septa however do not fuse until after birth, hence retaining right to left shunting of systemic venous and placental blood throughout the gestational process.

6. Outflow tract development.

   The formation of these tracts occurs due to the dense population by endocardial cells. These cells undergo epithelial to mesenchymal transformation, which is controlled by a protein known as Transforming Growth Factor β (TGF-β) and signalling via the Notch protein (Sylva et al., 2014). The transition of these cells allows the loss of cell adhesion molecules, and cell polarity, thus attaining the migration ability, hence invading tissue planes, in this case, allowing the formation of new tissues.

7. Cardiac valve formation

   The mesenchymal cells described briefly above, provide the foundation for the development of AV valves during the 5^{th} and 6^{th} weeks of gestation. These cells collaborate with cardiac neural crest cells, aiding in the correct septation of the outflow tracts, hence the formation of aortic and pulmonary valve leaflets, to be later developed into the valves at around the 7^{th} to 8^{th} weeks. Complete ventricular septation is dependent on the fusion of the muscular ventricular septum, the AV cushion tissues, and the outflow tract septum.

8. Vascularisation.

   The initial steps of vasculogenesis occurs at around day 18, when progenitor cells from the FHF form cardiac myoblasts, giving rise to paired dorsal aortae. At a later stage, around day 28, pharyngeal arches start to form. These arches sequentially form in a cranio-caudal order and connect the bilateral dorsal aortae to the aortic sac. These give rise to 5 paired arches eventually forming the main central and systemic vasculature.

9. Conduction system development.

The previously transformed epithelial to mesenchymal cells originating from the sinus venosus form the main coronary vessel. The coronary system forms its connection to the aorta via the invasion of arterial endothelial cells into the aorta. This allows the fully matured heart to be formed by day 50.

The above information was adapted from Yamagishi et al., 2009, Kloesel et al., 2016 and Mathew & Bordoni, 2022, as well as embryology books by Jonas, 2002; Park, 2014; Sadler, 2022 and Schoenwolf et al., 2014.

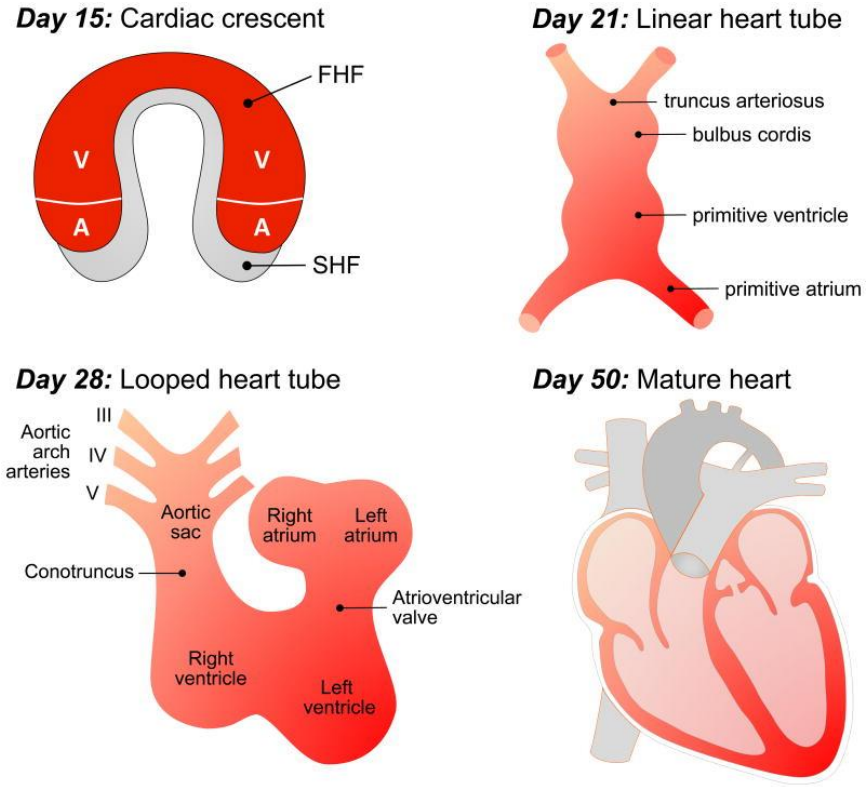


Figure 1.3.1. A schematic representation of the cardiac embryology development. By day 15 the first and second heart fields (FHF)(SHF) are specified, which will form segments of the linear heart tube and the arterial and venous poles respectively. By day 21, the linear heart tube is established due to the embryo cephalocaudal and lateral folding, consisting of the arterial and venous poles. By day 28 looping has been established towards the right, having the future cardiac regions being identified. Up until day 50, the septa, along with the valves are established allowing separation of the outflow tract and the chambers. (Kloesel et al., 2016)

On a molecular basis, cardiac development is dependent on several cellular processes including proliferation, specification, migration as well as morphogenesis, all of which being possible due to the pluripotency of the stem cells at these stages. A concept which is crucial during embryonic development is that of laterality, whereby the right and

left body axis are established. This phenomenon occurs relatively early on during embryonic development, at around the 3rd week of gestation, whereby the primitive streak establishes the cranial-caudal, left-right and medial-lateral axes. The primitive node also plays a crucial role in this, providing the organisational activity for the cranial end of the primitive streak. Signalling pathways involved in the body axis determination mostly form part of the TGF-β superfamily of proteins. Examples of these proteins include bone morphogenic proteins (BMPs), NODAL, growth and differentiation factors (GDFs), Wnt proteins as well as TGF-β (Weiss & Attisano, 2013). Proteins within the TGF-β family, other than being structurally related, generate intracellular signals via signalling molecules falling within the Suppressor of Mothers Against Decapentaplegic (SMAD) family (Wrana, 2013). Malfunctioning or inefficient signalling of the mentioned (amongst other) proteins would hence result in a spectrum of laterality defects, also referred to as heterotaxy (Kloesel et al., 2016).

The above-mentioned proteins play several roles throughout the embryo's cardiac development. As early as day 16, signalling pathways including BMP, Wnt/b-catenin, fibroblast growth factor and NODAL are involved in the triggering of epiblast differentiation into mesodermal cells. This allows the said mesodermal cells, once migrated to the adequate positioning, to differentiate further into the cardiogenic mesoderm. Other proteins such as vascular endothelial growth factor (VEGF) and their receptors are expressed in downstream processes, allowing the marking of cells committed to cardiogenic fate. These factors along with mesoderm posterior 1 (MESP1) transcription factor, account for the programming of cells towards their cardiogenic fate, and hence the development of the first and second heart fields (FHF & SHF). These factors, amongst others, allow for the accurate asymmetrical formation of the heart. Malfunctioning in any aspect of the signalling of the above-mentioned proteins thus results in complex congenital heart disease. Examples of which include pulmonary stenosis, TOF, DORV, double inlet left ventricle (DILV), transposition of the great arteries, amongst many others (Shiraishi & Ichikawa, 2012).

1.3.1 Congenital heart defects – A brief pathophysiological description.

Atrial septal defects (ASD) are one of the most commonly found CHD anomalies. According to the location of the defect within the atrial septum, these defects may be categorised according to the below table.

Table 1.3.1.1. Brief overview of different anomalies defined as atrial septal defects and their pathophysiology adapted from Bradley & Zaidi, 2020; Brida et al., 2022 and Webb & Gatzoulis, 2006.

Position of atrial septal defect	Pathophysiology
Patent foramen ovale	Leakage of venous blood through a persisting hole between the right and left atrium.
Ostium primum defect	Comprising approximately 15% of ASD cases. Persisting hole within the fossa ovalis superiorly and inferior to the atrioventricular valves.
Ostium secundum defect	Accounting for approximately 80% of ASD cases. Also located within the fossa ovalis, however typically has one or more defects within the septum primum.
Sinus venosus defect	Accounting for approximately 5-6% of ASD cases. Located within the mouth of the vena cava.
Conorary sinus defect	Accounts for <1% of ASD cases. Located in the unroofing tissue responsible for the separation of the coronary sinus from the atrium.
Common atrium	N/A

In terms of early developmental stages, the development of DORV may be traced to the heart tube looping at round 3-4 weeks post-conception. (Kirby, 2007) The embryonic formation of DORV arises through the inefficient lengthening of the heart tubes, thus resulting in the shortening of the heart tube, hence resulting in failure of outflow vessel alignment with the respective ventricles. Incorrect alignment of the aorta over the right ventricle leads to the final position causing both great vessels to arise out of the right ventricle, and hence the aorticopulmonary septum cannot properly form as a separation of the great vessels (Kloesel et al., 2016).

1.3.2 The underlying Genetics of CHD

Genetics always plays a crucial role in the development of diseases. In terms of CHD, changes within any genes encoding for transcription factors (TFs), cell signalling transducers, and chromatin modifiers are more likely to result in a phenotype (Williams

et al., 2019). Variations within these types of genes may lead to the interference of important processes occurring during cardiac development in neonates including cellular differentiation and patterning. Since most protein products from these genes have overlapping functional networks, the cause of disease is not as straight cut, and suggests a broad interacting network associated to disease pattern (Lage et al., 2012).

The combination of genetic heterogeneity and genetic diversity have resulted in variable expressivity resulting in 60% of CHD cases being unexplained (Zaidi & Brueckner, 2017). The resulting variable expressivity leads to different phenotypic products arising from the same genes, and hence variable penetrance resulting in individuals having a known phenotypic variant present with no disease. As a result, CHD tends to follow a non-Mendelian mode of inheritance, but rather is described in literature as being mediated by complex genetics (Williams et al., 2019). On the other hand, familial CHD variants may occur in several ways including autosomal dominant, recessive or X-linked which may be expressed with high penetrance and have variable clinical manifestations (Fahed et al., 2013). Whilst heterogeneity accounts for several phenotypes arising from the same gene, the same could be said in a reverse mannerwhereby, different genetic variants may result in identical physical cardiac malformations.

In most instances, the underlying genetics of CHD tends to be studied in 3 different cohorts. Primarily via trio targeted whole-genome or whole-exome sequencing the proband along with the two unaffected parents are analysed in search for *de novo* variants which may have arisen in the proband. Secondly familial studies are run whereby multiple members of a family undergo sequencing and phenotyping to identify any inheritance pattern and/or presence of phenotypic variants. Lastly cohort studies include the phenotypic and sequencing analysis of a large number of unrelated individuals as well as health control individuals in search of single gene or set gene enrichment in the affected sample. The studies which were performed on *de novo* and single genes have identified a high burden associated to genetic variation within predicted damaging genes associated to CHD typically highly expressed in the heart or involved in cardiac development. (Zaidi & Brueckner, 2017). Further evidence supporting the genetic contribution towards the development of CHD can be acquired from numerous sources. One of the most studied fields tends to be related to twin studies. Research has shown that there is a greater concordance of CHD in monozygotic rather than dizygotic twins, whilst evidence has proven that the occurrence of twins increases the chance of CHD (Herskind et al., 2013;

Wang et al., 2014). The fact that populations having high levels of consanguinity also have an increased incidence of CHD, shows a relation towards the recessive inheritance mode (Shieh et al., 2012). However, it is highly perplexing that a large proportion of CHD cases, being particularly quite severe, tend to occur in families having no other history of CHD.

One of the earliest causes of CHD to be identified are aneuploidies. Aneuploidies can typically be detected via karyotyping which allows the detection of chromosomal alterations less than 5-10 Mb. Trisomy 21, trisomy 18, trisomy 13, Turner syndrome, and Klinefelter syndrome along with their respective CHD were some of the initial aneuploidies to be identified through karyotyping (Diab et al., 2021). Aneuploidy results in a large number of dysregulated genes, which in turn affect development in a pleiotropic and severe manner. The percentage of CHD development in liveborn babies may vary according to the type of aneuploidy present, for instance trisomy 21 has a 35-50% of CHD development, whilst trisomy 13 and 18 have a 60-80% and monosomy X has a 33% probability (Zaidi & Brueckner, 2017). The specific types of CHD associated with aneuploidy encompasses a broad range of CHD phenotypes. Although certain aneuploidy are associated with specific CHD phenotypes, including trisomy 21 with atrioventricular septal defects, since aneuploidy disrupts a large number of genes, it is far more challenging to directly target the underlying developmental and genetic mechanism (Zaidi & Brueckner, 2017). Aneuploidies nowadays are detected prenatally via a non-invasive prenatal diagnostic screen. Once an aneuploidy is detected prenatally, upon birth a foetal echocardiogram may allow for early and accurate diagnosis of any cardiac anomalies. The risk of aneuploidies has been directly related to increased maternal age (Pierpont et al., 2018).

Recent critical advances in next generation sequencing (NGS) have allowed the biological understanding of CHD to be further understood. Particularly whole exome sequencing has allowed the identification of variations previously unidentifiable, namely *de novo* variation, variants without Mendelian inheritance, those with reduced penetrance as well as somatic alterations (Zaidi & Brueckner, 2017). Whilst large variations such as aneuploidies are typically discovered via karyotyping, comparative genomic hybridization (CHG), and fluorescent in situ hybridisation (FISH), different types of NGS as mentioned above, may be utilised for the detection of small genetic variations,

1.3.2.1 Copy number variants

Copy number variants (CNV)s refer to large structural variations within the DNA consisting of amplifications or deletions ranging from 1kb size to several megabases, hence leading to an altered dosage of the genes encompassed in the CNV (Zaidi & Brueckner, 2017; Williams et al., 2019). CNVs principally arise through inappropriate recombination due to flanking of region-specific repeat sequences or misaligned highly homologous genes (Williams et al., 2019).

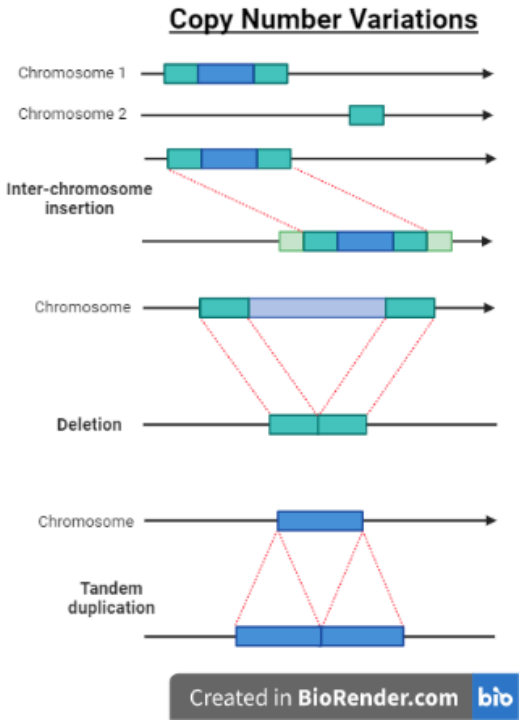


Figure 1.3.2.1.1. The different types of copy number variants (CNVs) including inter-chromosomal insertions, deletions, and tandem duplications. Created with BioRender.com

Generally, deletions tend to be more deleterious than amplifications/duplications due the sensitivity of gene dosage haploinsufficiency (Pierpont et al., 2018). CNVs (as a class) are commonly found within the general population having diverse potential mechanism of action and are not inherently pathological. Structural rearrangements such as CNVs tend to have a higher mutation rate in comparison to single base pair variations, and these CNVs may disrupt one, none, or many genomic elements (Costain et al., 2016). The more commonly found within a population the CNV, the less likely that it may have any pathogenic influence and tend to act as neutral variants or modifiers to disease susceptibility. Whilst rare CNVs tend to have a higher pathogenic influence and be associated to disease, especially in developmental disorders such as CHD. Throughout

the years, information has been accumulated on the different types of CNVs found in association to CHD. Typically, this association is identified via mapping of overlapping CNVs on a large quantity of patients with the aim of pinpointing a critical interval and single gene within this site having an association to CHD (Pierpont et al., 2018). Over the last decade, many novel CNVs have been identified in association with CHD and have nowadays been identified in enough patients to allow a definition of clinical features associated to those specific CNVs (Gelb & Chung, 2014). The greatest issue faced in terms of CNV is the characterisation of the phenotype, and whether it is caused by a single gene having pleotropic effects on multiple phenotypic aspects, or whether multiple genes are contributing towards the multiple phenotypic aspects (Pierpont et al., 2018).

An example of the most common CNV found within CHD cases is that of 22q11.2 deletion syndrome. Other than being a known cause for CHD, other associated phenotypes include phenotypes associated with DiGeorge syndrome, velocardiofacial syndromes and Takao conotruncal anomaly face syndrome, although phenotypes are known to vary, even within one family (Digilio et al., 2003). The above-mentioned syndromes, although highly overlapping, are not synonymous with one another, proven by the finding that approximately 10% of patients with the above-mentioned syndromes do not have a 22q11.2 deletion and not all patients with this deletion demonstrate classic features of the above-mentioned syndromes (Pierpont et al., 2018). 70% of the heart defects associated with 22q11.2 deletion may be accounted for by conotruncal malformations.

The second most common CNV arising syndrome is that of 1p36 deletion syndrome. Typical clinical features include dysmorphic faces, intellectual disabilities having varying severity, hypotonia, seizures, amongst others. Structural cardiac defects include ASD, VSD, Patent ductus arteriosus (PDA), valvular abnormalities, TOF, CoA, Ebstein's anomaly and infundibular stenosis of the right ventricle (Battaglia et al., 2008; Heilstedt et al., 2003). Cardiomyopathies are present in 27% of individuals whilst left ventricular noncompaction is present in 23%, and dilated cardiomyopathy in 4%.

Recent investigations have allowed the identification of multiple CNVs which contribute to isolated non-syndromic CHD. Specific lesions having large *de novo* CNVs identified include TOF, left-sided lesions, as well as other sporadic CHD cases (Edwards & Gelb, 2016). These studies estimate that 5-10% of sporadic non-syndromic CHD in

patients having a normal karyotype and FISH analyses can be attributed to have a rare CNV aetiology. Due to CNVs encompassing a range of genetic material, genes may be present within the CNV affected region. Due to this, some CNVs may encompass preciously identified CHD genes or cardiac developmental genes. An example of CNVs recurrently identified in CHD cases include the occurrence on chromosome 8p23.1, having an impact on the cardiac TF *GATA4*, chromosome 20p12.2 and 9q34.3 have an impact on the Notch signalling pathway, *NOTCH1* and *JAG1* (Greenway et al., 2009; Soemedi et al., 2012).

Other than the definition of novel CHD genes, CNV identification may be crucial to assess developmental networks via biological interactions and functional annotations found on bioinformatic repositories, whilst also providing information to gene-gene and protein-protein interactions (Soemedi et al., 2012).

1.3.2.2 Point mutations and Mendelian genetics

Studies performed on animal models, particularly mice, have allowed the discovery of transcription factors (TFs) and cofactors involved in CHD as well as their pathological role. TFs in CHD patients were also found to be enriched for *de novo* and loss of function (LOF) variants. (Williams et al., 2019). These TFs have also previously been described in relation to inherited forms of CHD (Prendiville et al., 2014; Zaidi & Brueckner, 2017). Although CHD is not classified as an inherited disorder having a high tendency of sporadic incidence, research over the years has allowed in depth insight towards Mendelian and inherited forms of CHD. Linkage analyses, positional cloning and targeted sequencing of selected candidate CHD genes are few methods which provided this insight, not only for the Mendelian genetics, but also for the identification of point mutations for the identification of CHD loci (Fahed et al., 2013). Contemporary strategies bypass the analysis of CHD loci and identify CHD mutations via next generation sequencing (NGS) at exome or genome level. Since this method allows the finding of Single nucleotide polymorphisms (SNPs) unrelated to CHD, various post-sequencing filtering must occur to identify rare/novel SNPs having predicted deleterious effects, being expressed during cardiac development (Fahed et al., 2013). Most of the genes primarily identified as being inherited in CHD form part of a group of cardiac TFs including *NKX2.5*, the GATA family, T-box factors namely *TBX5, TBX1*, and MEF2 factors (Prendiville et al., 2014; Zaidi & Brueckner, 2017). Variants within the above-

mentioned gene groups, particularly the specific genes *KX2-5, NKX2-6, GATA4, GATA5, GATA6, IRX4, TBX20,* and *ZIC3*, are inherited via a dominant nature and are predicted to reduce physiological quantities of the gene protein product. This occurs due to the variants causing inactivation of one allele, or LOF via disruption of DNA interaction, or perturbations in combination interactions between TFs and transcription cofactors (Williams et al., 2019). In addition to transcriptional regulators, other genes have been implicated in the Mendelian inherited CHD, including genes encoding for signalling molecules as well as cell structure components.

1.3.2.2.1 *NKX2*-5

Variants within this gene were one of the first identified inherited point mutations clearly shown to cause CHD. This gene encodes for a homeobox-containing transcription factor having a function in cardiac formation and development. It is expressed during the earliest cardiogenesis stages, having a direct influence on cardiomyocyte differentiation and proliferation (Williams et al., 2019). *NKX2.5* is the earliest identified myocardial progenitor cell marker in all species. Large pedigree evaluation including individuals with ASDs as well as abnormalities within their conduction systems, allowed the categorisation of this variant to underly both ASD and conduction defects (Zaidi & Brueckner, 2017). Nowadays this variant is known to have an influence on a wide spectrum of CHD including ASD, VSD, TOF, Hypoplastic left heart syndrome, transposition of the great arteries, DORV, amongst others (Wu et al., 2021). To date approximately 80 different variants have been identified within the *NKX2.5* gene including missense, synonymous and nonsense variants (Wu et al., 2021). Investigation of phenotype and penetrance of *NKX2.5* gene variants have shown a dependence on genetic background as well as interaction with variants in both mice and humans (Granados-Riveron et al., 2012).

1.3.2.2.2 GATA family

The GATA family encompasses a group of zinc finger TFs, having identified expression in cardiac development and thus play a role in cardiogenesis. Specifically, within this family *GATA4* is a zinc-finger TF being directly associated with *NKX2.5*. One of the earlier interactions identified between these genes was through the organism *Drosophila.* This interaction was found to play a crucial role in cardiac mesoderm

specification, hence variants pertaining to these genes had primarily been identified with complete heart tube formation failure (Bodmer, 1993). Variants in this gene have been associated with numerous variations including decreased transcriptional activity leading to BAV and VSD (Li et al., 2018). Variants in regulatory sequences for this gene, including *NEXN* have also been associated with the development of CHD (Yang et al., 2014). Variations in the noncoding region of this gene have also been associated with BAV, highlighting the importance of research into the noncoding and regulatory sequences of such developmental genes (Yang et al., 2017). Whilst *GATA4* is mostly implicated in haploinsufficiency within this gene, or known disease-causing variants present, another gene within this family, *GATA6*, has been implicated in familial and sporadic CHD. Presenting lesions within *GATA6* include Pulmonary vein stenosis (PVS), ASD, PDA, and persistent truncus arteriosus (Kodo et al., 2009). Other genes falling within this family, such as *GATA6* have also been found to cause severe OFT defects via interaction with other genes (Kodo et al., 2009; Maitra et al., 2010). On the other hand, variants within *GATA5* have only been recently investigated as an influencing factor to CHD. Rare sequence variants within this gene have been reported in instances with TOF, VSD, BAV and familial atrial fibrillation (Bonachea et al., 2014).

1.3.2.2.3 T-box family

This family encompasses a group of T-box protein (TBX) transcription factors being expressed through the developing heart and play a major role in cardiomyocyte identity. A member of this family *TBX5* is markedly expressed within both the developing forelimb buds and the heart, and therefore variants within this gene can be attributed to different phenotypes (Zaidi & Brueckner, 2017). Mice studies have identified an interaction of *TBX5* with both *GATA4* and *GATA6,* such that double heterozygous variants in *GATA6* lead to neonatal lethality whilst the same in *GATA4* results in severe cardiac malformations as well as embryonic lethality (Greulich et al., 2011). Variants in other family members such as *TBX20* were also identified in two families presenting with cardiac septation defects, dilated cardiomyopathies and mitral valve stenosis, whilst subsequently also being described in other cardiac malformations including TOF, truncus arteriosus and DORV (Huang et al., 2017).

1.4 Pulmonary hypertension & Eisenmenger syndrome.

Pulmonary hypertension (PH) is commonly found in conjunction with diagnosis of congenital heart disease (Brida & Gatzoulis, 2018). Pulmonary hypertension (PH), although representing a broad variety of disease entities, may be defined as the mean pulmonary arterial pressure being ≥ 25 mmHg at rest (Brida & Gatzoulis, 2018; Chen & Dai, 2015; Rosenzweig & Krishnan, 2021).

In terms of PH classification, as most cardiac diagnosis, there are scales of severity which may be attributed. Although the specific clinical classification has undergone numerous modifications throughout the years, they may still be subclassified into 5 major groups (Brida & Gatzoulis, 2018; Rose-Jones & Mclaughlin, 2015). The first classifying subgroup is that of pulmonary arterial hypertension (PAH), having an estimation of 15 to 25 cases per million, although being the least common form of PH, it is the most extensively investigated. As defined by the WHO, this subgroup was originally classified as primary PH attributing its major cause towards vasculopathy having a primary effect on the distal pulmonary arteries. Phenotypes within this cohort include disorders such as portal hypertension, congenital heart diseases, myeloproliferative disorders, along with others being idiopathic and/or familial. The familial aspect of this subcategory of PH may be predominantly attributed towards genetic variations found within the bone morphogenic protein receptor (BMPR2) gene (Rose-Jones & Mclaughlin, 2015). The second subcategory within this classification may be described as pulmonary venous hypertension (PVH) due to left heart disease (Oudiz, 2007). The most common cause of this PH subtype is chronic elevated left arterial pressure caused by left heart disease, being that of left ventricular diastolic or systolic dysfunction and valvular disease (Rose-Jones & Mclaughlin, 2015). In comparison to the previously discussed subgroup, relatively little information is available with regards to PVH. The third subcategory to fall within the classification of PH is that of PH due to lung disease/hypoxemia. It is typically seen in the setting of chronic obstructive pulmonary disease (COPD), interstitial lung disease (ILD) amongst others (Thabut et al., 2005). The fourth category is that of chronic thromboembolic PH (CTEPH), being the result of chronic pulmonary thromboembolic events. The final subgroup is that of PH due to miscellaneous causes having unclear pathogenesis. Some examples of diseases which may classify under this category include glycogen storage disease, sarcoid lung disease, and thyroid disorders.

Eisenmenger syndrome (ES) may be given several definitions, however despite these variations ES tends to result from the presence of large systemic to pulmonary shunts triggering the development of PAH and pulmonary vascular disease (PVD) (Brida & Gatzoulis, 2018). Upon the initial coining of the term ES in 1958 by Paul Hamilton Wood, it defined the condition of increased pulmonary arterial pressure and pulmonary vascular resistance in conjunction with a VSD and resultant shunt reversal with cyanosis. A study published in 2015 published by Chen & Dai, describes a modified consensus for the definition of ES containing the following patient criteria;

1. CHD diagnosis characterised by left to right shunt.
2. Presence of advanced pulmonary vascular disease at an early life stage along with absence of increased pulmonary flow.
3. Cyanotic congenital cardiac defects associated with particularly high pulmonary vascular resistance, exemplified by transposition of the great arteries.

Whereas an updated definition published in 2020 by Kaemmerer et al., describes ES as the association of PH with CHD caused by an initially large, non-restrictive intra/extra-cardiac communication alongside a systemic to pulmonary shunt resulting in; progressive vascular disease, central cyanosis, and shunt reversal.

Irrespective of definition, distinct features of ES include chronic cyanosis due to the multisystem involvement, this therefore includes the haematopoietic system having the presence of secondary erythrocytosis, thrombocytopenia, coagulation abnormalities, amongst others (Brida & Gatzoulis, 2018). As previously mentioned, defect closure tends to be contraindicated in terms of ES. This can be justified as the cardiac abnormality functions as a ‘relief valve’ towards the high pulmonary arterial pressure, hence maintaining systemic cardiac output via the right to left shunting, although at the expense of cyanosis (Brida & Gatzoulis, 2018). The symptoms typically present upon physical examination are central cyanosis with clubbing, RV heave and second heart sound, whilst more advanced cases tend to present with hepatomegaly and peripheral oedema (Rosenzweig & Krishnan, 2021).

1.4.1 Pulmonary Arterial Hypertension and Congenital Heart Disease

One of the most severe complications arising from congenital heart anomalies are those of the great vessels is pulmonary hypertension (Kaemmerer et al., 2018). Up to 30%

of adult and 75% of paediatric pulmonary arterial hypertension (PAH) cases have an association to the presence of CHD (Dimopoulos et al., 2014). The development of PAH in patients with congenital heart defects may arise due to the resulting increased pulmonary blood flow due to the systemic to pulmonary shunt (Zhu et al., 2018). This may hence also owe to intra/extracardiac shunts having unrestricted pressure/volume overload with regards to the pulmonary circulation. Hence resulting in induced stress, endothelial damage to the arteries and adverse pulmonary vessel remodelling (Brida & Gatzoulis, 2018). The diagnosis of PAH along with CHD aggravates the natural course of the underlying anomaly, being post-operative or post-interventional, and hence impacts the burden of the disease as well as its outcome. The high variety in CHD subtype impacts various categories of the disorder including clinical manifestation and outcomes, functionality, as well as prevalence (Kaemmerer et al., 2020). Hence the variance in CHD subtypes also largely impact PAH in accordance with the nature of the anomaly. This may be exemplified by the comparison of the development of early pulmonary vascular disease in early childhood in patients diagnosed with large unrestrictive post-tricuspid shunts, including VSDs, whilst patents with petricuspid defects including ASDs may exist for decades without the development of PVD (Rosenzweig & Krishnan, 2021). For instance, children having an increased risk for postoperative PAH are those with Down syndrome, complex CHD including Truncus arteriosus, transposition of the great arteries and AV canal defects, along with highly reactive PAH during the postoperative period (Rosenzweig & Krishnan, 2021).

The incidence of CHD associated with PAH has been on the decline in developed countries but still remains prevalent despite modern medicine advances. Eisenmenger syndrome also falls within this category, being developed in 3.5-7.1% of patients with CHD and PAH (Kaemmerer et al., 2020). In accordance with data acquired from two separate entities, the estimated prevalence of PAH associated with CHD was found to be between 10-11% of patients diagnosed with PH. This low percentage is said to increase as the number of patients surviving CHD increase and hence progress into adulthood. On the contrary, within the paediatric population patients with CHD attribute towards a larger portion of those diagnosed with PAH (Haworth & Hislop, 2009). The most recent value confirms the above, having an approximate prevalence of PAH in CHD patients at a wide range of 4.2 to 28% (Kaemmerer et al., 2020), 25% to 50% of these adults present with ES (Rosenzweig & Krishnan, 2021).

Within the past two decades, due to timely CHD diagnosis and cardiac surgical intervention, particularly when carried out at infancy, the survival of patients diagnosed with ES into adulthood has increased significantly, having a worldwide decline in ES incidence by 50% (Dimopoulos et al., 2010; Galie et al., 2008). In despite of this however, ES poses a significant problem in the case of patients having large shunts being unable to undergo reparative surgery prior to the development of pulmonary vascular disease (PVD) (Rosenzweig & Krishnan, 2021). In accordance with the ESC guidelines for management of grown-up CHD patients, general supportive therapies and measures which should be taken with ES patients include; regular assessments by PAH-CHD trained physicians, psychosocial support, maintenance of physical activity, regular immunisation against known pulmonary pathogens including influenza and pneumococcal infections, and in the case of females avoidance of pregnancy (Baumgartner et al., 2010). With regards to treatment of ES, the last decade has provided significant growth with regards to conventional and targeted PAH therapy for patients with ES. The majority of these treatments include digoxin, diuretics, and antiarrhythmics, although none of these having proven substantial improvement in ES survival (Rosenzweig & Krishnan, 2021). Generic treatment for PAH in patients with CHD has limited evidence for efficiency and safety, typically due to patients representing this cohort being either excluded from clinical trials, or underrepresented and inadequately characterised, this is when in comparison to data available on patients with other forms of PAH, such as idiopathic PAH (Kaemmerer et al., 2020).

1.4.2 Underlying genetics of PAH

PAH, being one of the major subgroups falling under pulmonary hypertension has been heavily researched throughout the years, particularly with regards to the genetic aetiology. Between 6-10% of PAH patients not associated with any other underlying disorders bay be attributed towards family history (Morrell et al., 2019). Therefore, genetic studies of PAH independently have identified eleven known genes increasing the risk of PAH development (Best et al., 2014; Chida et al., 2012; Kerstjens-Frederikse et al., 2013; Nasim et al., 2011). Many of these identified risk-causing genes encode for members within the signalling pathway involving the bone morphogenetic protein (BMP)/ transforming growth factor beta (TGF-β). These proteins play a significant role in the vasculogenesis and embryological heart development. Variants in other genes

within the family of TGF-β amongst other gene families including caveolin 1 (*CAV1)* SMAD members 4 and 9 (*SMAD4* & *SMAD9)*, the T-box family including *TBX4,* amongst others, have been identified as less frequent and/or rare cause of PAH (Best et al., 2014; Chida et al., 2012; Kerstjens-Frederikse et al., 2013; Nasim et al., 2011). The above-mentioned genes may be described as BMP receptor signalling intermediaries, and the sequencing of these genes allowed the confirmation of the role of BMPR2 in PAH. Variants in *CAV1* for instance were identified to physically colocalise BMP receptors, whilst other rare variants in this gene induced association with PAH. *KCNK3*, a gene encoding for potassium channels which contribute to membrane potential and hence determine pulmonary vascular tone. Variants in this gene were also identified via exome sequencing and linked to PAH development (Morrell et al., 2019). Loss of function (LOF) and deletions within *TBX4* were identified as the most common genetic cause in childhood PAH. (Levy et al., 2016) This hence suggests that PAH may be (at least) partially attributed as a developmental lung disease when presented in early life stages (Zhu et al., 2018).

Nowadays, it has been well established that between 70-80% of families with PAH cases may be attributed towards variants found within *BMPR2.* The most commonly identified gene associated with the cause of PAH is that of *BMPR2*. Variants within this gene were identified in approximately 70% of familial PAH cases and 25% of idiopathic PAH cases (Evans et al., 2016). Hence inducing the interest of investigation of the role of *BMPR2* in PAH. The BMPR2 protein is expressed on the surface of a high variety of cells, being particularly expressed on the pulmonary vascular endothelium. Here it forms a complex with another receptor as a response to circulating BMP ligands and coreceptors. These receptors and coreceptors required for the response and complex formation of BMPR2 are also found expressed in high quantities in the pulmonary endothelium. Therefore, high levels of this signalling within the pulmonary endothelium may contribute towards lung-specific effects of *BMPR2* variants. Unlike other members in the TGF-β family, BMP signalling through BMPR2 results in the inhibition of proliferation and migration of smooth muscle cells and endothelial cells, whilst preventing neointimal formation (Tatius et al., 2021). Loss of *BMPR2* also results in endothelial dysfunction and promotes endothelial to mesenchymal transition (Morrell et al., 2019) via the overactivation of the counter pathway via TGF-β (Tatius et al., 2021). Overactivation of TGF-β results in the increase of migration and proliferation of

endothelial and smooth muscle cells, and hence the progressive remodelling of vasculature (Hemnes & Humbert, 2017). Therefore, variants in this gene resulting in BMPR2 deficiency have been associated to the poor prognosis of PAH. Typical characteristics identified in patients expressing mutant *BMPR2* include early disease onset at an approximate age of 10 years, 35% greater pulmonary vascular resistance, 8 mmHg higher mean pulmonary arterial pressure, and overall poorer survival rate (Sztrymf et al., 2008). Whilst on the other hand, variants resulting in the increased activity of BMPR2 result in a decline in proliferation and cellular growth rate of the pulmonary vasculature (Tatius et al., 2021).

Other than the presence or absence of genetic variation within the *BMPR2* gene, penetrance is also a crucial factor to take into consideration. Unfortunately to date, the penetrance of disease phenotype with relation to this gene is still incomplete (Morrell et al., 2019). In terms of sex penetrance, it has been estimated that males only carry an approximate 14% penetrance, whilst females that of 42% (Larkin et al., 2012). Other than the epigenetic and environmental, alternative genetic factors may include the expression of the wild-type *BMPR2* provided by the unaffected allele, other genetic variants influencing TGF-β expression levels, as well as alternative splicing of *BMPR2* (Morrell et al., 2019). Studies performed on included pluripotent stem cells derived from unaffected *BMPR2* variant carriers suggested that factors such as genetic background, including variations in BMP pathway-modifying genes expression, may contribute towards penetrance (Gu et al., 2017).

As mentioned above, certain subtypes of PAH may be attributed towards genetic factors. This concept has been heavily studied in the recent years, shining light on the genetic interaction of both CHD and PAH. The most recent developments in this field having sequenced PAH-CHD patients, have indicated that the genetic contribution is minimal for known/candidate risk genes for PAH or CHD alone. Other than the known above-mentioned PAH risk genes, a novel variant *SOX17* was identified explaining up to 3.5% of cases out of 256 participants (Zhu et al., 2018). This gene has also been found to have an association with idiopathic PAH through other studies, but towards a lesser effect size. *SOX17* is a member of the SOX family of transcription factors being hight conserved and widely expressed in development (Corada et al., 2013). Subgroups of this family participate in the vasculogenesis processes as well as remodelling (Francois et al., 2010). For instance, during embryonic vasculature, *SOX17* is selectively expressed in arterial

endothelial cells. SOX17 induction interacts and inhibits the WNT/β-catenin signalling pathway via direct protein interaction between the carboxyl terminal domain of SOX17 and β-catenin, which are required for the transactivation of target genes (Zorn et al., 1999). Another gene recently identified not previously mentioned is the potassium channel gene *KCNK3*, which via exome sequencing was identified as a PAH risk gene.

Throughout the research performed in the field of genetics with regards to this disorder, not only have diagnosis developed, but treatment strategies also. Identification of genetic variants being the causative agents to disease, allows the analysis of targeted therapy. A strategy being recently approached tackles the recovery of *BMPR2* expression in PAH through gene therapy (Tatius et al., 2021). When performed in rat models, the induction of exogenous *BMPR2* expression via pulmonary endothelia targeting, allowed the restoration of BMPR2 protein levels within these human cells, and hence reduced the PAH phenotype in these animal models (Reynolds et al., 2012). Alternative therapy being researched involves the epigenetic modulation therapy targeting specific MicroRNA (miRNA) which disrupt the *BMPR2* pathway. Theoretically, administration of these specific miRNA may enhance the expression of other downregulated miRNA (Courboulin et al., 2011). Unfortunately, in terms of clinical application, due to the complexity of the lungs this method yet remains ineffective (Tatius et al., 2021). Another point of view which has been researched with regards to treatment via restoration of BMPR2 regulation is via the administration of BMP ligands. Administration of BMP ligands in vivo enabled the reversal effect of VEGFR inhibitor and increased *BMPR2* gene expression in PAH mouse models (Long et al., 2015). Although research has allowed the development of the above-mentioned drugs for the treatment of *BMPR2* mutation-induced PAH, external scientific reviews suggest further research and rigorous testing be performed for the monitoring of safety, effectiveness and possible side effects these drugs may incur on patients (Tatius et al., 2021).

1.5. Aim

The aim of the study is to investigate the potential genetic aetiology underlying complex cyanotic congenital heart disease in a proband having unaffected parents using a trio whole exome sequencing approach.

1.6. Objectives

The objectives of this study may be summarised into 5 main points;

1. Describe the salient clinical characteristics of the affected proband with cyanotic CHD.

2. To derive a list of candidate genes implicated in CHD through a systematic literature search.

3. To perform whole exome sequencing, alignment to human reference genome, variant calling, quality control, and variant filtering/prioritisation according to different Mendelian disease segregation models.

4. To annotate and interpret shortlisted variants according to ACMG/AMP guidelines.

5. To perform in-silico modelling of selected shortlisted missense or splice site variants having a likely clinical impact.

1.7 Rationale

It is anticipated that this research provides a preliminary insight into the role of genetic factors driving congenital heart disease in the local population, and thereby lay the foundation for larger genomic studies, translational research and personalised genomic-medicine driven approach. Understanding the genetic basis of the disease helps to identify patients and at-risk family members, facilitates early diagnosis and therefore better long-term outcome and leads to a better understanding of the disease.

Chapter 2: Methodology

2.1 Patient recruitment:

The affected proband and parents described and analysed in this study had approached the study supervisor via email, expressing interest to participate in research on the proband's unexplained complex congenital heart disease. No prior genetic studies had been conducted in the family, and the cause of the proband's phenotype was unascertained and could not be attributed to environmental exposures, infective or maternal complications during gestation. The study was subsequently approved by the institutional ethics review committee of the University of Malta (MED-2022-00328 Appendix C). The study protocol was in compliance with the Declaration of Helsinki and was also approved by the proband's consultant cardiologist and a consultant medical geneticist. All subjects gave written informed consent for their participation in the study and for genetic analysis.

Following ethical review and approval, the proband and parents were invited to participate in a brief interview. Clinical data relevant to the proband's diagnosis of congenital cyanotic heart disease was obtained from review of case notes, discussion with caring cardiologist and reports from surgical interventions.

2.2 DNA extraction:

DNA extraction was carried out on whole blood stored in K2-EDTA tubes using the commercial QIAamp® DNA Blood Midi Kit. A modified protocol of 'Purification of DNA from Whole blood (spin protocol) was applied'.

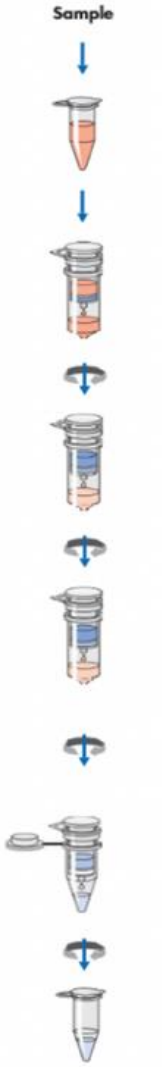
Initially, the samples had been equilibrated to room temperature and were mixed by being placed on a rotor for 20 minutes before the commencement of DNA extraction. Before the extraction, a heating block had been set to 70°C for use at a later stage. Centrifuge tubes of 15 mL capacity were labelled, and the buffers and protease necessary for this procedure were prepared according to the kit.

In the designated centrifuge tubes, 100µL of QIAGEN® protease was added to the base, followed by the pipetting of 1ml of the blood sample into the corresponding

tubes. To achieve a well-mixed solution, the tubes were then subjected to a vortex machine for fifteen seconds. Subsequently, 1.2mL of buffer AL was introduced into the centrifuge tubes, and homogenous mixing was achieved by inverting the tubes up to 15 times, followed by an additional 1-minute on the vortex machine to ensure homogeneity. The tubes were then subjected to a ten-minute incubation on the previously set heating block at 70°C. For efficient binding, 1mL of (96-100%) ethanol was pipetted into the centrifuge tubes, and the tubes were inverted 10 times, followed by fifteen-seconds on the vortex machine to guarantee a uniform homogenous solution.

The QIAamp® Midi columns were carefully positioned in the provided 15mL centrifuge tubes, ensuring not to dampen the rim. The solution from the previously prepared centrifuge tubes was then transferred to these columns. The tubes, now containing the columns and the transferred solution, underwent centrifugation at 3000rpm for three minutes. Following centrifugation, the columns were extracted from the tubes, the filtrate was discarded, and the columns were placed back into the same tubes. With caution to avoid moistening the rim, 2mL of buffer AW1 was pipetted into the QIAamp® Midi columns. Subsequently, the tubes were centrifuged at 5000rpm for one minute. After this centrifugation, the filtrate was retained, and 2mL of buffer AW2 was pipetted into the QIAamp® Midi columns, followed by further centrifugation at 5000rpm for fifteen minutes.

Upon completing this centrifugation, the QIAamp® Midi columns were transferred to new 15mL centrifuge tubes, and the tubes containing the filtrate were discarded. At this stage, the DNA from the blood samples was bound to the column. To elute the DNA, 200µL of room temperature buffer AE was pipetted into the QIAamp® Midi columns. Subsequently, the tubes were incubated at room temperature for five minutes, followed by centrifugation at 5000rpm for two minutes. Following the extraction process, the eluted DNA was preserved in 0.5 mL screw-capped tubes arranged in a 96-well storage format, specifically using Micronic®. Subsequently, these tubes were frozen for further processing.



Sample
400µL of thawed blood at room temperature, mixed on a rotator

Protein Digestion, cell lysis and binding
40µL and 400µL of protease K and Buffer AL were added to the sample respectively.
The mixture was homogonised and left to incubate at 56°C for 10 minutes.
400µL of ethanol was added and mixed. The mixture was then transferred into a QIAamp Mini spin column and centrifuged 8000 rpm x 1 minute.
The filtrate was discarded.

Washing
500 µL of Buffer AW1 was pipetted into the spin column and centrifuged 8000 rpm x 1 minute.
The spin column was transferred into a new collection tube and the filtrate was discarded.
Equal volume of Buffer AW2 was added centrifuged 14000 x 3 minutes.
The spin column was transferred into a new collection tube and the filtrate was discarded.

Elution
50µL of AE buffer was added into the spin column and incubated at room temperature for 5 minutes.
The spin column was centrifuged 8000 rpm x 1 minute.

Storage
The filtrate was pipetted in 0.5 mL screw-capped tubes in a 96-well storage format and stored in a freezer until further analysis.

Figure 2.2.1 A pictorial summary of the overall DNA extraction method being carried out via the QIAamp® DNA Blood Midi Kit.

2.3 DNA concentration quantification via UV spectrophotometry

The NanoDrop™ 2000 UV Spectrophotometer from ThermoFisher Scientific Inc, USA, was employed to quantify and assess the purity of the DNA samples that had been previously extracted. This step was crucial to verify that the extracted DNA met the required quality standards, ensuring its suitability for use in the subsequent stages of the study.

The UV spectrophotometer measures absorbance at wavelengths of 260 nm (A260) and 280 nm (A280), allowing the calculation of the A260/A280 ratio. A ratio falling between 1.70 and 2.00 is indicative of satisfactory DNA purity, enabling further analysis of the sample. The absorbance at 260 nm corresponds to nucleic acid concentration, while the absorbance at 280 nm reflects protein absorption, providing an indication of sample purity. The overall purity of DNA may be affected by specific proteins or solvents used during the DNA extraction process, potentially lingering through the final elution steps. The A260/A280 ratio is inversely proportional to protein concentration in the sample, meaning that a higher protein concentration will result in a lower ratio, indicating reduced purity. (Thermo Fischer Scientific, 2009)

The Beer-Lambert Law articulates the relationship between absorption, concentration (c), and path length (l), expressing that a sample's absorbance (A) is directly proportional to both concentration and path length. This relationship is mathematically represented as A = εcl, where c and l are as defined earlier, and ε represents the molar absorptivity. Additionally, absorption is influenced by two other factors: the light intensity of the sample and the blank. This influence is expressed by the equation $A = -\log(\frac{I}{I0})$, where Absorbance (A) is determined by the negative logarithm of the ratio of sample light intensity (I) to blank light intensity (I_0) (Parnis & Oldham, 2013).

The procedure for the DNA spectrophotometry described above is outlined as follows:

1. Prior to usage, the NanoDrop™ 2000 Spectrophotometer pedestal was wiped with a dry lint-free laboratory wipe to eliminate any residue from prior use.
2. Subsequently, 1.2 µL of elution buffer (buffer AE) was pipetted onto the pedestal to serve as the blank solution. The arm was gently lowered to perform the blanking

process. This generated a reference spectrum indicating a DNA concentration of approximately 0 ng/µL, ensuring the absence of functional errors.

3. The lower and upper pedestals were wiped with a fresh dry lint-free laboratory wipe before running each sample.
4. The DNA samples, obtained from the previous DNA extraction and stored in the freezer, were retrieved and thawed for approximately ten minutes.
5. For each DNA sample, 1.2 µL was pipetted onto the NanoDrop™ 2000 Spectrophotometer pedestal, and the arm was gently lowered.
6. The DNA sample's concentration in ng/µL, the A260/A280 ratio, and the spectral image were recorded.
7. Samples not achieving a DNA concentration value of ≥30 ng/µL and an A260/A280 ratio value between 1.70 and 2.00 underwent re-extraction and re-analysis until meeting these specified ranges.

2.4 Visualisation of DNA quality via Agarose Gel Electrophoresis

Agarose gel electrophoresis was employed on the previously extracted DNA samples as a confirmatory tool to assess the presence and integrity of DNA in these samples. This widely used biological laboratory technique is utilized for the separation of DNA, RNA, and proteins through a matrix under the influence of an applied electrical field. Molecules migrate based on factors such as molecular size, overall molecular charge, gel type, buffer ionic strength, and others.

Due to the overall negative charge of DNA, the molecules move from the negatively charged cathode towards the positively charged anode. The rate of migration is determined by the fragment's molecular weight, where the distance of migration is inversely proportional to the logarithm of the molecular weight. Therefore, the molecular weight of a fragment can be ascertained by comparing the distance it travels with that of a fragment of known weight (Yılmaz et al., 2012).

A 1% agarose gel was prepared in 100mL moulds, where 2 g of molecular biology grade agarose powder (Sigma-Aldrich®) was weighed and dissolved in 100mL of 1X Tris-Acetate-EDTA (TAE) buffer in a conical flask. The buffer consisted of 0.04 M Tris-Acetate and 1 mM EDTA. The mixture was gently swirled until fully dissolved, and the conical flask was sealed with plastic film, featuring a punctured hole to prevent pressure

buildup. The mixture was microwaved at 600 volts for approximately four minutes. Meanwhile, a gel cast was prepared on a flat surface, ensuring levelness using a spirit level.

After microwave heating, the plastic film was removed, and 10 µL of Ethidium bromide (EtBr) were pipetted into the conical flask with utmost care, given EtBr's DNA-binding fluorophore properties and toxic carcinogenic nature. The flask was gently swirled, and the solution was carefully decanted into the pre-levelled cast, taking care to prevent air-bubble formation. Any formed air bubbles were delicately removed using a sterile pipette tip to avoid interference with DNA migration in the gel. The combs were inserted into the gels and left for approximately twenty minutes to allow the gel to solidify and create wells.

After complete solidification, the combs were removed, and the gel was taken out from the cast, submerged in a Biometra Compact electrophoresis chamber containing fresh 1X TAE buffer solution. Various DNA samples (5 µL each) were pipetted into the wells, and one well contained 5 µL of λ Hind III size marker. The electrophoresis chamber was covered with the lid, and a voltage of 140 was applied for around twenty minutes until sufficient separation was observed. Subsequently, the gel was visualized under UV light using the BioDoc-It^{Tm} by UVP Gel Imager System, Thermo Fischer Scientific.

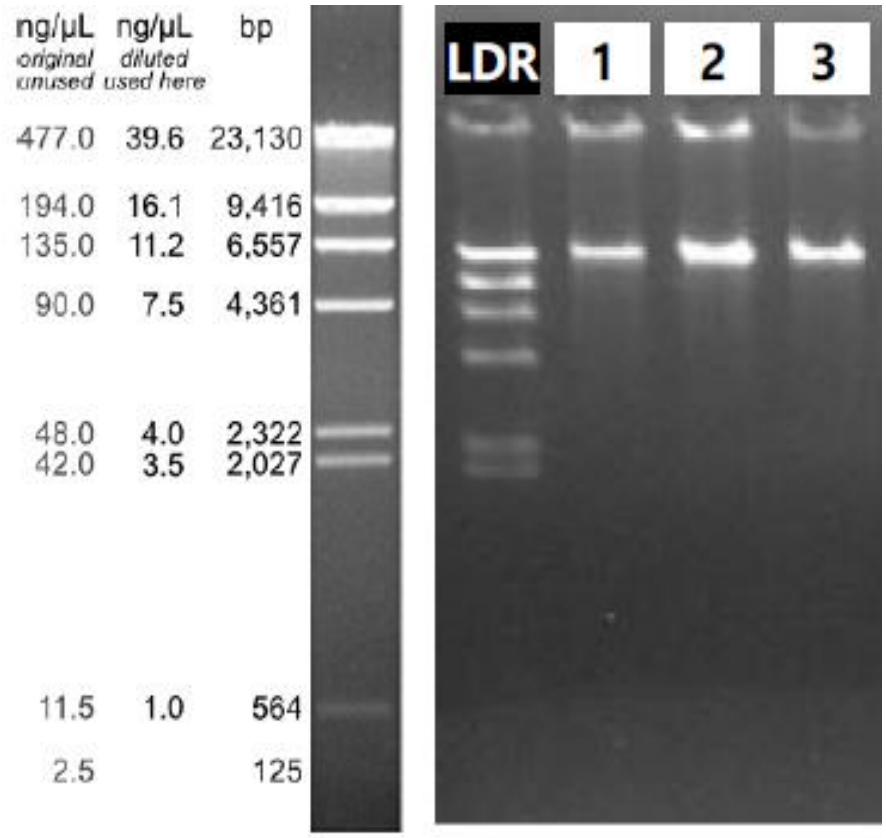


Figure 2.4.1. Figure of resulting 1% agarose gel electrophoresis of high molecular weight DNA prior to downstream whole genome sequencing and further processing. Low Density Ladder (LDR). Column 1 represents the proband, column 2 the mother and column 3 the father.

2.5 Gene panel selection

Six hundred thirty-five (635) genes were selected to be included in the gene panel which was utilised for the analysis of whole exome sequencing data. These 635 genes were chosen due to their implication in the phenotypes presented within the proband, including CHD, PAH, Hypertrophic osteoarthropathy (HOA) and situs inversus. The complete list is available in Appendix A.

2.6 Whole Exome Sequencing

The below **figure 2.6.1** shows an overview of the methodology workflow utilised for the acquisition of Exome Sequencing Data. The first two steps shown in the figure (genomic DNA extraction and fragmentation) along with the last step of alignment and data analysis was performed inhouse, whilst steps 3 till 5 (library preparation, cluster

amplification and sequencing) was outsourced to partner laboratories TheragenBio in South Korea, as described below.

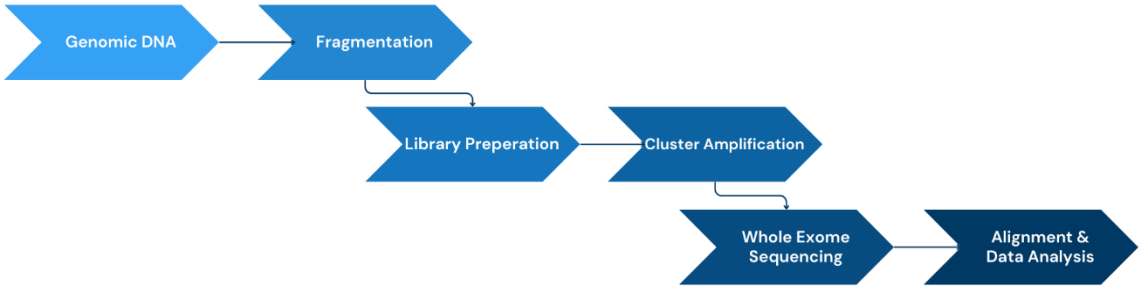


Figure 2.6.1. A figurative representation of the workflow performed for the acquisition of Whole Exome Sequencing (WES) data. Primarily the workflow starts with genomic DNA which undergoes fragmentation via sonication. The fragmented DNA further undergo the library preparation via the capture and denaturation via biotin oligonucleotides. The purified library undergoes cluster amplification for the subsequent whole exome sequencing (WES). The results from the WES are further aligned and the output data is analysed.

The SureSelect v6 (provided by Agilent®) was utilised for the performance of WES. This target capture design enables the targeting of 3GB of exome assembly hg19. The manufacturer's protocol was followed for the preparation of individual exonic DNA libraries from 300ng genomic DNA. The Agilent SureSelect XT All Exon v6 library capture kit was purchased through commercially available sources. This kit has an optimised design to target 99% of the coding regions found within RefSeq, HGMD and CCDS. This optimised design provides in depth coverage of targeted regions aiding in the analysis of protein-coding genomic regions (Agilent Technologies, 2015).

One of the earlier steps is that of ultra-sonication which allows the fragmentation of genomic DNA via the irradiation of a sample liquid with ultrasonic waves (>20kHz) at 150-200bp in size. The Illumina Q800R2 Sonicator was used for the fragmentation of DNA. These generated DNA fragments can be purified via Agencourt AMPure XP magnetic beads. The final steps of library preparation include the ligation of adaptors performed via the SureSelect XT kit including dNTP mix, polymerase and kinases including repeated PCR amplification. The adaptor-library undergoes a purification step as previously mentioned before a final amplification. The amplified adaptor-ligated DNA library is then hybridised utilising biotin-labelled capture oligonucleotides. The enrichment and purification of the captured sequences is performed via streptavidin-conjugated paramagnetic beads. Each library is then indexed via post-hybridisation amplification, following a final purification step.

Sequencing of the library was performed via the Illumina Novaseq 6000 sequencer at a depth of 50X outsourced to the laboratories of Thereagen Etex Bio Institute. This method of sequencing allows the latest high throughput technology via the utilisation of flow cells providing increased multiplexing per lane whilst being more cost effective. High throughput sequencing utilises the fundamentals of first-generation capillary electrophoresis sequencing. That being the application of DNA polymerase in repeated sequential cycles to a DNA template strand for the incorporation of fluorescently labelled deoxyribonucleotide triphosphates (dNTPs). Each extension cycle allows the identification of the added nucleotides by fluorescence excitation. High-throughput sequencing contrasts to the first-generation sequencing via the extension of the above description to millions of DNA fragments being sequenced in parallel.

2.7 Exome sequencing alignment and data analysis

An inhouse bioinformatic pipeline was applied for the analysis of the sequencing data. Default Illumina RTA pipeline parameters were applied for image analysis whilst CASAVA was used for base calling. Burrows-Wheeler transformation algorithm was applied for the alignment and mapping of the sequence reads to the human reference genome (UCSC hg19. NCBI build 37), read duplicates were eliminated utilising Picard (Li and Durbin, 2009). Tools utilised at the beginning of the pipeline on the FASTQ file include; FastQC (Cock et al., 2010). for the analysis of sequence data quality and Cutadapt (Martin, 2011) for the removal of adapter sequences. SAMTools was utilised for the calling of SNVs and indels with reference to public datasets such as dbSNP, 1000Genomes and GnomAD (Li et al., 2009). Next GATK Unified Genotyper was applied for the calling of SNPs. This tool utilises a Bayesian genotype likelihood model for the analysis of alleles and Phred-scaled confidence values (McKenna et al., 2010). Finally, the raw VCF files produced are uploaded to Franklin by Genoox along with the gene panel curated inhouse for the annotation of genetic variants.

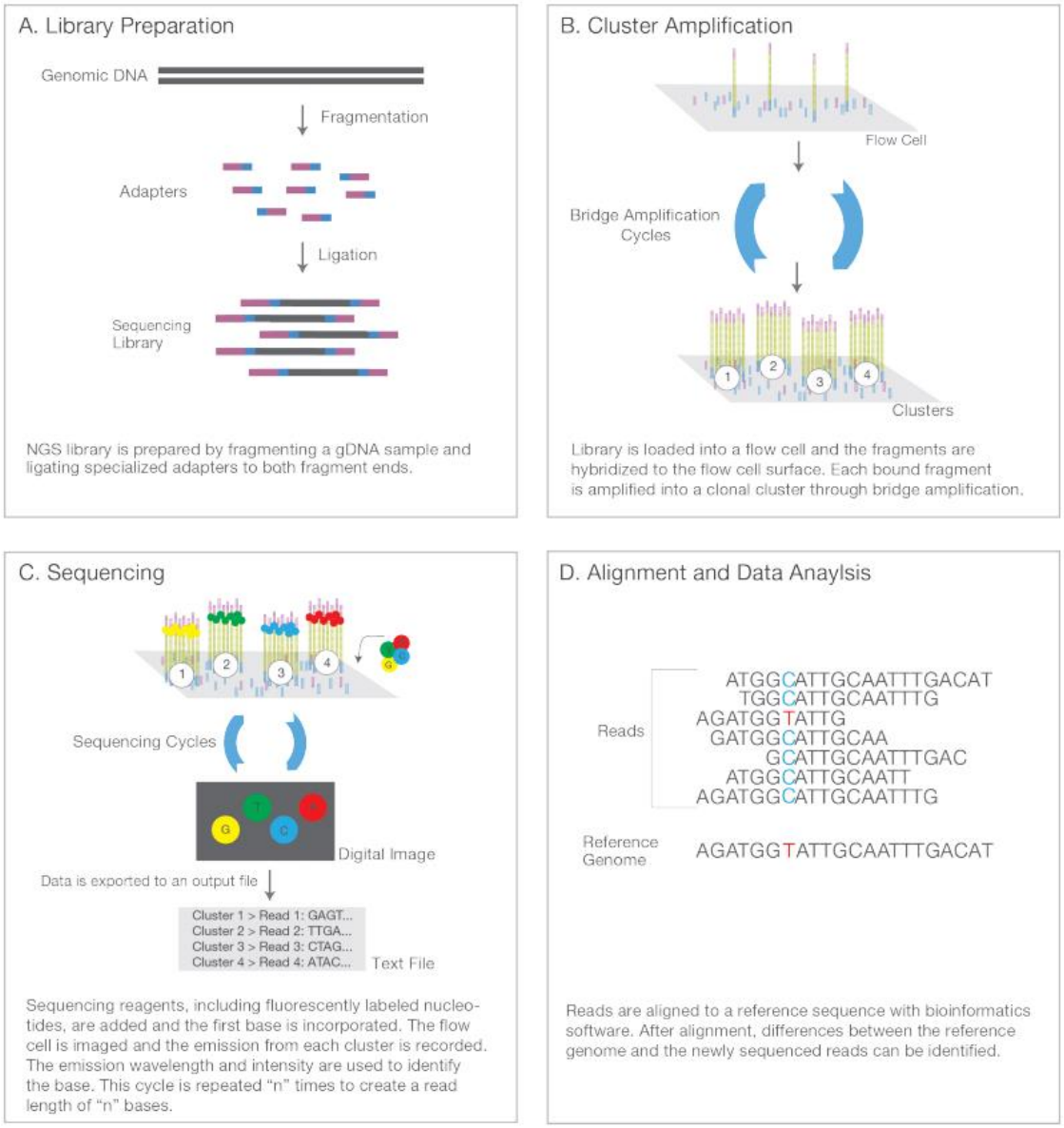


Figure 2.7.1. Summary of the whole exome sequencing (WES) workflow as described by Illumina. This includes four steps – A. Library preparation; B. Custer Amplification; C. Sequencing and D. Alignment and Data Analysis (Illumina product literature). The above steps A, B and C were outsourced to partner laboratories TheragenBio (South Korea), whilst step D was performed inhouse.

2.8. Bioinformatic workflow

FastQC v.011.7 was used for quality control of sequencing data, and Trimmomatic v.0.4.4 for trimming adapter sequences and low-quality sequences for fastq files. BWA-MEM v.0.7.17 was used for mapping and aligning sequencing reads to reference genome sequence using Burrow-Wheeler Algorithm. Samtools v1.8 was used to converting from SAM format to sorted, filtered, and indexed BAM files. Variant annotation was conducted using SnpEff (v4.3) to annotate genetic variants and predict

their effects. Qualimap 2.2.1 was applied to filter low-quality alignments and derive feature counts

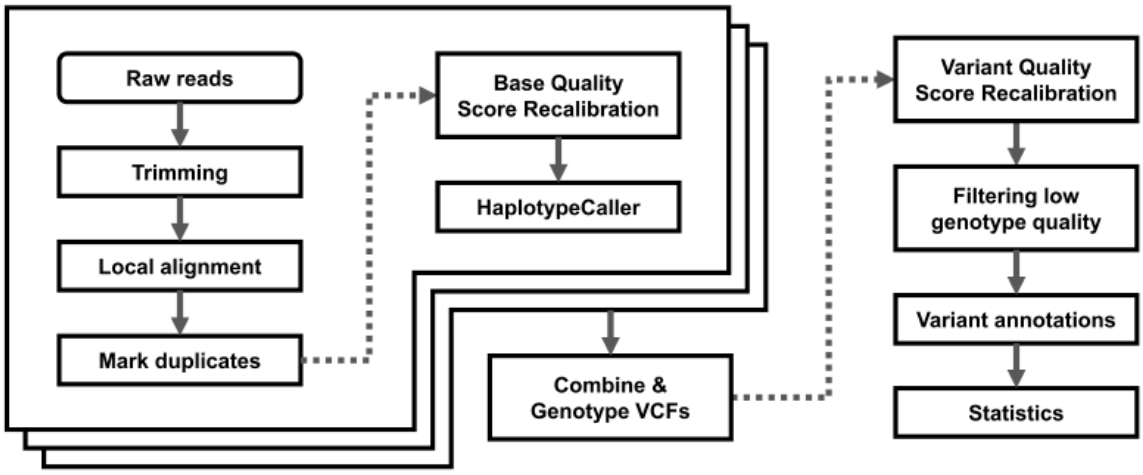


Figure 2.8.1. Bioinformatic analytical pipeline following GATK best practises recommendations.

2.9 Annotation of variant pathogenicity

The below steps were followed to determine and evaluate the pathogenicity of annotated variants.

1. Primarily the variants considered for pathogenic annotation included those having a confidence quality by depth at a medium to high level. This describes variants confidence according to their quality depth having been selected at a read of ≥10. This category ensures that the variants being annotated are not of a low quality.
2. Secondly the variants were further filtered according to genomic region. The regions being considered for pathogenic annotation where those of exonic regions and splice site regions namely the donor (+2), splice acceptor (-2) and splice region (+3->10). UTR at both the 3' and 5', as well as the intronic, intergenic, upstream and downstream regions were explored and reported, however were not included for variant pathogenicity annotation due to unlikely pathogenic effects arising from these regions.
3. A third filter aiding in the pathogenicity annotation of the variants within the trio is that referred to as effect. This filter allows the stratification of variants according to their protein altering affect. The effects of missense, stop gain, stop loss, start gain, start loss, frameshift and non-frameshift. These effects were selected in accordance with the ensemble database. The synonymous variants were explored however not

reported as part of the pathogenic annotation due to these variants not directly influencing protein structure and hence being less likely to lead to pathogenicity.

4. The variants were further filtered with respect to their pathogenicity annotation according to their frequency. The frequency explored was that of aggregated frequency which provides the frequency in accordance with multiple genetic frequency databases including 000 genomes, Exome Aggregation Consortium (ExAC), Exome Sequencing Project (ESP 6500), UK10K, GnomAD (Exome and Genome) amongst others. The frequency within these databases was selected at rare frequencies having <5% frequency.
5. The only variants considered for pathogenicity annotation were those having an LOF and sensitivity to missense mechanism of disease. The LOF filter is based on the gene constraint metric of observed/expected ratio in GnomAD. Genes are shown if pLoF o/e upper score <= 0.35, or if their pLI score >0.9, whilst the sensitivity to missense filter is based on the gene constraint metric of observed/expected ratio in GnomAD. Genes are shown if missense o/e upper score<=0.35.
6. The final determination with respect to the variant pathogenicity annotation was the American College of Medical Genetics/Association for Molecular Pathology (ACMG/AMP) classification described in section **2.10**.
7. All the variants were further annotated according to their *in-silico* predictors described in section **2.12**.

The above variant annotation is described including resulting values within section **3.2**.

2.10 ACMG/AMP classification

The American college of Medical Genetics (ACMG) and Genomics published their initial guidelines in 2000 with the aim of standardising the interpretation and reporting of sequence variation identified through clinical laboratory services. They primarily identified 5 categories by which any identified variants arising through clinical sequencing should be reported. The objectives of these categories were two-fold; primarily to provide a standardised framework for the interpretation and reporting of these variants amongst clinicians whilst also aiding in the education of physicians with respect to the variants being identified for the most accurate dissemination to the patients.

The five categories are summarised as follows;

1. The sequence variation identified has been previously reported and is a recognised cause of the disorder.

   Rigorous review of literature, locus-specific databases as well as central mutation databases (such as the Human Gene Mutation Database (HGMD), amongst others) should be undertaken prior to the reporting of a variant, so as to accurately assess that the variation is in fact causative of the disorder. In the absence of further functional studies, a concordance study within the family between phenotype and genotype is also of an acceptable criterion.
2. The sequence variation identified has not been previously reported but is expected to be causative of a disorder. Examples of which such variations include;
   a. Sequence variation resulting in the introduction of a stop codon or missense variation within the stop codon.
   b. Variation within the start codon (ATG).
   c. Sequence variation within the splice site.
   d. A shift in the mRNA reading frame via the deletion of one or more exons.
3. The sequence variation identified has not been previously reported and may or may not be causative of a disorder. Examples of such variations include;
   a. Splice consensus site variations.
   b. Variations which are likely to affect cryptic splice sites likely influencing transcription.
   c. Missense variations.

When validating category 3.c, a missense variation leading to a nonconservative substitution of an evolutionarily conserved amino acid, is likely to be causative of a disorder.

4. The sequence variation identified has not been previously reported and is unlikely to be causative of a disease.

   An example of which includes sequencing variation not resulting in the substitution of an amino acid, whilst also being unlikely to produce cryptic splice sites.
5. The sequence variation identified has not been previously reported and has been recognised as a neutral variant.

   Rigorous literature review of central mutation databases including HGMD and locus-specific databases should be undertaken to assess the available degree of certainty pertaining to the sequencing variation being previously reported as neutral.

These standards are always being revised since sequencing technology develops at a rapid pace. The Association for Molecular Pathology (AMP), College of American Pathologists (CAP), and ACMG collaborated during a workshop in 2013 that resulted in the creation of two scoring systems. The ACMG advises replacing the words "polymorphism" and "mutation" with the term "variant" due to the confusion and false assumptions around them. The following modifiers should be included when describing the term "variant" in order to provide an appropriate description (Richards, S. et al., 2015).

1. Pathogenic,
2. Likely Pathogenic,
3. Uncertain Significance,
4. Likely Benign, or
5. Benign.

While acknowledging that these modifiers may not be applicable to all human phenotypes, they do enable the establishment of a five-tiered system for the categorisation of variations pertinent to Mendelian diseases. This variant classification system is not specialised towards the categorisation and interpretation of multigenic non-Mendelian complex disorders or of somatic and/or pharmacogenomic variations. It is impertinent that one acknowledges that these guidelines were not intended for the identification of new genes in diseases, and this must be considered when utilising these guidelines. Along with the aforementioned modifiers, ACMG presents two sets of criteria: one for pathogenic/likely pathogenic variations and one for benign/likely benign variants (Richards, S. et al., 2015).

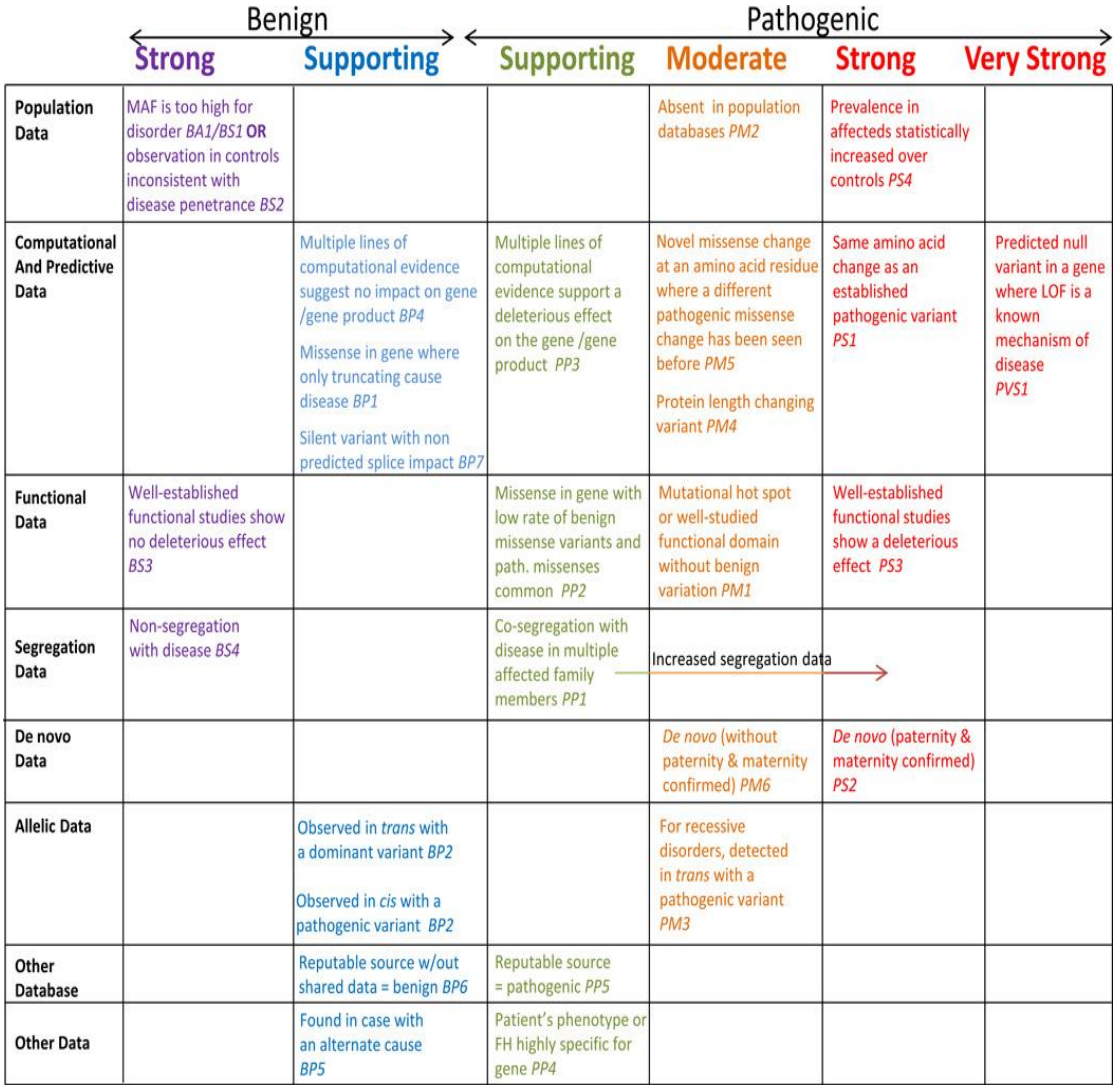
Variants following the Pathogenic/Likely Pathogenic classification may be stratified according to 4 criteria. On the other hand, variants following the Benign/Likely Benign classification may be stratified according to 3 criteria (**table 2.10.1**).

Table 2.10.1. Tabulated summary of the sub-stratification of both ACMG/AMP criteria (pathogenic/likely pathogenic, and benign/likely benign). (Richards, S. et al., 2015).

Pathogenic/Likely Pathogenic	**Benign/Likely Benign**
Very **S**trong evidence for **P**athogenicity (**PVS**1)	Stand-**a**lone evidence of **b**enign impact (**BA**1)
Strong Evidence of **P**athogenicity ranging from 1-4 (**PS**1-4)	**S**trong evidence of **b**enign impact ranging from 1-4 (**BS**1-4)
Moderate Evidence of **P**athogenicity ranging from 1-6 (**PM**1-6)	Supporting evidence of benign impact ranging from 1-7 (BP1-7)
Supporting evidence of **P**athogenicity ranging from 1-5 (PP1-5)	

Variants being described as PVS1 are referred to as null variants within a gene including but not limited to nonsense, frameshift and exon deletion variants, where the mechanism of disease is known and results in a loss of function (LOF). On the other hand, variants having an allele frequency of <5% as identified via 1000 genomes and ExAC are defined as 'stand-alone evidence of benign impact'. Variants of Uncertain Significance are classified either when the criteria are unmet or when the Benign and Pathogenic classification criteria are contradictory. (Richards, S. et al., 2015).

Figure 2.10.1 shows the Evidence Framework published by ACMG for the criteria of each abovementioned classifications.



	Benign		Pathogenic			
	Strong	Supporting	Supporting	Moderate	Strong	Very Strong
Population Data	MAF is too high for disorder *BA1/BS1* **OR** observation in controls inconsistent with disease penetrance *BS2*			Absent in population databases *PM2*	Prevalence in affecteds statistically increased over controls *PS4*	
Computational And Predictive Data		Multiple lines of computational evidence suggest no impact on gene /gene product *BP4* Missense in gene where only truncating cause disease *BP1* Silent variant with non predicted splice impact *BP7*	Multiple lines of computational evidence support a deleterious effect on the gene /gene product *PP3*	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before *PM5* Protein length changing variant *PM4*	Same amino acid change as an established pathogenic variant *PS1*	Predicted null variant in a gene where LOF is a known mechanism of disease *PVS1*
Functional Data	Well-established functional studies show no deleterious effect *BS3*		Missense in gene with low rate of benign missense variants and path. missenses common *PP2*	Mutational hot spot or well-studied functional domain without benign variation *PM1*	Well-established functional studies show a deleterious effect *PS3*	
Segregation Data	Non-segregation with disease *BS4*		Co-segregation with disease in multiple affected family members *PP1*	Increased segregation data →		
De novo Data				*De novo* (without paternity & maternity confirmed) *PM6*	*De novo* (paternity & maternity confirmed) *PS2*	
Allelic Data		Observed in *trans* with a dominant variant *BP2* Observed in *cis* with a pathogenic variant *BP2*		For recessive disorders, detected in *trans* with a pathogenic variant *PM3*		
Other Database		Reputable source w/out shared data = benign *BP6*	Reputable source = pathogenic *PP5*			
Other Data		Found in case with an alternate cause *BP5*	Patient's phenotype or FH highly specific for gene *PP4*			

Figure 2.10.1. The ACMG provided Evidence Framework regarding the classification of variants depending on their strength along with their associated Benign or Pathogenic classification (Richards, S. et al., 2015).

2.11. GnomAD database

Genome Aggregation Database (GnomAD) is a large publicly available database of 15,708 whole genomes and 125,748 exomes compiled from 141,456 individuals. This dataset contains 229.9 million variants of which 43,769 variants are high confidence loss of function (LoF), the majority including canonical transcripts of 16,694 genes. This dataset references genomic positions in accordance with GRCh37/hg19 reference sequence. The GnomAD dataset is representative of major global ethnicities grouped as; European, African, African/American, Asian, and Other. The other ethnic category groups all other ethnicities which do not fall into the above mentioned 4 major groups. The cohort of participants in the v2 of GnomAD are broad in terms of age range, whilst also excluding any duplicate individuals and first/second degree relatives. This assists in minimising the inflation of rare variants. The reference of GnomeAD in the population percentage for each variant seen in the results section, is representative of the variant according to the above description.

The observed/expected ratio allows the degree of intolerance to pLOF variation to be assessed in each gene, thus further estimating a confidence interval around this ratio. A low o/e value signifies stronger evolutionary selection for the particular variant class in comparison to a high o/e value. The o/e ratio was compared for missense variants across four genes of the GSC allowing the evaluation of evolutionary gene constraints. Both the observed and expected values depend on the gene and sample size simultaneously. (Karczewski et al., 2020)

The above description of GnomAD was adapted from the 7 papers published by the creators of GnomAD; Collins et al., 2020; Cummings et al., 2020; Karczewski et al., 2020; Minikel et al., 2020; Wang et al., 2020; Whiffin et al., 2020a; Whiffin et al., 2020b.

2.12 In silico prediction tools

2.12.1 Variant filtering – Franklin by Genoox

Franklin© by Genoox is an artificial intelligence (AI) tool for the annotation of variants utilising multiple in-silico predictors along with the ACMG/AMP guidelines (Einhorn, E. et al., 2019; Einhorn, Y. et al., 2018; Jackson et al., 2024a; Jackson et al., 2024b). Franklin© can be accessed through https://franklin.genoox.com. Till date there is no official publication released by Genoox describing the AI variant prioritisation engine behind Franklin©, however the website provides a detailed description behind the variant prioritisation engine, whilst the in-silico predictors utilised are heavily published. Franklin© provides variant interpretation at a 94.5% sensitivity and 96.6% specificity for the identification of variants as pathogenic or likely pathogenic (P/LP) as analysed by Mighton et al., in 2022. A publication by Genoox's Einhorn et al., in 2019 described the artificial intelligence-based variant classification engine (aiVCE) as being based on the ACMG/AMP standards and guidelines for sequence variant classification whilst putting these standards through an automated classification system. They further mention that the prediction models at the gene and rule level of the engine was based and built on various data sources including but not limited to; ClinVar, ClinGen, Uniprot,GnomAD, ExaC, Orphanet, amongst others. Franklin© also allows the strength of the evidence provided to be estimated according to different features including number of submitters along with the dates and type of submitters, and number of publications (Salfati et al., 2019). Franklin© allows the classification of variants into one of five categories; Benign (B), Likely benign(LB), Variant of uncertain significance (VUS), Likely pathogenic (LP) and Pathogenic (P). The status of the VUS is further confirmed according to the combination of in-silico predictor tools including REVEL, MetaLR, MutationTaster, MutationAssesor, FATHMM, SIFT, CADD and POLYPHEN. Franklin© further goes into the in-silico predictor for splice site prediction using dbscSNV, Ada and SpliceAI, along with GERP for the prediction of conserved regions, GenoCanyon. fitCons and ncER for whole genome functional annotation (Salfati et al., 2019). A recent study by Jackson et al., in 2024 compared the interpretation of variants present within an electronic health record between a clinical geneticist and Franklin© having a 94% concordance rate. The concordance rate of upgraded variants (benign > pathogenic) resulted in a percentage of 99% (Jackson et al.,

2024). As described by the Franklin© by Genoox website the advanced intelligence-driven engine is designed to prioritise and interpret variant data utilising multiple tools from diverse sources to target the most likely causal pathogenic variants. This engine provides information along with prioritisation of various variant types including single nucleotide polymorphisms (SNPs) and indels, copy number variants (CNVs), structural and compound variants. This priority engine relies on 5 core elements: (1) variant classification, (2) genotype association to clinical as well as phenotypical evidence, (3) gene/disease known inheritance models, and (4) the level of technical confidence for the variants in accordance with the variant caller. The algorithm also takes into consideration possible inheritance models pertaining to each variant within the variant prioritisation. This applies to single, trio as well as large family pedigree analyses. This information is accumulated from reputable published and curated sources allowing the consistency to be maintained within features of inheritance models.

The below is a description of the filters embedded in Franklin© as described in their website.

2.12.1.1 Singleton filters – SNVs or indels.

1. Phenotypes: Phenotypes of choice can be included into Franklin© filtering to only display variants having a known association to the chosen filter.
2. Gene properties:
    a. Gene inheritance: Known inheritance patterns of the genes according to their associated condition. This is curated from multiple sources including OMIM. MONDO, Orphanet, Decipher, GENCC amongst others.
    b. Mechanism of disease:
        i. Sensitivity to LOF – the sensitivity of the gene to loss of function mutations. This is based on the gene constraint metric observed/expected ratio present in GnomAD.
        ii. Sensitivity to Missense – the sensitivity of the gene to missense mutations. This is based on the gene constraint metric observed/expected ratio present in GnomAD.
    c. Public curated panels: panel based on known publicly curated panels including ACMG secondary findings and/or OMIM morbid genes.

3. Franklin classification: This allows the filtering of variants in accordance with Franklin© internal automated classification system. The internal VUS classification sub-categorises VUS into leaning-benign and leaning-pathogenic.
4. My organisation's classification: A personally curated classification may be uploaded and utilised.
5. Compare with: This filter allows the comparison of variants to other singleton samples within the internal case list.
    a. Shared variants:
        i. Identical – variants having similar zygosity.
        ii. Opposite – variants having opposite zygosity.
    b. Unique variants:
        i. Identical – variants having similar zygosity.
        ii. Opposite – variants having opposite zygosity.
    c. Panels: Any panel may be uploaded and used as a filter for variant selection within the panel. Multiple panels may be applied as combinational filters.
    d. Variant type: Filtering variants according to SNVs or indels.
    e. Region: Filtering results according to the variant region including exonic, intronic or UTR, ect.
    f. Effect: result filtering according to the effect of the variant including synonymous, missense, frameshift, ect.
    g. Regulatory: filtering according to Vista and Ensemble regulatory builds including enhancers and promoters.
    h. Chromosome: allows the filtering of variants according to chromosomal position.
    i. Zygosity: The filtering of variants according to Homozygous, Heterozygous, along with suspected compound.
    j. Allele balance (VAF).
    k. Frequency: In accordance with multiple population databases as well as internal frequency. The internal frequency includes variants present within the organization's cohort (controls uploaded to franklin). Aggregated frequency refers to the affrefation of several databases having a high weighting on GnomAD.

l. Confidence: Allows filtering according to the variant's technical confidence. The confidence value is calculated utilising multiple metrics including quality, depth, strand bias, ect.
m. Predictions: The filtering of variants according to their predictive results from various in-silico tools.
n. ClinVar evidence.
o. Variant callers.

2.12.1.2 Family specific filters

1. Inheritance.
   This allows for the filtering according to the calculated inheritance pattern of each variant. This does not refer to the inheritance pattern of the gene, however the inheritance of that variant within the family analysed.
   a. Single inheritance.
      i. Autosomal Dominant.
      ii. Autosomal Recessive.
      iii. X-linked recessive.
      iv. X-linked dominant.
      v. Y-linked.
   b. *De novo*: Franklin© states that some of these results may be due to sequencing errors and not attribute correctly to *de novo*.
   c. Compound heterozygote.
   d. Compound SNP/SV.
   e. Model strictness: Allows the control of the strictness/leniency of the inheritance model. When applied to the autosomal dominant inheritance pattern, 'lenient' may represent 'low penetrance' whilst 'strict' may represent 'complete penetrance'.
2. Family zygosity.

2.12.2 Sift Predictor

Sorting **I**ntolerant **f**rom **T**olerant (**SIFT**) is an algorithm designed for the computational prediction of amino acid substitution impact on protein function (Sim et

al., 2012). The initial SIFT algorithm was released as a website in 2001 providing end users with predictions on their variants (Ng & Henikoff, 2003). Since its publication, SIFT has been recognised as one of the standard tools for missense variation characterisation having recently been updated to include new features such as frameshift insertion/deletion tools (Hu & Ng, 2012), as well as independent data set metrics (Sim et al., 2012). It is imperative to note that SIFT is efficiently put into practise beyond the field of human research and human disease studies, and has also been successfully utilised in the analysis of missense variations within agricultural plans (Till et al., 2004; Till et al., 2007) along with various model organisms (Gharahkhani et al., 2011; Günther & Schmid, 2010; Guryev et al., 2004; Smits et al., 2004).

SIFT's performance was evaluated via two data sets obtained from UniProtKB, namely the human variation dataset (HumVar) and human divergence dataset (HumDiv) created by the authors of PolyPhen2 (Adzhubei et al., 2010) (described in section **2.12.3**).

Following the curation of the two datasets, they were mapped to various genomic databases including Ensembl, RefSeq as well as USCS Known ids via the UniProtKB mapping tool. Not all variants from the datasets could be linked or correlated, hence the complete quantity of variants utilised was fewer than the original dataset. Statistical calculations of the predictions by SIFT were defined according to; True positives (TP) being defined as disease causing variants having been predicted correctly as having an effect on protein function; False Positives (FP) are defined as neutral variants being incorrectly predicted to alter protein function; True Negatives (TN) are defined as neutral variants being correctly predicted as tolerated' and False Negatives (FN) are defined as variants being incorrectly predicted as tolerated and are in reality protein altering. Figure X shows the method in which the statistics are computed.

$$\text{Sensitivity} = TP/(TP + FN)$$
$$\text{Specificity} = TN/(TN + FP)$$
$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$
$$\text{Precision} = TP / (TP + FP)$$
$$\text{Negative predictive value (NPV)} = TN / (TN + FN)$$
$$\text{Matthews correlation coefficient (MCC)} = X / Y$$

where $X = [(TP \times TN) - (FP \times FN)]$ and $Y = SQRT[(TP + FP)\ (TP + FN)\ (TN + FP)\ (TN + FN)]$.

Figure 2.12.2.1. The statistical calculations taken into consideration as part of the SIFT algorithm for the determination of sensitivity, specificity, accuracy, precision, negative predictive value and the Matthews correlation coefficient.

The SIFT algorithm utilises sequence homology to compute the probability of an animo acid substitution having adverse effects on protein function. A crucial assumption with regards to SIFT algorithm is that evolutionary conserved regions are typically less tolerant to variants, hence the presence of amino acid substitutions and or insertions/deletions in said regions are more likely to influence function. The internal workflow within the SIFT algorithm is initiated with a query protein being searches against a protein database with the aim of obtaining homologous protein sequences. The sequence chosen is that having the appropriate diversity. The sequences are aligned and SIFT determines the composition of amino acids at that particular position, and computes a score (Ng & Henikoff, 2002). A SIFT score is obtained via a normalised probability of the observation of new amino acids at the position of the variation, the value of which ranges from 0-1. A SIFT value ranging between 0-0.05 is predicted to influence protein function (Ng & Henikoff, 2001).

The SIFT algorithm was further tested on external datasets independent of the training dataset having been the LacI, lysozyme and HIV protease substitutions (Ng, P. C. & Henikoff, 2001; Ng, Pauline C. & Henikoff, 2002). The SIFT algorithm was tested on the PolyPhen2's HumVar and HumDiv datasets. Since the number of comparative sequences chosen by SIFT is dependent on the protein database being used, the algorithm was tested on five protein databases namely Swiss-Prot, Swiss-Prot with Trembl, UniRef-50, UniRef-90 and UniRef-100, and the resulting prediction accuracies were measured. The results identified that the prediction accuracies were similar between databases,

however sensitivity and specificity varied according to the protein database used. Hence, considering these results, UniRef90 was considered for the pre-computed SIFT score due to its high coverage, sensitivity and balanced performance.

New features within the SIFT web server are those of (Sim et al., 2012) (1) genome-wide database of nonsynonymous variants predictions which is curated by (i) altering each coding base in the reference genome to the other three potential DNA bases, (ii) calculating the SIFT score for the resulting amino acid alterations and (iii) storing the amino acid changes and their corresponding predictions in a database. (2) Variant annotation is offered with dbSNP and/or 1000 genomes. (3) File format conversion is available since the first input format to SIFT was established prior to next generation sequencing, and hence has been updated to enable the conversion of formats including VCF, Pileup and GFF files. (4) Indel prediction tools are offered including indel annotation via VariantClassifier (Li & Stockwell, 2010), whilst a SIFT Indel prediction for frameshift indels has also been recently introduced.

2.12.3 PolyPhen-2

Polymorphism **P**henotyping (**P**oly**P**hen) server was first introduced in 2002 by Ramensky et al., and was described as having 3 primary uses namely, (i) as a web server for the annotation of functional nonsynonymous Single nucleotide polymorphisms nsSNPs, (ii) as a dataset of nsSNPs extracted from HGVbase, and lastly (iii) as an analytical tool for these datatypes for the predicted effect on protein structure and function. The initial version of PolyPhen was able to accept the amnio acid sequence of a protein, or SWALL database ID/acession number in conjunction with the sequence position at which the SNP is present and two amino acids characterising the polymorphism. The fully automated pipeline being run by PolyPhen requires the previously described input, and will internally (1) identify the nsSNPs in known genes after which (2) the substitution site is characterised according to sequence, this hence flows into (3) the profile analysis of homologous sequences which occurs via BLAST searching of a database, (4) the substitution site is then mapped onto a known protein's 3D structure also being done by BLAST sequence alignment of the query protein and a known protein from a database, (5) structural parameters are then utilised for the evaluation of amino acid substitution effect, (6) the residual contacts against 'critical

sites' are checked and finally (7) PolyPhen utilises predefined 'prediction rules' to determine whether an nsSNP is deleterious or otherwise (Ramensky et al., 2002).

PolyPhen-2 on the other hand, differs from its predecessor via the set of predictive features offered, the alignment pipeline and the classification method utilised (Adzhubei et al., 2010). The newer release utilises 3 structure-based in conjunction with 8 sequence-based predictive features, being selected for automatically (Adzhubei et al., 2010). Most of these features allow the comparison of a property from the wild-type allele against a property from the variant allele, hence defining an animo-acid replacement. The alignment pipeline in PolyPhen-2 utilises clustering algorithms for the selection of homologous sequences which are then analysed, and their multiple alignments are constructed and refined by this pipeline. The pipeline utilises Naïve Bayes classifiers to predict the functional significance of an allele replacement (Adzhubei et al., 2010).

The two datasets utilised for the training of PolyPhen2 are human variation dataset (HumVar) and human divergence dataset (HumDiv).
1.a. The HumDiv deleterious inventory was compiled utilising 3,155 annotated mutations known for being causative of Mendelian diseases in humans from UniProt.
1.b. On the other hand, the neutral HumDiv data set was curated via the comparison of 6,321 human proteins and their homologues in closely related mammals assumed to be non-damaging whilst identifying any altered amino acids.
2.a. Contrary to the HumDiv dataset, the HumVar deleterious dataset contained 13,032 variants from UniProt annotated as disease causing in humans, regardless of their Mendelian origin or otherwise.
2.b. The HumVar dataset was curated via 8,946 human nsSNPs not having been annotated as disease causing.

Comparison of PolyPhen-2's performance against PolyPhen via receiver operating curves comparing true positive rates against false negative rates, had a substantially superior result. At a false positive rate of 20%, PolyPhen-2 obtains a true positive prediction at 92% on HumDiv and 73% on HumVar (Adzhubei et al., 2010).

PolyPhen-2 has been utilised more recently in additional applications namely for the; identification of rare alleles causative of Mendelian diseases (Bamshad et al., 2011), scanning of alleles being potentially medically actionable within a human's genome (Ashley et al., 2010) as well as aiding in the deep sequencing of large populations by

providing profiling towards rare variations identified (Tennessen et al., 2012). The latest publication by Adzhubei et al., in 2013 describes protocols for the upgraded utilisation of PolyPhen-2's web interface. These include; the prediction of a single-residue substitution/reference SNP; the batch mode analysis of a large number of SNPs; and the search of precomputed predictions of the Whole human exome sequence in a database. The latest publication defines PolyPhen-2 as a software and web-based tool that predicts the impact of amino acid substitutions on human protein stability and function. It uses structural and evolutionary considerations, annotates single-nucleotide polymorphisms (SNPs), maps coding SNPs to gene transcripts, and estimates the probability of missense mutations being damaging. Key features include a robust alignment pipeline, machine-learning classification, and integration with the UCSC Genome Browser.

2.12.4 MutPred

MutPred was initially developed to target two aspects which had not been targeted by the available in silico mutation predictors of the time (including SIFT, PolyPhen, PANTHER, LS-SNP, SNAP, PMUT and CanPredict). The two aspects mentioned above are namely; the improvement of characterisation and identification of variation attributes moving beyond sequence composition, structure and evolutionary conservation for more accurate classification and hypothesis yield with regards to molecular mechanisms of disease, along with the development of new computational approaches for the improvement of accuracy with regards to classification when similar attributes and training sets are utilised (Li et al., 2009). The initial version of MutPred described in the 2009 publication by Li et al., aimed at the inclusion of various structural and functional properties of proteins to prove that gain and loss of such properties being predicted provide a substantial classification accuracy, hence enabling the underlying biochemical cause of disease.

As mentioned above, the development of MutPred was targeted at utilising various attributed from protein sequence in classification. The major attributes may be grouped into three major classes (table **2.11.4.1**); (i) protein structure and dynamics predictions, (ii) functional properties predictions, (iii) amino acid sequence and evolutionary information. For the algorithm to efficiently discriminate between neutral

polymorphisms and those being associated with disease, support vector machine (SVM) and random forest (RF) classifiers were compared and applied to the algorithm.

Table 2.12.4.1. Protein attributes taken into consideration by MutPred for more accurate protein sequence classification prediction along with the reference from which were taken directly by (Li et al., 2009).

Attributes to protein structure and dynamics	Reference
Secondary structure	(Rost, 1996)
Solvent Accessibility	(Rost, 1996)
Transmembrane helices	(Krogh et al., 2001)
Coiled-coil structures	(Delorenzi & Speed, 2002)
Stability	(Capriotti et al., 2005)
B-Factors	(Radivojac et al., 2004)
Intrinsic Disorders	(Peng et al., 2006)
Attributes to functional properties	
DNA binding residues	(Ahmad et al., 2004)
Catalytic residues	(Li et al., 2009)
Calmodulin binding sites	(Radivojac et al., 2006)
Phosphorylation sites	(Iakoucheva et al., 2004)
Methylation sites	(Fogel, 2006)
Ubiquitination sites	(Radivojac et al., 2010)

The initial MutPred builds on the previously available SIFT algorithm, allowing improvement with respect to the classification accuracy of human disease variation. It was proposed to accurately and reliably hypothesis the molecular foundation of disease for around 11% of previously known disease-causing variations. MutPred2 being published in 2020 by Pejaver et al., claimed to address the challenges within *in silico* variant predictors. These challenges are namely; (i) the lack of information regarding the potential mechanisms being affected by the query variation, and (ii) the mapping of predicted pathogenic substitutions onto protein feature annotations (Schnoes et al., 2009; Schnoes et al., 2013). Both of these challenges result in the lack of a model which explicitly models the type of chain in regional structure and function (Schnoes et al., 2009; Schnoes et al., 2013). The extended algorithm of MutPred2 quantifies amino acid substitution pathogenicity whilst describing the affect on phenotype via a broad repertoire model of amino acid sequence alteration resulting in structural and functional variation (Pejaver et al., 2020). The methodology of MutPred2 allows the estimation of pathogenic missense variants within the human genome, along with the identification of molecular

signatures associated with data sets containing both Mendelian disease variants as well as *de novo* mutations in individuals diagnosed with neurodevelopmental disorders (Pejaver et al., 2020). High-scoring variants from these datasets were prioritised and their functional roles were experimentally validated.

MutPred2 is a method and software based on machine learning that combines genetic and molecular data to probabilistically assess the pathogenicity of amino acid substitutions. It offers a general prediction of pathogenicity and a ranked list of specific molecular alterations that may impact the phenotype. The methodology utilised in the algorithm includes the estimation of prior and posterior probabilities, aiding in the interpretation of pathogenicity and molecular alteration scores. MutPred2 currently models various structural and functional properties, such as secondary structure, signal peptide, transmembrane topology, catalytic activity, macromolecular binding, post-translational modifications (PTMs), metal binding, and allostery (Pejaver et al., 2020). The algorithm's pathogenicity model was trained on 53,180 pathogenic and 206,946 neutral variants obtained from Human Gene Mutation Database (HGMD) (Stenson et al., 2020), SwissVar (Mottaz et al., 2010), dbSNP (Sherry et al., 2001) along with pairwise alignments. The algorithm's inferring molecular mechanism model was trained from a combination of datasets to ensure effective genetic and molecular data integration.

To overcome this challenge, we created MutPred2, a tool designed to deduce the structural, functional, and phenotypic implications of coding variants. MutPred2 enhances pathogenicity prediction and identifies potential molecular alterations by modelling the impact of variants on local protein structure and function, employing a suggested ranking approach. Additionally, the algorithm's ability to assign specific molecular impacts enables the quantification of molecular signatures within various datasets, such as different disease classes, specific diseases, healthy populations, subpopulations, and more. A notable function of MutPred2 is its capability to predict more pathogenic *de novo* missense variations in cases in comparison to controls, hence proving accuracy and specificity. It is remarkable to note that brain-relevant information which would increase the prediction performance was not exploited, the individual filtering step included the removal of genes common to both affected cases and controls. In comparison with the tools recommended in the ACMG/AMP Standards and Guidelines (Richards et al., 2015), MutPred2 compares positively via stringent independent tests performed.

2.12.5 FATHMM

Functional **A**nalysis **T**hrough **H**idden **M**arkov **M**odels (**FATHMM**) is a software and server providing a species-independent having optional species-specific weighting, for the prediction of functional effects of protein missense variants (Shihab et al., 2013). The authors of FATHMM built off the work of SIFT (Ng & Henikoff, 2001), PANTHER (Thomas et al., 2003), (Calabrese et al., 2009) along with the HMMER3 software (Eddy, 2009), to enhance the computational prediction of functional effects resulting from amino acid substitutions by employing hidden Markov models (Shihab et al., 2013).

The model presents two methods of analysis, one being a species-independent and/or unweighted method, whilst the other being a weighted/ species-specific method. The first unweighted method utilises an iterative search protocol whereby the automatic collection and alignment of homologous sequences is performed. The multiple sequence alignment result is then utilised for the interrogation of homologous sequences and sequence conservation within protein domains of protein families. The addition of this domain-based analysis allows the cohesive overview of important evolutionary and structural constraints which might have been missed via the utilisation of an automatically collected homologous sequence alignment (Shihab et al., 2013). On the other hand, the weighted method as its name implies, utilises 'pathogenicity weights' being derived from the relative frequencies of deleterious and neutral amino acid substitution maps on conserved protein domains. The weighted method for human variations outperformed performance accuracies of traditional methods including SIFT, PolyPhen, SNPs&GO, MutPred and PANTHER. FATHMM has therefore demonstrated that it may be efficiently applied to any high-throughput large-scale genomic dataset whilst providing the added benefit over other available tools of phenotypic outcome associations.

The databases for mutations utilised in FATHMM were curated from HGMD, UniProt, VariBench, SwissVar and Hicks et al., 2011. The functional consequence of protein function is predicted via the input of the submission query to the server/software, the protein domains are annotated in conjunction with the search for homologous sequences within databases. The protein domain and hidden Markov model (HMM) are only extracted if the domain assignment is deemed significant, having an e-value of <0.01 whilst also having the amino acid substitution mapping onto a match state within the model. An important assumption being made within the model according to calculated

probability is that; an amino acid probability reduction when comparing the wild-type to the mutant residue is indicative of a potentially deleterious effect on protein function, whilst an increase in amino acid probability would be indicative of a favourable substitution. Hence the model further assumes that large reductions in amino acid probability is directly proportional to larger effects, hence being a similar proportion of small amino acid probability to small effect. Following the above, the species-specific pathogenicity weights are applied to the previously-obtained results. The molecular as well as phenotypic consequences of amino acid substitutions are then annotated via domain-centric ontologies (de Lima Morais et al., 2011). The phenotypic consequences of amino acid substitutions are also annotated via the extension of these mappings onto various ontologies including; human, mammalian, and plant phenotype ontologies. The above-described predictions were finally evaluated according to calculations of true positives and true negatives with respect to accuracy, precision, sensitivity, specificity, negative predictive value, ant the Matthew's correlation coefficient. (Shihab et al., 2013[b]).

The pathogenicity weights were not directly utilised in the training of pathogenic sequence/ variant recognition. The weights are computationally capable of recognising protein domain's reaction to missense mutations (Shihab et al., 2013). The outperformance of this algorithm to traditional in silico prediction models reaffirms FATHMM's ability to identify critical structural and/or evolutionary constraints via HMMs curated manually which represent the alignment of conserved protein domains (Shihab et al., 2013). FATHMM has also been adapted to incorporate a cancer-specific model for the functional analysis of driver variants. This adaptation scored better in comparison to other available tools in terms of performance accuracy with regards to distinguishing between driver variants and other germ-line variants (being both neutral and deleterious) (Shihab et al., 2013[a]). In a later publication in 2014, Shihab et al., further incorporated a disease-specific weighting system into the previous FATHMM algorithm. The previous FATHMM including other available algorithms were not specifically designed to discriminate between disease-specific nsSNPs and other non-specific/functional variants (Shihab et al., 2014). Hence, the additional disease-specific weighting utilised 17 different disease categories, which in comparison to traditional prediction algorithms enabled a reduction in the quantity of false positive values, along with an investigation capable of prioritising nsSNPs (Shihab et al., 2014).

A mentioned limitation to the disease-specific methodology of FATHMM is that, in extreme cases, dominating pathogenicity weights may bias/exaggerate the variant effect. This might come into play upon the prioritisation of variants in proteins and their domains having strong associations with the disease concept under investigation. In this case, amino acid probability could be dominated by the pathogenicity weight, hence biasing the prediction (Shihab et al., 2014). When these weights are inkling towards a disease concept, neutral polymorphisms within various regions of a protein and its domains would be misclassified as 'damaging' (Shihab et al., 2014).

2.12.6 GERP

Genomic **E**volutionary **R**ate **P**rofiling (**GERP**) is initially described in 2010 by Davydov et al., where it was described as a statistical, biologically transparent algorithm for the identification of element constraints. GERP measures deficits in nucleotide substitutions at regions with high resolution and measures such deficits as 'rejected substitutions'. These rejected substitutions are utilised for the ranking and characterisation of constraint elements as they reflect the intensity of previous purifying selections. GERP has pushed towards novel direct estimates of constraints, in comparison to previously available models which do not offer this (Cooper et al., 2005).

The below description of the model has been adapted from Cooper et al., written in 2005 which was the first description of the GERP algorithm/framework.

This model analyses multiple sequence alignment from the human genome capturing approximately 3.85 neutral substitutions per site. The model works via the analysis at 4 levels namely; (i) Alignment, (ii) Neutral rate estimation and tree construction, (iii) identification of constrained elements, (iv) false positive rate identification, (v) clustering of constrained elements, and finally (vi) ultra conserved elements. The human and non-human sequencing data utilised in this model was an expanded version of the data generated by Thomas et al., in 2003. GERP takes various inputs into consideration, one input is the global multiple-sequence alignment whereby the orthologous bases are aligned as accurately as possible from each species. Another input is a species tree with branched lengths which allows the estimate of relative contribution in terms of neutral divergence for each species/ancestry lineage within the

tree. It is important to note that within the alignment, a compression takes place where the sequence has no gaps, hence ensuring consistence between coordinates of alignment and annotation, further ensuring accurate comparisons of constrained element annotations against other sequence features including exons and repeats. The alignment is hence associated with two vectors, one related to the observed evolutionary rates and another to expected evolutionary neutrality rates.

The first step of the internal processing is that of (i) sequence alignment. This entails the multiple sequence alignment which ensures high-scoring local alignments allow the adequate representation of global alignment. Non-human sequences are compared to human sequences, but primarily each sequence group was reordered and reorientated for the local alignment chains to be monotonic. In parallel to this shuffling, sequences lacking detectable similarity to the alternate sequence group are clipped and deleted. The second step of (ii) tree construction and neutral rate estimation followed a previously described method by Cooper et al., in 2003. This methodology enables the estimation of neutral divergence between closely related species, and hence subsequent exploration of rate estimates against relative tree branch-length. To obtain a source of aligned neutral DNA, all alignment regions that included clearly constrained elements in the human sequence were removed from the uncompressed global alignment.

The subsequent step (iii) in the model is that of constrained element identification. Candidate constrained elements are further defined via the analysis of alignment position stretches having ratios of observed to expected rates below a certain threshold. It is noteworthy to mention that a decrease in the observed to expected ratio threshold is directly related to an increase in stringency of element identification, resulting in fewer, smaller elements having a higher score per base. This model enables the merging of candidate elements across few intervening bases which do not meet the ratio criteria due to certain functional elements having unconstrained or weakly constrained bases. ‘rejected substitutions’ hence are formed via the sum of individual site differences scored between the observed and expected rates. Candidates which fail to meet the rejected substitution threshold are hence eliminated, and the remaining are interpreted as legitimate constrained elements. The majority of the results are based on identifying potential constrained elements using a threshold observed-to-expected ratio of 1 and the merging of tolerances. To capture smaller elements with higher per-base scores, an observed-to-expected threshold of 0 is employed. Conversely, to identify larger and

cumulatively higher-scoring elements, the merging tolerance is increased. For identifying large non-exonic constrained regions overlapping coding exons, positions within coding exons are disregarded, hence neither positively nor negatively impacting the scoring of constrained elements are considered. This allows a constrained element to overlap a coding exon without influencing its score.

Another crucial step as part of the GERP framework is that of (iv) false positive rate discovery. This is acquired via the analysis of permutations being randomly generated from alignment. Constrained elements identified via these permutations are representative of the discovery process and function towards the aggregate distribution of observed and expected evolutionary rates. Alignments having been found to reside in unambiguous constrained regions were excluded for the assessment of false discovery rate in neutral DNA. These alignments were defined by having constrained element scoring ≥25 rejected substitutions, hence excuding around 50,000 alignment columns. Additionally (v) constrained element clustering was obtained via analysis of the correlation between repeat and constrained element density, calculated via linear regression models. These densities were defined by employing successive, non-overlapping windows with a width of 25-kb across the entire length of the human sequence. Finally (vi) ultra-conserved elements were identified via the normalisation of the rejected substitutions score of each constrained element. These elements were then individually compared with an external genome database.

(Cooper et al., 2005)

GERP++ as described by Davydov et al., in 2010 provides (like GERP) a novel bottom-up method utilising rejected substitution as a constraint metric. However GERP++ utilises a more robust statistical likelihood estimation procedure for the computation of evolutionary rates, hence resulting in a 100-fold reduced turnover time. The updated version of this algorithm also introduced a novel criteria for grouping of constrained positions into constrained elements via statistical computations. This newer improved method predicts that a larger fraction of the human genome is found within constrained elements due to a very low false positive rate and the annotation of far fewer yet longer elements. (Davydov et al., 2010)

2.12.7 REVEL predictor

Rare **E**xome **V**ariant **E**nsemble **L**earner (**REVEL**) is an ensemble method for the pathogenicity prediction of rare missense variants (Ioannidis et al., 2016). REVEL incorporates individual, previously developed prediction tools including MutPred, FATHMM, VEST, PolyPhen, SIFT, PROVEAN, MutationAssessor, MutationTaster, LRT, GERP, SiPhy, phyloP, and phastCons. This model was trained on recently identified disease missense variants and rare neutral missense variants, which had not been included in the training dataset for its constituent predictors (Ioannidis et al., 2016). Along with REVEL, two large independent tests set formulated by recently discovered pathogenic and benign variants were developed to aid in the application of REVEL to newly identified variants by next generation sequencing.

The algorithm random forest was formulated via the R 'random forest' package having 1000 binary classification trees (Breiman, 2001; Hastie et al., 2009). Each split within the random forest had 4 features allocated at random. To combat the imbalance between the quantity of neutral training variants and disease training variants available, an equal number of both was selected for the generation of the bootstrapped training set for each tree within the forest (Ioannidis et al., 2016). Disease variants were gathered from the HGMD version 2015.2 (Stenson et al., 2014). The variants chosen for from this database were restricted to only missense disease variants dated from August 1st 2012, hence minimising overlap with previously trained feature components within the REVEL random forest. Additional missense exome sequencing variants were gathered from ESP (Tennessen et al., 2012), ARIC (The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators.1989) and 1000 genomes (Abecasis et al., 2012) databases of European-American and African-American populations. Any variants which has been used to previously train the individual components features comprising REVEL for both disease ant neutral variants, specifically MutPred, PolyPhen-2, MutationTaster, FATHMM and VEST.

The previously-mentioned features that compromise REVEL comprise 18 individual pathogenicity predictor scores being derived from 13 predictive features. For the purpose of this study, MutPred scores were re-computed utilising UniProt canonical protein sequences and when unavailable, Ensemble canonical transcripts. 16 scores were obtained from dbNSFP of which eight were functional prediction scores and another eight

were conservation scores. Upon the presence of multiple protein isoform associations between PolyPhen-2, FATHMM and PROVEAN, the average score of all isoforms was taken into consideration. (Ioannidis et al., 2016).

Two independent test sets were setup, having no overlap over either the REVEL training set, or the training sets utilised within the individual REVEL components. The first test consisted of disease variants acquired from SwissVar, whilst the second test consisted of pathogenic, likely pathogenic, benign and likely benign variants found within ClinVar according to the ACMG guidelines (Ioannidis et al., 2016). To eliminate any overlap between the two tests, ant variants being present within both SwissVar and ClinVar were excluded. Upon the application of REVEL to the two independent test sets, it had the best performance in comparison to individual tools and the ensemble methods including MetaSVM, MetaLR, KGGSeq, Condel, CADD, DANN, and Eigen. (Ioannidis et al., 2016)

The individual features comprising REVEL were characterised and correlated amongst each other. The conservation scores along with the functional scores being utilised were found to be highly correlated, whilst FATHMM was lowly correlated to all other scores, and SIFT and MutPred were also moderately correlated to all other scores (Ioannidis et al., 2016). The identified 5 most critical features within the REVEL random forest were FATHMM, VEST, MutationAssessor, MutPred, and PolyPhen-2. The most favouring result is that REVEL acquired the best performance in terms of distinguishing pathogenic from rate neutral variants having an allele frequency of <0.5%. (Ioannidis et al., 2016)

A REVEL score can range from 0-1 being a reflection of the proportion of trees within the random forest having classified the variant as pathogenic. The distribution of REVEL scores was similar to the reported exome sequencing variants having only a slight shift towards higher scores for all exome sequencing variants (Ioannidis et al., 2016). Key limitations of REVEL is the reliance on assertion of pathogenicity from pre-existing databases, which might be less accurate. Additionally, these pathogenicity assertations may have been based on predictions from particular tools including SIFT and PolyPhen-2 which could result in inflated performance of these predictors and the ensemble scores produced from their results. (Ioannidis et al., 2016)

2.12.8 BayesDel

BayesDel was developed as a measure for the algorithm PERCH (**P**olymorphism **E**valuation, **R**anking and **C**lassification for a **H**eritable trait) (Feng, 2017), whereby BayesDel was utilised for the measure of deleteriousness utilising Bayesian methodology (Lewinger et al., 2007). The below description of BayesDel was adapted from Feng's publication in 2017. BayesDel was curated with the goal of achieving an accuracy similar to state-of-the-art having a Bayes factor output. This measure combines multiple deleterious predictors to give a singular score output. This tool was designed for the analyses of large-scale variants, hence the predictors utilised were those readily available for exome or genome-wide annotations including; Polyphen2, FATHMM, SIFT, Mutation Taster, Mutation Assessor, PhyloP, GERP++ and SiPhy.

The deleteriousness predictors utilised in this framework included maximum and minor allele frequency across populations within ExAc version 0.3 and the 1000 genomes project phase 3. This approach utilises the naïve Bayesian approach which assumes all predictors are mutually exclusive. The predictors utilised here however are not mutually exclusive as they typically measure the same characteristics of variants including sequence conservation and amino acid substitution physical properties. Weighting was hence applied to the naïve Bayesian model to alleviate the independence assumption. The rationale behind this is that the correlated predictors are given weights for them to jointly form a unit contribution. A likelihood ratio is then empirically estimated for each score value, this is the result of the in the probability of pathogenic variants divided by the probability of benign variants. This calculation would be too computationally heavy for each score, hence the score ranges for the calculation of likelihood ratio are divided into bins. To enhance the stability of the estimation, neighbouring bins are combined to ensure that the probability among benign variants (being the denominator of the likelihood ratio) has a value of at least 0.02. Each bin is represented by its mean value, and a curve depicting likelihood ratios based on score values is constructed. This curve is then smoothed through least-squares fitting of polynomials to segments of the data. Utilizing this curve, any given query score can be converted into a likelihood ratio via linear interpolation. (Feng, 2017) For the total weights to be obtained the model is optimised via a controlled random search algorithm for the 'local mutation' modification (Kaelo & Ali, 2006). The modification of 'local mutation' allows for the controlled random search

algorithm to be more robust for the acquisition of the global maximum score via the reduction of evaluation quantities (Kaelo & Ali, 2006).

The training model comprised of variants taken from ClinVar and UniProtKB whilst those from the ENIGMA dataset were excluded to avoid any overlap between the testing and training datasets. The total number of variants for the training dataset was of 39978 neutral variants and 39395 pathogenic variants (Feng, 2017). Should some component scores be missing during the application, the model alters the weights of missing components to 0 whilst normalising the remaining weights always assuring the sum of weights is 1. Should too many scores be missing such that the sum of weights prior to normalisation is <0.5, BayesDel is then calculated from a genome-wide deleteriousness score exemplified by CADD (Kircher et al., 2014; Feng, 2017).

For the testing of BayesDel, it was compared to other deleteriousness scores including each of its individual components and other combinatory methods such as CADD, MetaLR and MetaSVM (Feng, 2017). The dataset for the evaluation of BayesDel was chosen meticulously to prevent the three circularity types typically hindering deleterious prediction tools described by Grimm et al., in 2015. The criteria for the formation of the BayesDel testing dataset was four-folded. (1) The variants included in the training dataset should not overlap with variants in the testing dataset; (2) genes having variants identified as mostly pathogenic or neutral should be excluded; (3) allele frequency should not define benign variants or the presence of homozygous alternative allele genotypes since allele frequencies are incorporated in BayesDel; (4) during the definition of pathogenic variants, deleteriousness tools should be eliminated from the variant classification scheme (Feng, 2017).

The advantages of this developed deleteriousness score can be summarised in two points. Primarily BayesDel in combination with the other scores compromising the PERCH algorithm is more accurate than other testing methods, including the combined scores. Secondly, the fact that the naïve Bayesian approach is utilised allows missing values to be treated naturally, hence no imputation is necessary, limiting the introduction of bias which could hinder the usage of the model. A limitation of BayesDel is the over-fitting potential of known variants. This may result in different performance between known and novel variants. This limitation is not one belonging solely to BayesDel, rather all ensemble models face this difficulty. (Feng, 2017)

Additional studies performed on BayesDel, amongst other in silico missense variation predictor tools, also confirmed the reliability of this model (Pejaver et al., 2022). BayesDel was amongst one of the suggested models for the recommendation of PP3 and BP4 ACMG criteria. Pejaver et al., recommended the usage of a single tool for genome-wide clinical laboratory missense variant determination having a strong level of evidence for PP3 and BP4. This recommendation allows the strength maximisation of applied evidence, whilst minimising the quantity of false positive predictions in these categories (Pejaver et al., 2022). An additional recommendation for the classification for allele frequency data, BayesDel was amongst the recommended *in silico* tools due to its exclusion of direct allele frequency (Pejaver et al., 2022).

2.12.9 Genocanyon.

Unlike other previously-described in silico variant predictor tools, GenoCanyon acquires its name from the canyon-like plots generated by this tool. GenoCyanyon was first described by Lu et al., 2015 as a whole-genome annotation tool which utilises comparative genomic conservation scores and biochemical signals extracted from the ENCODE project (The ENCODE (ENCyclopedia Of DNA Elements) Project.2004). The prediction score generated by GenoCanyon is hence the posterior probability of a genomic position being functional. In comparison with other available methods, GenoCanyon additionally to measuring the deleteriousness of a given variant, also measures the functional potential of each given genomic location. This tool annotates the whole-genome via the performance of unsupervised statistical learning via 22 experimental and computational annotations, hence inferring the functional potential of each position onto the human genome. Thus, GenoCanyon allows for the prediction of functional regions along with a generalisable framework.

The below described statistical model and estimations were described in detail by Lu et al., in 2015. The statistical model was curated from 22 different annotations corresponding to conservation score or biochemical activity of which included 2 genomic conservation measures, 2 open chromatin indicators, 8 histone modifications, 10 transcription factor binding site peaks. These were selected as annotations since their functional impacts had been well studied and are simpler models. This model does not

include DNA methylation due to the unavailability of modelling of functional impact of methylation for gene silencing mechanisms. Genomic data for the mentioned 22 annotations were downloaded from USCS Genome Browser. GERP (Cooper et al., 2005) and PhyloP (Pollard et al., 2010) were selected as the conservation measures due to their normal distribution. PhyloP46way was selected as it provides a comprehensive view over the conserved signal via the small phylogenetic distance. The model contains 49 parameters which were estimated on the GWAS catalogue (Welter et al., 2014) containing 13,070 SNPs being unique and significant to GWAS studies. Each SNP had an interval between 500bp upstream and 499bp downstream marked, hence aligning to the 13,070 intervals, each spanning 1k bp. Due to the large bp coverage, the whole collection was representative enough for the distribution and annotation learning for both functional and non-functional groups. Substantial tests were conducted to conclude that the GWAS-loci-based dataset compiled through this study spanning roughly 13,000,000 base pairs (bp), contains enough functional elements to consider accurate parameter estimation. This compiled dataset is also general enough to not allow genome heterogeneity to heavily impact estimation. To check for sensitivity of the model to the perturbation in annotation data, the model was re-fitted multiple times after removing several mis-fitted annotations.

Prediction scores were calculated for the entire human genome, having 33% predicted to be functional when utilising a 0.5% cutoff for functionality definition (Lu et al., 2015). Contrasting evidence has been published in literature about the percentage functionality of the human genome ranging from 4.5% (Lindblad-Toh et al., 2011; Meader et al., 2010; Parker et al., 2009; Ward & Kellis, 2012a; Ward & Kellis, 2012b) to 80% ENCyclopedia Of DNA Elements) Project.2004) in different literature. The prediction by GenoCanyon represents a mixed probability involving multiple tissues, having a result lying in the middle of the two opposing values, 33%.

Although GenoCanyon was not designed as a variant classifier, pathogenic variants are still enriched for a prediction score. GenoCanyon was found to have high sensitivity but low specificity. This can be attributed towards the fact that GenoCanyon measures the functional potential of genomic locations, and not the tolerability of specific variants. However, many tolerated synonymous SNPs in a gene eventually become ‘false-positives’, resulting in the low specificity. In addition, many of the known ‘benign’ variants were included as by-products of association studies, having their properties investigated due to lying in candidate regions of disease pathways,

hence further giving insight to the high mean prediction score of benign variants. On the contrary, the underlying region of a variant being proven as pathogenic from associate experimentation, would directly relate to functions in disease. Thus, the high sensitivity of GenoCanyon hints at the positivity of its prediction ability.

Thus, GenoCanyon is variable from variant classifiers, in that it measures functional potential of genomic locations rather than pathogenicity of selected variants, hence a high score does not necessarily indicate deleteriousness. GenoCanyon may also serve as a conservative tool for noise reduction in sequencing experimentation where variants distribute themselves throughout the entire genome. Examples of this include rare variants association studies, whereby the GenoCanyon may be utilised to filter SNPs and reduce 2/3rd of the tests, as more than this fraction of the human genome is likely to be functional. GenoCanyon also provides a useful tool in the prediction of functional potential at each nucleotide. This can also be utilised in association studies, where genetic variants are used as markers to capture signals for nearby regions. In this case, the mean prediction score for the surrounding region of an SNP may be acquired via GenoCanyon, whilst traditional variant classifiers are not capable of providing said information.

Since this model utilises unsupervised learning, it does not suffer from the biased knowledge of non-coding DNA, allowing the model to be generalised in many directions. Since the annotations utilised in this model were adapted from ENCODE and were further clustered across multiple cell lines, the functional regions predicted by GenoCanyon are in fact the union of functional elements from various cell types. Additionally, the model can be further extended to alternative species to humans. As functional elements in model species are typically studied in greater detail, this tool utilised for different species may benefit species comparison and help detect functional orthologs in humans. The conversion of biochemical annotations into binary variables also removes the need for signal strength information. The inclusion of such information along with further annotation would improve the specificity of such a model. A final important note, relates to the model assuming the leading role of genetic function hence treating conservation measures and biochemical signals as consequences.

The above description of GenoCanyon was adapted from Lu et al publication in 2015.

2.12.10 CADD score

Combined **A**nnotated **D**ependent **D**epletion (**CADD**) is a framework which integrates various diverse annotations into a single quantitative score objectively being first described by Kircher et al., in 2014. They described CADD as a general framework for the integration of various genomic annotations along with the scoring of any possible human SNVs or small insertions/deletions (indels). The foundations of this framework is to contrast the annotations of stimulated variants relative to fixed/virtually fixed derived alleles in humans. Naturally, deleterious variants tend to be depleted due to natural selection in fixed variants (Kimura, 1983), whilst this does not depict itself in stimulated variants. Hence CADD measures deleteriousness, allowing a strong correlation between both pathogenicity and molecular functionality of variants. CADD was implemented as a support vector machine trained for differentiation between 14.7 million alleles derived at high frequency from 14.7 million stimulated variants (Kircher et al., 2014). Approximately 95% of these alleles were fully fixed within humans, having <5% being virtually fixed polymorphisms according to the 1000 genomes project variant catalogue (Abecasis et al., 2012). The compiled variant-by-annotation matrix compiled from various conservation metrics, regulator information, transcription data, amongst others, resulted in 29.4 million variants, half being fixed or virtually fixed whilst half being stimulated variants, along with 63 distinct annotations (Kircher et al., 2014).

The model's validity was assessed via the construction of a series of univariate models contrasting observed and stimulated variants through the utilisation of 63 annotations as individual predictors. Approximately all models scored significantly and consistently with expectations. This was exemplified via the identification of a 20-fold depletion in nonsense variants, 2-fold depletion in missense variants and no depletion in intergenic/upstream/downstream variants. On the other hand, nonsense and missense variants present in close proximity to the cDNAs had larger depletion than those near the ends. Conservation metrics were found to be amongst the strongest individual genome-wide annotation (Kircher et al., 2014). Correlations between annotations and the value of additional interaction terms between annotations was also examined through CADD. The correlation of annotations were found to be positive, along with statistical significance being allocated towards interactions. However relatively few interacting

pairs showed substantial improvement towards the additive model. The Support Vector Machine (SVM) was trained with linear kernel features derived from the previously described 63 annotations, having a limited quantity of interaction terms (Kircher et al., 2014). A strong correlation was identified between independently trained, non-overlapping observed and stimulated variants.

Phred-like scores (Ewing & Green, 1998), referred to as 'C-scores' were based on the rank of the C-score of each variant in relation to all 8.6 billion SNV possibilities, having a range from 1-99. These scores were utilised for simplification of interpretation (Kircher et al., 2014). The proportion of all possible substitutions according to their given scaled C-score which has its own specific functional consequence. The resulting C-scores from highest to lowest proportion were; nonsense variants, missense and canonical splice site variants, whilst intergenic variants were lowest (Kircher et al., 2014). 76% of potential SNVs having C-score >20 were non-coding, whilst 74% of potential missense and 18% of potential nonsense SNVs resulted in C-scores <20. C-scores allow the compilation of essential information between and within functional categories of variants. C-scores were also compared to levels of genetic diversity, resulting in a negative correlation towards derived allele frequency having been identified through the 1000 genomes project (Abecasis et al., 2012; Kircher et al., 2014). CADD was tested primarily for the functional and disease-relevant variation within five (5) contexts namely the genes (1) *MLL2* and (2) *HBB*, (3) pathogenic variants curated in the NIH ClinVar database, (4) C-score correlation to somatic cancer mutations in p53, and (5) two enhancer and one promoter saturation mutagenesis. The analysis of these collective contexts demonstrates CADD's quantitative deleteriousness, pathogenicity, and molecular functionality prediction in both protein-altering as well as regulatory through various experimental and disease contexts (Kircher et al., 2014).

CADD was further tested for its evaluation towards variants within exome and/or genome-wide studies. *De novo* exome variants identified in children having autism spectrum disorder (ASD) and intellectual disability (ID) along with their unaffected siblings as controls. This included 88 nonsense, 1015 missense, 359 synonymous, 32 canonical splice site, and 150 other variants, including indels. The variants identified in affected children proved to be significantly more deleterious than those in unaffected controls. CADD was further tested for its pathogenic variant ranking, where the C-score distribution within the genomes was examined for representative populations. Here

CADD highly ranked known disease-causing variants within the whole spectrum of variation in personal genomes. CADD was found to have considerable superiority over the alternative protein-based and conservation metrics with respect to ranking of known pathogenic variants in the spectrum of variants within personal genomes (Kircher et al., 2014).

CADD scores were finally analysed for SNPs identified via GWAS for complex traits, comparing the resulting scores control SNPs matched for allele frequency and genotype array availability. C-scores were found to be significantly higher in GWAS SNPs than in controls. Although extensive control of properties such as gene-body effect, gene expression level, conservation and regulatory element overlap, the high C-score difference persisted. Studies suggest that GWAS-identified SNPs, typically those associated with lead SNPs from large studies, are enriched for causal variants. This is consistent with previously identified GWAS enrichments for individual annotations (ENCODE project consortium, 2012; Kircher et al., 2014).

In summary, CADD allows the integration of diverse genetic variation annotations into a single score. Some practical and conceptual advantages to CADD which provide a major advantage to genetic studies of human diseases include; (1) the objective merge of individual annotations into a single value, allowing improved performance; (2) the incorporation of expansions to existing annotations along with entirely novel annotations; (3) the combination of the specificity of subset-relevant functional metrics with the generality of conservation-based metrics (Kircher et al., 2014).

Some limitations to CADD include; (1) C-scores measure variation reduction, hence correlating with deleteriousness, however these are also affected by local variation rate, background selection, gene conservation bias, amongst others, hence limiting accuracy. (2) C-scores are reflective of variants having a given annotation pattern being visible to selection, but this may not capture hence differences in selective intensity. (3) Due to a lack of ‘gold standard’ data particularly within non-coding regions of the genome, the annotation must be limited to the data being utilised for training. (4) Upon the development of CADD, it was not possible for the relationship between CADD-estimated deleteriousness and variant pathogenicity likelihood to be calibrated precisely.

The most recent version of CADD is v1.4 which along with supporting the human genome build GRCh38, also includes simplified variant lookup, extended documentation, an application program interface along with improved mechanisms for the integration of CADD scores into other tools/applications (Rentzsch et al., 2019).

2.12.11 MutationTaster

MutationTaster was developed in 2010 by Schwarz et al., where it was described as a free web-based application for the rapid evaluation of the potential of DNA variations to be disease-causative. This tool integrates information from various biomedical databases such as UniProt, SwissProt, Ensemble, NCBI and HapMap.

Analyses performed by MutationTaster include evolutionary conservation, splice-site alterations, loss of protein features, along with alterations to mRNA quantity. The model compares an alteration to known SNPs, and if the alteration should be found in that region, the rs ID from dbSNP is linked with the HapMap frequency. According to the type of alteration query submitted to MutationTaster, it selects between 3 prediction models; (1) one aiming at synonymous (silent) or intronic alterations; (2) an alternative model aiming at single amino acid alterations; (3) and finally a model aimed at alterations resulting in complex variations to amino acid sequence. Should one of the above alterations be identified in at least one HapMap population, it is hence classified as a polymorphism. The naïve Bayes classifier is utilised above the previously described for evaluation of results, hence allowing the prediction of disease potential of a said ‘change’. MutationTaster analyses potential splice site changes by utilising an intrinsically installed version of NNSplice (Reese et al., 1997). A short sequence of the wild type and ‘mutant’ sequences are compared, allowing the classification of the splice alteration as lost, gained, splicing likelihood increase or decrease. Along with the described prediction, the genomic position of the splice site, the prediction score for the mutant splice site generated by NNSplice and the sequence snippet are also provided as results. MutationTaster also utilised a locally installed polyadenylation signal prediction tool polyadq for the prediction of polyadenylation affects on protein products. This model also performed analysis on the Kozak consensus sequence crucial to translation initiation. It further

checks for any alterations caused to protein features, and the length of the resulting protein.

As previously mentioned, one of the analyses performed by MutationTaster is that regarding conservation. This conservation analysis is performed via the alignment of the human amino acid sequence in question to amino acid sequences and/or the nucleotide sequence homologues to ten other species. The values resulting from this analysis can be via three statuses of evolutionary conservation namely; *all identical*, *(partly) conserved (*having similar amino acids between the compared sequences*)*, or *not conserved.* Should no homologous gene be found or alignment be acquired, this is also stated by MutationTaster. The restriction of conservation to ten species was intentional, as in the analysis it was identified that additional species did not result in considerable influence on prediction accuracy, rather resulted in the decreased speed of MutationTaster.

Splice site analysis is also performed via a local version of NNSplice as previously mentioned. An approximate length of 60 bases around the alteration of interest is utilised in comparison to the wild-type sequence. Should the alteration to this sequence result in any changes to the splice site, a display will be shown of the effect of the splice site. The results depicted to splice site changes are only considered in cases having at least one of the two sequences yielding a prediction score of ≥0.5 via NNSplice.

The MutationTaster database stores all human SwissProt protein features and tests whether an amino acid alteration affects protein features directly or indirectly. This algorithm also further analyses whether the resulting protein will be affected in length, including any nonsense-mediated mRNA decay. This nonsense-mediated mRNA decay border is determined according to 50bp upstream of the last intron-exon junction, and hence this regions is analysed in terms of any premature codon termination at the 5' end, leading to the nonsense-mediated mRNA decay. Should the algorithm identify that the variation leads to nonsense-mediated mRNA decay, it will immediately be flagged as a 'disease mutation'.

MutationTaster2 is the updated version of the web-based software described above as described by Schwarz et al., in 2014. This updated version includes all publicly available SNPs and indels from the 1000 genomes project along with disease

variants adapted from ClinVar and HGMD public. MutationTaster2 aimed at reducing the quantity of false positive splice-site predictions. This was achieved via the consideration of loss/decreasing strength of slice sites only at existing intron-exon borders. A further improvement to the previous MutationTaster is the ability for analysis of sequence alterations spanning the intron-exon junctions which likely perturb typical splicing, hence resulting in considerable pathogenic potential. Furthermore, the more recent version has a substantial increase in speed via the production of protein-conservation analysis from the results of BLASTP being implemented into the internal search for amino acid sequences. MutationTaster2 also includes a dedicated query engine for the user-friendly analysis of NGS results. Here VCF files may be uploaded and various parameters may be adjusted including the confinement of considerations to homozygous variants alone, and the filtering for previously-identified polymorphisms. GeneDistiller, a candidate-gene search engine is also integrated into the MutationTaster2 allows users to identify the most likely candidate genes alongst potentially deleterious variants. The web interface of MutationTaster2 also includes the availability of single queries utilising chromosomal positions. The identified limitation to MutationTaster2 is its lack of application towards intragenic variants.

The latest documented version of MutationTaster by Steinhaus et al., 2021 is that referred to as MutationTaster2021. The major change in this version from its predecessors is the abandonment of the Bayes classifier and the utilisation of Random Forest models tailored to different types of variants. In the training of this model, balance accuracy was focused on, referring to equal prediction quality for both benign and deleterious variants. The number of false positive predictions were further enforced in this recent updated version. Although the previous filtering against common polymorphisms aided in this reduction of false positives, a large quantity of rare/population-specific variants remained false positive values. The likely cause of this is that the false positives tend to have a higher percentage of phylogenetic conservation in comparison to frequent polymorphisms being previously used as the training data. Hence the training data now includes benign intragenic variants identified in GnomAD having at least one homozygous carrier. With regards to updates for deleterious variant identification, previous training cases included intragenic variants taken from HGMD and ClinVar. Whereas in the updated version MutationTaster2021, the analysis was restricted to the variants labelled as ‘DM’ in HGMD or ‘pathogenic/likely pathogenic’

in ClinVar. Any ClinVar variants having conflicting labels were excluded and variants identified in both training datasets were excluded. Dedicated prediction models for the two UTR regions were also set up for the provision of better predictions. Although these have lower accuracies in comparison to the model for non-coding variants, they have still provided better results for UTR variants. The previous versions of MutationTaster utilised NNSplice for the prediction of variant effect on splicing. Analyses in the new version identified higher accuracy from MaxEntScan (Yeo, 2004). An important note is that MutationTaster2 along with MutationTaster2021 do not search for cryptic splice sites being activated by DNA variants due to this yielding too many false positive predictions.

2.12.12 HOPE

Have (y)**O**ur **P**rotein **E**xplained (**HOPE**) was firstly described by Venselaar et al., in 2010 is a web application for the fully automated analysis of genetic variants. This application was built on the strengths of previously available software and servers including PolyPhen, SIFT and ALAMUT due to their high reliability in the interpretation of mutational effects. HOPE takes a protein 3D structure centred approach. Via this approach, it acquires information regarding a protein's 3D structure from data sources including UniProt. Each protein's data is stored in a PostgreSQL-based information system. A decision scheme is then applied for the processing of this data and the production of variant effect on 3D structure and functionality of the protein. The web-interface generates a full report about the variant and its effects on the protein for simple results analysis. HOPE is intended for use by life scientists, who may not routinely utilise protein structure or bioinformatics as part of their research.

The input for HOPE web interface is the protein sequence. This inputted sequence is utilised as a query for BLAST searches against the Protein Data Bank (PDB) and UniProt databases. The UniProt search allows the acquisition of the protein's accession code, being utilised later to obtain the DAS-predictions. As an alternative input, one can input the accession code directly in comparison to the entire sequence. The PDB BLAST search is imperative for the ability to have a protein structure and/or template structure for the homology modelling. The PDB-file utilised by HOPE would be 100% identical to

the submitted sequence, whilst also analysing the structure for analysis according to the resolution, experimental method of acquisition and the length of protein covered in PDB. Homology modelling between the variant and the provided sequence is performed utilising the Twinset version of YASARA (Krieger et al., 2002), containing automatic homology modelling script having the only input requirement being the sequence. WHAT IF web services (Hekkelman et al., 2010) are utilised for the analysis of the structure of the protein of interest. In the absence of a 3D structure or a modelling template, HOPE must base its conclusions off of sequence related data instead of structural information. UniProt is utilised for acquisition of features which can be mapped onto the sequence. This information includes the location of transmembrane domains, active sites, secondary structures, mofits, domains, sequence variants along with experimental information.

The data obtained from multiple sources are stored on the PosrgreSQL database system. The same sequence obtained from different sources such as those from UniProt and PDB, might have some variation between one another. For instance, sequences from UniProt tend to include the signal peptide, whilst PDB lacks this information. For this reason, the acquired sequences are aligned utilising ClustalW on the database. Examples of the data-type of the protein features stored in the system include; contacts, variable features, fixed features and variants. Contacts describe the interaction of a residue with another component exemplified by ligands and different bonds. Variable features include those of which have a value, exemplified by torsion angle. Fixed features may be exemplified via residues located within a domain or motif. Variants refer to mutations or variations within a sequence at a known position, being exemplified by splice site variants, SNPs, ect. Any request is stored on disk for the duration of a month, to ensure that should a similar request be submitted, the information is prepared. Monthly, each system is scrapped to ensure that results are not issued on out-dated information.

The scheme of which HOPE makes decision utilises all collected information in combination with previous knowledge about wild-type and mutated amino acid properties. Such properties include charge, size and hydrophobicity, amongst others which enable the prediction of the effect of the variant on the protein's structure and function. This scheme can be split into six (6) analyses, and hence their corresponding six (6) outputs with respect to the effect the variant has on the residue, these include: Contacts; structural domains; modifications; variants; conservation; and amino acid properties. These six aspects are concluded by HOPE separately. The information utilised

by HOPE may be acquired from several sources such as DAS-server and UniProt. HOPE is trained to identify that data generated through experimentation and calculations based on 3D coordinates may be more accurate than in-silico models predict, hence tanking the information according to source and utilising the most accurate source. The ranking is performed having WHAT IF calculations preferred, seconded by UniProt annotations and finally DAS predictions.

Currently HOPE is useful and reliable for the analysis of point mutations, as this website does not currently address more complex submissions. Some results provided by HOPE may result from the simple combination of calculation, literature data and general knowledge of the protein's structure-function relationship.

2.12.13 Dynamut2

The initial DynaMut web server was described by Rodrigues et al., in 2018 as the 'dynamic component to mutation analysis'. DynaMut implements the computational approach of normal mode analysis (NMA). NMA is utilised for the generation of possible motions, hence allowing valuable insight into the motion of proteins, and hence their conformation. The integration of NMA in DynaMut was established due to studies providing the utilisation of NMA for protein structure analysis including functional relationships (Grant et al., 2006) and prediction of the effects of SNVs on protein stability (Frappier & Najmanovich, 2014). DynaMut implements and integrates previously well-established normal mode approaches (NMAs) with their developed graph-based signatures allowing the prediction of protein stability amongst variation (Rodrigues et al., 2018). The integration of NMAs in DynaMut is via two different approaches being Bio3D (Grant et al., 2006) and ENCoM (Frappier & Najmanovich, 2014). These two integrations to DynaMut provide rapid and simplified access to protein motion analysis. In addition, this web server also enables the analysis of a variant's impact on a protein's stability and dynamics via the resulting vibrational entropy changes. The integration of these two contrasting approaches along with the combination of additional well-established methodologies and wild-type residue characteristics allows DynaMut to provide an accurate assessment. DynaMut2 varies from the previous version DynaMut on two accounts; the input type and the processing time (Rodrigues et al., 2021).

Taking the above into consideration, DynaMut was trained on 2297 randomly selected variants taken from a previously established S2648 dataset (Dehouck et al., 2009; Pandurangan et al., 2017; Pires et al., 2014a; Pires et al., 2014b) having been derived from the ProTherm database (Kumar et al., 2006). The above-mentioned dataset includes 2648 different SNVs in 131 globular proteins having experimentally determined original structure as well as variant protein stability. A blind set was also compiled from the S2648 dataset, having been previously utilised in literature (Dehouck et al., 2009; Pandurangan et al., 2017; Pires et al., 2014a; Pires et al., 2014b), hence enabling the comparison of method performance for variants impacting folding and free energy. These datasets have been reported for comparisons in performance with regards to the prediction of changes in free folding energy (ΔΔG) (Khan & Vihinen, 2010; Potapov et al., 2009; Thiltgen & Goldstein, 2012). DynaMut considers a hypothetical reverse mutation, whereby the ΔΔG of a variant from the wild-type to the mutant should be equivalent to the negative ΔΔG of

the reverse (the mutant to the wild-type protein) (Pandurangan et al., 2017; Pires & Ascher, 2016; Pires & Ascher, 2017; Thiltgen & Goldstein, 2012). The predictive model of DynaMut was trained including the hypothetical reverse mutation utilising 4594 variants and the blind test utilised 702 SNVs. DynaMut combines the effects of variants on protein stability, ΔΔG and dynamics, via Bio3D, ENCoM and DUET. These datasets were supplied as evidence for the training using Random Forest algorithm from the Python library. The combination of these predictors allows an optimised and vigorous predictor. DynaMut2 in comparison to the previous version, does not utilise the previously published dataset, to minimise uncertainty pertaining to quality and biological significance of the modelled variant. The final dataset hence comprised 4633 variants (Rodrigues et al., 2021). The dataset for multiple point mutations utilised 1323 entries from ProTherm, however the majority of these entries comprised double and triple mutants, hence the final dataset comprised 1098 entries having been randomly split into training and test sets of 872 and 227 entries respectively (Rodrigues et al., 2021). Changes in Gibbs free energy of folding can occur due to various factors. Therefore arpeggio (Jubb et al., 2017) is utilised for the calculation of the quantity of hydrophobic contacts in the wild-type residue and contact potential scores taken from AAINDEX database (Kawashima & Kanehisa, 2000). The random forest was further utilised in the updated version (DynaMut2) for the prediction of ΔΔG for both the single and multiple mutation predictor (Rodrigues et al., 2021).

DynaMut may be utilised by end users for the analysis of protein dynamics and/or the effect of SNVs on protein dynamics and stability. The input for the two types of analysis varies. The input for the analysis of the SNVs effect on protein dynamics and stability is that of protein structure via PDB format file or the PDB four-letter accession code. Protein dynamics analysis can be performed via the input of two input options. The ‘single mutation’ option utilises a PDB file or accession code, the SNV specified via a string of the wild-type residue one-letter codes, the corresponding residue number and the variant residue one-letter code. The ‘mutation list’ option requires the input of a file containing numerous variants for batch processing. Both of these inputs require the end-user to specify the chain identifier pertaining to the variant. DynaMut2 on the other hand can be utilised in 3 different ways, in comparison to the previous two analyses allowed on DynaMut. DynaMut2 predicts the ΔΔG for single point mutations, multiple point mutations, along with the analysis of protein dynamics according to NMAs (Rodrigues et

al., 2021). For the additional prediction allowed in the updated version of multiple mutation analysis, the input as in the 'single mutation' is required, along with commas separating one mutation from another (Rodrigues et al., 2021).

The prediction of effect of single point mutation was evaluated via the Pearson's correlation coefficient. The results for Pearson's correlation and the non-redundant independent test set outperformed all other methods compared (DynaMut1, SDM, mCDM, DUET, ENCoM, Maestro, I-mutant and MUpro). The performance of the stabilising and destabilising mutations were further analysed independently, due to the natural imbalance between these two groups. The resulting Pearson's correlation result (0.62 and 0.51 respectively) was also reflected by DynaMut2's ability to correctly classify stabilising and destabilising variants, further outperforming previous approaches (Rodrigues et al., 2021). DynaMut2 was also investigated in terms of potential bias to predictive performance, hence being tested on the O2567 dataset (Caldararu et al., 2020). DynaMut2 resulted in significantly higher performance in comparison to other approached with respect to mutations on buried residues. Small deterioration was identified with regards to mutations on exposed residues, however still comparable to alternative approaches (Rodrigues et al., 2021). DynaMut2 has shown to outperform other methods for prediction of changes in stability caused by single point mutations (Rodrigues et al., 2021). DynaMut2 is also significantly faster than its precursor DynaMut, enabling a shorter turnover rate for larger analysis and structures.

2.12.14 MetaDome

MetaDome was first described by Wiel et al., in 2019 as a freely available web server utilising the creator's concept of meta-domains for the utilisation of population-based and pathogenic variation datasets information without the requirement of a bioinformatics intermediary. MetaDome was initiated due to the findings of Wiel et al., in 2017 who further elaborated the concept of homologous proteins via multiple sequence alignment for the location of equivalent positions between protein sequences. They integrated this application for homologous Pfam protein domain relationships within the human genome. Through this integration, they identified that 71-72% of deleterious missense variants from HGMD and ClinVar are found within regions translated to a Pfam

protein domain. This further lead to the observation that pathogenic variants found within comparable domain positions tend to be paired with the absence of population-based variation, and vice versa (Wiel et al., 2017). Due to these findings, the utilisation of variant information for homologous protein domains was termed “meta-domains”. Thus, the initiation of the concept to enable easy access to this information towards the genetic community, allowing the creation of MetaDome. MetaDome processes the gene input, with the choice for transcript selection and provides protein domain and pathogenic variant annotation whilst generating a ‘tolerance landscape’ for the resulting proteins from the gene input, aiding in the visualisation of regional tolerance to genetic variation. This web server also utilises homologous protein domain regions for the aggregation of population based as well as pathogenic variants identified across the genome which have been aligned at the same position for the domain in the specified gene of interest.

The below description of MetaDome was extracted from Wiel et al., 2019.

MetaDome was developed in Python v3.5.1 (Rossum, 2012) using the Flask framework v0.12.4 (Ronacher, A., 2010) for the web server. The software’s architecture is based off of the domain-driven design paradigm (Evans, 2004).For the smooth deployment of MetaDome the application is containerised via Docker v17.12.1 for 5 facets; (i) the flask application, (ii) PostgreSQL database for the storage of mapped databases, (iii) Celery task queue management for the facilitation of larger tasks, (iv) Redis for the storage of the results and finally (v) RabbitMQ for the mediation between clients and workers. MetaDome utilises population and clinically relevant SNVs acquired from genetic variation databases. GnomAD was used for the acquisition of population variation via the selection of all synonymous, nonsense and missense variants passing the PASS filter criteria (Lek et al., 2016). ClinVar was utilised for the acquisition of pathogenic variants. MetaDome uses an auto-generated and storage for the complete mapping of genomic, protein position and all domain annotations in a PostgreSQL database. These maps are created for each protein-coding translation within the GENCODE Basic set for human canonical and isoform Swiss-Prot protein sequences via Protein-Protein BLAST (Camacho et al., 2009). Sequences excluded from the database were those missing a start and stop codon and sequences resulting as unidentical between cDNA and GENCODE translation. The formulated database also tabulates the transcript information, retaining global information on genes having identical sequence matches within Swiss-Prot. For each identical match between translation and Swiss-Prot,

alignment of the two sequences via ClustalW2 is performed. Furthermore, each nucleotide's genomic position is mapped onto the protein position, tabulated, and stored. The annotation tools utilised in MetaDome are InterProScan (Finn et al., 2017) and Pfam-A (Finn et al., 2016).

Meta-domains utilised by MetaDome refer to homologous Pfam protein domains being annotated via InterproScan consisting of at least two homologous within the human genome. Multiple Sequence Alignments are generated via a three (3) step process; (i) all sequences for the domain instances are retrieved, (ii) the Pfam HMM corresponding to the Pfam identifier annotated by InterproScan are retrieved and (iii) HMMER (Finn et al., 2015) is utilised to align the sequences. The resulting file produced from the above process is retrieved by the MetaDome web server whenever a user requests meta-domain information for any position of interest. The file is used via the previously described mapping database to obtain the corresponding genomic position for each residue. the corresponding GnomAD and/or ClinVar variation is retrieved using the genomic position acquired previously. Genetic tolerance is computed via the nonsynonymous over synonymous ratio. This score is based on the observed missense and synonymous variation within GnomAD and is further corrected for sequence composition by accounting for the background possible missense and synonymous variants according to the codon table.

MetaDome can be utilised for the annotation of positions in a protein or in a protein's domain. Thus, the server requires access to genomic positional information along with protein sequence and protein domain information. This can be acquired via the mapping of GENCODE gene translations to UniProtKB/Swiss-Prot entries. This mapping occurs per-position and in correspondence to protein domains or genomic regions. For the creation of the mapping database 19728 human genes were linked to 33492 Swiss-Prot human canonical or isoform sequences. Every protein-coding transcript was incorporated into the database, hence resulting in 3334 Pfam domains.

MetaDome server hence allows the combination of resources and information from various fields including genomics and proteomics for a greater population and pathogenic variation analysing power via the transposition of these variants to homologues protein domains. This information transfer is achieved through per-position mapping between GENCODE and Swiss-Prot databases. MetaDome is particularly

effective if a requested variant is present within a protein domain having homologues. MetaDome increases the resolution of genetic tolerance at a single amino acid via the variation aggregation over protein domain homologues. MetaDome further allows the identification of variants which affect protein domain functionality by being annotated through the whole human genome, thus identifying potentially disease-causing variants. It is important to note regarding the aggregation of genetic variation in this manner, leads to the loss of specific context including haplotype information and/or interactions with other proteins. Aggregation performed in MetaDome via meta-domains only encapsulates general biological/molecular functions attributed to the domains.

2.12.15 Missense3D

Missense3D was first described by Ittisoponpisan et al., in 2019 as an online application to an inhouse pipeline for the structural assessment of missense variants. Missense3D answers the question of whether predictions provided by *in silico* models are obtained utilising 3D models similar to experimental models, and accounts for the accuracy of these models. To achieve this, stereochemical effects resulting from missense variants via PDB coordinates are compared to homology-predicted structures being generated via a range of sequence identities between queries and templates.

The below described intricacies of Missense3D functionality was adapted from the original paper by Ittisoponpisan et al., in 2019.

The first step of the Missense3D pipeline includes the data compilation, whereby 606 human protein structures were obtained from MolPorbity top8000 database of high-quality coordinates. 1965 deleterious missense variants and 2134 neutral obtained from Humsavar, ClinVar and ExAC were mapped onto the previously obtained structures. Prior to the release of Missense3D, the authors describe using Phyre2 to yield 54% of the human proteome residues confidently by predicted models (Ittisoponpisan et al., 2019). The next step in the pipeline is the prediction of models based on the previous maps using Phyre2. A mutant structure is then generated from the wild-type coordinates utilising SCWRL2 (Krivov et al., 2009). Side chains of the target residue and those of any residue within 5Å from the target residue were excluded from the coordinates. The excluded side chain of the target residue was replaced via the mutant side chain. The mutant coordinates

were generated via the reintroduction of the neighbouring residue's wild-type side chains, and further repackaged via SCWRL4. Structural analysis between the mutant and wild type structures is then performed as a final step of the pipeline to identify whether the substitutions structural consequence is expected as damaging in terms of stability of the protein, or otherwise. 17 features were considered for the structural analysis, in accordance with well-established principles of protein conformation and studies on structural consequences of disease-associated substitutions (Al-Numair & Martin, 2013; Bhattacharya et al., 2017; Gao et al., 2015; Kucukkal et al., 2015; Yue et al., 2005; Yue et al., 2006). To correct for any errors in modelling via SCWRL4 and within the predicted structure, 1 Å was added to the standard distances on three structural features using distance (including hydrogen, disulfide bonds and salt bridges). These 17 features include; disulphide bond breakage; buried proline introduced; clashes; buried hydrophilic introduced; buried charge introduced; buried charge switch; alteration to secondary structure; replacement of a buried charge; disallowed phi/psi; buried glycine replaced; buried H-bond breakage; buried salt bridge breakage; cavity altered; buried/exposed switch' Cis Pro replaced; glycine in a bend; exposed hydrophobic introduced.

A true positive (TP) (in Missense3D) is a disease-associated variant having been identified as damaging structural impact. This was selected since the features are designed to identify a large disruption in the folded structure. On the other hand, a false positive (FP) was identified to be a neutral missense variant resulting in structurally damaging via the pipeline's analysis. The true positive rate (TPR) is lacking as it does not take into consideration a missense variant which may result in disease via affecting features such as residues critical for function, or ligand binding. The false positive rate (FPR) on the other hand is likely to be overestimated due to structurally damaging proteins may not necessarily result in disease should the corresponding gene be non-essential or haplosufficient. Additionally, some missense variants may also be disease associated, but not yet identified, hence resulting in the overestimated FPR. Internal analysis was performed on each of the 17 previously mentioned features and their ability to distinguish between disease-associated and neutral variants, according to the fraction of the TPR over FPR. The feature cis pro replaced was identified as not significant despite having a good TPR to FPR ratio due to few observations limiting the tests power – having proline cis peptides being rare within proteins (Jabs et al., 1999) and only 7 disease-associated and 4 neural within the training dataset. The feature 'exposed hydrophobic introduced' was

also proved not to be effective for the identification of structural disease variant alerts having a TPR/FPR of 0.6. The most effective feature was found to be the 'breaking a disulfide bond' having a TPR/FPR of 25.3. Manual inspection was followed for the features results to further confirm reliability. Introducing a proline buried residue also showed to be a highly discriminating feature between disease-causing missense variants and neutral missense variants. Overall, 40.1% of the disease-associated and 11.4% of the neutral variants were identified to have at least 1 out of the 16 structurally damaging changes. When compared to FoldX (Van Durme et al., 2011), the overall TPR was the same, however the mutants generated by FoldX had a greater FPR, hence having a poorer TPR/FPR ratio. More than 91% of the variants were predicted to have similar effects regardless of the tool utilised; SCWRL4 or FoldX.

Rare, common, and unknown neutral missense variants' structural consequences were analysed based on their minor allele frequencies (MAF). Out of which 273 were common, 1150 were rare and 311 were unknown (according to MAF). The resulting FPRs for the respective groups were 5.9%, 11.6% and 15.8% respectively, whilst when run via SIFT resulted in substantially higher FPRs: 29.3%, 48.2% and 51.1% respectively. This false positive rate could be attributed towards the chance that many rare variants are currently assigned as neutral but may prove to be disease associated (Cirulli & Goldstein, 2010; Ittisoponpisan et al., 2017). The same may apply for variants considered having uncertain clinical significance, resulting in a possible association to disease manifestation (Ittisoponpisan & David, 2018).

As previously mentioned, the above-described pipeline for the structural analysis of missense variants was made freely available in the format of a web server Missense3D. This web server has two possible inputs; the position on the protein sequence or the position on the 3D structure. For the first input type of position on protein sequence, the user must provide the UniProt ID of the query protein along with its amino acid position on the protein sequence, the wild-type residue and its substitution. The PDB coordinate file and chain identifier should be specified. With this input, Missense3D can automatically generate the UniProt to PDB residue mapping. The second input type requires the user to upload a 3D coordinate file by either PDB code specification or a coordinate set. Following this, the amino acid position on the 3D structure, the wild-type residue, substitution, and chain identifiers within the 3D coordinate file is specified. Although Missense3D accepts predicted models from any server, it does not guarantee

correct mapping of a residue having sequence-based numbering onto coordinates. The resulting report generated by Missense3D includes a breakdown of the structural features changes thought to be damaging brought about by the substituted amino acid. The web server has a turnaround time of around 3 minutes. Missense3D is designed to model structural consequences of missense variants, in comparison to readily available global conformation of protein structure prediction server, including Phyre2. Therefore, Missense3D Is able to quantitatively describe the potential of structurally damaging missense variants in both homology-predicted and experimental structures. A limitation to Missense3D is the lack of comparison between Phyre2 results and other modelling servers such as I-Tasser (Yang et al., 2015) and Rosetta (Ó Conchúir et al., 2015). Missense3D does not take into consideration the disruption of protein-protein, protein-DNA and protein-small ligand interactions.

2.12.16 Aminode

Aminode is a webtool developed by Chang et al., 2018 to aid in the routine and rapid inference of evolutionarily constrained regions (ERCs). Aminode is pre-loaded with analytical results from comparison of the whole human proteome to the proteome of 62 vertebrate species. This webtool allows the immediate search and download of the relative rate of amino acid substitution and ERC maps of human protein profiles.

Aminode performs comparative analysis of multiple protein homologues in the context of evolutionary relationship to calculate the relative amino acid substitution rate of protein and further analyse ERCs. The input available includes the amino acid sequence of protein homologues and the phylogenetic tree describing the evolutionary relationship. The analysis of the human proteome can be via two routes: (1) a pre-computed analysis allowing the retrieval of the human proteome cross-analysed with 62 vertebrate species available in the Ensembl genome browser. Or (2) custom analysis via the submission of a protein's amino acid sequence and (optionally) their phylogenetic tree allowing ERCs identification via customisable parameters.

The multiple sequence alignments are obtained from Mutalin (Corpet, 1988), and sequences containing gaps in >50% of aligned proteins are eliminated. The Hartigan algorithm builds the framework for the best fit calculation of a given tree according to the maximum parsimony approach (Hartigan, 1973). For each position in the multiple

alignment, a substitution score (SS) is calculated as the sum of all node substitution scores at that position. The SS number divided by the number of informative sequences provides a relative substitution score. For the execution of the human proteome analysis, the protein sequence and phylogenetic tree of 63 species were downloaded from the Ensembl genome browser.

Results issued by Aminode contain evolutionary constrained region analyses for human proteins having a minimum of two (2) vertebrate orthologs annotated in Ensemble. The pre-generated output provided by Aminode are a visual representation of the relative rate of amino acid substitution. This is visualised by a line plotted over the multiple sequence alignment. Local minima are indicative of regions having low rates of substitution relative to the surrounding protein regions. Local maxima are indicative of regions having high rates of substitution. Hence, valleys in the resulting graph are representative of regions being more evolutionarily constrained than regions in the peaks. The predicted ERCs are marked by yellow bars above the multiple sequence alignment.

2.12.17 VarSEAK

VarSEAK is a variant interpretation software available online and as a downloadable software which allows the interpretation of variants, particularly splice site variants. This software was developed by JSI medical systems GmbH, having the latest ReadMe version 2.1 released in February 2022. The splice site prediction algorithm was trained on a dataset adapted from P Pollastro & Rampone, 2003 and P Pollastro & Rampone, 2002 consisting of approximately 200,000 splice sites adapted from GRHCh37 and 300,000 false splice sites adapted from HS3D dataset. This resulting combined algorithm was further validated for accuracy utilising the dataset described by Leman et al 2018. Only GT-splice sites were considered for 5' splice sites whilst the much rarer GC-splice sites will always receive the class 3 (unknown splicing effect) if they have a likelihood of being affected by a variant and a class 1 if they are unaffected by the GT 5' donor splice site rules. Within this dataset, the splice site prediction algorithm has a 96.41% accuracy.

The information required for the processing of variants are;

- The Gene.
- The transcript (should no transcript be given, the longest transcript for this gene will automatically be used).
- The variant being either;
    - c.-HGVS nomenclature.
    - Sequence (20-150 bases having the variant in the centre).

VarSEAK results can be divided into 7 sections:

- General information pertaining to the gene, transcript and the variant including the chromosome, strand, start and end position, exon number and cDNA length.
- Sequence graph and legend – the HGSV nomenclature along with critical authentic and/or potential splice sites (marked with triangles). Cryptic splice sites are highlighted in pink. If no splice sites are identified within 30 bp of the variants, the up/downstream variants are also flagged.
- The overall predicted splice site class is always the highest occurring class resulting from both 3' and 5' splice sites;
    - Class 1 – no splicing effect.
    - Class 2 – likely no splicing effect.
    - Class 3 – unknown splicing effect.
    - Class 4 – likely splicing effect
    - Class 5 – Splicing effect.
- Relevant splice site positions with their respective SSP classes and scores. The below information is provided for each listed position:
    - Score – predicted non/functionality likelihood of the splice site (-100%-100%). Splice sites having unknown functionality score 0%.
    - ΔScore (delta score) – difference between the score on the reference sequence and the variant sequence splice site.
    - MaxEntScan: The ENT score from MaxEntScan adapted from Yeo & Burge, 2003.
    - ΔMaxEntScan: difference between the MaxEntScan on the reference sequence and the variant sequence splice site.
- Prediction details corresponding to both the 5' and 3' splice sites.

- Public database information including the rs Number, varSEAK classification, ClinVar clinical significance, and GnomAD AF.

Chapter 3 – Results

3.1 Clinical phenotype of the proband under investigation

The proband under investigation is a 24-year-old Caucasian male, who was diagnosed in infancy with complex cyanotic congenital heart disease. He is the second in a sibship of three offspring born to non-consanguineous Caucasian parents. He was born by normal vaginal delivery following an uneventful term pregnancy. No delay in early motor, cognitive or developmental milestones were recorded. He exhibited typical growth and development in infancy and childhood having weighed 5.6 kg, 64cm length and 39cm head circumference at 2 ½ months and 6.7kg, 63.5cm length and 63.5cm head circumference at 4 months. No relevant family history related to congenital cardiovascular defects was reported. Echocardiography at 2 months old revealed dextrocardia and situs inversus with double outlet right ventricle and subpulmonary stenosis having a 4m/s gradient. Just under the age of 2 years the proband underwent his first intervention. This involved intracardiac exploration on cardiopulmonary bypass with the insertion of a 5mm left-sided modified Blalock-Taussig (LMBT) shunt from the brachiocephalic artery to the left pulmonary artery as a palliative procedure plan for potential eventual biventricular repair. Following this intervention, the spontaneous improvement of subpulmonary stenosis was observed, along with the subsequent development of pulmonary hypertension and Eisenmenger physiology. In addition to the complex congenital heart defect described above, the proband was also diagnosed with advanced secondary open angle glaucoma at the age of 17 years as well as secondary hypertrophic osteoarthropathy (HOA) at the age of 20 years. HOA is characterised by fibrovascular proliferation driven by hypoxia triggered growth factors manifesting as disabling arthralgia and arthritis, digital clubbing and periostitis of tubular bones, usually, but not always as a consequence of chronic pulmonary disease. The HOA manifested in the proband as recurrent fever of unknown origin, arthralgias, bone pain, recurrent knee and ankle effusions and digital clubbing. A graphical summary of the main cardiac, skeletal and ocular phenotypes in the proband is provided in figure 3.1.1. The proband is under the care of a consultant cardiologist specialising in congenital cardiac diseases in adulthood with routine echocardiography and regular follow ups. A list of the proband's medication is provided hereunder. The proband is prescribed several antihypertensive

medications including two used specifically for pulmonary hypertension macitentan (endothelin 1 receptor blocker) and sildenafil (PDE5 inhibitor). Enalapril and propranolol are part of treatment of systemic hypertension whilst colchicine was prescribed with targeted relief of bone pain (caused by the HOA).

- Sildenafil 20mg tds
- Macitentan 10mg dly
- Propranolol 20mg bd
- Enalapril 10mg nocte
- Warfarin
- Bumetanide 1mg dly
- Colchicine 500 μg daily (started 27/5/2022)
- Glaucoma eyedrops
- Intermittent iron supplements

At the time of the proband's and his family's self-referral for participation in genetic research, no genetic analysis was ever performed on the proband along with the family in question.

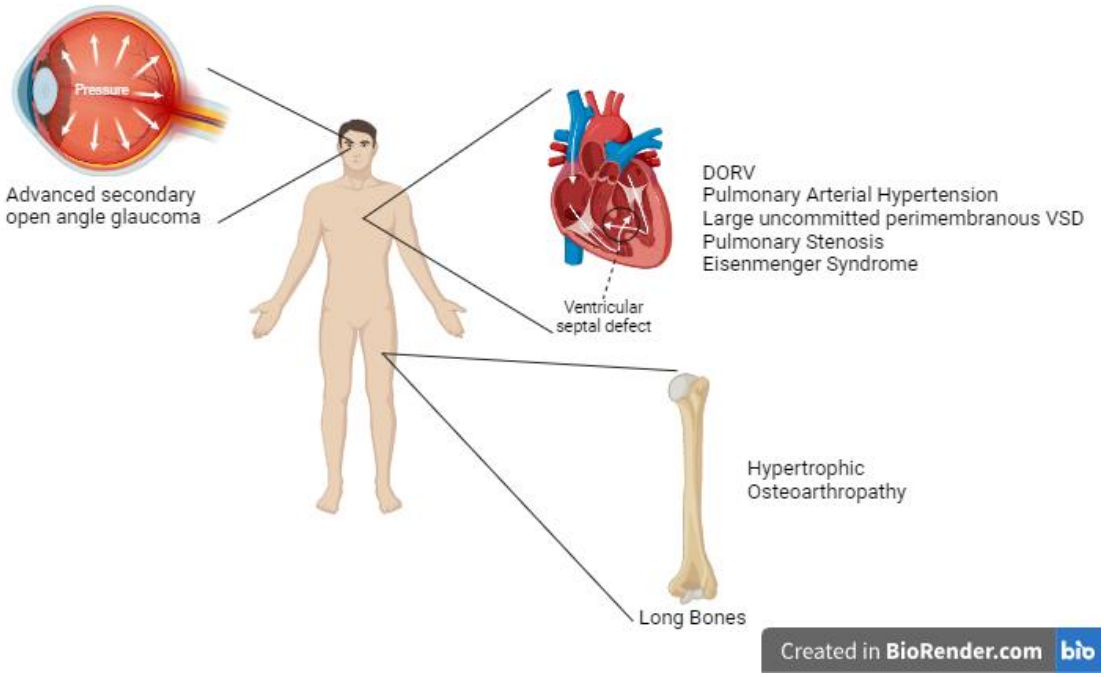


Figure 3.1.1. A graphical summary of the main cardiac, skeletal and ocular phenotypes within the proband. The ocular phenotype consists of advanced secondary open angle glaucoma. The skeletal phenotype is present within the long bones, graphically represented by the humerus. The proband was clinically diagnosed with hypertrophic osteoarthropathy. The cardiac phenotypes include double outlet right ventricle (DORV), Pulmonary arterial hypertension, Large uncommitted perimembranous ventricular septal defect (VSD), pulmonary stenosis and Eisenmenger syndrome.

The self-reported history of the family under investigation can be briefly visualised in the pedigree shown below (figure **3.1.2**).

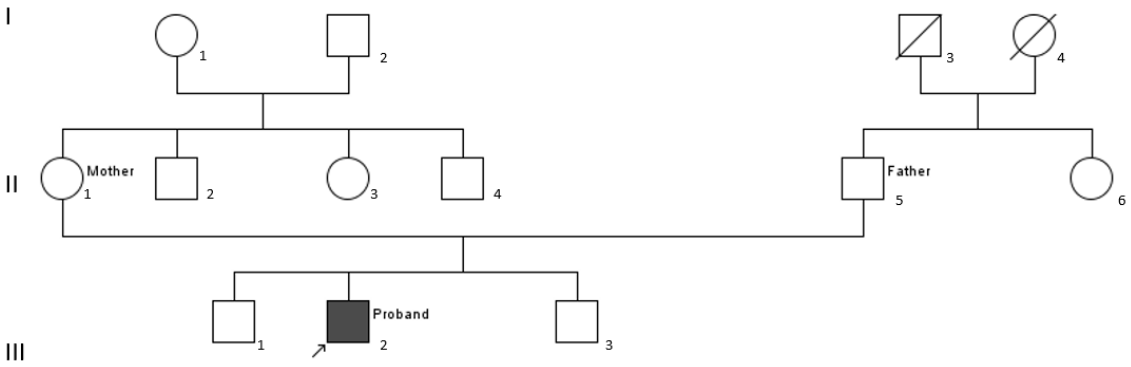


Figure 3.1.2. A three-generation pedigree of the family included in this study. The proband (subject III.2) has a complex congenital heart defect with DORV and a large VSD and Eisenmenger syndrome. No relevant family history was available. This study performed trio-WES on the proband (III.2), his father (II.5) and mother (II.1).

3.2 Variant filtering prioritisation strategy.

The Variant call files (VCF) from the three sequenced individuals were analysed using Franklin by Genoox portal. Each VCF was filtered to derive all variants in the 635 gene panel implicated in CHD, PAH, HOA and situs inversus curated via a systematic literature search (Appendix A).

- A total of 2,075 variants in 427 loci were identified in the proband, of which 1,870 were SNVs and 205 indels.
- A total of 2,128 variants in 447 loci were identified in the mother, of which 1,887 were SNVs and 241 indels.
- A total of 2,118 variants in 427 loci were identified in the father, of which 1,892 were SNVs and 226 indels.

The above values include all coding/non-coding variants located in exons, exon-intron junctions and UTRs detected in the trio within the curated gene panel (Appendix A). These values are summarised graphically in **figure 3.2.1**.

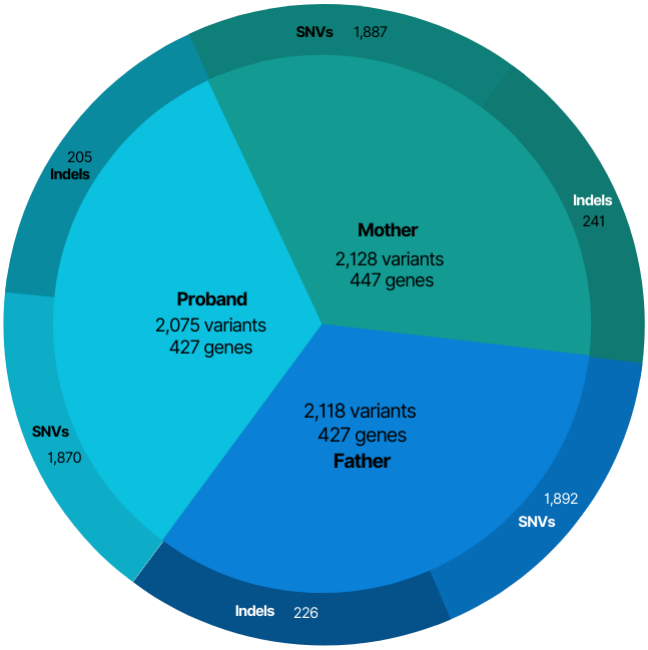


Figure 3.2.1. Pie chart showing the number of variants in the number of genes per individual in the trio along with the subdivision of these variants according to sub-type – SNVs or Indels. The proband having 2,075 variants present in 427 genes, of which 1,870 SNVs & 205 Indels. The mother having 2,128 variants in 447 genes, of which 1,887 SNVs & 241 Indels. The father having 2,118 variants in 427 genes, of which 1,892 SNVs & 226 Indels.

The sequential variant filtering and prioritisation strategy applied to the trio WES dataset split into 6 steps is outlined next. The funnel plot depicted in **figure 3.2.2** is a graphical representation of the variant filtering strategy outlined next.

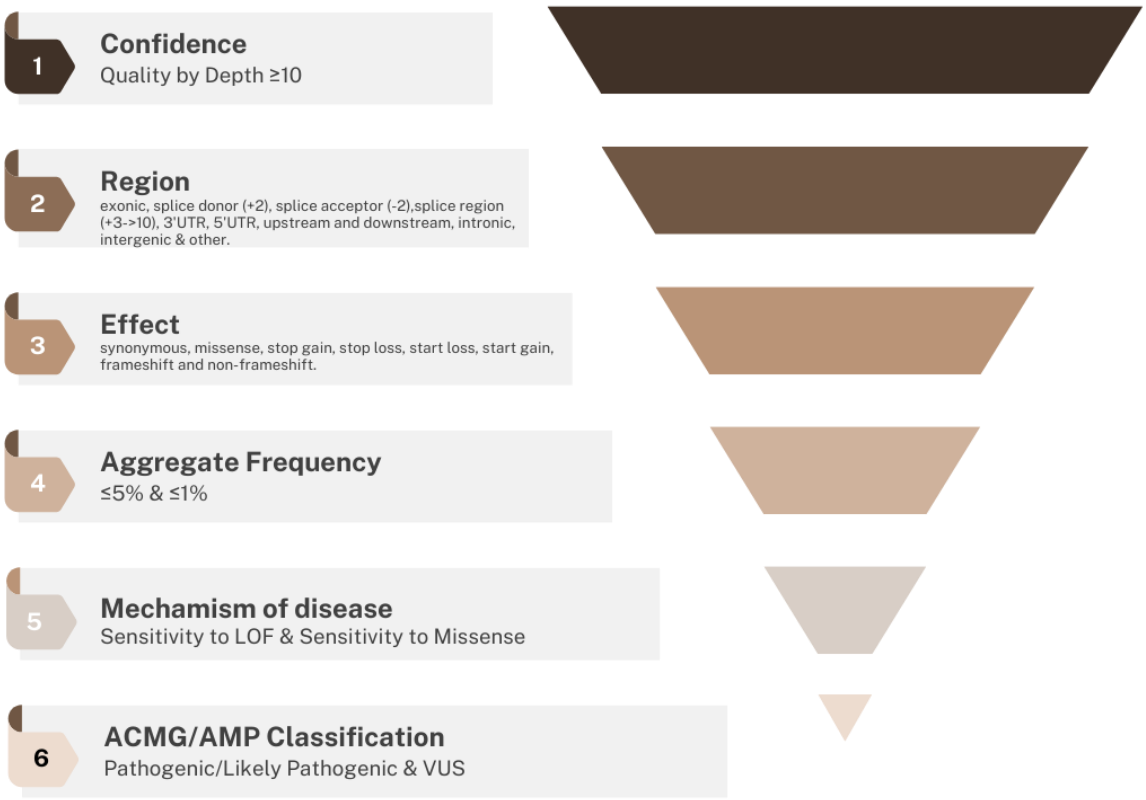


Figure 3.2.2. Funnel plot showing the sequential variant filtering prioritisation strategy applied to the trio WES dataset. The first step being filtering according to conference having a quality and depth of the reads (≥10). The second filter being according to region, including exonic, splice donor (+2), splice acceptor (-2), splice region (+3->10), 3'UTR, 5'UTR, upstream and downstream, intronic, intergenic, and other. The third filter step being that of effect of the variant on protein further sub-filtering the variants according to synonymous, missense, stop gain, stop loss, start gain, start loss, frameshift and non-frameshift. The fourth filtering step being according to the variant's aggregate frequency, including those at a value of ≤5% and ≤1%. The fifth filtering step being according to resulting variant's mechanism of disease, sub-categorised according to sensitivity to LOF and sensitivity to missense. And the final filtering step being that according to ACMG/AMP classification hence further sub-categorised according to the variants' result as pathogenic/likely pathogenic (P/LP) and Variants of uncertain significance (VUS).

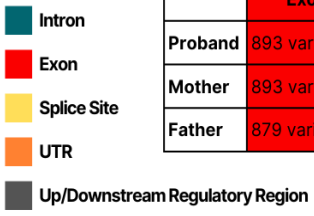
3.2.1 Step 1: Filtering by Confidence

The first (1st) step consisted of filtering according to call confidence, which was sub filtered via quality by depth, thus eliminating any reads having a depth ≤10. The addition of this filter allowed the shortlisting of the below variants.

- In the proband, 2,055 variants in 426 genes were identified, of which 1,857 were classified as SNVs and the remaining 198 variants were classified as Indels.
- In the mother, 2,113 variants were identified in 446 genes, of which 1,880 were classified as SNVs whilst the remaining 233 were classified as indels.
- In the father, 2,040 variants were identified in 420 genes, of which 1,830 were classified as SNVs whilst the remaining 198 variants were classified as Indels.

All the consecutive steps of variant filtering are carried over to the next step, so the filtering remains applicable from one step to the next and is hence cumulative.

A subdivision of the detected variants according to genomic regions in each individual is shown in **figure 3.2.3.**



	Exon, splice sites	UTRs	Upstream downstream	Intronic, intergenic other
Proband	893 variants in 311 genes	103 variants in 77 genes	9 variants in 5 genes	1,050 variants in 314 genes
Mother	893 variants in 326 genes	119 variants in 80 genes	9 variants in 7 genes	1,092 variants in 339 genes
Father	879 variants in 301 genes	80 variants in 100 genes	9 variants in 5 genes	1,052 variants in 327 genes

Figure 3.2.1.1. Visual representation of the results gathered from region filtering showing the individual results for the proband, mother, and father according to the region filter applied being colour coded. The legend on the left side shows the colour coding according to genomic region intron (green), exons (red), splice sites (yellow), UTRs (orange) and upstream and downstream regulatory regions (grey).

3.2.2 Step 2: Filtering by Region

In this second (2nd) step, we prioritised variants located in coding (exons) regions and variants in canonical splice donor/acceptor regions (±2 bases) along with splice region variants (±3 – 10 bases of the canonical splice donor/acceptor). We selected to focus on variants within coding or splice regions due to these having a higher likelihood of being associated to disease due to;

a) A direct functional impact on protein structure, function and/or stability, and b) The high level of conservation of coding/splice sites across species and, c) The higher relevance of these regions to disease due to their link to protein function and disease.

3.2.3 Step 3: Filtering by Effect

In the third (3rd) filtering step, variants in the coding and splice regions were further prioritised to identify protein-altering variants. For each gene, filtering was performed with reference to the canonical transcript in the Ensembl database. A subdivision of the functional impact of these exonic and splice variants is show in **table 3.2.3.1** and **figure 3.2.3.1.**

In all subsequent filtering steps, we retained protein-altering missense variants and variants that alter the reading frame (nonsense, frameshift indels and non-frameshift indels) due to their likelihood towards clinical relevance over synonymous variants.

Table 3.2.3.1. Summary of number of variants and the respective number of genes for each member of the trio (proband, mother and father) according to the filters; synonymous variants, missense variants, stop gain, stop loss, start gain and start loss, frameshift and non-frameshift.

	Synonymous		**Missense**		**Stop gain, stop loss, start gain, start loss**		**Frameshift & non-frameshift**	
	Variants	**Genes**	**Variants**	**Genes**	**Variants**	**Genes**	**Variants**	**Genes**
Proband	574	253	430	220	135	104	143	107
Mother	571	260	416	225	120	92	129	97
Father	546	244	854	294	123	98	133	101

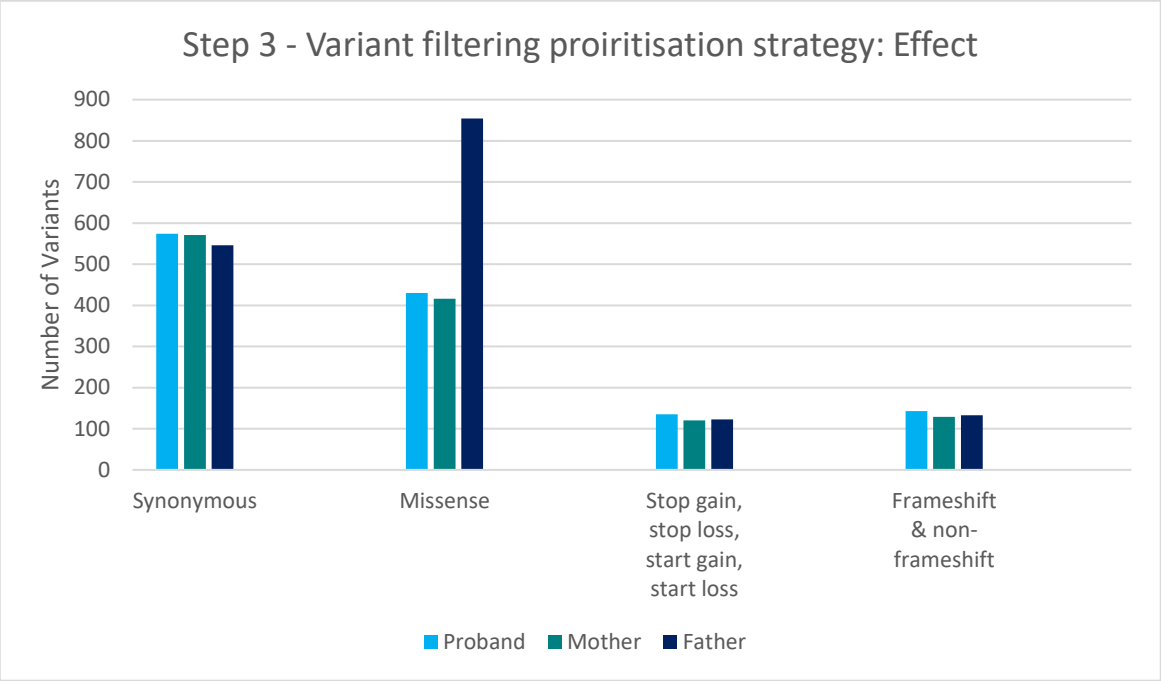


Figure3.2.3.1. Graphical representation of the 3rd step of the variant filtering algorithm's resulting variants. The bar chart shows the number of variants per person of the trio (Proband, Mother and Father) according to the filter of protein altering variants including; Synonymous, Missense, Stop Gain, Stop Loss, Start Gain, Start Loss, Frameshift and non-frameshift.

3.2.4 Step 4: Filtering by Allele Frequency

In the fourth (4^{th}) filtering step, a frequency filter was implemented. Frequency filtering aims to shortlist and prioritise genomic variants that are absent or found at very low frequency in reference genomic datasets that are derived from large-scale population genomic research initiatives, such as GnomAD or 1000 genomes project. The rationale for selecting rare variants is further supported by the hypothesis of this research, which aims at the investigation of a possible monogenic/Mendelian aetiology in the proband. Rare *de novo* deleterious variants can theoretically drive the development of CHD phenotypes as observed in the family under investigation. The aggregate datasets used are described in the methods. Two frequency filters were applied in succession: an aggregate frequency ≤5% and an aggregate frequency of ≤1%. The number of variants stratified by effect and frequency threshold is show in **table 3.2.4.1** and **figure 3.2.4.1.**

Table 3.2.4.1 Summary of number of variants and the respective number of genes for each member of the trio (proband, mother and father) according to the selected filters; Aggregated frequency ≤5%, Aggregated Frequency ≤1%, Aggregated Frequency ≤5% Missense Variants, Aggregated Frequency ≤5% stop gain, stop loss, start loss, start gain variants and Aggregated Frequency ≤5% frameshift and non-frameshift variants.

	Allele Frequency ≤1%		**Allele Frequency ≤5% Missense Variants**		**Allele Frequency ≤5% stop gain, stop loss, start loss, start gain variants**		**Allele Frequency ≤5% frameshift and non-frameshift Variants**	
	Variants	**genes**	**Variants**	**genes**	**Variants**	**genes**	**Variants**	**genes**
Proband	33	20	53	35	14	13	15	14
Mother	26	25	47	41	10	10	11	11
Father	31	17	45	31	9	7	10	8

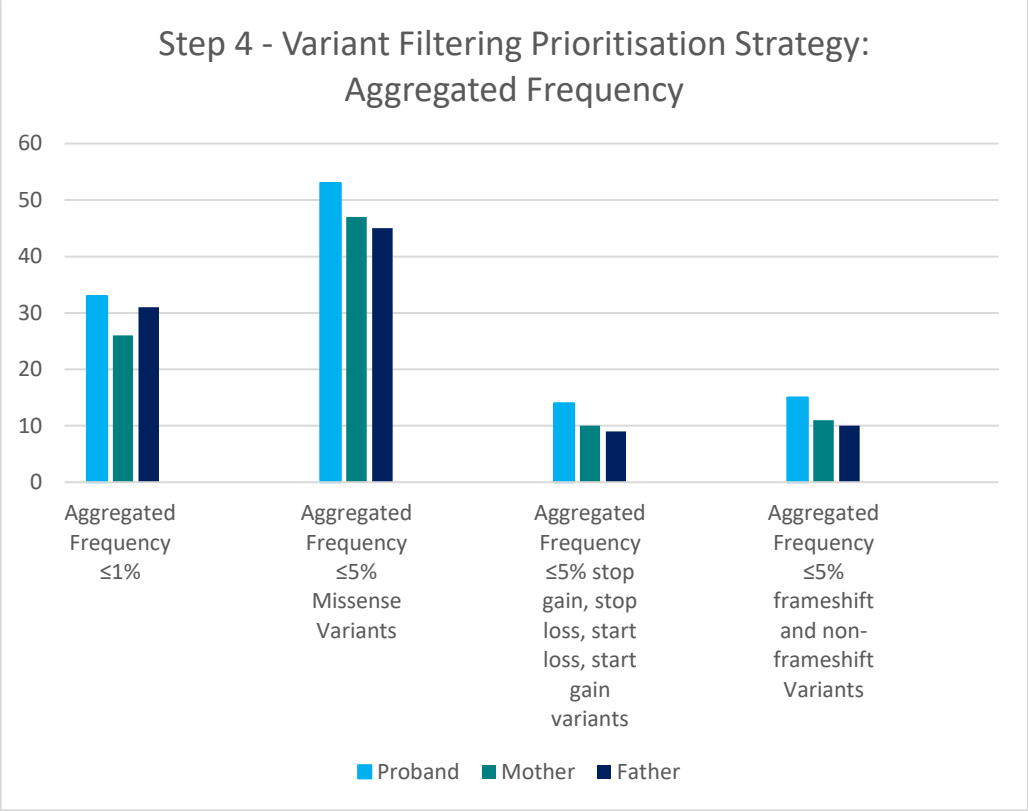


Figure 3.2.4.1 Graphical representation of the 4th step of the variant filtering algorithm's resulting variants. The bar chart shows the number of variants per person if the trio (proband, mother and father) according to the filter of aggregated frequency at ≤1% along with the respective previous protein altering variant filter and at ≤1%.

3.2.5. Step 5: Filtering by Mechanism of Disease.

In the fifth (5th) sequential filtering step, we applied gene-level rules to assess the sensitivity of a gene to missense variation or LOF. Gene sensitivity to LOF describes how a gene responds to mutations thus result in loss/reduction of the encoded protein function. Gene sensitivity to missense variants refers to the degree of which its function is affected by single amino acid changes. Genes vary in their tolerance to variants that reflect their biological role. The sensitivity to LOF filter is based on the gene constraint metric of observed/expected ratio in GnomAD. Genes are shown if pLoF o/e (observed/expected ratio) upper score <= 0.35, or if their pLI (probability LOF intolerance) score >0.9, whilst the sensitivity to missense filter is based on the gene constraint metric of o/e in GnomAD. Genes are shown if missense o/e upper score <= 0.35. Gene missense sensitivity is defined as missense variation in a gene with a low rate of benign missense variants and for which missense variants are a common cause of disease. Unlike the previous filters where typically the number of variants and genes do not overlap and are independent between filters, for sensitivity to LOF and missense, variants and genes may overlap in both situations. The number of variants stratified according to mechanism of disease can be visualised in the below **table 3.2.5.1** and **figure 3.2.5.1.**

Table 3.2.5.1. Summary of number of variants and their respective genes for each member of the trio (proband, mother and father) according to the filters selected; Mechanism of disease subcategorised into the Sensitivity to LOF and Missense, Sensitivity to LOF alone and Sensitivity to Missense alone.

	Mechanism of Disease: Sensitivity to LOF & Missense		**Mechanism of Disease: Sensitivity to LOF**		**Mechanism of Disease: Sensitivity to Missense**	
	Variants	**Genes**	**Variants**	**Genes**	**Variants**	**Genes**
Proband	22	17	22	17	1	1
Mother	21	20	21	20	0	0
Father	13	11	13	11	0	0

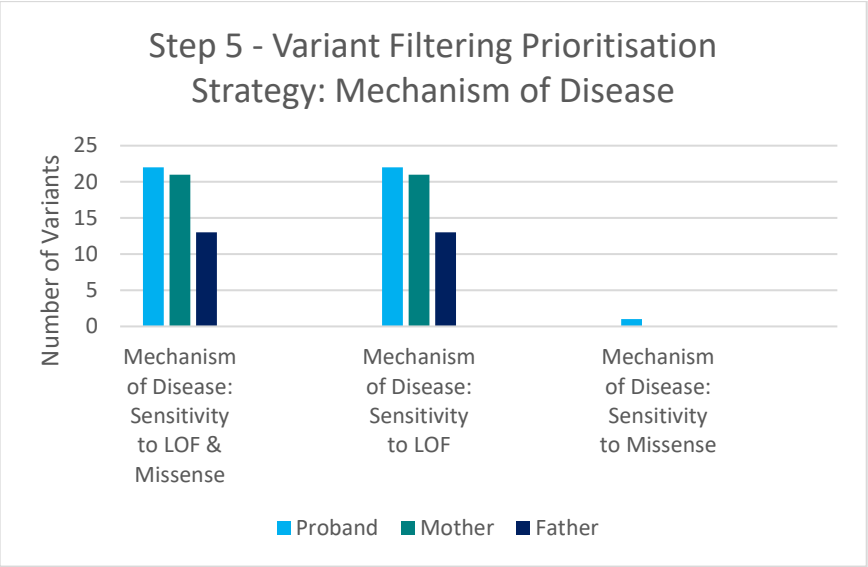


Figure 3.2.5.1. Clustered bar chart showing the 5th step of the variant filtering algorithm. The bar chart shows the number of variants per person in the trio (proband, mother and father) according to the filter of mechanism of disease subcategorised into the Sensitivity to LOF and Missense, Sensitivity to LOF alone and Sensitivity to Missense alone.

3.2.6 Step 6: Filtering by ACMG/AMP classification

The sixth (6^{th}) and final step of the variant filtering strategy is that of ACMG/AMP classification, whereby the variants are selected according to their attributed ACMG/AMP classification ranging from pathogenic/likely pathogenic (P/LP), benign/likely benign (B/LB) and variants of uncertain significance (VUS).

One (1) variant in one (1) gene (*BMPR2* NM_001204.6 p.Arg491Trp) was identified in the proband and ranked as pathogenic according to ACMG/AMP criteria. This variant was detected in the monoallelic (heterozygous) state and was absent in the parents, and thus was considered *de novo*. The functional impact of this variant is outlined in **section 3.4.1** below. **Table 3.2.6.2** summarise the key features of this variant.

In addition to the pathogenic *BMPR2* p.Arg491Trp missense variant, four (4) VUS were also shortlisted in the proband. These are:

- *RLF* NM_012421.4 p.Arg38Cys;
- *AHSA1* NM_012111.3 c.691-10_691-9delTT;
- *GNAQ* NM_002072.4 c.736-6_736-5dupTT;
- *KLHL3* NM_017415.3 c .527-9_527-8delTT

These variants in *AHSA1, GNAQ* and *KLHL3* are *de novo* and thus unique to the affected proband. The *RLF* pArg38Cys variant was detected in the heterozygous state in both the affected proband and the unaffected mother. The biological significance of these shortlisted candidate variants is described in further sections.

No P/LP variants were identified in the parents. Seven (7) VUS were identified in the parents (5 in the mother, 3 in the father) of the proband summarised in table 3.2.6 below. These VUS were all in the monoallelic state and all except one (*RLF* p.Arg38Cys) were not detected in the affected proband, and hence were not considered further.

Table 3.2.6.1. Summary of the shortlisted variants for each member of the trio (proband, mother and father) according to the filters selected; ACMG/AMP classification according to Pathogenic/Likely Pathogenic (P/LP), and Variants of Uncertain Significance (VUS).

	ACMG/AMP: P/LP	**ACMG/AMP: VUS**
Proband	*BMPR2* p.Arg491Trp	Variant 1: *RLF* p.Arg38Cys Variant 2: *AHSA1* c.691-10_691-9delTT Variant 3: *GNAQ* c.736-6_736-5dupTT Variant 4: *KLHL3* c .527-9_527-8delTT
Mother	N/A	Variant 1: *FBN2* p.Ala2761Val Variant 2: *PBRM1* p.Tyr148Phe Variant 3: *GLI3* p.Asn1031Lys Variant 4: *RLF* p.Arg38Cys Variant 5: *EDNRA* p.Ile327Met
Father	N/A	Variant 1: *DCHS1* p.Ala486Val Variant 2: *CACNA1C* p.Ala28Thr

The above-described sequential filtering strategy therefore allowed the prioritisation of deleterious variants within the trio WES dataset from all the identified variants in all the gene panel, to be funnelled down to the most clinically relevant variants in their respective genes being described in respective sections. The below funnel plot (**figure 3.2.6.1**) depicts the above-described variant filtering strategy applied to the aggregate data of each individual forming part of the trio.

Table 3.2.6.2 summarises the variants surviving the filtering prioritisation strategy in the trio following the above-described variant filtering algorithm. One variant in *BMPR2* was classified as P/LP by ACMG/AMP consensus criteria were identified in the proband and is therefore considered *de novo*. No P/LP variants were identified in the parents. In the mother and the father, five (5) and two (2) variants respectively classified as VUS were identified.

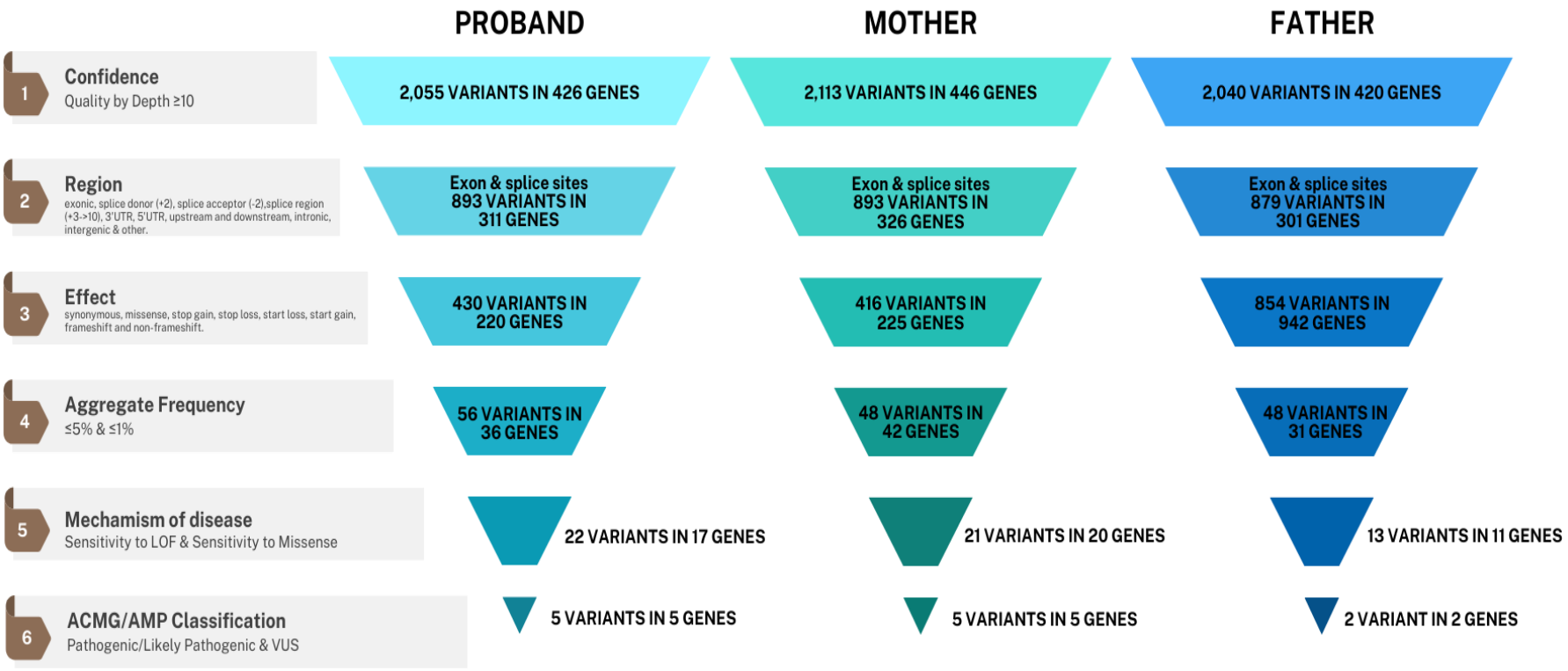


Figure 3.2.6.1. Figurative representation of the above figure 3.2.2 variant filtering strategy showing the values resulting at each step for each member of the trio; the proband, the mother and the father. Each sequential step from 1 to 6 shows the filtering of the variants according to various factors including confidence, region, effect, aggregate frequency, mechanism of disease and ACMG/AMP classification. The values presented in each individual's funnel plots are sequential and represent the variants prioritised at each filtering step.

Table 3.2.6.2. One variant surviving the above-described filtering prioritisation strategy categorised as pathogenic/likely-pathogenic by the ACMG/AMP consensus criteria in the trio. The table shows the properties of the identified variant present only within the proband, including the variation type, position, dbSNP, transcript, amino acid change, exon, zygosity, and effect of the variants. The table also shows the prevalence of the variants in databases including 1000 genomes, ExAc (all), GnomAD (exomes) and GnomAD (genome).Additional properties presented include ACMG/AMP criteria for the classification as P/LP, aggregated and internal frequency and different in-silico predictors along with their values. The scores pertaining to the predictions of each in-silico predictor is described in further in the methods and values found in the supplementary material. "Del" refers to "deleterious".

Gene	Position (hg19)	dbSNP	Transcript	Aa change	Exon	Zygosity	Effect	1000 genomes	ExAC (All)	GnomAD (Exome)	GnomAD (Genome)	ACMG/AMP
BMPR2	Chr2:203417496	rs137852746	NM_001204.6	p.Arg491Trp	11	Het	Missense	N/A	N/A	N/A	N/A	Pathogenic PS4, PM1, PP2, PM2, PM5, PP3, PP5
	Aggregated Frequency	**Internal frequency**	**SIFT Pred**	**POLYPHEN2 Pred**	**MUTTASTER Pred**	**MUTASSESSOR Pred**	**FATHMM Pred**	**GERP**	**REVEL Pred**	**PHRED CADD**	**BayesDel**	**Genocanyon**
	N/A	N/A	Del	Del	Del	Hi	Del	4.58	Del	32	Del (Strong) (0.56)	Del (0.94)

Table 3.2.6.3 Variants surviving the above-described filtering prioritisation strategy categorised as Variants of uncertain significance (VUS) by the ACMG/AMP consensus criteria in the trio. The table shows the properties of the identified variants including the variation type, position, dbSNP, transcript, amino acid change, exon, zygosity, and effect of the variants. The table also shows the prevalence of the variants in databases including 1000 genomes, ExAc (all), GnomAD (exomes) and GnomAD (genome) along with the aggregated and internal frequency along with in silico predictors such as Splice AI for the variants found in splice regions and REVEL pred and BayesDel for missense variants.

Gene	Position (hg19)	dbSNP	Transcript	AA	Nucleotide	Exon	Zygosity	Region	1000 genomes
RLF	Chr1:40627183	rs147792979	NM_012421.4	p.Arg38Cys	c.112C>T	1	Het	Exonic	0.000399361
AHSA1	Chr14:77934394	rs34989956	NM_012111.3		c.691-10_691-9delTT	6	Het	Splice Region	N/A
GNAQ	Chr9:80343587	rs5898555	NM_002072.4		c.736-6_736-5dupTT	5	Het	Splice Region	0.00219649
KLHL3	Chr5:137013350	rs112292887	NM_017415.3		c.527-9_527-8delTT	5	Het	Splice Region	N/A
	ExAC (All)	**GnomAD (Exome)**	**GnomAD (Genome)**	**ACMG/AMP**	**Aggregated Frequency**	**Internal Sample Count**	**Splice AI**	**REVEL Pred**	**BayesDel**
RLF	0.001664236	0.001718084	0.001435	VUS - PM2, PP2, BP4	0.001685649	0%	N/A	Benign (moderate)	Benign (supporting)
AHSA1	0.018350931	N/A	N/A	VUS - PM2	3.00E-05	4.34%	Benign	N/A	N/A
GNAQ	0.011886968	0.024861103	0.003388	VUS - BP6	0.003522065	1.37%	N/A	N/A	N/A
KLHL3	0.013843614	0.021972589	0.001161	VUS - Criteria Unmet	0.001313062	0.45%	Benign	N/A	N/A

3.3. Sequential variant prioritisation – Application of Mendelian inheritance models

An additional filtering technique was also applied to the trio WES dataset. In addition to the sequential steps outlined above, we incorporated analyses according to different Mendelian segregation models.

The following models were considered:

a. Autosomal Recessive (AR): The proband/affected individual carries a pathogenic variant on both alleles (homozygous) for which both parents are heterozygotes and are thus only carriers. No variants were identified having an autosomal recessive pattern of inheritance in the trio.
b. X-linked recessive: A pathogenic variant in a gene on chromosome X causes the phenotype to be expressed in males, whilst females must carry the mutation on both X chromosomes in order to be affected. No variants were identified having an X-linked recessive pattern of inheritance in the trio.
c. X-linked dominant: A pathogenic variant in a gene on chromosome X causes the disease in a heterozygous manner. In this case only one copy of the allele is sufficient when inherited from a parent having the disorder. No variants were identified having an X-linked dominant pattern of inheritance in the trio.
d. Y linked: A pathogenic variant in a gene on chromosome Y causes the disease and is only present in males. No variants were identified having a Y-linked pattern of inheritance in the trio.
e. Compound heterozygous: the proband/individual carries two different heterozygous mutations on the same gene, each being inherited in a heterozygous manner from each healthy parent. No variants were identified having a compound heterozygous pattern of inheritance in the trio.
f. Autosomal Dominant: The proband/affected individual carries a pathogenic variant on a single allele. Commonly only one of the parents is affected, thus also being heterozygous for the pathogenic variant. The other parent shows the reference genotype. One variant was identified in the trio WES dataset having an autosomal dominant pattern of transmission. This is the *RLF* NM_012421.4 p.Arg38Cys which was identified in the heterozygous state in both the mother and the proband. As

outlined earlier, the mother was unaffected, hence this substitution variant was not considered further.

g. *De novo*: A variant that is present only on the proband – not present in any of the parents. Four (4) variants were identified in the trio WES dataset having a *De novo* pattern of transmission. These are *AHSA1* NM_012111.3 c.691-10_691-9delTT, *BMPR2* NM_001204.6 p.Arg491Trp, *KLHL3* NM_017415.3 c .527-9_527-8delTT and *GNAQ* NM_002072.4 c.736-6_736-5dupTT. One of which (*BMPR2*) being characterised as pathogenic by ACMG/AMP criteria.

Table 3.3.1 summarises the variants identified in this final step of inheritance modelling.

Table 3.3.1. Summary of the variants identified in the trio WES dataset all of which being present in the heterozygous state. The features of the genes tabulated include the position of the genes, their dbsNP, transcript ID, amino acid change for missense variants, nucleotide change, pattern of inheritance (Autosomal dominant (AD) and *de novo*, exon and region (splice region or exonic). Additional features for each gene include their presence in genomic datasets including ExAc, and GnomAD, along with their ACMG/AMP classification and in silico predictions including the Splice AI for splice varaints and REVEL and BayesDel for missense variants. Aggregated frequency is also defined for each variant along with their internal sample count, which accounts for ethnically matched controls.

Gene	Position (hg19)	dbSNP	Transcript	AA	Nucleotide	Inheritance	Exon	Region	1000 genomes
BMPR2	Chr2:203417496	rs137852746	NM_001204.6	p.Arg491Tryp	c.1471C>T	*De novo*	11	Exonic	N/A
RLF	Chr1:40627183	rs147792979	NM_012421.4	p.Arg38Cys	c.112C>T	AD	1	Exonic	0.000399361
AHSA1	Chr14:77934394	rs34989956	NM_012111.3		c.691-10_691-9delTT	*De novo*	6	Splice Region	N/A
GNAQ	Chr9:80343587	rs5898555	NM_002072.4		c.736-6_736-5dupTT	*De novo*	5	Splice Region	0.00219649
KLHL3	Chr5:137013350	rs112292887	NM_017415.3		c.527-9_527-8delTT	*De novo*	5	Splice Region	N/A
	ExAC (All)	**GnomAD (Exome)**	**GnomAD (Genome)**	**ACMG/AMP**	**Aggregated Frequency**	**Internal Frequency**	**Splice AI**	**REVEL Pred**	**BayesDel**
BMPR2	N/A	N/A	N/A	Pathogenic PS4, PM1, PP2, PM2, PM5, PP3, PP5	N/A	N/A	N/A	Deleterious	Deleterious (strong)
RLF	0.001664236	0.001718084	0.001435	VUS - PM2, PP2, BP4	0.001685649	0%	N/A	Benign (moderate)	Benign (supporting)
AHSA1	0.018350931	N/A	N/A	VUS - PM2	3.00E-05	4.34%	Benign	N/A	N/A
GNAQ	0.011886968	0.024861103	0.003388	VUS - BP6	0.003522065	1.37%	N/A	N/A	N/A
KLHL3	0.013843614	0.021972589	0.001161	VUS - Criteria Unmet	0.001313062	0.45%	Benign	N/A	N/A

3.4 Description of identified shortlisted variants.

This section will provide an in-depth description of the shortlisted deleterious (ranked as P/LP/VUS) variants in all individuals within the trio. This variant description will include a comprehensive analysis of the identified variants including *in-silico* modelling and predictor tools to assess the effect of the individual variants on protein structure and stability.

3.4.1.*BMPR2 p.Arg491Trp*

The *BMPR2* p.Arg491Trp (chromosome 2, position 203417496, C➔T) variant was detected in the affected proband in the heterozygous state and absent in both parents. This is consistent with a *de-novo* pattern of transmission.

The variant is classified as pathogenic according to ACMG-AMP consensus criteria based on the following parameters:

a. PP5 – Classified as pathogenic on ClinVar, associated with Idiopathic and/or Familial Pulmonary Arterial Hypertension. The variant has multiple consistent submissions on ClinVar, last reviewed January 2024.
b. PM5 - Alternative substitutions at the same amino acid residue determined to be pathogenic. 2 pathogenic alternative variants (p.Arg491Leu and p.Arg491Gln) have been described and classified as pathogenic.
c. PP3 - Multiple lines of computational evidence support a deleterious effect on the gene or gene product (conservation, evolutionary, splicing impact. The variant has a MetaRNN score of 0.939.
d. 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. The variant is located in a domain which is a mutational hotspot, having 17 amino acids in total, of which 13 have been described as pathogenic variants, and 4 as VUS, with no benign variant.
e. The variant is absent from GnomAD Genomes and GnomaAD exomes datasets. In addition, it has not been detected in other population datasets, including the Turkish Variome, Iranome,and GenomeAsia datasets.

This gene encodes a member of the bone morphogenetic protein (BMP) receptor family of transmembrane serine/threonine kinases. The ligands of this receptor are members of the TGF-β superfamily. BMPs are involved in endochondral bone formation and embryogenesis. These proteins transduce their signals through the formation of heteromeric complexes of two different types of serine (threonine) kinase receptors: type I receptors of about 50-55 kD and type II receptors of about 70-80 kD. Mutations in this gene have been associated with primary pulmonary hypertension, both familial and fenfluramine-associated, and with pulmonary venoocclusive disease.

This *de novo* missense variant detected in the proband lies in a protein kinase domain (residues205-496). The variant lies in an evolutionary conserved region, close to residues that are intolerant to variation. The evolutionary conservation of the protein can be visualised in **figure 3.4.1.1**. The tolerance landscape of the protein along with the respective domains can be visualised in **figure 3.4.1.2**.

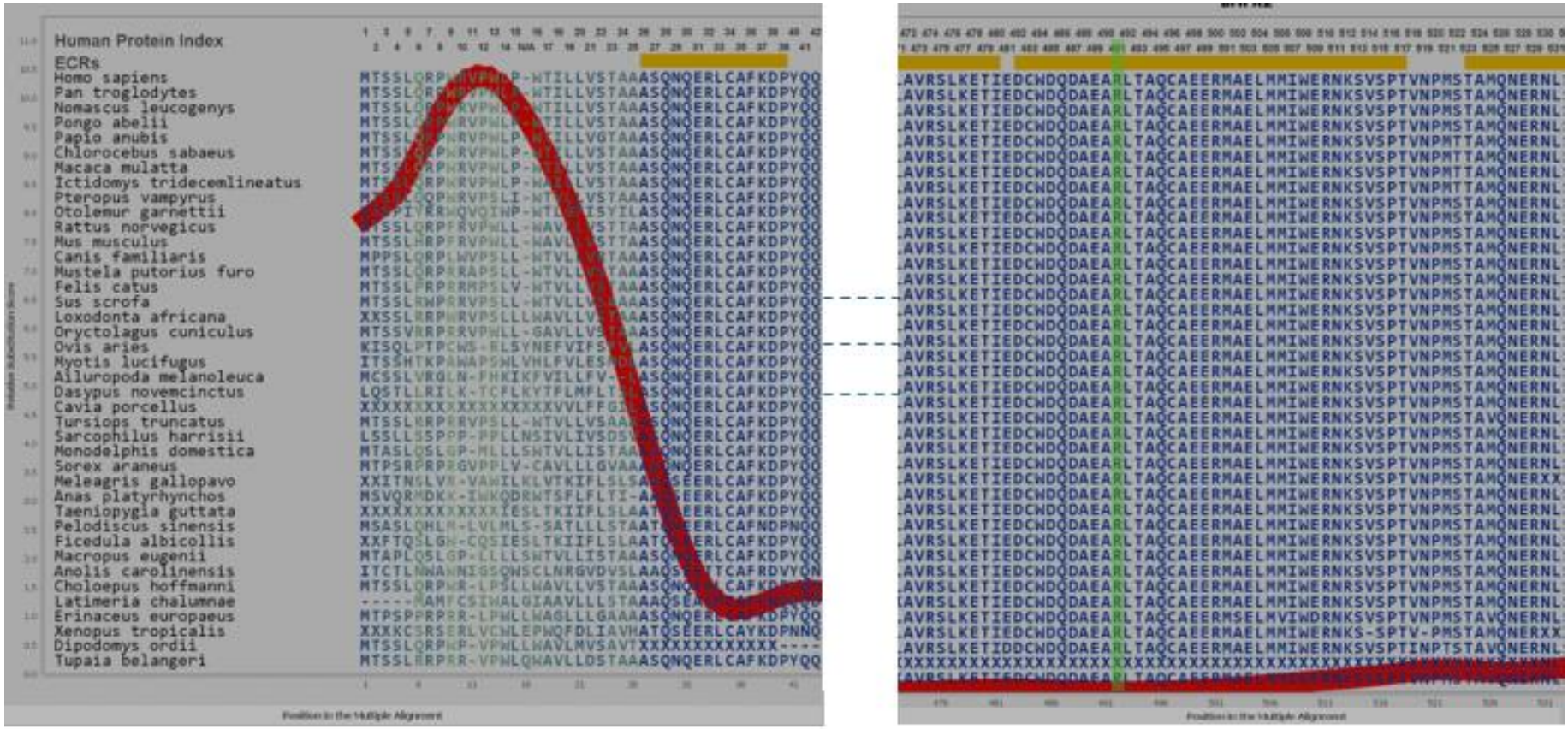


Figure 3.4.1.1.The Multiple sequence alignment shows the first 41 residues of BMPR2 along with the region flanking the variant of interest which is highlighted by the green vertical line. The red line depicts evolutionary constraints, with local maxima indicating a protein sequence region having relatively low evolutionary constraints, whilst the local minima indicate a protein sequence region with high evolutionary constraints. The orange horizontal bars represent the evolutionary constrained regions (ECRs). Figure adapted from aminode.org (Kevin T. Chang, Junyan Guo, Alberto di Ronza & Marco Sardiello. Aminode: Identification of Evolutionary Constraints in the Human Proteome. Scientific Reports 8:1357 (2018)).

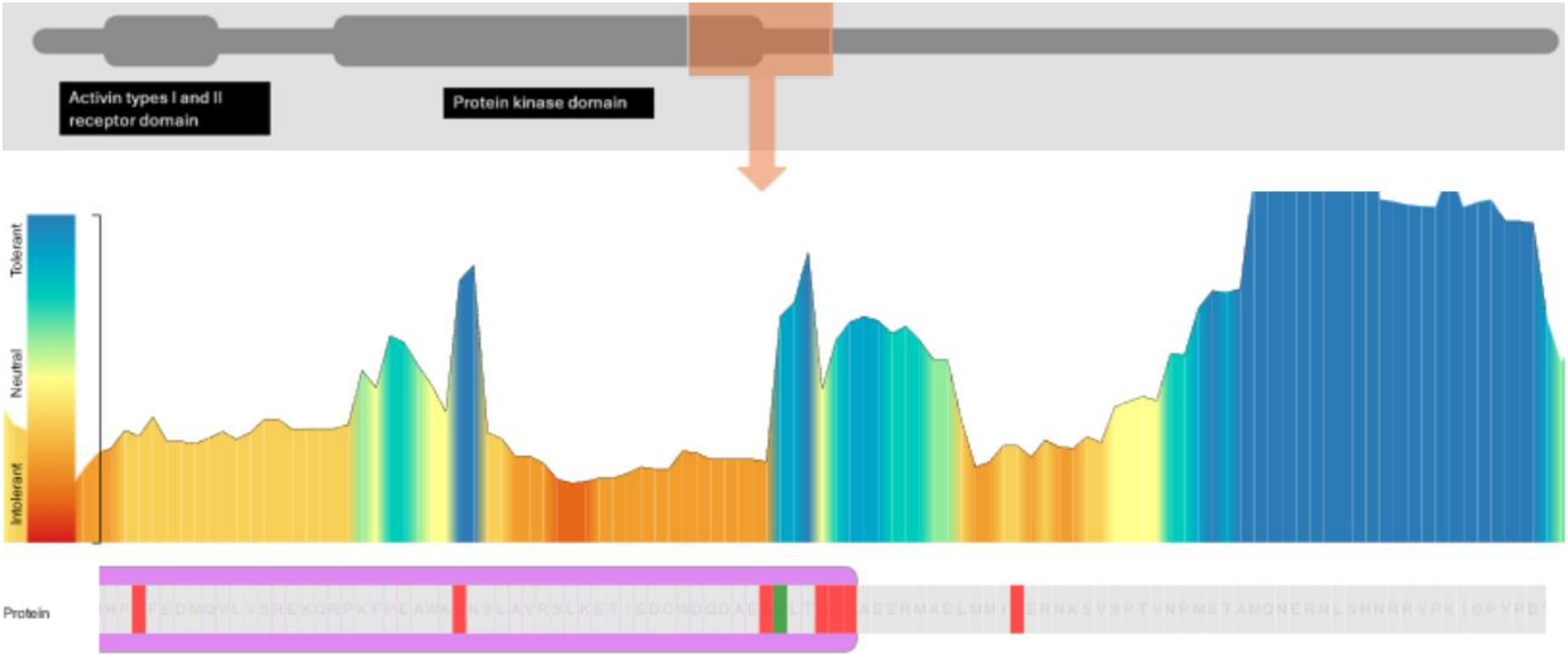


Figure 3.4.1.2. Tolerance landscape of BMPR2 along with the protein sequence. Position 491 is highlighted in green as the wild type R. The far left of the figure shows a legend for the heat map further displayed horizontally along the protein sequence. The orange box represents the boundary which is presented in the protein sequence shown at the bottom of the figure. Figure adapted from MetaDome (Pathogenicity analysis of genetic variants through aggregation of homologous human protein domains Human Mutation. 2019; 1-9. 10.1002/humu.23798).

Molecular modelling of the *BMPR2* p.Arg491Trp was implemented using structure derived from the Protein Data Bank (Accession code 3g2f). The wild-type Arginine and variant tryptophan amino acids differ in size, charge, and hydrophobicity. The charge of the buried wild-type residue is lost by this substitution. The variant Trp residue is larger and substitutes a positively charged Arg with a neutral Trp residue. Furthermore, the variant Trp is more hydrophobic than the wild-type Arg.

The wild-type residue forms a hydrogen bond with Glutamic Acid at position 386, Glutamine at position 403 and Aspartic Acid at position 485, and a salt bridge with Glutamic Acid at position 386 and Aspartic Acid at position 487.The difference in size and hydrophobicity alters these interactions.

A summary of the physiochemical properties of both the wild-type and the variant can be found in **table 3.4.1.1** below.

Table 3.4.1.1. Summary of the physiochemical properties of the wild-type amino acid BMPR2 and the Arg491Trp variant. The physiochemical properties summarised are the clash score, residue charge, interactions, salt bridges, and Van Der Waal interactions.

	Wild type – Arginine 491	**Variant – Tryptophan 491**
Clash Score	Local clash score 24.10	Local clash score 44.50
Residue Charge	Buried positively charged residue	Uncharged residue
H-bond Interactions	Arg491 – Glu386 – 3 Angstroms. Arg491 – Gln403 – 5.4 Angstroms. Arg491 – Asp485 – 7.1 Angstroms. Arg491 – Gln486 -9.5 Angstroms.	Weak H bonds between Trp491 and Glu489 and ASP487.
Salt bridge	Arg491 – Asp487 – 3.6 Angstroms. Arg491 – Glu386 – 13 Angstroms	N/A.
Van Der Waal interactions	N/A	Hydrophobic Van Der Waal Interactions between W491and GLN-403 and ALA-488.

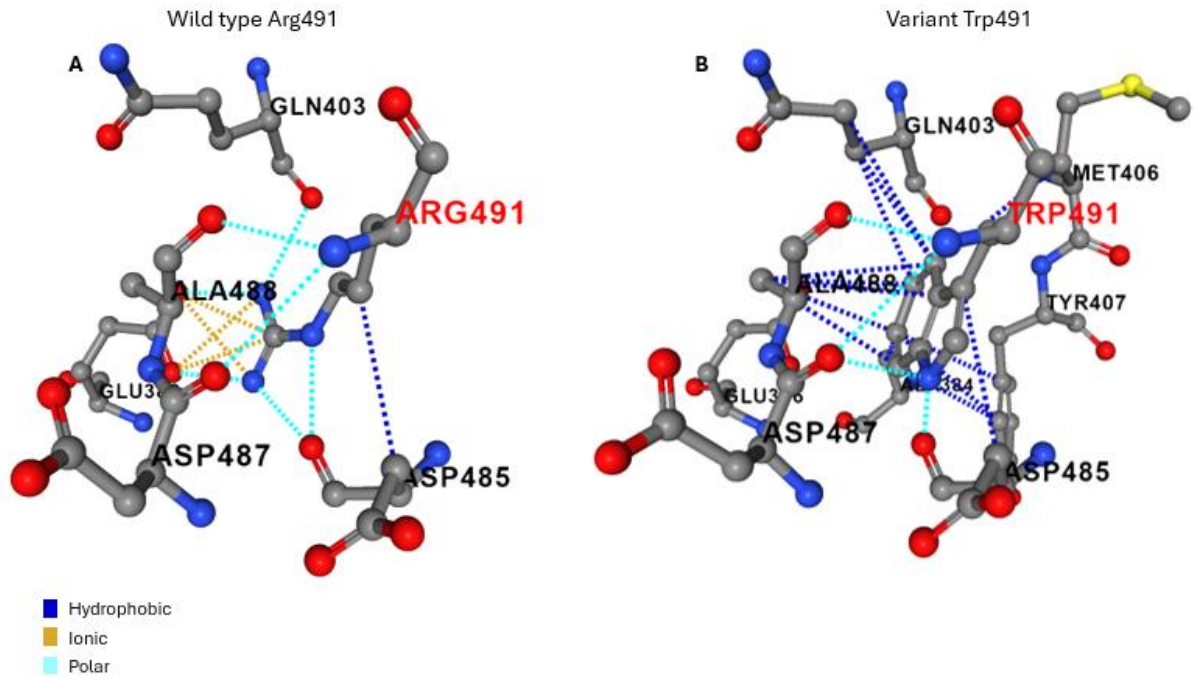


Figure 3.4.1.3. Figure A (computed via MutPred2) depicts the wild-type Arginine at position 491 within the BMPR2 protein along with the surrounding interacting amino acids. Figure B depicts the variant Tryptophan at position 491 in the BMPR2 protein along with the surrounding interacting amino acids. The interactions between the amino acid residues are colour coded according to the key. (Pejaver V, Urresti J, Lugo-Martinez J, Pagel KA, Lin GN, Nam H, Mort M, Cooper DN, Sebat J, Iakoucheva LM, Mooney SD, Radivojac P. Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. Nat. Commun. 11, 5918 (2020))

The structural consequences of the *BMPR2* p.Arg491Trp variant were explored using Missense3D. The substitution is predicted to be structurally damaging based on the following three criteria:

a. Stearic clash. The mutant structure has a MolProbity clash score ≥ 30 and the increase in clash score is > 18 compared to the wild type. This substitution triggers clash alert. The local clash score for wild type is 24.10 and the local clash score for mutant is 44.50.
b. Buried H-bond breakage. The substitution breaks all sidechain / sidechain H-bond(s) and/or side-chain / main-chain H-bond(s) formed by the wild type buried residue.
c. Buried charge replaced. This substitution replaces a buried charged residue (ARG, RSA 4.8%) with an uncharged residue (TRP).

The below figure 3.4.1 4 shows the wild type Arg at position 491 and the mutant Trp as predicted by Missense3D. The above-mentioned clash score can be visualised in figure 3.4.1.5 showing the different rotamers of the tryptophan variant at position 491 in the BMPR2 p.Arg491Trp protein.

Analysis using the Dynamut2 webserver predicts the substitution to be destabilising (ΔΔGStability -1.23 kcal/mol).

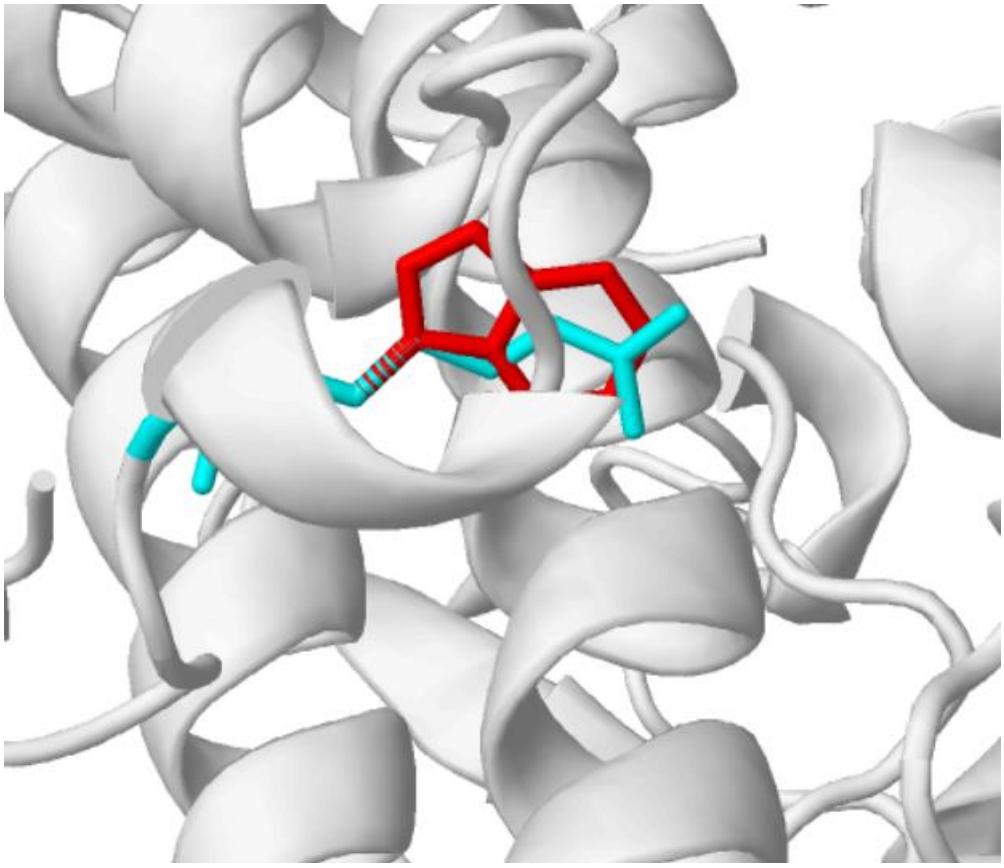


Figure 3.4.1.4. In silico model (computed by PyMol) of Arginine in blue and Tryptophan in Red at position 491 in the BMPR2 protein. (The PyMOL Molecular Graphics System, Version 3.0 Schrödinger, LLC).

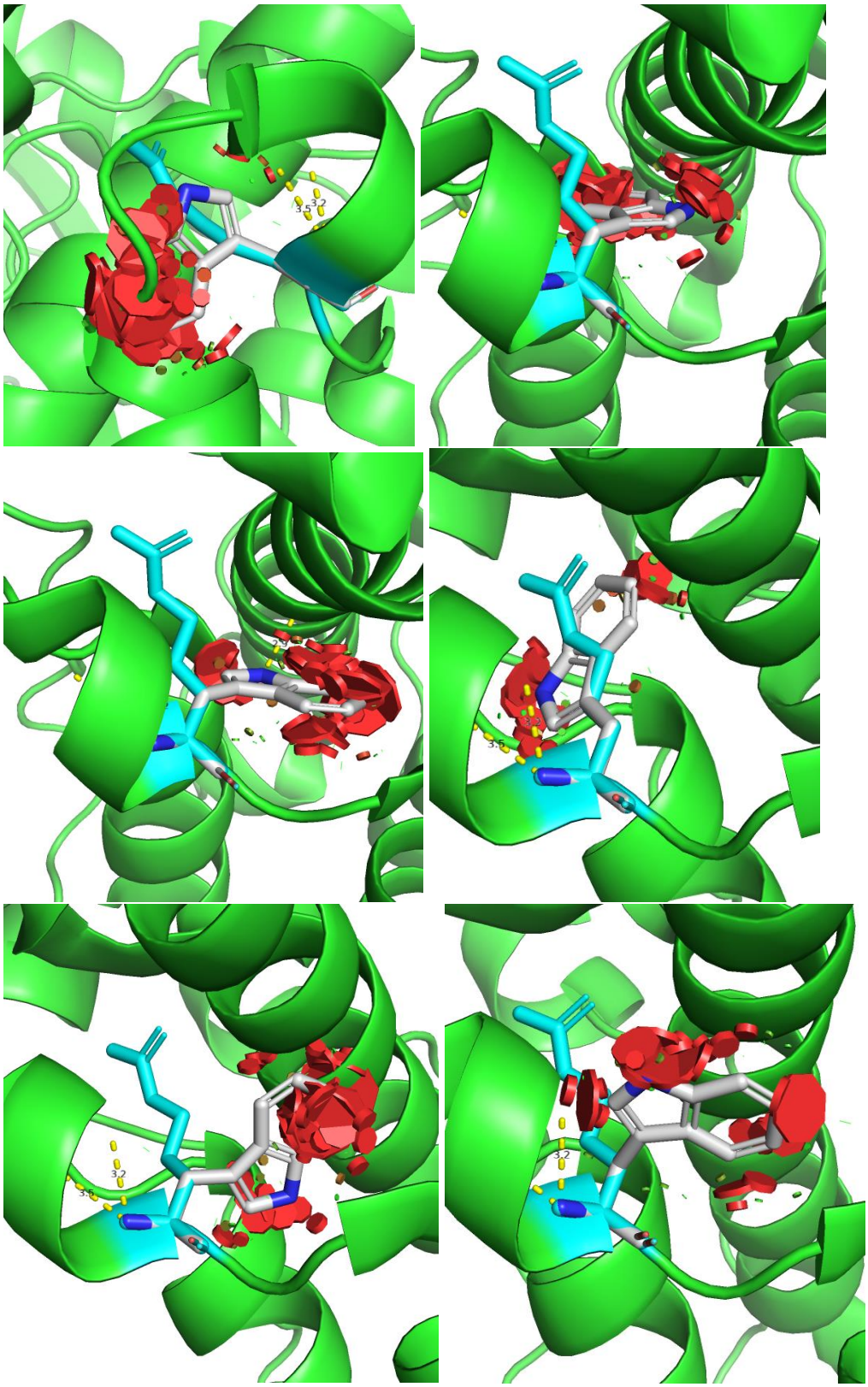


Figure 3.4.1.5. The different rotamers for the mutagenesis of Arginine at position 491 within the protein BMPR2 to Tryptophan. The green structure depicts the BMPR2 The red disks represent clashes of the mutated residue with the surrounding protein molecule. (The PyMOL Molecular Graphics System, Version 3.0 Schrödinger, LLC).

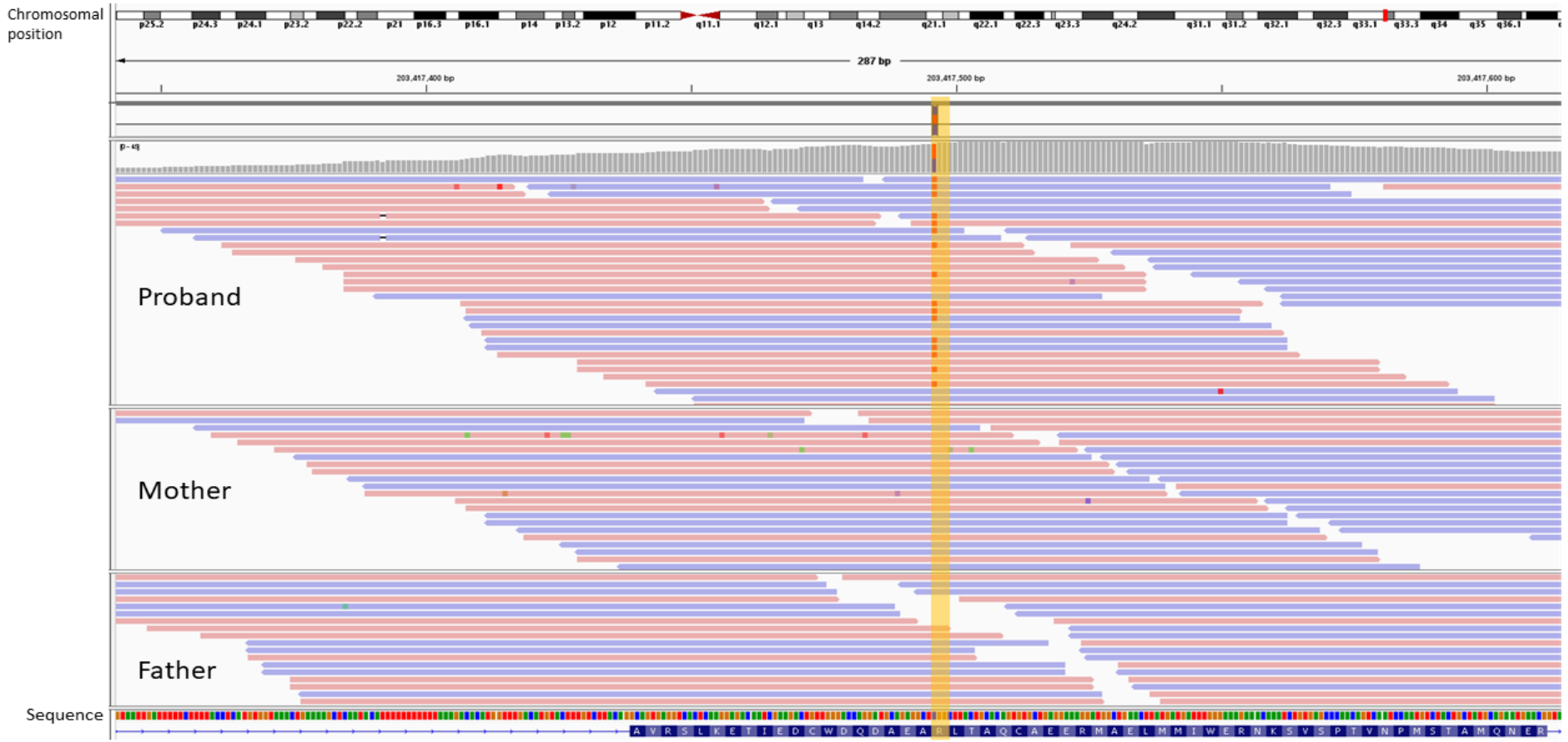


Figure 3.4.1.6. Graphical representation of the sequencing data for *BMPR2* for the trio participants; proband, mother and father respectively. At position 491 the Arginine residue can be visualised in the bottom blue bar. The red markers in this area present within the sequence for the proband in approximately half of the sequencing reads, consistent with a heterozygous genotype. The yellow vertical bar shows Arginine at position 491 and the identified variant in the proband aligned to the mother and the father sequence at the same position. The absence of red markers in the mother and the father sequence show that this variant is only present within the proband. Figure adapted from Integrative Genomics Viewer (IGV) (James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. Integrative Genomics Viewer. Nature Biotechnology 29, 24–26 (2011). A public access version is also available: PMC3346182).

3.4.2.*GNAQ*

The *GNAQ* gene is responsible for the production of Guanine nucleotide-binding protein G(q) subunit alpha (Gαq) which is localised in various parts of the cell including the cell membrane as a lipid anchor and the Golgi apparatus (Tsutsumi et al., 2009). Gaq forms part of the trimeric Guanine nucleotide binding protein (GNBP). Upon activation by the GPCR ligand, Gaq releases and binds GDP and GTP simultaneously, dissociating from the trimeric protein complex, hence activating further downstream pathways including the RAS-MEK-ERK, HIPPO-YAP,5 and, indirectly, mTOR. The upregulation of these pathways have been associated to the aberrant endothelial cellular growth and function (Comi et al., 2016).

The variant identified NM_002072.4, c.736-6_736-5dupTT is a double T insertion within a poly T tract at intron 5/exon 6 boundary, 5 bases proximal to the canonical AG acceptor splice site on chromosome 9. In silico predictors show that this duplication does not alter splicing activity at the 3' acceptor splice site as represented by **figure 3.4.2.1**. The variant has a dbSNP ID rs5898555. The variant was identified in a *de novo* manner within the proband and is classified as a VUS according to ACMG/AMP criteria, due to the singular BP6 characterisation, thus being borderline likely benign however not fulfilling the criteria. This variant is supported by 4 ClinVar entries ranking it as benign.

. This variant had an internal frequency of 1.37 % in ethnically matched control individuals hence further contributing towards the rationale of this variant as benign within the affected proband.

In view of the non-deleterious role of this splice variant, it has been concluded as benign and thus having no role in the proband's phenotype.

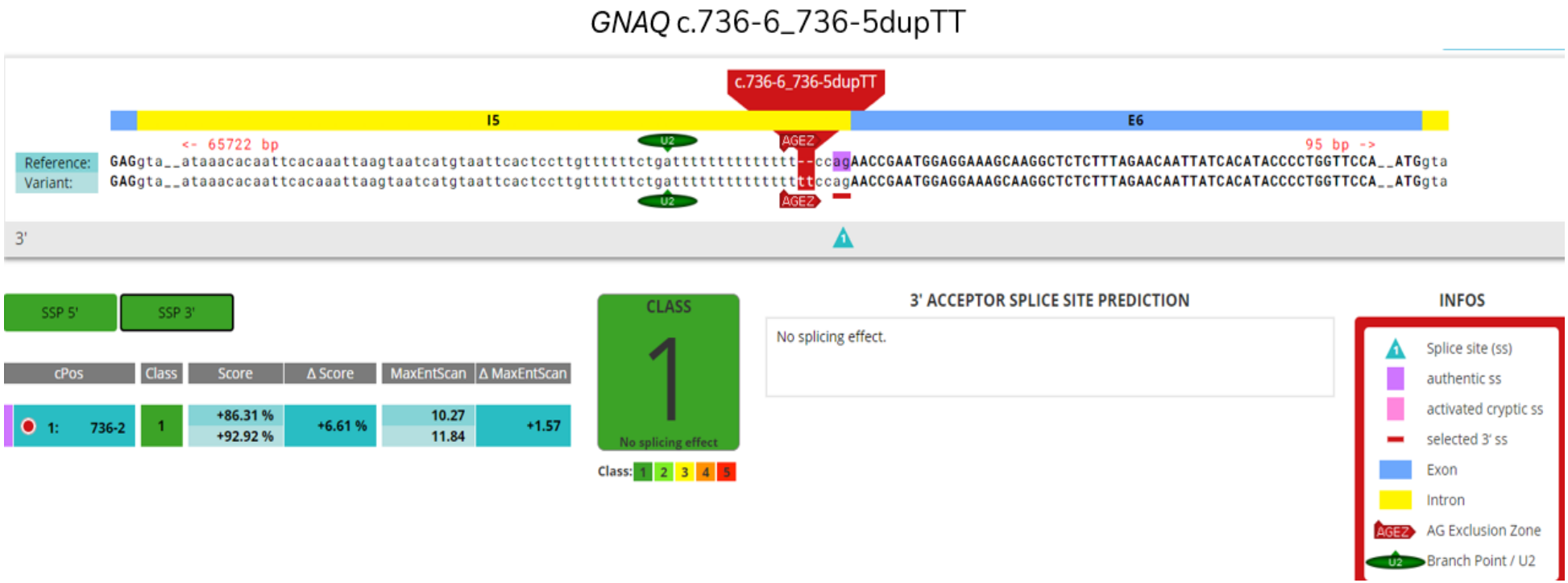


Figure 3.4.2.1. In silico splice site variant effect prediction for the *GNAQ* variant via the tool varSEAK. The variant of interest can be visualised by the red highlighted region along with the nucleotide labelling. This variant is a double T insertion within the poly T tract of intron 5, 5 bases away from the canonical AG splice acceptor site. The predicted class is 1 relating to no splicing effect. The bottom right shows a legend for all the symbols in the figure. (varSEAK, JSI Medical Systems)

3.4.3.*KLHL3*

The *KLHL3* gene is responsible for the production of Kelch-like protein 3 and has been identified within the cytoplasm of animal cells (Louis-Dit-Picard et al., 2012).

The variant identified NM_017415.3, c c.527-9_527-8delTT is a double T deletion within a poly T tract at intron 5/exon 6 boundary, 8 bases proximal to the canonical AG acceptor splice site on chromosome 5. In silico predictors show that this duplication does not alter splicing activity at the 3' acceptor splice site as represented by **figure 3.4.3.1**. The variant has a dbSNP ID rs112292887. The variant was identified in a *De novo* manner within the proband and is classified as a VUS according to ACMG/AMP criteria, due to not fulfilling any criteria for ACMG/AMP classification the criteria. This variant has no ClinVar entries.

SpliceAI ranked this variant as benign. This variant had an internal frequency of 0.45% in ethnically matched control individuals whilst also being present at low frequency (<1%) within aggregate datasets.

In view of the non-deleterious role of this splice variant, it has been concluded as benign and thus having no role in the proband's phenotype.

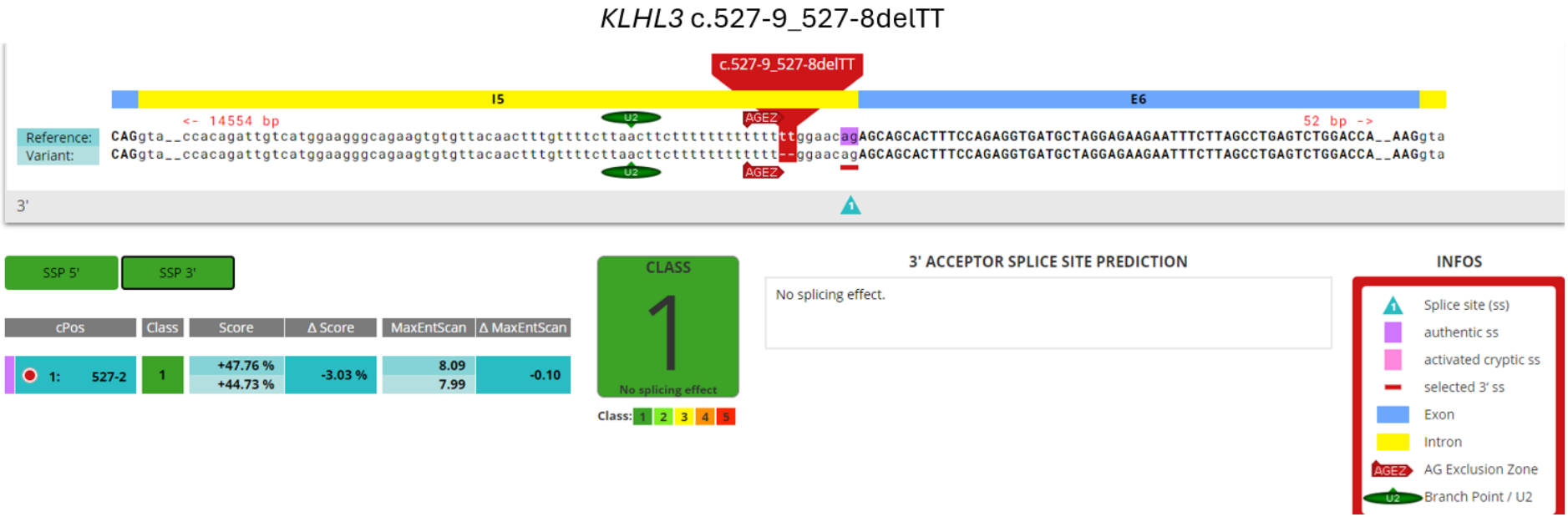


Figure 3.4.3.1. In silico splice site variant effect prediction for the *KLHL3* variant via the tool varSEAK. The variant of interest can be visualised by the red highlighted region along with the nucleotide labelling. This variant is a double T deletion within the poly T tract of intron 5, 8 bases away from the canonical AG splice acceptor site. The predicted class is 1 relating to no splicing effect. The bottom right shows a legend for all the symbols in the figure. (varSEAK, JSI Medical Systems)

3.4.4. *RLF*

The *RLF* gene is responsible for the production of the Zinc finger protein RLF (rearranged L-myc fusion gene protein) within the nucleus. The protein product is responsible for DNA binding and transcription factor activity. (Gaudet P et al., 2011)

The variant identified NM_012421.4, p.Arg38Cys is a missense variant in exon 1 on chromosome 1. In silico predictors rank this missense variant as deleterious. The variant has a dbSNP ID rs147792979. This variant was classified as pathogenic according to ACMG/AMP criteria. The variant was identified in the heterozygous state in the affected proband and the unaffected mother whilst being absent in ethnically matched datasets.

In view of the presence of this variant in the unaffected mother and affected proband, it is unlikely that this variant is deleterious due to the absence of the phenotype in the mother.

3.4.5. *AHSA1*

The *AHSA1* gene is responsible for the production of activator of 90 kDa heat shock protein ATPase homolog 1, an intracellular protein responsible for the activation of heat shock proteins.

The variant identified NM_012111.3 , c.691-10_691-9delTT is a double T deletion within a poly T tract at intron 6/exon 7 boundary, 9 bases proximal to the canonical AG acceptor splice site on chromosome 14. In silico predictors show that this duplication does not alter splicing activity at the 3' acceptor splice site as represented by **figure 3.4.5.1**. The variant has a dbSNP ID rs34989956. The variant was identified in a *De novo* manner within the proband and is classified as a VUS according to ACMG/AMP criteria, due to the PM2 categorisation for ACMG/AMP classification. This variant has no ClinVar entries.

SpliceAI ranked this variant as benign. This variant had an internal frequency of 4.34% which accounts for ethnically matched control individuals whilst also being present at low frequency (<1%) within aggregate frequencies.

In view of the non-deleterious role of this splice variant, it has been concluded as benign and thus having no role in the proband's phenotype.

Figure3.4.5.1. In silico splice site variant effect prediction for the *AHSA1* variant via the tool varSEAK. The variant of interest can be visualised by the red highlighted region along with the nucleotide labelling. This variant is a double T deletion within the poly T tract of intron 5, 9 bases away from the canonical AG splice acceptor site. The predicted class is 1 relating to no splicing effect. The bottom right shows a legend for all the symbols in the figure. (varSEAK, JSI Medical Systems)

Chapter 4 – Discussion.

4.1. Summary of Key Features

In this study, we have applied trio WES to identify deleterious genome variation of a proband with complex cyanotic CHD and Eisenmenger physiology. Using a sequential sequencing filtering/prioritisation strategy, a *de novo* deleterious *BMPR2* missense variant was identified. This variant is implicated in CHD and pulmonary hypertension.

The genetic aetiology of the proband's phenotypes (CHD, PAH and HOA) were investigated via a trio WES genetic analysis. Out of a 635 gene panel only one variant in *BMPR2* survived the variant filtering prioritisation strategy. This variant was found in the heterozygous state in the affected proband alone (absent in the parents), thus being consistent with a *de novo* pattern of transmission. Consequently, variants in this gene have been implicated in embryonic cardiac malformations and are further associated with the development of PAH as further discussed below.

The proband in question, developed PAH secondary to the CHD, which was identified at around 2 years of age. The primary genetic finding reported here - *BMPR2* p.Arg491Trp - is consistent with the clinical diagnosis in the proband. Literature has also supported the hypothesis of eventual PAH development in cases having *BMPR2* variants (Kim et al., 2017). These findings therefore suggest a possible genetic aetiology partly driving the PAH phenotype present within the affected proband.

4.2. BMPR2 p.Arg491Trp

BMPR2 encodes for the bone morphogenic protein II receptor, which forms part of the transforming growth factor beta (TGF-β) cell signalling family. BMPR2 is a single pass type I membrane protein. The protein kinase domain at which the mutation at position 491 is present is cellularly found within the cytoplasm. The BMPs were first discovered in rats as factors involved in the induction of ectopic cartilage formation (Miyazono et al., 2010). This TGF-β family possess serine/threonine kinase activity hence forming heteromeric complexes within type I membrane bound receptors. This

complex formation thus initiates phosphorylation of the type I receptor and further downstream SMAD or mitogen-activated protein kinases (MAPKs) (Massagué & Chen, 2000). BMPs form the largest group within the TGF-β family. The BMP plays critical roles both at the embryogenesis stage of life, and subsequently in adulthood. BMP is typically antagonised by various interacting proteins such as Noggin and Chordin (Massagué & Chen, 2000). This antagonistic effect induces further neural markers within the ectoderm and initiate conversion of the ventral mesoderm to dorsal tissue within explants of the gastrula ventral marginal zone (Massagué & Chen, 2000). Amongst some of the roles of BMPs include skeletal development, bone homeostasis as well as tissue regeneration via the signal transduction pathway mentioned above (Kang et al., 2004).

BMPR2 has been initially identified as the chondrogenic and osteogenic differentiation regulating factor, however its roles in early embryogenesis have been explored (Beppu et al., 2000). The expression of *BMPR2* (analysed within rodents) is relatively low during the initial heart developmental stages but gradually increase along embryo development particularly within the anterior telencephalon, branchial arches, tail tip mesoderm and limb bud. At later embryonic developmental stages the expression of BMPR2 is exceptionally high within the neuroectoderm towards the mouth anlagen (Danesh et al., 2009). BMP signalling is involved in the enhancement of endothelial specification, venous differentiation and angiogenesis during embryo development, hence having a crucial role in vascular homeostasis (Dyer et al., 2014; Zhang & Bradley, 1996). BMPR2, as a component to the BMP signalling transduction, has been found to be predominantly expressed within the vascular endothelia and smooth muscle layer of the pulmonary vasculature within typical lungs, whilst being under expressed in the airway and arterial smooth muscle (Atkinson et al., 2002). Within humans, *BMPR2* is expressed in microvascular endothelial cells, umbilical vein endothelial cells and aortic endothelial cells (Finkenzeller et al., 2012), thus highlighting the imperative role of *BMPR2* in vascular development. BMPR2 has been identified in vascular development through model organisms such as zebra fish where BMP2-BMPR2 mediated signalling is involved in the regulation of angiogenesis from the zebrafish axial vein. This hence demonstrates that BMPR2-dependent signalling promotes endothelial cellular proliferation and angiogenesis (Wiley et al., 2011). Human pulmonary arterial endothelial cell (HPAEC) survival and further proliferation via the ERK1/2 activation leads to endothelial cell migration. This is induced by the canonical WNT signalling pathway and

in turn the RhoA-Rac1 pathway, primarily activated via BMPR2 (de Jesus Perez et al., 2009). Studies performed on *BMPR2* knockout mice exhibited an inclination of PAH (Lee et al., 1998). Additional to the relation of BMPR2 to PAH, it has also been proven to mediate signal transduction in skeletal development (Katagiri et al., 1990; Katagiri et al., 1994; Wu et al., 2010). *BMPR2* is involved in osteoblast differentiation, whilst also being a mediator to bone formation and skeletal development during osteosclerosis and fracture healing (Garimella et al., 2007; Onishi et al., 1998). As previously mentioned, the BMPR2 protein is involved in bone development. BMPR2 and its ligands bring about the differentiation of mesenchymal stem cells towards osteoblasts, hence contributing directly towards the maturation of osteoblasts.

BMPR2 has been extensively described in literature as the primary gene associated with primary pulmonary arterial hypertension (PAH) (Ghigna et al., 2016; Higasa et al., 2017; Liu et al., 2012; Rudarakanchana et al., 2002; Wang, H. et al., 2014; Wang, X. et al., 2019; Yang et al., 2018). The *BMPR2* gene consists of 13 exons coding for four domains. Genetic variation to *BMPR2* may lead to missense, frameshift, nonsense, truncation as well as splice site variations, which may result in the loss of *BMPR2*-mediated signalling (Morrell, 2010). This gene typically follows an autosomal dominant inheritance pattern (Lane et al., 2000). Along with PAH, variants within the *BMPR2* gene have been reported in patients presenting with chronic obstructive pulmonary disease (COPD), hereditary haemorrhagic telangiectasia (HHT), prostatic neoplasms, colorectal cancer, as well as obesity (Kim et al., 2000; Morrell, 2006; Park et al., 2010; Rigelsky et al., 2008; Schleinitz et al., 2011). *BMPR2* variants are responsible for 75-90% of familial PAH (Evans et al., 2016) and 3.5-40% of sporadic cases (Girerd et al., 2016). A hypothesis on why *BMPR2* mutation carriers eventually develop PAH relates to the hyperactivation of the TGF-β signalling due to a decrease in BMP signal transduction, hence causing hyperproliferation of the smooth muscle cells in the pulmonary arterioles. This was further backed up by the identification of BMP signalling activation resulting in inhibition of smooth muscle cell proliferation. This underlying TGF-β signalling cascade in relation to BMP activity requires further understanding (Kim et al., 2017).

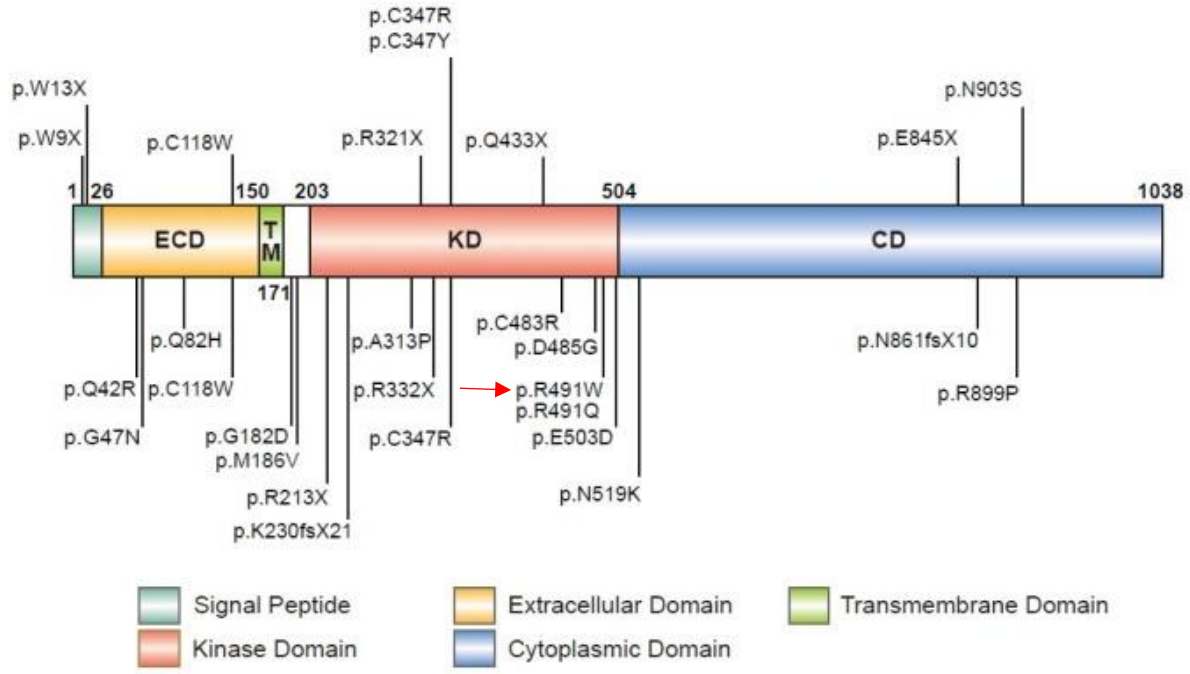


Figure 4.2.1. Adapted from Kim et al.(2017). *BMPR2* gene showing experimentally verified BMPR2 variants. The figure indicates the domains of the BMPR2 protein. Patient-derived cells functionally validated BMPR2 pathogenic variants are indicated above the gene, whilst BMPR2 pathogenic variants validated via in vitro functional assays are indicated below the gene. The red arrow indicates the R491W variant within the protein kinase domain. (Kim et al., 2017).

The penetrance of variants within *BMPR2* range from 14% in males and 42% in females (Larkin et al., 2012). Other genes having variants known to cause rare PAH include *KCNK3, ACVRL1, ENG, CAV1* along with the SMAD family (Austin et al., 2012; Chaouat et al., 2004; Shintani et al., 2009; Trembath, 2001). 298 variants in *BMPR2* have been identified within independent PAH patients (Machado et al., 2006; Machado et al., 2009). A study by (Liu et al., 2012) performed genetic screening on 305 Chinese PAH patients. Their study identified 21 missense mutations within *BMPR2* confined to exons 2, 3, 6, 8, 9, 11, and 12. The Kinase domain spanning along exons 5-11 was identified to harbour 13 discrete PAH related missense variations amongst the 305 participants, contributing towards the majority of the missense variants identified throughout their study.

The variant identified in the proband (p.R491W) is not a novel variant and has been mentioned in literature being first published on ClinVar in 2016 as the most recent publication on ClinVar in November 2023. Deng et al., in 2000 described the presence of R491W mutation in 3 families diagnosed with PPH. They also performed conservation analysis within this gene region and all type II TGF-β superfamily receptors, showing

that arginine at position 491 is conserved within many species. Deng et al., in 2000 further describe the mechanism by which they associate the mutations in *BMPR2* to primary pulmonary hypertension (PPH). They exclude the likelihood of these mutations acting in a dominantly negative way via the inhibition of the apoptotic result of TGF-β pathway, along with the exclusion of the entire knockout of the BMP-signalling pathway. Deng et al, suggest that due to the BMP pathway resulting in apoptosis in certain cell types, a partial block in this pathway's signal transmission might result in a reduced proliferative effect with the likely cause of either the dominantly negative protein interactions, or haploinsufficiency of *BMPR2* causing reduced signal transmission. Both mechanisms would result in a partial block of the BMP signal transmission, thus likely causing PPH. The previously mentioned study by Liu et al, in 2012 further described Arg 491 as one the key residues having essential catalytic activity. Rudarakanchana et al., in 2002 performed in vitro analysis of *BMPR2* variants, including p.R491W and further attributed this mutation towards a near complete abolition of SMAD pathway signalling. Higasa et al., 2017 also identified the p.R491W variant within 9 families having heritable PAH and upon further conservation analysis and in silico predictor modelling of the variant, classified this variant as damaging to protein function.

Recent studies have further correlated point mutations at Arg491 to be pathogenic and have described this gene region as a 'newly identified hot-spot for pulmonary arterial hypertension' (Chaikuad et al., 2019; Lyu et al., 2020). In 2020, Lyu et al. researched the largest *BMPR2* rare variant profile in 670 Chinese PAH patients, along with the parallel screening of these rare *BMPR2* variants in a reference population of10,508 participants. This study concluded that these variants had a significantly higher population prevalence in PAH diagnosed patients in comparison to the reference population. This was particularly significant for the amino acid Arginine at position 491 which was entirely absent from the reference population, but presented as a mutational hot-spot within the PAH study population. This study hence concluded that missense *BMPR2* variants identified within the PAH study population were found to be distributed at a higher proportion within the extracellular ligand-binding domain whilst the majority of identified *BMPR2* rare variants contributed to loss-of-function or splicing. In comparison to the previously described study, the study in 2019 by Chaikuad et al. analysed the structural consequences of PAH-associated missense mutation in the intracellular domain of BMPR2. Through their analysis, they identified that the most frequent and severely

destabilising variants were missense variants buried within the kinase C-lobe causing instability and misfolding within the catalytic domain. R491W was amongst one of the severely destabilising variants identified through this study, predicted to introduce severe steric clashes. This variant was found to disrupt the R491-Glu386 salt bridge along with the R491-Asp485 hydrogen bonds. These findings were also investigated within this study as can be visualised in section 3.4.1, table 3.4.1.1.

4.3 BMPR2 and Congenital Heart Disease (CHD)

Variants in the *BMPR2* gene have been associated with development of PAH throughout literature, with respect to primary, idiopathic, and hereditary forms of pulmonary arterial hypertension (PPH, IPAH & HPAH). The variant p.Arg491Trp identified in the proband in this case has also been cited in literature, as can be seen in **table 4.3.2**. However, the literature primarily correlates the variant with PPH, IPAH and HPAH phenotypes. Only one study in 2019 by Larrañaga-Moreira et al., identified this variant in a patient having previously diagnosed CHD.

Variants in *BMPR2* have also been described in literature in patients with CHD and PAH phenotypes. The first paper to investigate the correlation between variants in *BMPR2* and PAH-CHD phenotypes was published by Roberts et al. in 2004. This study was based on a cohort of 40 adults and 66 children with PAH-CHD and the results further supported previous mouse models by Danesh et al. in 2009. This study provided the first correlation between the development of PAH in patients with CHD and a genomic driver from variants in *BMPR2*.

Roberts et al., 2004 investigated the genomic correlation of *BMPR2* variants to patients phenotypically presenting with PAH and CHD. The study screened for variants in *BMPR2* in a cohort of 40 adults and 66 children with PAH-CHD (Eisenmenger syndrome). 5 out of 6 of the paediatric population within this study having *BMPR2* variants also had Down syndrome. Out of the adult population, only one had Down syndrome. The phenotypes correlating to the PAH-CHD included patent ductus arteriosus (PDA), transposition of the great arteries (TGA), partial anomalous pulmonary venous return (PAPVR), atrial and ventricular septal defects (ASD/VSD), atrioventricular canal (AVC) along with rare lesions having systemic-to-pulmonary shunts. This study identified 6 novel variants 3 of which were identified in 3 out of 4 adults (case 1 – case

3) phenotypically presenting with AVC complete type C (AVC-C) and the other 3 (case 4 – case 6) in children. The details pertaining to these variants identified can be visualised in **table 4.3.1**.

Table 4.3.1. Clinical findings and phenotypes of the patients having *BMPR2* variants in the study by Roberts et al., 2004. The table shows the patients' age at initial PHD diagnosis, whether the CHD has been repaired, the patients' sex and their diagnosed CHD phenotype. The second half of the table describes the *BMPR2* variant identified in these patients including the exon, nucleic acid change and amino acid change. AVC complete type C (AVC-C).

Patient	**Case 1**	**Case 2**	**Case 3**	**Case 4**	**Case 5**	**Case 6**
Age at PAH diagnosis years	1	16	5	3	2	19
CHD repair	No	No	Yes	No	No	Yes
Sex	F	F	M	F	M	M
CHD phenotype	AVC-C	AVC-C	AVC-C	ASD/PDA	ASD/PDA/PAPVR	AW and VSD
***BMPR2* variant**						
Exon	2	3	3	5	11	2
Nucleic acid change	125A>G	304A>G	319T>C	556A>G	1509A>C	140G>A
Amino acid change	p.Q42R	p.T102A	p.S107P	p.M186V	p.E503D	p.G47N

The findings supported results of previous mouse models by Danesh et al., in 2009. 75% of the *BMPR2* variants were present in the adult cohort having AV canals. Further literature has also supported the correlation between variants in *BMPR2* and the development of PAH along with the presence of CHD. The clinical relevance of *BMPR2* variants to the PAH-CHD phenotype can be further emphasised by table 5.3.2 which summarises the papers correlating variants in *BMPR2* with this phenotype. Although variants in *BMPR2* have been associated with the development of PAH in patients with CHD, this is not typically observed in the majority of cases. A recent study in 2020 performed by Welch & Chung only identified 7 out of 258 (2.7%) patients presenting with PAH and CHD. The presence of variants in *BMPR2* in cases with both PAH and CHD has also been correlated to the development of PAH to pulmonary vascular disease (PVD) in patients with CHD (Liu et al., 2016). This publication was the first to report the correlation between *BMPR2* variants in adults and children with PAH-CHD in whom the PAH is resulting from pulmonary vascular obstructive disease. This study thus suggested

that genetic drivers can predispose to PAH in patients with CHD in addition to the increased flow in the pulmonary circulation.

Table 4.3.2. Summary of available literature correlating the *BMPR2* p.Arg491Trp variant to the Pulmonary Arterial Hypertension (PAH) phenotype. The table shows the title of the paper, published year, author and the phenotype reported in the individual having the BMPR2 p.Arg491Trp variant. All except one paper described the presence of this variant with the primary, hereditary, or idiopathic PAH (PPH, HPAH & IPAH). The paper by Larrañaga-Moreira et al in 2019 was the only identified paper to report this variant in conjunction with both PAH and CHD phenotypes.

Reference	Year	Author	Phenotype Reported
Familial Primary Pulmonary Hypertension (Gene *PPH1*) Is Caused by Mutations in the Bone Morphogenetic Protein Receptor–II Gene.	2000	Zemin Deng, Jane H. Morse, Susan L. Slager, Nieves Cuervo, Keith J. Moore, George Venetos, Sergey Kalachikov, Eftihia Cayanis, Stuart G. Fischer, Robyn J. Barst, Susan E. Hodge, and James A. Knowles	PPH
Altered growth responses of pulmonary artery smooth muscle cells from patients with primary pulmonary hypertension to transforming growth factor-beta(1) and bone morphogenetic proteins	2001	N W Morrell, X Yang, P D Upton, K B Jourdan, N Morgan, K K Sheares, R C Trembath	PPH
Bone morphogenetic protein receptor-II mutation Arg491Trp causes malignant phenotype of familial primary pulmonary hypertension	2004	Jing Zhicheng , Lu Lihe, Han Zhiyan, Cheng Xiansheng, Zou Yubao, Yang Yuejin, Hui Rutai	HPAH
Clinical outcomes of pulmonary arterial hypertension in carriers of BMPR2 mutation	2007	Benjamin Sztrymf, Florence Coulet, Barbara Girerd, Azzedine Yaici, Xavier Jais, Olivier Sitbon, David Montani, Rogério Souza, Gerald Simonneau, Florent Soubrier, Marc Humbert	IPAH & HPAH
Clinical Outcomes of Pulmonary Arterial Hypertension in Carriers of BMPR2 Mutation	2008	Benjamin Sztrymf, Florence Coulet, Barbara Girerd, Azzedine Yaici, Xavier Jais, Olivier Sitbon, David Montani, Roge´rio Souza, Gerald Simonneau, Florent Soubrier, and Marc Humbert	IPAH & HPAH

clinical implications of determining BMPR2 mutation status in a large cohort of children and adults with pulmonary arterial hypertension	2008	Erika B Rosenzweig, Jane H Morse, James A Knowles, Kiran K Chada, Amar M Khan, Kari E Roberts, Jude J McElroy, Nicole K Juskiw, Nicole C Mallory, Stuart Rich, Beverly Diamond, Robyn J Barst	HPAH
Bone morphogenetic protein signalling in heritable versus idiopathic pulmonary hypertension	2009	Laurence Dewachter, Serge Adnot, Christophe Guignabert, Ly Tu, Elisabeth Marcos, Elie Fadel, Marc Humbert, Philippe Dartevelle, Gérald Simonneau, Robert Naeije, and Saadia Eddahibi1	HPAH & IPAH
Identities and frequencies of BMPR2 mutations in Chinese patients with idiopathic pulmonary arterial hypertension	2010	H Wang, Q-Q Cui, K Sun, L Song, Y-B Zou, X-J Wang, L Jia, X Liu, S Gao, C-N Zhang, R-T Hui	IPAH
Hemodynamic and clinical onset in patients with hereditary pulmonary arterial hypertension and *BMPR2* mutations	2011	Nicole Pfarr, Justyna Szamalek-Hoegel, Christine Fischer, Katrin Hinderhofer, Christian Nagel, Nicola Ehlken, Henning Tiede, Horst Olschewski, Frank Reichenberger, Ardeschir HA Ghofrani, Werner Seeger, and Ekkehard Grünig	HPAH
Molecular genetics and clinical features of Chinese idiopathic and heritable pulmonary arterial hypertension patients	2012	D. Liu, Q-Q. Liu, M. Eyries, W-H. Wu, P. Yuan, R. Zhang, F. Soubrier, Z-C. Jing	IPAH
Hemodynamic and genetic analysis in children with idiopathic, heritable, and congenital heart disease associated pulmonary arterial hypertension	2013	Nicole Pfarr, Christine Fischer, Nicola Ehlken, Tabea Becker-Grünig, Vanesa López-González, Matthias Gorenflo, Alfred Hager, Katrin Hinderhofer, Oliver Miera, Christian Nagel, Dietmar Schranz, and Ekkehard Grünig	IPAH

Defective cellular trafficking of the bone morphogenetic protein receptor type II by mutations underlying familial pulmonary arterial hypertension	2015	Anne John, Praseetha Kizhakkedath, Lihadh Al-Gazali, Bassam R.Ali	HPAH
Bone Morphogenetic Protein Receptor Type 2 Mutation in Pulmonary Arterial Hypertension	2016	Cathelijne E. van der Bruggen, Chris M. Happé, Peter Dorfmüller, Pia Trip, Onno A. Spruijt, Nina Rol, Femke P. Hoevenaars, Arjan C. Houweling, Barbara Girerd, Johannes T. Marcus, Olaf Mercier, Marc Humbert, M. Louis Handoko, Jolanda van der Velden, Anton Vonk Noordegraaf, Harm Jan Bogaard, Marie-José Goumans and Frances S. de Man	IPAH & HPAH
A burden of rare variants in *BMPR2* and *KCNK3* contributes to a risk of familial pulmonary arterial hypertension	2017	Koichiro Higasa, Aiko Ogawa, Chikashi Terao, Masakazu Shimizu, Shinji Kosugi, Ryo Yamada, Hiroshi Date, Hiromi Matsubara & Fumihiko Matsuda	HPAH
Clinical and genetic characteristics of pulmonary arterial hypertension in Lebanon	2018	Ossama K. Abou Hassan, Wiam Haidar, Georges Nemer, Hadi Skouri, Fadi Haddad, and Imad BouAkl	PAH
Genetic analyses in a cohort of 191 pulmonary arterial hypertension patients	2018	Hang Yang, Qixian Zeng, Yanyun Ma, Bingyang Liu, Qianlong Chen, Wenke Li, Changming Xiong and Zhou Zhou	PAH
Classification of Pulmonary Arterial Hypertension by Genetic and Familial Testing	2019	José M. Larrañaga-Moreira, Pedro J. Marcos-Rodríguez, Isabel Otero-González, María J. Paniagua-Martín, María G. Crespo-Leiro, Roberto Barriales-Villa	PAH-CHD

The features of rare pathogenic *BMPR2* variants in pulmonary arterial hypertension: Comparison between patients and reference population	2020	Zi-Chao Lyu, Lan Wang, Jian-Hui Lin, Su-Qi Li, Dan-Chen Wu, Tian-Yu Lian, Shao-Fei Liu, Jue Ye, Xin Jiang, Xiao-Jian Wang & Zhi-Cheng Jing.	HPAH
Prevalence and clinical features of bone morphogenetic protein receptor type 2 mutation in Korean idiopathic pulmonary arterial hypertension patients: The PILGRIM explorative cohort	2020	Jang, A. Y., Kim, B., Kwon, S., Seo, J., Kim, H. K., Chang, H., Chang, S., Cho, G., Rhee, S. J., Jung, H. O., Kim, K., Seo, H. S., Kim, K. H., Shin, J., Lee, J. S., Kim, M., Lee, Y. J., & Chung, W.	HPAH
Genetic Evaluation in a Cohort of 126 Dutch Pulmonary Arterial Hypertension Patients	2020	Lieke M van den Heuvel, Samara M A Jansen, Suzanne I M Alsters, Marco C Post, Jasper J van der Smagt, Frances S Handoko-De Man, J Peter van Tintelen, Hans Gille, Imke Christiaans, Anton Vonk Noordegraaf, HarmJan Bogaard, Arjan C Houweling	IPAH

Table 4.3.3. Summary of available literature correlating variants in *BMPR2* to patients diagnosed with pulmonary arterial hypertension (PAH) and/or congenital heart disease (CHD), although not all papers report the mutation of BMPR2 in PAH-CHD patients. The table shows the reference of the paper, year,*BMPR2* variant and phenotype. (CHD-PAH -AVC-C) congenital heart disease - complete atrial ventricular canal defect type C. ASD: atrial septal defect. PDA: patent ductus arteriosus. PAPVR - partial anomalous pulmonary venous return. AW- aortopulmonary window. ASD: atrial septal defect; PDA: patent ductus arteriosus. APAH associated Pulmonary Arterial Hypertension.

Paper Title	Year	*BMPR2* Variant	Phenotype
Clinical and genetic characteristics of pulmonary arterial hypertension in Lebanon.(Abou Hassan et al., 2018)	2018	p.Q6*(1)	Large ASD & PAH
		p.N126S (1)	PAH
		p.R491W (2)	PAH
		p.S775N (3)	VSD, ASD, PAH
The Genetic Epidemiology of Pediatric Pulmonary Arterial Hypertension.(Haarman et al., 2020)	2020	p.(Trp16*)	HPAH
		p.(Pro134Leufs*18)	HPAH
		c.530-?_c.621+?del	HPAH
		p.(Arg491Trp)	HPAH
		p.(Tyr314Serfs*11)	HPAH
		p.(Arg899*)	HPAH
Transforming growth factor-beta receptor mutations and pulmonary arterial hypertension in childhood.(Harrison et al., 2005)	2005	p.W16X	IPAH
		exon 5/6/7	IPAH
Genetic analyses in a cohort of children with pulmonary hypertension. (Levy et al., 2016)	2016	N/A	IPAH & FPAH
BMPR2 mutation is a potential predisposing genetic risk factor for congenital heart disease associated pulmonary vascular disease. (Liu et al., 2016)	2016	c.-33A>G	CHD-PAH
		c.-212insC	CHD-PAH
		c.-310 A>G	CHD-PAH
		c.79T>G, p.S27A	CHD-PAH
		c.145A>G p.S49G	CHD-PAH
		c.276A>C p.Q92H	CHD-PAH
		c.344T>G p.F115C	CHD-PAH
		c.383C>T p.T128I	CHD-PAH
		c.180_182del p.61Sdel	CHD-PAH
		c.529+13A>C	CHD-PAH

		c.536G>A p.R179H	CHD-PAH
		c.711G>A p.R237H	CHD-PAH
		c.907C>T p.R303C	CHD-PAH
		c.1042G>A p.V348I (7)	CHD-PAH
		c.1196C>G p.S399X	CHD-PAH
		c.1310A>G p.Q437R	CHD-PAH
		c.1310A>G p.Q437R	CHD-PAH
		c.2495C>G p.S832C	CHD-PAH
A novel ***BMPR2*** gene mutation associated with exercise-induced pulmonary hypertension in septal defects. (Möller et al., 2010)	2010	Y589C 2	CHD-PAH - unrepaired VSD
		S775N	CHD-PAH - closed VSD
Sequencing of mutations in the serine/threonine kinase domain of the bone morphogenetic protein receptor type 2 gene causing pulmonary arterial hypertension. (Mutlu et al., 2016)	2016	p.C347Y	IPAH
Hemodynamic and genetic analysis in children with idiopathic, heritable, and congenital heart disease associated pulmonary arterial hypertension. (Pfarr et al., 2013)	2004	c.419-?_621 + ?del	IPAH
		c.1297C > T (p.Q433X)	IPAH
		c.1472 G > A (p.R491Q)	IPAH
		c.2668DelA (p.R890GfsX6) *	IPAH
		c.419-10 T > C *#	IPAH
		c.1-?_76 + ?del	CHD-PAH
BMPR2 mutations in pulmonary arterial hypertension with congenital heart disease. (Roberts et al., 2004)	2004	p.Q42R c.125A>G	CHD-PAH -AVC-C
		p.T102A c.304A>G	CHD-PAH -AVC-C
		p.S107P c.319T>C	CHD-PAH -AVC-C
		p.M186V c.556A>G	CHD-PAH ASD/PDA
		p.E503D c.1509A>C	CHD-PAH ASD/PDA/PAPVR
		p.G47N c.140G>A	CHD-PAH AW and VSD
Improvement of pulmonary arterial hypertension following medication and shunt closure in a ***BMPR2*** mutation carrier with atrial septal defect. (Suzuki et al., 2017)	2017	c.535_547delCGTAAACAAGGTCinsATG	CHD-PAH ASD

The Efficacy of a Genetic Analysis of the ***BMPR2*** Gene in a Patient with Severe Pulmonary Arterial Hypertension and an Atrial Septal Defect Treated with Bilateral Lung Transplantation. (Tatebe et al., 2017)	2017	c.2474A>G p.Tyr825Cys	CHD-PAH ASD
Clinical characterization of pediatric pulmonary hypertension: complex presentation and diagnosis. (van Loon et al., 2009)	2009	N/A	IPAH (2) & FPAH (1)
Rare variant analysis of 4241 pulmonary arterial hypertension cases from an international consortium implicates FBLN2, PDGFD, and rare *de novo* variants in PAH. (Zhu et al., 2021)	2021	N/A	(209) IPAH (108) FPAH (13) APAH

4.4. Limitations

The findings of this study must be interpreted in the context of some limitations:

1. WES approaches restrict the discovery to coding regions of the genome, which consist of 1-2% of the genome. Thus, this study did not detect variants in non-coding intronic, regulating or deep intronic variants that may be relevant to disease.
2. Trio WES is not able of detecting mosaicism in the proband and parents, especially if the level of mosaicism is low in blood-derived DNA.
3. WES is not suitable for the detection of complex structured variants (inversion, translocations).
4. Phenotypic heterogeneity. The same genetic variant may lead to different phenotypes, particularly for rare diseases having variable expression and incomplete penetrance.
5. Studies focusing on a single trio have limited power to detect rare variants, or to generalise findings to an affected population.
6. The approach utilised in this study does not account for environmental, epigenetic or polygenic factors which have been implicated in CHD.

4.5. Future work

Future work in this area of genetic research holds substantial promise for further elucidating the intricate mechanisms underlying congenital heart disease (CHD) and pulmonary arterial hypertension (PAH), particularly concerning the role of *BMPR2* variants. First and foremost, expanding the sample size to include more patients with similar phenotypic manifestations could provide a more comprehensive understanding of the prevalence and penetrance of the identified *de novo* variant in *BMPR2* among individuals with CHD and its association with PAH. The gene panel curated for the purpose of this study could be further utilised for population screening of patients diagnosed with CHD and PAH and/or general screening of samples within the biobank and/or in patients having suffered sudden cardiac death. Further future work could

potentially include conducting functional studies to elucidate the molecular mechanisms by which this variant contributes to the pathogenesis of both CHD and PAH. The identified variant could also be used for population screening in diagnosed CHD patients. Moreover, exploring potential therapeutic targets based on the genetic pathways implicated by the *BMPR2* variant could pave the way for personalized treatment strategies tailored to individuals with this genetic predisposition. Additionally, longitudinal studies tracking the clinical progression of CHD and PAH in individuals harbouring the *BMPR2* variant could offer insights into disease trajectory and inform prognostic considerations. Collaborative efforts integrating genomic data with clinical outcomes across diverse populations could also enhance our understanding of genotype-phenotype correlations and facilitate the development of precision medicine approaches for managing patients with CHD and associated PAH. Finally, advancements in genetic sequencing technologies, such as whole-genome sequencing and single-cell sequencing, may further refine our understanding of the genetic architecture underlying these complex diseases, opening avenues for novel therapeutic interventions and improved patient care.

Chapter 5. References

Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T., & McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature, 491(7422), 56-65. 10.1038/nature11632

Abu-Harb, M., Hey, E., & Wren, C. (1994). Death in infancy from unrecognised congenital heart disease. *Archives of Disease in Childhood, 71*(1), 3-7. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1029901/

ACMG recommendations for standards for interpretation of sequence variations. (2000). *Genetics in Medicine, 2*(5), 302-303. 10.1097/00125817-200009000-00009

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., & Sunyaev, S. R. (2010a). A method and server for predicting damaging missense mutations. *Nature Methods, 7*(4), 248-249. 10.1038/nmeth0410-248

Agilent Technologies. (2015). SureSelect Human All Exon V6 - Bridging the Gap, USA, Retrieved from https://www.agilent.com/cs/library/datasheets/public/SureSelect%20V6%20DataSheet%205991-5572EN.pdf

Ahmad, S., Gromiha, M. M., & Sarai, A. (2004). Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics (Oxford, England), 20*(4), 477-486. 10.1093/bioinformatics/btg432

Al-Numair, N. S., & Martin, A. C. R. (2013). The SAAP pipeline and database: tools to analyze the impact and predict the pathogenicity of mutations. *BMC Genomics, 14 Suppl 3*(Suppl 3), S4. 10.1186/1471-2164-14-S3-S4

Alvarez-Curto, E., Inoue, A., Jenkins, L., Raihan, S. Z., Prihandoko, R., Tobin, A. B., & Milligan, G. (2016a). Targeted Elimination of G Proteins and Arrestins Defines Their Specific Contributions to Both Intensity and Duration of G Protein-coupled Receptor Signaling. *The Journal of Biological Chemistry, 291*(53), 27147-27159. 10.1074/jbc.M116.754887

An integrated encyclopedia of DNA elements in the human genome. (2012). *Nature, 489*(7414), 57-74. 10.1038/nature11247

Anderson, R. H., McCarthy, K., & Cook, A. C. (2001). Double outlet right ventricle. *Cardiology in the Young, 11*(3), 329-344. 10.1017/S1047951101000373

Anderson, R., McCarthy, K., & Cook, A. (2001). Double outlet right ventricle. Cardiology in the Young, 11(3), 329-344. doi:10.1017/S1047951101000373

Aselton, P., Jick, H., Milunsky, A., Hunter, J. R., & Stergachis, A. (1985). First-trimester drug use and congenital disorders. *Obstetrics and Gynecology, 65*(4), 451-455.

Ashley, E. A., Butte, A. J., Wheeler, M. T., Chen, R., Klein, T. E., Dewey, F. E., Dudley, J. T., Ormond, K. E., Pavlovic, A., Morgan, A. A., Pushkarev, D., Neff, N. F., Hudgins, L., Gong, L., Hodges, L. M., Berlin, D. S., Thorn, C. F., Sangkuhl, K., Hebert, J. M., Altman, R. B. (2010). Clinical assessment incorporating a personal genome. *Lancet (London, England), 375*(9725), 1525-1535. 10.1016/S0140-6736(10)60452-7

Atkinson, C., Stewart, S., Upton, P. D., Machado, R., Thomson, J. R., Trembath, R. C., & Morrell, N. W. (2002). Primary pulmonary hypertension is associated with reduced pulmonary vascular expression of type II bone morphogenetic protein receptor. *Circulation, 105*(14), 1672-1678. 10.1161/01.cir.0000012754.72951.3d

Austin, E. D., Ma, L., LeDuc, C., Berman Rosenzweig, E., Borczuk, A., Phillips, J. A., Palomero, T., Sumazin, P., Kim, H. R., Talati, M. H., West, J., Loyd, J. E., & Chung, W. K. (2012). Whole exome sequencing to identify a novel gene (caveolin-1) associated with human pulmonary arterial hypertension. *Circulation. Cardiovascular Genetics, 5*(3), 336-343. 10.1161/CIRCGENETICS.111.961888

B.G Johansson. (1972). Agarose Gel Electrophoresis Scandinavian Journal of Clinical and Laboratory Investigation.10.3109/00365517209102747

Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. A., & Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews. Genetics, 12*(11), 745-755. 10.1038/nrg3031

Barrett, C., & Richens, A. (2003). Epilepsy and pregnancy: Report of an Epilepsy Research Foundation Workshop. *Epilepsy Research, 52*(3), 147-187. 10.1016/s0920-1211(02)00237-1

Basel-Salmon, L., Ruhrman-Shahar, N., Orenstein, N., Goldberg, Y., Gonzaga-Jauregui, C., Shuldiner, A. R., Sukenik-Halevy, R., Maya, I., Magal, N., Hagari, O., Azulay, N., Lidzbarsky, G. A., & Bazak, L. (2021). When phenotype does not match genotype: importance of "real-time" refining of phenotypic information for exome data interpretation. *Genetics in Medicine: Official Journal of the American College of Medical Genetics, 23*(1), 215-221. 10.1038/s41436-020-00938-5

Battaglia, A., Hoyme, H. E., Dallapiccola, B., Zackai, E., Hudgins, L., McDonald-McGinn, D., Bahi-Buisson, N., Romano, C., Williams, C. A., Brailey, L. L., Zuberi, S. M., & Carey, J. C. (2008). Further Delineation of Deletion 1p36 Syndrome in 60 Patients: A Recognizable Phenotype and Common Cause of Developmental Delay and Mental Retardation. *Pediatrics, 121*(2), 404-410. 10.1542/peds.2007-0929

Baumgartner, H., Bonhoeffer, P., De Groot, N. M. S., de Haan, F., Deanfield, J. E., Galie, N., Gatzoulis, M. A., Gohlke-Baerwolf, C., Kaemmerer, H., Kilner, P., Meijboom, F., Mulder, B. J. M., Oechslin, E., Oliver, J. M., Serraf, A., Szatmari, A., Thaulow, E., Vouhe, P. R., & Walma, E. (2010). ESC Guidelines for the management of grown-up congenital heart disease. *European Heart Journal, 31*(23), 2915-2957. 10.1093/eurheartj/ehq249

Becerra, J. E., Khoury, M. J., Cordero, J. F., & Erickson, J. D. (1990). Diabetes mellitus during pregnancy and the risks for specific birth defects: a population-based case-control study. *Pediatrics, 85*(1), 1-9.

Beppu, H., Kawabata, M., Hamamoto, T., Chytil, A., Minowa, O., Noda, T., & Miyazono, K. (2000). BMP type II receptor is required for gastrulation and early development of mouse embryos. *Developmental Biology, 221*(1), 249-258. 10.1006/dbio.2000.9670

Bercovich, D., Elimelech, A., Zlotogora, J., Korem, S., Yardeni, T., Gal, N., Goldstein, N., Vilensky, B., Segev, R., Avraham, S., Loewenthal, R., Schwartz, G., & Anikster, Y. (2008). Genotype–phenotype correlations analysis of mutations in the phenylalanine hydroxylase (PAH) gene. *Journal of Human Genetics, 53*(5), 407-418. 10.1007/s10038-008-0264-4

Bhattacharya, R., Rose, P. W., Burley, S. K., & Prlić, A. (2017). Impact of genetic variation on three dimensional structure and function of proteins. *PloS One, 12*(3), e0171355. 10.1371/journal.pone.0171355

Blalock, A., & Taussig, H. B. (1984). Landmark article May 19, 1945: The surgical treatment of malformations of the heart in which there is pulmonary stenosis or pulmonary atresia. By Alfred Blalock and Helen B. Taussig. *Jama, 251*(16), 2123-2138. 10.1001/jama.251.16.2123

Blue, G. M., Kirk, E. P., Giannoulatou, E., Sholler, G. F., Dunwoodie, S. L., Harvey, R. P., & Winlaw, D. S. (2017). Advances in the Genetics of Congenital Heart Disease: A Clinician's Guide. *Journal of the American College of Cardiology, 69*(7), 859-870. 10.1016/j.jacc.2016.11.060

Bodmer, R. (1993). The gene tinman is required for specification of the heart and visceral muscles in Drosophila. *Development (Cambridge, England), 118*(3), 719-729. 10.1242/dev.118.3.719

Bonachea, E. M., Zender, G., White, P., Corsmeier, D., Newsom, D., Fitzgerald-Butt, S., Garg, V., & McBride, K. L. (2014). Use of a targeted, combinatorial next-generation sequencing approach for the study of bicuspid aortic valve. *BMC Medical Genomics, 7*, 56. 10.1186/1755-8794-7-56

Botto, L. D., Lin, A. E., Riehle-Colarusso, T., Malik, S., & Correa, A. (2007). Seeking causes: Classifying and evaluating congenital heart defects in etiologic studies. *Birth Defects Research. Part A, Clinical and Molecular Teratology, 79*(10), 714-727. 10.1002/bdra.20403

Botto, L. D., Lynberg, M. C., & Erickson, J. D. (2001). Congenital heart defects, maternal febrile illness, and multivitamin use: a population-based study. *Epidemiology (Cambridge, Mass.), 12*(5), 485-490. 10.1097/00001648-200109000-00004

Botto, L. D., Mulinare, J., & Erickson, J. D. (2000). Occurrence of congenital heart defects in relation to maternal mulitivitamin use. *American Journal of Epidemiology, 151*(9), 878-884. 10.1093/oxfordjournals.aje.a010291

Bouma, B. J., & Mulder, B. J. M. (2017). Changing Landscape of Congenital Heart Disease. *Circulation Research, 120*(6), 908-922. 10.1161/CIRCRESAHA.116.309302

Bower, C., Stanley, F., Connell, A. F., Gent, C. R., & Massey, M. S. (1992). Birth defects in the infants of aboriginal and non-aboriginal mothers with diabetes in Western Australia. *The Medical Journal of Australia, 156*(8), 520-524. 10.5694/j.1326-5377.1992.tb121410.x

Bracken, M. B. (1986). Drug use in pregnancy and congenital heart disease in offspring. *The New England Journal of Medicine, 314*(17), 1120. 10.1056/NEJM198604243141717

Bradley, E. A., & Zaidi, A. N. (2020). Atrial Septal Defect. *Cardiology Clinics, 38*(3), 317-324. 10.1016/j.ccl.2020.04.001

Bradley, T. J., Karamlou, T., Kulik, A., Mitrovic, B., Vigneswaran, T., Jaffer, S., Glasgow, P. D., Williams, W. G., Van Arsdell, G. S., & McCrindle, B. W. (2007). Determinants of repair type, reintervention, and mortality in 393 children with

double-outlet right ventricle. *The Journal of Thoracic and Cardiovascular Surgery, 134*(4), 967-973.e6. 10.1016/j.jtcvs.2007.05.061

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5-32.

Brida, M., Chessa, M., Celermajer, D., Li, W., Geva, T., Khairy, P., Griselli, M., Baumgartner, H., & Gatzoulis, M. A. (2022). Atrial septal defect in adulthood: a new paradigm for congenital heart disease. *European Heart Journal, 43*(28), 2660-2671. 10.1093/eurheartj/ehab646

Buskens, E., Grobbee, D. E., Frohn-Mulder, I. M., Wladimiroff, J. W., & Hess, J. (1995). Aspects of the aetiology of congenital heart disease. *European Heart Journal, 16*(5), 584-587. 10.1093/oxfordjournals.eurheartj.a060960

Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P. L., & Casadio, R. (2009). Functional annotations improve the predictive score of human disease-related mutations in proteins. *Human Mutation, 30*(8), 1237-1244. 10.1002/humu.21047

Caldararu, O., Mehra, R., Blundell, T. L., & Kepp, K. P. (2020). Systematic Investigation of the Data Set Dependency of Protein Stability Predictors. *Journal of Chemical Information and Modeling*, *60*(10), 4772-4784. 10.1021/acs.jcim.0c00591

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics, 10*, 421. 10.1186/1471-2105-10-421

Canfield, M. A., Honein, M. A., Yuskiv, N., Xing, J., Mai, C. T., Collins, J. S., Devine, O., Petrini, J., Ramadhani, T. A., Hobbs, C. A., & Kirby, R. S. (2006). National estimates and race/ethnic-specific variation of selected birth defects in the United

States, 1999-2001. *Birth Defects Research. Part A, Clinical and Molecular Teratology, 76*(11), 747-756. 10.1002/bdra.20294

Capriotti, E., Fariselli, P., & Casadio, R. (2005). I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Research, 33*(Web Server issue), 306. 10.1093/nar/gki375

Chang, K. T., Guo, J., di Ronza, A., & Sardiello, M. (2018). Aminode: Identification of Evolutionary Constraints in the Human Proteome. *Scientific Reports, 8,* 1357. 10.1038/s41598-018-19744-w

Chaikuad, A., Thangaratnarajah, C., von Delft, F., & Bullock, A. N. (2019). Structural consequences of BMPR2 kinase domain mutations causing pulmonary arterial hypertension. *Scientific Reports, 9*(1), 18351. 10.1038/s41598-019-54830-7

Chaouat, A., Coulet, F., Favre, C., Simonneau, G., Weitzenblum, E., Soubrier, F., & Humbert, M. (2004). Endoglin germline mutation in a patient with hereditary haemorrhagic telangiectasia and dexfenfluramine associated pulmonary arterial hypertension. *Thorax, 59*(5), 446-448. 10.1136/thx.2003.11890

Chen, R., Im, H., & Snyder, M. (2015). Whole-Exome Enrichment with the Agilent SureSelect Human All Exon Platform. *Cold Spring Harbor Protocols, 2015*(7), pdb.prot083659. 10.1101/pdb.prot083659

Cirulli, E. T., & Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews. Genetics, 11*(6), 415-425. 10.1038/nrg2779

Clarren, S. K., & Smith, D. W. (1978). The fetal alcohol syndrome. *The New England Journal of Medicine, 298*(19), 1063-1067. 10.1056/NEJM197805112981906

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research, 38*(6), 1767-1771. 10.1093/nar/gkp1137

Collins, R. L., Brand, H., Karczewski, K. J., Zhao, X., Alföldi, J., Francioli, L. C., Khera, A. V., Lowther, C., Gauthier, L. D., Wang, H., Watts, N. A., Solomonson, M., O'Donnell-Luria, A., Baumann, A., Munshi, R., Walker, M., Whelan, C. W., Huang, Y., Brookings, T., Genome Aggregation, D. C. (2020). A structural variation reference for medical and population genetics. *Nature, 581*(7809), 444-451. 10.1038/s41586-020-2287-8

Comi, A. M., Sahin, M., Hammill, A., Kaplan, E. H., Juhász, C., North, P., Ball, K. L., Levin, A. V., Cohen, B., Morris, J., Lo, W., & Roach, E. S. (2016). Leveraging a Sturge-Weber Gene Discovery: An Agenda for Future Research. *Pediatric Neurology, 58,* 12-24. 10.1016/j.pediatrneurol.2015.11.009

Connelly, M. S., Webb, G. D., Somerville, J., Warnes, C. A., Perloff, J. K., Liberthson, R. R., Puga, F. J., Collins-Nakai, R. L., Williams, W. G., Mercier, L. A., Huckell, V. F., Finley, J. P., & McKay, R. (1998). Canadian Consensus Conference on Adult Congenital Heart Disease 1996. *The Canadian Journal of Cardiology, 14*(3), 395-452.

Cooper, G. M., Brudno, M., Green, E. D., Batzoglou, S., & Sidow, A. (2003). Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. *Genome Research, 13*(5), 813-820. 10.1101/gr.1064503

Cooper, G. M., Stone, E. A., Asimenos, G., Green, E. D., Batzoglou, S., & Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research, 15*(7), 901-913. 10.1101/gr.3577405

Corpet, F. (1988). Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Research, 16*(22), 10881-10890. 10.1093/nar/16.22.10881

Costain, G., Silversides, C. K., & Bassett, A. S. (2016). The importance of copy number variation in congenital heart disease. *Npj Genomic Medicine, 1*(1), 16031. 10.1038/npjgenmed.2016.31

Cousins, L. (1991). Etiology and prevention of congenital anomalies among infants of overt diabetic women. *Clinical Obstetrics and Gynecology, 34*(3), 481-493.

Cowan, J. R., & Ware, S. M. (2015). Genetics and Genetic Testing in Congenital Heart Disease. *Clinics in Perinatology, 42*(2), 373-393. 10.1016/j.clp.2015.02.009

Cummings, B. B., Karczewski, K. J., Kosmicki, J. A., Seaby, E. G., Watts, N. A., Singer-Berk, M., Mudge, J. M., Karjalainen, J., Satterstrom, F. K., O'Donnell-Luria, A. H., Poterba, T., Seed, C., Solomonson, M., Alföldi, J., Alföldi, J., Armean, I. M., Banks, E., Bergelson, L., Cibulskis, K., . . . Genome Aggregation, D. C. (2020). Transcript expression-aware annotation improves rare variant interpretation. *Nature, 581*(7809), 452-458. 10.1038/s41586-020-2329-2

Czeizel, A. E. (1998). Periconceptional folic acid containing multivitamin supplementation. *European Journal of Obstetrics, Gynecology, and Reproductive Biology, 78*(2), 151-161. 10.1016/s0301-2115(98)00061-x

Czeizel, A. E., Rockenbauer, M., Olsen, J., & Sørensen, H. T. (2000). Oral phenoxymethylpenicillin treatment during pregnancy. Results of a population-based Hungarian case-control study. *Archives of Gynecology and Obstetrics, 263*(4), 178-181. 10.1007/s004040050277

Czeizel, A. E., Rockenbauer, M., Sørensen, H. T., & Olsen, J. (2001). A population-based case-control teratologic study of ampicillin treatment during pregnancy. *American Journal of Obstetrics and Gynecology, 185*(1), 140-147. 10.1067/mob.2001.113907

Dai, W. S., Hsu, M. A., & Itri, L. M. (1989). Safety of pregnancy after discontinuation of isotretinoin. *Archives of Dermatology, 125*(3), 362-365.

Danesh, S. M., Villasenor, A., Chong, D., Soukup, C., & Cleaver, O. (2009). BMP and BMP receptor expression during murine organogenesis. *Gene Expression Patterns : GEP, 9*(5), 255-265. 10.1016/j.gep.2009.04.002

Das, S., Abecasis, G., & Fuchsberger, C. (2015). Minimac4: A next generation imputation tool for mega reference panels. Abstract 1278W. Paper presented at the *The 65th Annual Meeting of the American Society of Human Genetics,*

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., & Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Computational Biology*, *6*(12), e1001025. 10.1371/journal.pcbi.1001025

de Jesus Perez, V. A., Alastalo, T., Wu, J. C., Axelrod, J. D., Cooke, J. P., Amieva, M., & Rabinovitch, M. (2009). Bone morphogenetic protein 2 induces pulmonary angiogenesis via Wnt–β-catenin and Wnt–RhoA–Rac1 pathways. *The Journal of Cell Biology, 184*(1), 83-99. 10.1083/jcb.200806049

de Lima Morais, D. A., Fang, H., Rackham, O. J. L., Wilson, D., Pethica, R., Chothia, C., & Gough, J. (2011). SUPERFAMILY 1.75 including a domain-centric gene

ontology method. *Nucleic Acids Research, 39*(Database issue), 427. 10.1093/nar/gkq1130

Dehouck, Y., Grosfils, A., Folch, B., Gilis, D., Bogaerts, P., & Rooman, M. (2009). Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics (Oxford, England), 25*(19), 2537-2543. 10.1093/bioinformatics/btp445

Delorenzi, M., & Speed, T. (2002). An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. *Bioinformatics (Oxford, England), 18*(4), 617-625. 10.1093/bioinformatics/18.4.617

Dencker, B. B., Larsen, H., Jensen, E. S., Schønheyder, H. C., Nielsen, G. L., & Sørensen, H. T. (2002). Birth outcome of 1886 pregnancies after exposure to phenoxymethylpenicillin in utero. *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases, 8*(4), 196-201. 10.1046/j.1469-0691.2002.00368.x

Deng, Z., Morse, J. H., Slager, S. L., Cuervo, N., Moore, K. J., Venetos, G., Kalachikov, S., Cayanis, E., Fischer, S. G., Barst, R. J., Hodge, S. E., & Knowles, J. A. (2000). Familial Primary Pulmonary Hypertension (Gene PPH1) Is Caused by Mutations in the Bone Morphogenetic Protein Receptor–II Gene. *American Journal of Human Genetics, 67*(3), 737-744. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1287532/

Diab, N. S., Barish, S., Dong, W., Zhao, S., Allington, G., Yu, X., Kahle, K. T., Brueckner, M., & Jin, S. C. (2021). Molecular Genetics and Complex Inheritance of Congenital Heart Disease. *Genes, 12*(7)10.3390/genes12071020

Digilio, M. C., Angioni, A., De Santis, M., Lombardo, A., Giannotti, A., Dallapiccola, B., & Marino, B. (2003). Spectrum of clinical variability in familial deletion 22q11.2: from full manifestation to extremely mild clinical anomalies. *Clinical Genetics, 63*(4), 308-313. 10.1034/j.1399-0004.2003.00049.x

Dolbec, K., & Mick, N. W. (2011). Congenital Heart Disease. *Emergency Medicine Clinics of North America, 29*(4), 811-827. 10.1016/j.emc.2011.08.005

Dolk Helen, Loane Maria, Garne Ester, & European Surveillance of Congenital Anomalies (EUROCAT) Working Group. (2011). Congenital Heart Defects in Europe. *Circulation, 123*(8), 841-849. 10.1161/CIRCULATIONAHA.110.958405

Dyer, L. A., Pi, X., & Patterson, C. (2014). The role of BMPs in endothelial cell function and dysfunction. *Trends in Endocrinology and Metabolism: TEM, 25*(9), 472-480. 10.1016/j.tem.2014.05.003

Ebadi, A., Spicer, D. E., Backer, C. L., Fricker, F. J., & Anderson, R. H. (2017). Double-outlet right ventricle revisited. *The Journal of Thoracic and Cardiovascular Surgery, 154*(2), 598-604. 10.1016/j.jtcvs.2017.03.049

Eddy, S. R. (2009). A new generation of homology search tools based on probabilistic inference. *Genome Informatics. International Conference on Genome Informatics, 23*(1), 205-211.

Edwards, J. J., & Gelb, B. D. (2016). Genetics of congenital heart disease. *Current Opinion in Cardiology, 31*(3), 235-241. 10.1097/HCO.0000000000000274

Edwards, M. J. (1998). Apoptosis, the heat shock response, hyperthermia, birth defects, disease and cancer. Where are the common links? *Cell Stress & Chaperones, 3*(4), 213-220. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC312966/

Einhorn, E., Kamshov, A., Lev, O., Einhorn, M., Paz-Yaacov, N., Shami, K., & Gross, S. (2019). Implementation of gene-specific ClinGen variant classification recommendations using artificial intelligence: Frequency thresholds.

Einhorn, Y. K., Paz-Yaacov, N., Einhorn, M., Harrison, S., & Yaron, Y. (2019). Reinterpretation of Sequence Variants Using Artificial Intelligence — Results of 2 Benchmarking Experiments. 2019. https://www.genoox.com/wp-content/uploads/2022/06/ACMG2019_Reinterpretation-Benchmarking_v3.pdf

Einhorn, Y., Lev, O., Einhorn, M., Trabelsi, A., Paz-Yaacov, N., & Gross, S. J. (2018). Benchmarking an automated Variant Classification Engine (aVCE) algorithm using ClinVar: Results of a time-capsule experiment. Paper presented at the *ACMG Annual Clinical Genetics Meeting*,

Ekure, E. N., Adeyemo, A., Liu, H., Sokunbi, O., Kalu, N., Martinez, A. F., Owosela, B., Tekendo-Ngongang, C., Addissie, Y. A., Olusegun-Joseph, A., Ikebudu, D., Berger, S. I., Muenke, M., Han, Z., & Kruszka, P. (2021). Exome Sequencing and Congenital Heart Disease in Sub-Saharan Africa. *Circulation. Genomic and Precision Medicine, 14*(1), e003108. 10.1161/CIRCGEN.120.003108

Ellesøe, S. G., Workman, C. T., Bouvagnet, P., Loffredo, C. A., McBride, K. L., Hinton, R. B., van Engelen, K., Gertsen, E. C., Mulder, B. J. M., Postma, A. V., Anderson, R. H., Hjortdal, V. E., Brunak, S., & Larsen, L. A. (2018). Familial co-occurrence of congenital heart defects follows distinct patterns. *European Heart Journal, 39*(12), 1015-1022. 10.1093/eurheartj/ehx314

ENCODE project consotrium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature, 489*(7414), 57-74. 10.1038/nature11247

Erikssen, G., Liestøl, K., Seem, E., Birkeland, S., Saatvedt, K. J., Hoel, T. N., Døhlen, G., Skulstad, H., Svennevig, J. L., Thaulow, E., & Lindberg, H. L. (2015). Achievements in Congenital Heart Defect Surgery. *Circulation, 131*(4), 337-346. 10.1161/CIRCULATIONAHA.114.012033

Eriksson, U. J., & Simán, C. M. (1996). Pregnant diabetic rats fed the antioxidant butylated hydroxytoluene show decreased occurrence of malformations in offspring. *Diabetes, 45*(11), 1497-1502. 10.2337/diab.45.11.1497

Espinoza-Lewis, R. A., & Wang, D. (2012). MicroRNAs in heart development. *Current Topics in Developmental Biology, 100,* 279-317. 10.1016/B978-0-12-387786-4.00009-9

Evans, E. (2004). *Domain-driven design: tackling complexity in the heart of software*. Addison-Wesley Professional.

Evans, J. D. W., Girerd, B., Montani, D., Wang, X., Galiè, N., Austin, E. D., Elliott, G., Asano, K., Grünig, E., Yan, Y., Jing, Z., Manes, A., Palazzini, M., Wheeler, L. A., Nakayama, I., Satoh, T., Eichstaedt, C., Hinderhofer, K., Wolf, M., . . . Morrell, N. W. (2016). BMPR2 mutations and survival in pulmonary arterial hypertension: an individual participant data meta-analysis. *The Lancet. Respiratory Medicine, 4*(2), 129-137. 10.1016/S2213-2600(15)00544-5

Ewing, B., & Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research, 8*(3), 186-194.

Fahed, A. C., Gelb, B. D., Seidman, J. G., & Seidman, C. E. (2013). Genetics of Congenital Heart Disease. *Circulation Research, 112*(4), 707-720. 10.1161/CIRCRESAHA.112.300853

Feng, B. (2017). PERCH: A Unified Framework for Disease Gene Prioritization. *Human Mutation, 38*(3), 243-251. 10.1002/humu.23158

Ferrara, A., Kahn, H. S., Quesenberry, C. P., Riley, C., & Hedderson, M. M. (2004). An increase in the incidence of gestational diabetes mellitus: Northern California, 1991-2000. *Obstetrics and Gynecology, 103*(3), 526-533. 10.1097/01.AOG.0000113623.18286.20

Ferreira, F., Azevedo, L., Neiva, R., Sousa, C., Fonseca, H., Marcão, A., Rocha, H., Carmona, C., Ramos, S., Bandeira, A., Martins, E., Campos, T., Rodrigues, E., Garcia, P., Diogo, L., Ferreira, A. C., Sequeira, S., Silva, F., Rodrigues, L., Vilarinho, L. (2021). Phenylketonuria in Portugal: Genotype–phenotype correlations using molecular, biochemical, and haplotypic analyses. *Molecular Genetics & Genomic Medicine, 9*(3), e1559. 10.1002/mgg3.1559

Finkenzeller, G., Hager, S., & Stark, G. B. (2012). Effects of bone morphogenetic protein 2 on human umbilical vein endothelial cells. *Microvascular Research, 84*(1), 81-85. 10.1016/j.mvr.2012.03.010

Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., Chang, H., Dosztányi, Z., El-Gebali, S., Fraser, M., Gough, J., Haft, D., Holliday, G. L., Huang, H., Huang, X., Letunic, I., Lopez, R., Lu, S., Marchler-Bauer, A., . . . Mitchell, A. L. (2017). InterPro in 2017—beyond protein family and domain annotations. *Nucleic Acids Research, 45*(Database issue), D190-D199. 10.1093/nar/gkw1107

Finn, R. D., Clements, J., Arndt, W., Miller, B. L., Wheeler, T. J., Schreiber, F., Bateman, A., & Eddy, S. R. (2015). HMMER web server: 2015 update. *Nucleic Acids Research, 43*(Web Server issue), W30-W38. 10.1093/nar/gkv397

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J., & Bateman, A. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research, 44*(Database issue), D279-D285. 10.1093/nar/gkv1344

Fogel, G. (2006). Conference report - 2005 IEEE symposium on computational intelligence in bioinformatics and computational biology (IEEE CIBCB 2005). *Computational Intelligence Magazine, IEEE, 1*, 43-44. 10.1109/MCI.2006.1626495

Frappier, V., & Najmanovich, R. J. (2014). A coarse-grained elastic network atom contact model and its use in the simulation of protein dynamics and the prediction of the effect of mutations. *PLoS Computational Biology, 10*(4), e1003569. 10.1371/journal.pcbi.1003569

Fu, F., Li, R., Li, Y., Nie, Z. -., Lei, T., Wang, D., Yang, X., Han, J., Pan, M., Zhen, L., Ou, Y., Li, J., Li, F. -., Jing, X., Li, D., & Liao, C. (2018). Whole exome sequencing as a diagnostic adjunct to clinical testing in fetuses with structural abnormalities. *Ultrasound in Obstetrics & Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology, 51*(4), 493-502. 10.1002/uog.18915

Gao, M., Zhou, H., & Skolnick, J. (2015). Insights into Disease-Associated Mutations in the Human Proteome through Protein Structural Analysis. *Structure (London, England: 1993), 23*(7), 1362-1369. 10.1016/j.str.2015.03.028

Garg, V., Kathiriya, I. S., Barnes, R., Schluterman, M. K., King, I. N., Butler, C. A., Rothrock, C. R., Eapen, R. S., Hirayama-Yamada, K., Joo, K., Matsuoka, R.,

Cohen, J. C., & Srivastava, D. (2003). GATA4 mutations cause human congenital heart defects and reveal an interaction with TBX5. *Nature, 424*(6947), 443-447. 10.1038/nature01827

Garimella, R., Kacena, M. A., Tague, S. E., Wang, J., Horowitz, M. C., & Anderson, H. C. (2007). Expression of bone morphogenetic proteins and their receptors in the bone marrow megakaryocytes of GATA-1(low) mice: a possible role in osteosclerosis. *The Journal of Histochemistry and Cytochemistry: Official Journal of the Histochemistry Society, 55*(7), 745-752. 10.1369/jhc.6A7164.2007

Gaudet P, Livstone MS, Lewis SE, Thomas PD. Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. Brief Bioinform. 2011 Sep;12(5):449-62. doi: 10.1093/bib/bbr042. Epub 2011 Aug 27. PMID: 21873635; PMCID: PMC3178059.

Gedikbasi, A., Oztarhan, K., Gul, A., Sargin, A., & Ceylan, Y. (2008). Diagnosis and prognosis in double-outlet right ventricle. *American Journal of Perinatology, 25*(7), 427-434. 10.1055/s-0028-1083840

Geiger, J. M., Baudin, M., & Saurat, J. H. (1994). Teratogenic risk with etretinate and acitretin treatment. *Dermatology (Basel, Switzerland), 189*(2), 109-116. 10.1159/000246811

Gelb, B. D. (2015). History of Our Understanding of the Causes of Congenital Heart Disease. *Circulation. Cardiovascular Genetics, 8*(3), 529-536. 10.1161/CIRCGENETICS.115.001058

Gelb, B. D., & Chung, W. K. (2014). Complex Genetics and the Etiology of Human Congenital Heart Disease. *Cold Spring Harbor Perspectives in Medicine, 4*(7), a013953. 10.1101/cshperspect.a013953

Gharahkhani, P., O'Leary, C. A., Kyaw-Tanner, M., Sturm, R. A., & Duffy, D. L. (2011). A non-synonymous mutation in the canine Pkd1 gene is associated with autosomal dominant polycystic kidney disease in Bull Terriers. *PloS One, 6*(7), e22455. 10.1371/journal.pone.0022455

Ghigna, M., Guignabert, C., Montani, D., Girerd, B., Jaïs, X., Savale, L., Hervé, P., Montpréville, V. T. d., Mercier, O., Sitbon, O., Soubrier, F., Fadel, E., Simonneau, G., Humbert, M., & Dorfmüller, P. (2016). BMPR2 mutation status influences bronchial vascular changes in pulmonary arterial hypertension. *European Respiratory Journal, 48*(6), 1668-1681. 10.1183/13993003.00464-2016

Gillette, P. C. (1998). Genetic and environmental risk factors of major cardiovascular malformations: The baltimore–washington infant study: 1981–1989 edited by Charlotte Ferencz, Adolfo Correa–Villasenor, Christopher A. Loffredo, P. David Wilson, Futura Publishing Company, Inc., Armonk, N.Y. (1998) 463 pages, illustrated, $95.00 ISBN: 1044–4157. *Clinical Cardiology, 21*(11), 867-868. https://doi.org/10.1002/clc.4960211122

Gillum, R. F. (1994). Epidemiology of congenital heart disease in the United States. *American Heart Journal, 127*(4 Pt 1), 919-927. 10.1016/0002-8703(94)90562-2

Girerd, B., Montani, D., Jaïs, X., Eyries, M., Yaici, A., Sztrymf, B., Savale, L., Parent, F., Coulet, F., Godinas, L., Lau, E. M., Tamura, Y., Sitbon, O., Soubrier, F., Simonneau, G., & Humbert, M. (2016). Genetic counselling in a national referral centre for pulmonary hypertension. *The European Respiratory Journal, 47*(2), 541-552. 10.1183/13993003.00717-2015

Glessner, J. T., Bick, A. G., Ito, K., Homsy, J., Rodriguez-Murillo, L., Fromer, M., Mazaika, E., Vardarajan, B., Italia, M., Leipzig, J., DePalma, S. R., Golhar, R., Sanders, S. J., Yamrom, B., Ronemus, M., Iossifov, I., Willsey, A. J., State, M. W., Kaltman, J. R., . . . Chung, W. K. (2014). Increased frequency of *de novo* copy number variants in congenital heart disease by integrative analysis of single nucleotide polymorphism array and exome sequence data. *Circulation Research, 115*(10), 884-896. 10.1161/CIRCRESAHA.115.304458

Go, A. S., Mozaffarian, D., Roger, V. L., Benjamin, E. J., Berry, J. D., Borden, W. B., Bravata, D. M., Dai, S., Ford, E. S., Fox, C. S., Franco, S., Fullerton, H. J., Gillespie, C., Hailpern, S. M., Heit, J. A., Howard, V. J., Huffman, M. D., Kissela, B. M., Kittner, S. J., . . . Turner, M. B. (2013). Heart Disease and Stroke Statistics—2013 Update. *Circulation, 127*(1), e6-e245. 10.1161/CIR.0b013e31828124ad

Granados-Riveron, J. T., Pope, M., Bu'lock, F. A., Thornborough, C., Eason, J., Setchfield, K., Ketley, A., Kirk, E. P., Fatkin, D., Feneley, M. P., Harvey, R. P., & Brook, J. D. (2012). Combined mutation screening of NKX2-5, GATA4, and TBX5 in congenital heart disease: multiple heterozygosity and novel mutations. *Congenital Heart Disease, 7*(2), 151-159. 10.1111/j.1747-0803.2011.00573.x

Grant, B. J., Rodrigues, A. P. C., ElSawy, K. M., McCammon, J. A., & Caves, L. S. D. (2006). Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics (Oxford, England), 22*(21), 2695-2696. 10.1093/bioinformatics/btl461

Graziani, F., & Delogu, A. B. (2016). Evaluation of Adults With Congenital Heart Disease. *World Journal for Pediatric & Congenital Heart Surgery, 7*(2), 185-191. 10.1177/2150135115623285

Grech, L., Mifsud, A., Caruana, M., & Carbonaro, F. (2017). A Case of Advanced Glaucoma with Increased Episcleral Venous Pressure in a 17-Year-Old with Eisenmenger Syndrome. *Case Reports in Ophthalmological Medicine, 2017*, 5808047. 10.1155/2017/5808047

Green, D. M., Zevon, M. A., Lowrie, G., Seigelstein, N., & Hall, B. (1991). Congenital Anomalies in Children of Patients Who Received Chemotherapy for Cancer in Childhood and Adolescence. *New England Journal of Medicine, 325*(3), 141-146. 10.1056/NEJM199107183250301

Greenway, S. C., Pereira, A. C., Lin, J. C., DePalma, S. R., Israel, S. J., Mesquita, S. M., Ergul, E., Conta, J. H., Korn, J. M., McCarroll, S. A., Gorham, J. M., Gabriel, S., Altshuler, D. M., de Lourdes Quintanilla-Dieck, M., Artunduaga, M. A., Eavey, R. D., Plenge, R. M., Shadick, N. A., Weinblatt, M. E., Seidman, C. E. (2009). *De novo* copy number variants identify new genes and loci in isolated sporadic tetralogy of Fallot. *Nature Genetics, 41*(8), 931-935. 10.1038/ng.415

Greulich, F., Rudat, C., & Kispert, A. (2011). Mechanisms of T-box gene function in the developing heart. *Cardiovascular Research, 91*(2), 212-222. 10.1093/cvr/cvr112

Grimm, D. G., Azencott, C., Aicheler, F., Gieraths, U., MacArthur, D. G., Samocha, K. E., Cooper, D. N., Stenson, P. D., Daly, M. J., Smoller, J. W., Duncan, L. E., & Borgwardt, K. M. (2015). The evaluation of tools used to predict the impact of

missense variants is hindered by two types of circularity. *Human Mutation, 36*(5), 513-523. 10.1002/humu.22768

Groselj, U., Tansek, M. Z., Kovac, J., Hovnik, T., Podkrajsek, K. T., & Battelino, T. (2012). Five novel mutations and two large deletions in a population analysis of the phenylalanine hydroxylase gene. *Molecular Genetics and Metabolism, 106*(2), 142-148. 10.1016/j.ymgme.2012.03.015

Guldberg, P., Mallmann, R., Henriksen, K. F., & Güttler, F. (1996). Phenylalanine hydroxylase deficiency in a population in Germany: mutational profile and nine novel mutations. *Human Mutation, 8*(3), 276-279. 10.1002/(SICI)1098-1004(1996)8:3<276::AID-HUMU14>3.0.CO;2-#

Günther, T., & Schmid, K. J. (2010). Deleterious amino acid polymorphisms in Arabidopsis thaliana and rice. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik, 121*(1), 157-168. 10.1007/s00122-010-1299-4

Guryev, V., Berezikov, E., Malik, R., Plasterk, R. H. A., & Cuppen, E. (2004). Single nucleotide polymorphisms associated with rat expressed sequences. *Genome Research, 14*(7), 1438-1443. 10.1101/gr.2154304

Hagay, Z. J., Weiss, Y., Zusman, I., Peled-Kamar, M., Reece, E. A., Eriksson, U. J., & Groner, Y. (1995). Prevention of diabetes-associated embryopathy by overexpression of the free radical scavenger copper zinc superoxide dismutase in transgenic mouse embryos. *American Journal of Obstetrics and Gynecology, 173*(4), 1036-1041. 10.1016/0002-9378(95)91323-8

Hartigan, J. A. (1973). Minimum Mutation Fits to a Given Tree. *Biometrics, 29*(1), 53-65. 10.2307/2529676

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer.

Heart Disease and Stroke Statistics—2021 Update. (2021). *Circulation, 143*(8), e254-e743. 10.1161/CIR.0000000000000950

Hedermann, G., Hedley, P. L., Thagaard, I. N., Krebs, L., Ekelund, C. K., Sørensen, T. I. A., & Christiansen, M. (2021). Maternal obesity and metabolic disorders associate with congenital heart defects in the offspring: A systematic review. *PloS One, 16*(5), e0252343. 10.1371/journal.pone.0252343

Heilstedt, H. A., Ballif, B. C., Howard, L. A., Kashork, C. D., & Shaffer, L. G. (2003). Population data suggest that deletions of 1p36 are a relatively common chromosome abnormality. *Clinical Genetics, 64*(4), 310-316. 10.1034/j.1399-0004.2003.00126.x

Hekkelman, M. L., te Beek, T. A. H., Pettifer, S. R., Thorne, D., Attwood, T. K., & Vriend, G. (2010). WIWS: a protein structure bioinformatics Web service collection. *Nucleic Acids Research, 38*(Web Server issue), W719-W723. 10.1093/nar/gkq453

Hennermann, J. B., Vetter, B., Wolf, C., Windt, E., Bührdel, P., Seidel, J., Mönch, E., & Kulozik, A. E. (2000). Phenylketonuria and hyperphenylalaninemia in eastern Germany: a characteristic molecular profile and 15 novel mutations. *Human Mutation, 15*(3), 254-260. 10.1002/(SICI)1098-1004(200003)15:3<254::AID-HUMU6>3.0.CO;2-W

Hernández-Díaz, S., Werler, M. M., Walker, A. M., & Mitchell, A. A. (2000). Folic acid antagonists during pregnancy and the risk of birth defects. *The New England Journal of Medicine, 343*(22), 1608-1614. 10.1056/NEJM200011303432204

Herskind, A. M., Almind Pedersen, D., & Christensen, K. (2013). Increased prevalence of congenital heart defects in monozygotic and dizygotic twins. *Circulation, 128*(11), 1182-1188. 10.1161/CIRCULATIONAHA.113.002453

Hicks, S., Wheeler, D. A., Plon, S. E., & Kimmel, M. (2011). Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. *Human Mutation, 32*(6), 661-668. 10.1002/humu.21490

Higasa, K., Ogawa, A., Terao, C., Shimizu, M., Kosugi, S., Yamada, R., Date, H., Matsubara, H., & Matsuda, F. (2017). A burden of rare variants in BMPR2 and KCNK3 contributes to a risk of familial pulmonary arterial hypertension. *BMC Pulmonary Medicine, 17*, 57. 10.1186/s12890-017-0400-z

Hillert, A., Anikster, Y., Belanger-Quintana, A., Burlina, A., Burton, B. K., Carducci, C., Chiesa, A. E., Christodoulou, J., Đorđević, M., Desviat, L. R., Eliyahu, A., Evers, R. A. F., Fajkusova, L., Feillet, F., Bonfim-Freitas, P. E., Giżewska, M., Gundorova, P., Karall, D., Kneller, K., . . . Blau, N. (2020). The Genetic Landscape and Epidemiology of Phenylketonuria. *American Journal of Human Genetics, 107*(2), 234-250. 10.1016/j.ajhg.2020.06.006

Hoffman, J. I. E., & Kaplan, S. (2002). The incidence of congenital heart disease. *Journal of the American College of Cardiology, 39*(12), 1890-1900. 10.1016/s0735-1097(02)01886-7

Hoffman, J. I. E., Kaplan, S., & Liberthson, R. R. (2004). Prevalence of congenital heart disease. *American Heart Journal, 147*(3), 425-439. 10.1016/j.ahj.2003.05.003

Houyel, L., Khoshnood, B., Anderson, R. H., Lelong, N., Thieulin, A., Goffinet, F., & Bonnet, D. (2011). Population-based evaluation of a suggested anatomic and clinical classification of congenital heart defects based on the International Paediatric and Congenital Cardiac Code. *Orphanet Journal of Rare Diseases, 6*, 64. 10.1186/1750-1172-6-64

Hu, J., & Ng, P. C. (2012). Predicting the effects of frameshifting indels. *Genome Biology, 13*(2), R9. 10.1186/gb-2012-13-2-r9

Huang, R., Wang, J., Xue, S., Qiu, X., Shi, H., Li, R., Qu, X., Yang, X., Liu, H., Li, N., Li, Y., Xu, Y., & Yang, Y. (2017). TBX20 loss-of-function mutation responsible for familial tetralogy of Fallot or sporadic persistent truncus arteriosus. *International Journal of Medical Sciences, 14*(4), 323-332. 10.7150/ijms.17834

Hutson, M. R., & Kirby, M. L. (2009). Double Outlet Right Ventricle. In F. Lang (Ed.), *Encyclopedia of Molecular Mechanisms of Disease* (pp. 543-545). Springer Berlin Heidelberg. 10.1007/978-3-540-29676-8_511

Iakoucheva, L. M., Radivojac, P., Brown, C. J., O'Connor, T. R., Sikes, J. G., Obradovic, Z., & Dunker, A. K. (2004). The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Research, 32*(3), 1037-1049. 10.1093/nar/gkh253

Increased Frequency of *De novo* Copy Number Variants in Congenital Heart Disease by Integrative Analysis of Single Nucleotide Polymorphism Array and Exome

Sequence Data. (2014). *Circulation Research*, *115*(10), 884-896. 10.1161/CIRCRESAHA.115.304458

Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., Cannon-Albright, L. A., Teerlink, C. C., Stanford, J. L., Isaacs, W. B., Xu, J., Cooney, K. A., Lange, E. M., Schleutker, J., Carpten, J. D., . . . Sieh, W. (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *American Journal of Human Genetics*, *99*(4), 877-885. 10.1016/j.ajhg.2016.08.016

Ittisoponpisan, S., & David, A. (2018). Structural Biology Helps Interpret Variants of Uncertain Significance in Genes Causing Endocrine and Metabolic Disorders. *Journal of the Endocrine Society*, *2*(8), 842-854. 10.1210/js.2018-00077

Ittisoponpisan, S., Alhuzimi, E., Sternberg, M. J. E., & David, A. (2017). Landscape of Pleiotropic Proteins Causing Human Disease: Structural and System Biology Insights. *Human Mutation*, *38*(3), 289-296. 10.1002/humu.23155

Ittisoponpisan, S., Islam, S. A., Khanna, T., Alhuzimi, E., David, A., & Sternberg, M. J. E. (2019). Can Predicted Protein 3D Structures Provide Reliable Insights into whether Missense Variants Are Disease Associated? *Journal of Molecular Biology*, *431*(11), 2197-2212. 10.1016/j.jmb.2019.04.009

Jabs, A., Weiss, M. S., & Hilgenfeld, R. (1999). Non-proline cis peptide bonds in proteins. *Journal of Molecular Biology*, *286*(1), 291-304. 10.1006/jmbi.1998.2459

Jackson, S., Freeman, R., Noronha, A., Jamil, H., Chavez, E., Carmichael, J., Ruiz, K. M., Miller, C., Benke, S., Perrot, R., Hockley, M., Murphy, K., Casillan, A., Radanovich, L., Deforest, R., Nunes, M. E., Galarreta-Aima, C., Sidlow, R.,

Einhorn, Y., & Woods, J. (2024). Applying data science methodologies with artificial intelligence variant reinterpretation to map and estimate genetic disorder prevalence utilizing clinical data. *American Journal of Medical Genetics. Part A*, 10.1002/ajmg.a.63505

Jacobs, E. G., Leung, M. P., & Karlberg, J. (2000). Distribution of symptomatic congenital heart disease in Hong Kong. *Pediatric Cardiology, 21*(2), 148-157. 10.1007/s002469910025

Janssen, P. A., Rothman, I., & Schwartz, S. M. (1996). Congenital malformations in newborns of women with established and gestational diabetes in Washington State, 1984-91. *Paediatric and Perinatal Epidemiology, 10*(1), 52-63. 10.1111/j.1365-3016.1996.tb00026.x

Jenkins, K. J., Correa, A., Feinstein, J. A., Botto, L., Britt, A. E., Daniels, S. R., Elixson, M., Warnes, C. A., & Webb, C. L. (2007). Noninherited risk factors and congenital cardiovascular defects: current knowledge: a scientific statement from the American Heart Association Council on Cardiovascular Disease in the Young: endorsed by the American Academy of Pediatrics. *Circulation, 115*(23), 2995-3014. 10.1161/CIRCULATIONAHA.106.183216

Jick, H., Holmes, L. B., Hunter, J. R., Madsen, S., & Stergachis, A. (1981). First-trimester drug use and congenital disorders. *Jama, 246*(4), 343-346.

Jin, S. C., Homsy, J., Zaidi, S., Lu, Q., Morton, S., DePalma, S. R., Zeng, X., Qi, H., Chang, W., Sierant, M. C., Hung, W., Haider, S., Zhang, J., Knight, J., Bjornson, R. D., Castaldi, C., Tikhonoa, I. R., Bilguvar, K., Mane, S. M., Brueckner, M. (2017). Contribution of rare inherited and *de novo* variants in 2,871 congenital heart disease probands. *Nature Genetics, 49*(11), 1593-1601. 10.1038/ng.3970

Jonas, R. (2002). *Comprehensive surgical management of congenital heart disease*. CRC press.

Jones, K. L., & Smith, D. W. (1973). Recognition of the fetal alcohol syndrome in early infancy. *Lancet (London, England), 302*(7836), 999-1001. 10.1016/s0140-6736(73)91092-1

Joziasse, I. C., van der Smagt, J. J., Poot, M., Hochstenbach, R., Nelen, M. R., van Gijn, M., Dooijes, D., Mulder, B. J. M., & Doevendans, P. A. (2009). A duplication including GATA4 does not co-segregate with congenital heart defects. *American Journal of Medical Genetics. Part A, 149A*(5), 1062-1066. 10.1002/ajmg.a.32769

Jubb, H. C., Higueruelo, A. P., Ochoa-Montaño, B., Pitt, W. R., Ascher, D. B., & Blundell, T. L. (2017). Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. *Journal of Molecular Biology, 429*(3), 365-371. 10.1016/j.jmb.2016.12.004

Kaelo, P., & Ali, M. M. (2006). Some Variants of the Controlled Random Search Algorithm for Global Optimization. *Journal of Optimization Theory and Applications, 130*(2), 253-264. 10.1007/s10957-006-9101-0

Källén, K. (1999). Maternal smoking and congenital heart defects. *European Journal of Epidemiology, 15*(8), 731-737. 10.1023/a:1007671631188

Kang, Q., Sun, M. H., Cheng, H., Peng, Y., Montag, A. G., Deyrup, A. T., Jiang, W., Luu, H. H., Luo, J., Szatkowski, J. P., Vanichakarn, P., Park, J. Y., Li, Y., Haydon, R. C., & He, T. -. (2004). Characterization of the distinct orthotopic bone-forming activity of 14 BMPs using recombinant adenovirus-mediated gene delivery. *Gene Therapy, 11*(17), 1312-1320. 10.1038/sj.gt.3302298

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., . . . MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature, 581*(7809), 434-443. 10.1038/s41586-020-2308-7

Kasahara, H., Lee, B., Schott, J., Benson, D. W., Seidman, J. G., Seidman, C. E., & Izumo, S. (2000). Loss of function and inhibitory effects of human CSX/NKX2.5 homeoprotein mutations associated with congenital heart disease. *The Journal of Clinical Investigation, 106*(2), 299-308. 10.1172/JCI9860

Katagiri, T., Yamaguchi, A., Ikeda, T., Yoshiki, S., Wozney, J. M., Rosen, V., Wang, E. A., Tanaka, H., Omura, S., & Suda, T. (1990). The non-osteogenic mouse pluripotent cell line, C3H10T1/2, is induced to differentiate into osteoblastic cells by recombinant human bone morphogenetic protein-2. *Biochemical and Biophysical Research Communications, 172*(1), 295-299. 10.1016/s0006-291x(05)80208-6

Katagiri, T., Yamaguchi, A., Komaki, M., Abe, E., Takahashi, N., Ikeda, T., Rosen, V., Wozney, J. M., Fujisawa-Sehara, A., & Suda, T. (1994). Bone morphogenetic protein-2 converts the differentiation pathway of C2C12 myoblasts into the osteoblast lineage. *The Journal of Cell Biology, 127*(6 Pt 1), 1755-1766. 10.1083/jcb.127.6.1755

Kawashima, S., & Kanehisa, M. (2000). AAindex: Amino Acid index database. *Nucleic Acids Research, 28*(1), 374. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC102411/

Keltner, J. L., Gittinger, J. W., Miller, N. R., & Burder, R. M. (1987). A red eye and high intraocular pressure. *Survey of Ophthalmology, 31*(5), 328-336. 10.1016/0039-6257(87)90117-2

Khan, S., & Vihinen, M. (2010). Performance of protein stability predictors. *Human Mutation, 31*(6), 675-684. 10.1002/humu.21242

Khanna, T., Hanna, G., Sternberg, M. J. E., & David, A. (2021). Missense3D-DB web catalogue: an atom-based analysis and repository of 4M human protein-coding genetic variants. *Human Genetics, 140*(5), 805-812. 10.1007/s00439-020-02246-z

Kim, I. Y., Lee, D. H., Ahn, H. J., Tokunaga, H., Song, W., Devereaux, L. M., Jin, D., Sampath, T. K., & Morton, R. A. (2000). Expression of bone morphogenetic protein receptors type-IA, -IB and -II correlates with tumor grade in human prostate cancer tissues. *Cancer Research, 60*(11), 2840-2844.

Kim, M., Park, S. Y., Chang, H. R., Jung, E. Y., Munkhjargal, A., Lim, J., Lee, M., & Kim, Y. (2017). Clinical significance linked to functional defects in bone morphogenetic protein type 2 receptor, BMPR2. *BMB Reports, 50*(6), 308-317. 10.5483/BMBRep.2017.50.6.059

Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge University Press.

Kirby, M. L. (2007). *Cardiac development*. Oxford University Press.

Kirby, M. L., & Kirby, M. L. (2007). *Cardiac Development*. Oxford University Press.

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics, 46*(3), 310-315. 10.1038/ng.2892

Kleinert, S., Sano, T., Weintraub, R. G., Mee, R. B., Karl, T. R., & Wilkinson, J. L. (1997). Anatomic features and surgical strategies in double-outlet right ventricle. *Circulation, 96*(4), 1233-1239. 10.1161/01.cir.96.4.1233

Kloesel, B., DiNardo, J. A., & Body, S. C. (2016). Cardiac Embryology and Molecular Mechanisms of Congenital Heart Disease – A Primer for Anesthesiologists. *Anesthesia and Analgesia, 123*(3), 551-569. 10.1213/ANE.0000000000001451

Kodo, K., Nishizawa, T., Furutani, M., Arai, S., Yamamura, E., Joo, K., Takahashi, T., Matsuoka, R., & Yamagishi, H. (2009). GATA6 mutations cause human cardiac outflow tract defects by disrupting semaphorin-plexin signaling. *Proceedings of the National Academy of Sciences of the United States of America, 106*(33), 13933-13938. 10.1073/pnas.0904744106

Kousseff, B. G. (1999). Diabetic embryopathy. *Current Opinion in Pediatrics, 11*(4), 348-352. 10.1097/00008480-199908000-00014

Kreile, M., Lubina, O., Ozola-Zalite, I., Lugovska, R., Pronina, N., Sterna, O., Vevere, P., Konika, M., Malniece, I., & Gailite, L. (2020). Phenylketonuria in the Latvian population: Molecular basis, phenylalanine levels, and patient compliance. *Molecular Genetics and Metabolism Reports, 25,* 100671. 10.1016/j.ymgmr.2020.100671

Krieger, E., Koraimann, G., & Vriend, G. (2002). Increasing the precision of comparative models with YASARA NOVA--a self-parameterizing force field. *Proteins, 47*(3), 393-402. 10.1002/prot.10104

Krivov, G. G., Shapovalov, M. V., & Dunbrack, R. L. (2009). Improved prediction of protein side-chain conformations with SCWRL4. *Proteins, 77*(4), 778-795. 10.1002/prot.22488

Krogh, A., Larsson, B., von Heijne, G., & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology, 305*(3), 567-580. 10.1006/jmbi.2000.4315

Kucukkal, T. G., Petukh, M., Li, L., & Alexov, E. (2015). Structural and physico-chemical effects of disease and non-disease nsSNPs on proteins. *Current Opinion in Structural Biology, 32*, 18-24. 10.1016/j.sbi.2015.01.003

Kuehl, K. S., & Loffredo, C. A. (2003). Population-based study of l-transposition of the great arteries: possible associations with environmental factors. *Birth Defects Research. Part A, Clinical and Molecular Teratology, 67*(3), 162-167. 10.1002/bdra.10015

Kuehl, K. S., Loffredo, C. A., & Ferencz, C. (1999). Failure to diagnose congenital heart disease in infancy. *Pediatrics, 103*(4 Pt 1), 743-747. 10.1542/peds.103.4.743

Kumar, M. D. S., Bava, K. A., Gromiha, M. M., Prabakaran, P., Kitajima, K., Uedaira, H., & Sarai, A. (2006). ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Research, 34*(Database issue), 204. 10.1093/nar/gkj103

Lage, K., Greenway, S. C., Rosenfeld, J. A., Wakimoto, H., Gorham, J. M., Segrè, A. V., Roberts, A. E., Smoot, L. B., Pu, W. T., Pereira, A. C., Mesquita, S. M., Tommerup, N., Brunak, S., Ballif, B. C., Shaffer, L. G., Donahoe, P. K., Daly, M.

J., Seidman, J. G., Seidman, C. E., & Larsen, L. A. (2012). Genetic and environmental risk factors in congenital heart disease functionally converge in protein networks driving heart development. *Proceedings of the National Academy of Sciences of the United States of America, 109*(35), 14035-14040. 10.1073/pnas.1210730109

LaHaye, S., Corsmeier, D., Basu, M., Bowman, J. L., Fitzgerald-Butt, S., Zender, G., Bosse, K., McBride, K. L., White, P., & Garg, V. (2016). Utilization of Whole Exome Sequencing to Identify Causative Mutations in Familial Congenital Heart Disease. *Circulation. Cardiovascular Genetics, 9*(4), 320-329. 10.1161/CIRCGENETICS.115.001324

Lane, K. B., Machado, R. D., Pauciulo, M. W., Thomson, J. R., Phillips, J. A., Loyd, J. E., Nichols, W. C., & Trembath, R. C. (2000). Heterozygous germline mutations in BMPR2, encoding a TGF-beta receptor, cause familial primary pulmonary hypertension. *Nature Genetics, 26*(1), 81-84. 10.1038/79226

Lanzoni, M., Morris, J., Garne, E., Loane, M., & Kinsner-Ovaskainen, A. (2017). *European Monitoring of Congenital Anomalies JRC-EUROCAT Report on Statistical Monitoring of Congenital Anomalies (2006 -2015).* (). Luxembourg: Publications Office of the European Union.

Larkin, E. K., Newman, J. H., Austin, E. D., Hemnes, A. R., Wheeler, L., Robbins, I. M., West, J. D., Phillips, J. A., Hamid, R., & Loyd, J. E. (2012). Longitudinal analysis casts doubt on the presence of genetic anticipation in heritable pulmonary arterial hypertension. *American Journal of Respiratory and Critical Care Medicine, 186*(9), 892-896. 10.1164/rccm.201205-0886OC

Lee, K. M., Tsai, K. Y., Wang, N., & Ingber, D. E. (1998). Extracellular matrix and pulmonary hypertension: control of vascular smooth muscle cell contractility. *The American Journal of Physiology, 274*(1), 76. 10.1152/ajpheart.1998.274.1.H76

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., . . . MacArthur, D. G. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature, 536*(7616), 285-291. 10.1038/nature19057

Lenke, R. R., & Levy, H. L. (1980). Maternal phenylketonuria and hyperphenylalaninemia. An international survey of the outcome of untreated and treated pregnancies. *The New England Journal of Medicine, 303*(21), 1202-1208. 10.1056/NEJM198011203032104

Levy, H. L., Guldberg, P., Güttler, F., Hanley, W. B., Matalon, R., Rouse, B. M., Trefz, F., Azen, C., Allred, E. N., de la Cruz, F., & Koch, R. (2001). Congenital heart disease in maternal phenylketonuria: report from the Maternal PKU Collaborative Study. *Pediatric Research, 49*(5), 636-642. 10.1203/00006450-200105000-00005

Lewinger, J. P., Conti, D. V., Baurley, J. W., Triche, T. J., & Thomas, D. C. (2007). Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. *Genetic Epidemiology, 31*(8), 871-882. 10.1002/gepi.20248

Li, B., Krishnan, V. G., Mort, M. E., Xin, F., Kamati, K. K., Cooper, D. N., Mooney, S. D., & Radivojac, P. (2009). Automated inference of molecular mechanisms of

disease from amino acid substitutions. *Bioinformatics, 25*(21), 2744-2750. 10.1093/bioinformatics/btp528

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics, 25*(14), 1754-1760. 10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England), 25*(16), 2078-2079. 10.1093/bioinformatics/btp352

Li, J. H., Mazur, C. A., Berisa, T., & Pickrell, J. K. (2021). Low-pass sequencing increases the power of GWAS and decreases measurement error of polygenic risk scores compared to genotyping arrays. *Genome Research, 31*(4), 529-537. 10.1101/gr.266486.120

Li, K., & Stockwell, T. B. (2010). VariantClassifier: A hierarchical variant classifier for annotated genomes. *BMC Research Notes, 3*, 191. 10.1186/1756-0500-3-191

Li, R., Fu, F., Yu, Q., Wang, D., Jing, X., Zhang, Y., Li, F., Li, F., Han, J., Pan, M., Zhen, L., Li, D., & Liao, C. (2020). Prenatal exome sequencing in fetuses with congenital heart defects. *Clinical Genetics, 98*(3), 215-230. https://doi.org/10.1111/cge.13774

Li, R., Xu, Y., Wang, J., Liu, X., Yuan, F., Huang, R., Xue, S., Li, L., Liu, H., Li, Y., Qu, X., Shi, H., Zhang, M., Qiu, X., & Yang, Y. (2018). GATA4 Loss-of-Function Mutation and the Congenitally Bicuspid Aortic Valve. *American Journal of Cardiology, 121*(4), 469-474. 10.1016/j.amjcard.2017.11.012

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., Ward, L. D., Lowe, C. B., Holloway, A. K., Clamp, M., Gnerre, S., Alföldi, J., Beal, K., Chang, J., Clawson, H., . . . Kellis, M. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. *Nature, 478*(7370), 476-482. 10.1038/nature10530

Linn, S., Schoenbaum, S. C., Monson, R. R., Rosner, B., Stubblefield, P. G., & Ryan, K. J. (1982). No association between coffee consumption and adverse outcomes of pregnancy. *The New England Journal of Medicine, 306*(3), 141-145. 10.1056/NEJM198201213060304

Liu, D., Liu, Q. -., Eyries, M., Wu, W. -., Yuan, P., Zhang, R., Soubrier, F., & Jing, Z. -. (2012). Molecular genetics and clinical features of Chinese idiopathic and heritable pulmonary arterial hypertension patients. *European Respiratory Journal, 39*(3), 597-603. 10.1183/09031936.00072911

Liu, X., Yagi, H., Saeed, S., Bais, A. S., Gabriel, G. C., Chen, Z., Peterson, K. A., Li, Y., Schwartz, M. C., Reynolds, W. T., Saydmohammed, M., Gibbs, B., Wu, Y., Devine, W., Chatterjee, B., Klena, N. T., Kostka, D., Bentley, K. L. d. M., Ganapathiraju, M. K., Lo, C. W. (2017). The complex genetics of hypoplastic left heart syndrome. *Nature Genetics, 49*(7), 1152-1159. 10.1038/ng.3870

Loh, P., Danecek, P., Palamara, P. F., Fuchsberger, C., A Reshef, Y., K Finucane, H., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G. R., Durbin, R., & L Price, A. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics, 48*(11), 1443-1448. 10.1038/ng.3679

Louis-Dit-Picard, H., Barc, J., Trujillano, D., Miserey-Lenkei, S., Bouatia-Naji, N., Pylypenko, O., Beaurain, G., Bonnefond, A., Sand, O., Simian, C., Vidal-Petiot, E.,

Soukaseum, C., Mandet, C., Broux, F., Chabre, O., Delahousse, M., Esnault, V., Fiquet, B., Houillier, P., Jeunemaitre, X. (2012). KLHL3 mutations cause familial hyperkalemic hypertension by impairing ion transport in the distal nephron. *Nature Genetics, 44*(4), 456-3. 10.1038/ng.2218

Lu, Q., Hu, Y., Sun, J., Cheng, Y., Cheung, K., & Zhao, H. (2015). A Statistical Framework to Predict Functional Non-Coding Regions in the Human Genome Through Integrated Analysis of Annotation Data. *Scientific Reports, 5*, 10576. 10.1038/srep10576

Luxán, G., D'Amato, G., & de la Pompa, J. L. (2016). Intercellular Signaling in Cardiac Development and Disease: The NOTCH pathway. In T. Nakanishi, R. R. Markwald, H. S. Baldwin, B. B. Keller, D. Srivastava & H. Yamagishi (Eds.), *Etiology and Morphogenesis of Congenital Heart Disease: From Gene Function and Cellular Interaction to Morphology* (). Springer.

Lyu, Z., Wang, L., Lin, J., Li, S., Wu, D., Lian, T., Liu, S., Ye, J., Jiang, X., Wang, X., & Jing, Z. (2020). The features of rare pathogenic BMPR2 variants in pulmonary arterial hypertension: Comparison between patients and reference population. *International Journal of Cardiology, 318*, 138-143. 10.1016/j.ijcard.2020.06.068

MacGrogan, D., Münch, J., & de la Pompa, J. L. (2018). Notch and interacting signalling pathways in cardiac development, disease, and regeneration. *Nature Reviews. Cardiology, 15*(11), 685-704. 10.1038/s41569-018-0100-2

Machado, R. D., Aldred, M. A., James, V., Harrison, R. E., Patel, B., Schwalbe, E. C., Gruenig, E., Janssen, B., Koehler, R., Seeger, W., Eickelberg, O., Olschewski, H., Elliott, C. G., Glissmeyer, E., Carlquist, J., Kim, M., Torbicki, A., Fijalkowska, A.,

Szewczyk, G., Trembath, R. C. (2006). Mutations of the TGF-beta type II receptor BMPR2 in pulmonary arterial hypertension. *Human Mutation, 27*(2), 121-132. 10.1002/humu.20285

Machado, R. D., Eickelberg, O., Elliott, C. G., Geraci, M. W., Hanaoka, M., Loyd, J. E., Newman, J. H., Phillips, J. A., Soubrier, F., Trembath, R. C., & Chung, W. K. (2009). Genetics and genomics of pulmonary arterial hypertension. *Journal of the American College of Cardiology, 54*(1 Suppl), S32-S42. 10.1016/j.jacc.2009.04.015

Mahle, W. T., Martinez, R., Silverman, N., Cohen, M. S., & Anderson, R. H. (2008). Anatomy, echocardiography, and surgical approach to double outlet right ventricle. *Cardiology in the Young, 18 Suppl 3*, 39-51. 10.1017/S1047951108003284

Maitra, M., Koenig, S. N., Srivastava, D., & Garg, V. (2010). Identification of GATA6 Sequence Variants in Patients With Congenital Heart Defects. *Pediatric Research, 68*(4), 281-285. 10.1203/PDR.0b013e3181ed17e4

Marcel Martin. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal,* (17) https://journal.embnet.org/index.php/embnetjournal/article/view/200/479

Martin, A. R., Atkinson, E. G., Chapman, S. B., Stevenson, A., Stroud, R. E., Abebe, T., Akena, D., Alemayehu, M., Ashaba, F. K., Atwoli, L., Bowers, T., Chibnik, L. B., Daly, M. J., DeSmet, T., Dodge, S., Fekadu, A., Ferriera, S., Gelaye, B., Gichuru, S., . . . Zingela, Z. (2021). Low-coverage sequencing cost-effectively detects

known and novel variation in underrepresented populations. *The American Journal of Human Genetics, 108*(4), 656-668. 10.1016/j.ajhg.2021.03.012

Martin, C. L., & Warburton, D. (2015). Detection of Chromosomal Aberrations in Clinical Practice: From Karyotype to Genome Sequence. *Annual Review of Genomics and Human Genetics, 16*(1), 309-326. 10.1146/annurev-genom-090413-025346

Martínez-Frías, M. L., Bermejo, E., Rodríguez-Pinilla, E., & Frías, J. L. (2004). Risk for congenital anomalies associated with different sporadic and daily doses of alcohol consumption during pregnancy: a case-control study. *Birth Defects Research. Part A, Clinical and Molecular Teratology, 70*(4), 194-200. 10.1002/bdra.20017

Massagué, J., & Chen, Y. (2000). Controlling TGF-β signaling. *Genes & Development, 14*(6), 627-644. 10.1101/gad.14.6.627

Mathew, P., & Bordoni, B. (2022). Embryology, Heart. *StatPearls* (). StatPearls Publishing.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research, 20*(9), 1297-1303. 10.1101/gr.107524.110

Meader, S., Ponting, C. P., & Lunter, G. (2010). Massive turnover of functional sequence in human and other mammalian genomes. *Genome Research, 20*(10), 1335-1343. 10.1101/gr.108795.110

Mendis Shanthi, Puska Pekka, & Norrving Bo. (2011). *Global Atlas of Cardiovascular Disease Prevention and Control.* (). Geneva: World Health Organization in collaboration with the World Heart Federation and the World Stroke Organization. https://web.archive.org/web/20140817123106/http://whqlibdoc.who.int/publications/2011/9789241564373_eng.pdf?ua=1

Mighton, C., Smith, A. C., Mayers, J., Tomaszewski, R., Taylor, S., Hume, S., Agatep, R., Spriggs, E., Feilotter, H. E., Semenuk, L., Wong, H., Lazo de la Vega, L., Marshall, C. R., Axford, M. M., Silver, T., Charames, G. S., Di Gioacchino, V., Watkins, N., Foulkes, W. D., . . . Lerner-Ellis, J. (2022). Data sharing to improve concordance in variant interpretation across laboratories: results from the Canadian Open Genetics Repository. *Journal of Medical Genetics, 59*(6), 571-578. 10.1136/jmedgenet-2021-107738

Mills, J. L., & Graubard, B. I. (1987). Is moderate drinking during pregnancy associated with an increased risk for malformations? *Pediatrics, 80*(3), 309-314.

Minikel, E. V., Karczewski, K. J., Martin, H. C., Cummings, B. B., Whiffin, N., Rhodes, D., Alföldi, J., Trembath, R. C., van Heel, D. A., Daly, M. J., Alföldi, J., Armean, I. M., Banks, E., Bergelson, L., Cibulskis, K., Collins, R. L., Connolly, K. M., Covarrubias, M., Cummings, B. B., . . . Genome Aggregation, D. C. (2020). Evaluating drug targets through human loss-of-function genetic variation. *Nature, 581*(7809), 459-464. 10.1038/s41586-020-2267-z

Mirkes, P. E., Cornel, L. M., Park, H. W., & Cunningham, M. L. (1997). Induction of thermotolerance in early postimplantation rat embryos is associated with increased resistance to hyperthermia-induced apoptosis. *Teratology, 56*(3), 210-219. 10.1002/(SICI)1096-9926(199709)56:3<210::AID-TERA4>3.0.CO;2-4

Miyazono, K., Kamiya, Y., & Morikawa, M. (2010). Bone morphogenetic protein receptors and signal transduction. *Journal of Biochemistry, 147*(1), 35-51. 10.1093/jb/mvp148

Mone, F., Eberhardt, R. Y., Morris, R. K., Hurles, M. E., McMullan, D. J., Maher, E. R., Lord, J., Chitty, L. S., Giordano, J. L., Wapner, R. J., & Kilby, M. D. (2021). COngenital heart disease and the Diagnostic yield with Exome sequencing (CODE) study: prospective cohort study and systematic review. *Ultrasound in Obstetrics & Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology, 57*(1), 43-51. 10.1002/uog.22072

Moore, L. L., Singer, M. R., Bradlee, M. L., Rothman, K. J., & Milunsky, A. (2000). A prospective study of the risk of congenital defects associated with maternal obesity and diabetes mellitus. *Epidemiology (Cambridge, Mass.), 11*(6), 689-694. 10.1097/00001648-200011000-00013

Morrell, N. W. (2006). Pulmonary hypertension due to BMPR2 mutation: a new paradigm for tissue remodeling? *Proceedings of the American Thoracic Society, 3*(8), 680-686. 10.1513/pats.200605-118SF

Morrell, N. W. (2010). Role of bone morphogenetic protein receptors in the development of pulmonary arterial hypertension. *Advances in Experimental Medicine and Biology, 661*, 251-264. 10.1007/978-1-60761-500-2_16

Moskowitz, I. P., Wang, J., Peterson, M. A., Pu, W. T., Mackinnon, A. C., Oxburgh, L., Chu, G. C., Sarkar, M., Berul, C., Smoot, L., Robertson, E. J., Schwartz, R., Seidman, J. G., & Seidman, C. E. (2011). Transcription factor genes Smad4 and Gata4 cooperatively regulate cardiac valve development. [corrected]. *Proceedings*

of the National Academy of Sciences of the United States of America, *108*(10), 4006-4011. 10.1073/pnas.1019025108

Mottaz, A., David, F. P. A., Veuthey, A., & Yip, Y. L. (2010). Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. *Bioinformatics (Oxford, England)*, *26*(6), 851-852. 10.1093/bioinformatics/btq028

Muntau, A. C., Röschinger, W., Habich, M., Demmelmair, H., Hoffmann, B., Sommerhoff, C. P., & Roscher, A. A. (2002). Tetrahydrobiopterin as an alternative treatment for mild phenylketonuria. *The New England Journal of Medicine*, *347*(26), 2122-2132. 10.1056/NEJMoa021654

Ng, P. C., & Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome Research*, *11*(5), 863-874. 10.1101/gr.176601

Ng, P. C., & Henikoff, S. (2002). Accounting for human polymorphisms predicted to affect protein function. *Genome Research*, *12*(3), 436-446. 10.1101/gr.212802

Ng, P. C., & Henikoff, S. (2003). SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research*, *31*(13), 3812-3814. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC168916/

Nielsen, G. L., Nørgard, B., Puho, E., Rothman, K. J., Sørensen, H. T., & Czeizel, A. E. (2005). Risk of specific congenital abnormalities in offspring of women with diabetes. *Diabetic Medicine: A Journal of the British Diabetic Association*, *22*(6), 693-696. 10.1111/j.1464-5491.2005.01477.x

Nora, J. J. (1968). Multifactorial inheritance hypothesis for the etiology of congenital heart diseases. The genetic-environmental interaction. *Circulation, 38*(3), 604-617. 10.1161/01.cir.38.3.604

Nora, J. J., Dodd, P. F., McNamara, D. G., Hattwick, M. A., Leachman, R. D., & Cooley, D. A. (1969). Risk to offspring of parents with congenital heart defects. *Jama, 209*(13), 2052-2053.

Ó Conchúir, S., Barlow, K. A., Pache, R. A., Ollikainen, N., Kundert, K., O'Meara, M. J., Smith, C. A., & Kortemme, T. (2015). A Web Resource for Standardized Benchmark Datasets, Metrics, and Rosetta Protocols for Macromolecular Modeling and Design. *PloS One, 10*(9), e0130433. 10.1371/journal.pone.0130433

O'Donnell, A., & Yutzey, K. E. (2020). Mechanisms of heart valve development and disease. *Development (Cambridge, England), 147*(13), dev183020. 10.1242/dev.183020

Oladunjoye, O., Piekarski, B., Baird, C., Banka, P., Marx, G., Del Nido, P. J., & Emani, S. M. (2019). Repair of double outlet right ventricle: Midterm outcomes. *The Journal of Thoracic and Cardiovascular Surgery, ,* S0022-0. 10.1016/j.jtcvs.2019.06.120

Olsen, J., Overvad, K., & Frische, G. (1991). Coffee consumption, birthweight, and reproductive failures. *Epidemiology (Cambridge, Mass.), 2*(5), 370-374. 10.1097/00001648-199109000-00011

Onishi, T., Ishidou, Y., Nagamine, T., Yone, K., Imamura, T., Kato, M., Sampath, T. K., ten Dijke, P., & Sakou, T. (1998). Distinct and overlapping patterns of localization of bone morphogenetic protein (BMP) family members and a BMP

type II receptor during fracture healing in rats. *Bone, 22*(6), 605-612. 10.1016/s8756-3282(98)00056-8

Oster, M. E., Lee, K. A., Honein, M. A., Riehle-Colarusso, T., Shin, M., & Correa, A. (2013). Temporal trends in survival among infants with critical congenital heart defects. *Pediatrics, 131*(5), 1502. 10.1542/peds.2012-3435

Pandurangan, A. P., Ochoa-Montaño, B., Ascher, D. B., & Blundell, T. L. (2017). SDM: a server for predicting effects of mutations on protein stability. *Nucleic Acids Research, 45*(W1), W229-W235. 10.1093/nar/gkx439

Pang, K., Meng, H., Hu, S., Wang, H., Hsi, D., Hua, Z., Pan, X., & Li, S. (2017). Echocardiographic Classification and Surgical Approaches to Double-Outlet Right Ventricle for Great Arteries Arising Almost Exclusively from the Right Ventricle. *Texas Heart Institute Journal*, *44*(4), 245-251. 10.14503/THIJ-16-5759

Parikh, R. S., Desai, S., & Kothari, K. (2011). Dilated episcleral veins with secondary open angle glaucoma. *Indian Journal of Ophthalmology, 59*(2), 153-155. 10.4103/0301-4738.77045

Park, M. K. (2014). *Pediatric cardiology for practitioners E-Book*. Elsevier Health Sciences.

Park, S. W., Hur, S. Y., Yoo, N. J., & Lee, S. H. (2010). Somatic frameshift mutations of bone morphogenic protein receptor 2 gene in gastric and colorectal cancers with microsatellite instability. *APMIS: Acta Pathologica, Microbiologica, Et Immunologica Scandinavica, 118*(11), 824-829. 10.1111/j.1600-0463.2010.02670.x

Parker, S. C. J., Hansen, L., Abaan, H. O., Tullius, T. D., & Margulies, E. H. (2009). Local DNA topography correlates with functional noncoding regions of the human genome. *Science (New York, N.Y.), 324*(5925), 389-392. 10.1126/science.1169050

Parnis, J. M., & Oldham, K. B. (2013). Beyond the Beer–Lambert law: The dependence of absorbance on time in photochemistry. *Journal of Photochemistry and Photobiology A: Chemistry, 267*, 6-10. 10.1016/j.jphotochem.2013.06.006

Pejaver, V., Byrne, A. B., Feng, B., Pagel, K. A., Mooney, S. D., Karchin, R., O'Donnell-Luria, A., Harrison, S. M., Tavtigian, S. V., Greenblatt, M. S., Biesecker, L. G., Radivojac, P., Brenner, S. E., & ClinGen Sequence Variant Interpretation Working Group. (2022). Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. *American Journal of Human Genetics, 109*(12), 2163-2177. 10.1016/j.ajhg.2022.10.013

Pejaver, V., Urresti, J., Lugo-Martinez, J., Pagel, K. A., Lin, G. N., Nam, H., Mort, M., Cooper, D. N., Sebat, J., Iakoucheva, L. M., Mooney, S. D., & Radivojac, P. (2020). Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. *Nature Communications, 11*(1), 1-13. 10.1038/s41467-020-19669-x

Peng, K., Radivojac, P., Vucetic, S., Dunker, A. K., & Obradovic, Z. (2006). Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics, 7*, 208. 10.1186/1471-2105-7-208

Pettersen, J. C. (1977). Birth defects and drugs in pregnancy, O. P. Heinonen, D. Slone and S. Shapiro. Publishing Sciences Group, Inc., Littleton, Massachusetts, 1977. 516 pp. Price unstated. *American Journal of Medical Genetics, 1*(1), 120-121. https://doi.org/10.1002/ajmg.1320010113

Pierpont, M. E., Basson, C. T., Benson, D. W., Gelb, B. D., Giglia, T. M., Goldmuntz, E., McGee, G., Sable, C. A., Srivastava, D., & Webb, C. L. (2007). Genetic basis for congenital heart defects: current knowledge: a scientific statement from the American Heart Association Congenital Cardiac Defects Committee, Council on Cardiovascular Disease in the Young: endorsed by the American Academy of Pediatrics. *Circulation, 115*(23), 3015-3038. 10.1161/CIRCULATIONAHA.106.183056

Pierpont, M. E., Brueckner, M., Chung, W. K., Garg, V., Lacro, R. V., McGuire, A. L., Mital, S., Priest, J. R., Pu, W. T., Roberts, A., Ware, S. M., Gelb, B. D., Russell, M. W., & null, n. (2018). Genetic Basis for Congenital Heart Disease: Revisited: A Scientific Statement From the American Heart Association. *Circulation, 138*(21), e653-e711. 10.1161/CIR.0000000000000606

Pires, D. E. V., & Ascher, D. B. (2016). mCSM-AB: a web server for predicting antibody-antigen affinity changes upon mutation with graph-based signatures. *Nucleic Acids Research, 44*(W1), 469. 10.1093/nar/gkw458

Pires, D. E. V., & Ascher, D. B. (2017). mCSM-NA: predicting the effects of mutations on protein-nucleic acids interactions. *Nucleic Acids Research, 45*(W1), W241-W246. 10.1093/nar/gkx236

Pires, D. E. V., Ascher, D. B., & Blundell, T. L. (2014). DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Research, 42*(Web Server issue), 314. 10.1093/nar/gku411

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., & Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research, 20*(1), 110-121. 10.1101/gr.097857.109

Pollastro, P, & Rampone, S. (2003). HS3D: Homo Sapiens Splice Site Data Set. https://iris.unisannio.it/handle/20.500.12070/1267#.XisN6eDF-HY.mendeley.

Pollastro, Pasquale, & Rampone, S. (2002). HS3D, A dataset of homo sapiens splice regions, and its extraction procedure from a major public database. International Journal of Modern Physics C, 13(8), 1105–1117. https://doi.org/10.1142/S0129183102003796

Potapov, V., Cohen, M., & Schreiber, G. (2009). Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. *Protein Engineering, Design & Selection: PEDS, 22*(9), 553-560. 10.1093/protein/gzp030

Prendiville, T., Jay, P. Y., & Pu, W. T. (2014). Insights into the genetic structure of congenital heart disease from human and murine studies on monogenic disorders. *Cold Spring Harbor Perspectives in Medicine, 4*(10), a013946. 10.1101/cshperspect.a013946

Puri, R. D. (2015). Fetal Dysmorphology. *Journal of Fetal Medicine, 3*(2), 151-159. 10.1007/s40556-015-0057-8

Radius, R. L., & Maumenee, A. E. (1978). Dilated episcleral vessels and open-angle glaucoma. *American Journal of Ophthalmology, 86*(1), 31-35. 10.1016/0002-9394(78)90010-7

Radivojac, P., Obradovic, Z., Smith, D. K., Zhu, G., Vucetic, S., Brown, C. J., Lawson, J. D., & Dunker, A. K. (2004). Protein flexibility and intrinsic disorder. *Protein Science: A Publication of the Protein Society, 13*(1), 71-80. 10.1110/ps.03128904

Radivojac, P., Vacic, V., Haynes, C., Cocklin, R. R., Mohan, A., Heyen, J. W., Goebl, M. G., & Iakoucheva, L. M. (2010). Identification, analysis, and prediction of protein ubiquitination sites. *Proteins, 78*(2), 365-380. 10.1002/prot.22555

Radivojac, P., Vucetic, S., O'Connor, T. R., Uversky, V. N., Obradovic, Z., & Dunker, A. K. (2006). Calmodulin signaling: analysis and prediction of a disorder-dependent molecular recognition. *Proteins, 63*(2), 398-410. 10.1002/prot.20873

Ramensky, V., Bork, P., & Sunyaev, S. (2002). Human non-synonymous SNPs: server and survey. *Nucleic Acids Research, 30*(17), 3894-3900. 10.1093/nar/gkf493

Ramos-Arroyo, M. A., Rodriguez-Pinilla, E., & Cordero, J. F. (1992). Maternal diabetes: the risk for specific birth defects. *European Journal of Epidemiology, 8*(4), 503-508. 10.1007/BF00146367

Ray, J. G., O'Brien, T. E., & Chan, W. S. (2001). Preconception care and the risk of congenital anomalies in the offspring of women with diabetes mellitus: a meta-analysis. *QJM: Monthly Journal of the Association of Physicians, 94*(8), 435-444. 10.1093/qjmed/94.8.435

Reece, E. A., & Wu, Y. K. (1997). Prevention of diabetic embryopathy in offspring of diabetic rats with use of a cocktail of deficient substrates and an antioxidant. *American Journal of Obstetrics and Gynecology, 176*(4), 790-798. 10.1016/s0002-9378(97)70602-1

Reece, E. A., Homko, C. J., & Wu, Y. K. (1996). Multifactorial basis of the syndrome of diabetic embryopathy. *Teratology, 54*(4), 171-182. 10.1002/(SICI)1096-9926(199610)54:4<171::AID-TERA1>3.0.CO;2-4

Reese, M. G., Eeckman, F. H., Kulp, D., & Haussler, D. (1997). Improved splice site detection in Genie. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology, 4*(3), 311-323. 10.1089/cmb.1997.4.311

Rentzsch, P., Schubach, M., Shendure, J., & Kircher, M. (2021). CADD-Splice—improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Medicine, 13*(1), 31. 10.1186/s13073-021-00835-9

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research, 47*(D1), D886-D894. 10.1093/nar/gky1016

Reva, B., Antipin, Y., & Sander, C. (2007). Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biology, 8*(11), R232. 10.1186/gb-2007-8-11-r232

Reva, B., Antipin, Y., & Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Research, 39*(17), e118. 10.1093/nar/gkr407

Richards, C. S., Bale, S., Bellissimo, D. B., Das, S., Grody, W. W., Hegde, M. R., Lyon, E., & Ward, B. E. (2008). ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. *Genetics in Medicine: Official Journal of the American College of Medical Genetics, 10*(4), 294-300. 10.1097/GIM.0b013e31816b5cae

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., & Rehm, H. L. (2015). Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine : Official Journal of the American College of Medical Genetics, 17*(5), 405-424. 10.1038/gim.2015.30

Rigelsky, C. M., Jennings, C., Lehtonen, R., Minai, O. A., Eng, C., & Aldred, M. A. (2008). BMPR2 mutation in a patient with pulmonary arterial hypertension and suspected hereditary hemorrhagic telangiectasia. *American Journal of Medical Genetics. Part A, 146A*(19), 2551-2556. 10.1002/ajmg.a.32468

Roberts KE, McElroy JJ, Wong WP, Yen E, Widlitz A, Barst RJ, Knowles JA, Morse JH. BMPR2 mutations in pulmonary arterial hypertension with congenital heart disease. Eur Respir J. 2004 Sep;24(3):371-4. doi: 10.1183/09031936.04.00018604. PMID: 15358693.

Rodrigues, C. H. M., Pires, D. E. V., & Ascher, D. B. (2021). DynaMut2: Assessing changes in stability and flexibility upon single and multiple point missense mutations. *Protein Science : A Publication of the Protein Society, 30*(1), 60-69. 10.1002/pro.3942

Rodrigues, C. H., Pires, D. E., & Ascher, D. B. (2018). DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic Acids Research, 46*(W1), W350-W355. 10.1093/nar/gky300

Rohit, M., & Rajan, P. (2020). Approach to Cyanotic Congenital Heart Disease in Children. *Indian Journal of Pediatrics, 87*(5), 372-380. 10.1007/s12098-020-03274-3

Rosenberg, L., Mitchell, A. A., Shapiro, S., & Slone, D. (1982). Selected birth defects in relation to caffeine-containing beverages. *Jama, 247*(10), 1429-1432.

Rossum, V. (2012). Python Tutorial. *Python Softw Found, 42*, 1. https://doi.org/10.1111/j.1094-348X.2008.00203_7.x

Rost, B. (1996). PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods in Enzymology, 266,* 525-539. 10.1016/s0076-6879(96)66033-9

Rostad, H., & Sørland, S. J. (1981). Atrial septal defects of secundum type in patients less than 40 years of age. A follow-up study. *Acta Medica Scandinavica. Supplementum, 645,* 29-35. 10.1111/j.0954-6820.1981.tb02598.x

Rothman, K. J., Fyler, D. C., Goldblatt, A., & Kreidberg, M. B. (1979). Exogenous hormones and other drug exposures of children with congenital heart disease. *American Journal of Epidemiology, 109*(4), 433-439. 10.1093/oxfordjournals.aje.a112701

Roulston, A., Marcellus, R. C., & Branton, P. E. (1999). Viruses and apoptosis. *Annual Review of Microbiology, 53,* 577-628. 10.1146/annurev.micro.53.1.577

Rouse, B., & Azen, C. (2004). Effect of high maternal blood phenylalanine on offspring congenital anomalies and developmental outcome at ages 4 and 6 years: the importance of strict dietary control preconception and throughout pregnancy. *The Journal of Pediatrics, 144*(2), 235-239. 10.1016/j.jpeds.2003.10.062

Rudarakanchana, N., Flanagan, J. A., Chen, H., Upton, P. D., Machado, R., Patel, D., Trembath, R. C., & Morrell, N. W. (2002). Functional analysis of bone morphogenetic protein type II receptor mutations underlying primary pulmonary hypertension. *Human Molecular Genetics, 11*(13), 1517-1525. 10.1093/hmg/11.13.1517

Sadler, T. W. (2022). *Langman's medical embryology*. Lippincott Williams & Wilkins.

Salfati, E. L., Spencer, E. G., Topol, S. E., Muse, E. D., Rueda, M., Lucas, J. R., Wagner, G. N., Campman, S., Topol, E. J., & Torkamani, A. (2019). Re-analysis of whole-exome sequencing data uncovers novel diagnostic variants and improves molecular diagnostic yields for sudden death and idiopathic diseases. *Genome Medicine, 11*, 83. 10.1186/s13073-019-0702-2

Samrén, E. B., van Duijn, C. M., Christiaens, G. C., Hofman, A., & Lindhout, D. (1999). Antiepileptic drug regimens and major congenital abnormalities in the offspring. *Annals of Neurology, 46*(5), 739-746.

Sanger, F., & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology, 94*(3), 441-448. 10.1016/0022-2836(75)90213-2

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America, 74*(12), 5463-5467. 10.1073/pnas.74.12.5463

Scanlon, K. S., Ferencz, C., Loffredo, C. A., Wilson, P. D., Correa-Villaseñor, A., Khoury, M. J., & Willett, W. C. (1998). Preconceptional folate intake and malformations of the cardiac outflow tract. Baltimore-Washington Infant Study Group. *Epidemiology (Cambridge, Mass.), 9*(1), 95-98.

Schaefer-Graf, U. M., Buchanan, T. A., Xiang, A., Songster, G., Montoro, M., & Kjos, S. L. (2000). Patterns of congenital anomalies and relationship to initial maternal fasting glucose levels in pregnancies complicated by type 2 and gestational diabetes. *American Journal of Obstetrics and Gynecology, 182*(2), 313-320. 10.1016/s0002-9378(00)70217-1

Schleinitz, D., Klöting, N., Böttcher, Y., Wolf, S., Dietrich, K., Tönjes, A., Breitfeld, J., Enigk, B., Halbritter, J., Körner, A., Schön, M. R., Jenkner, J., Tseng, Y., Lohmann, T., Dressler, M., Stumvoll, M., Blüher, M., & Kovacs, P. (2011). Genetic and evolutionary analyses of the human bone morphogenetic protein receptor 2 (BMPR2) in the pathophysiology of obesity. *PloS One, 6*(2), e16155. 10.1371/journal.pone.0016155

Schnoes, A. M., Brown, S. D., Dodevski, I., & Babbitt, P. C. (2009). Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Computational Biology, 5*(12), e1000605. 10.1371/journal.pcbi.1000605

Schnoes, A. M., Ream, D. C., Thorman, A. W., Babbitt, P. C., & Friedberg, I. (2013). Biases in the experimental annotations of protein function and their effect on our understanding of protein function space. *PLoS Computational Biology, 9*(5), e1003063. 10.1371/journal.pcbi.1003063

Schoenwolf, G. C., Bleyl, S. B., Brauer, P. R., & Francis-West, P. H. (2014). *Larsen's human embryology*. Elsevier Health Sciences.

Schott, J. J., Benson, D. W., Basson, C. T., Pease, W., Silberbach, G. M., Moak, J. P., Maron, B. J., Seidman, C. E., & Seidman, J. G. (1998). Congenital heart disease caused by mutations in the transcription factor NKX2-5. *Science (New York, N.Y.), 281*(5373), 108-111. 10.1126/science.281.5373.108

Schwarz, J. M., Cooper, D. N., Schuelke, M., & Seelow, D. (2014). MutationTaster2: mutation prediction for the deep-sequencing age. *Nature Methods, 11*(4), 361-362. 10.1038/nmeth.2890

Schwarz, J. M., Rödelsperger, C., Schuelke, M., & Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. *Nature Methods, 7*(8), 575-576. 10.1038/nmeth0810-575

Seelow, D., Schwarz, J. M., & Schuelke, M. (2008). GeneDistiller—Distilling Candidate Genes from Linkage Intervals. *Plos One, 3*(12), e3874. 10.1371/journal.pone.0003874

Seldon, W. A., Rubinstein, C., & Fraser, A. A. (1962). The incidence of atrial septal defect in adults. *British Heart Journal, 24*(5), 557-560. 10.1136/hrt.24.5.557

Shaw, G. M., O'Malley, C. D., Wasserman, C. R., Tolarova, M. M., & Lammer, E. J. (1995). Maternal periconceptional use of multivitamins and reduced risk for conotruncal heart defects and limb deficiencies among offspring. *American Journal of Medical Genetics, 59*(4), 536-545. https://doi.org/10.1002/ajmg.1320590428

Sheffield, J. S., Butler-Koster, E. L., Casey, B. M., McIntire, D. D., & Leveno, K. J. (2002). Maternal diabetes mellitus and infant malformations. *Obstetrics and Gynecology, 100*(5 Pt 1), 925-930. 10.1016/s0029-7844(02)02242-1

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research, 29*(1), 308-311. 10.1093/nar/29.1.308

Shieh, J. T. C., Bittles, A. H., & Hudgins, L. (2012). Consanguinity and the risk of congenital heart disease. *American Journal of Medical Genetics. Part A, 158A*(5), 1236-1241. 10.1002/ajmg.a.35272

Shihab, H. A., Gough, J., Cooper, D. N., Day, I. N. M., & Gaunt, T. R. (2013). Predicting the functional consequences of cancer-associated amino acid

substitutions. *Bioinformatics (Oxford, England), 29*(12), 1504-1510. 10.1093/bioinformatics/btt182

Shihab, H. A., Gough, J., Mort, M., Cooper, D. N., Day, I. N. M., & Gaunt, T. R. (2014). Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. *Human Genomics, 8*(1), 11. 10.1186/1479-7364-8-11

Shintani, M., Yagi, H., Nakayama, T., Saji, T., & Matsuoka, R. (2009). A new nonsense mutation of SMAD8 associated with pulmonary arterial hypertension. *Journal of Medical Genetics, 46*(5), 331-337. 10.1136/jmg.2008.062703

Shiraishi, I., & Ichikawa, H. (2012). Human heterotaxy syndrome – from molecular genetics to clinical features, management, and prognosis –. *Circulation Journal: Official Journal of the Japanese Circulation Society, 76*(9), 2066-2075. 10.1253/circj.cj-12-0957

Sifrim, A., Hitz, M., Wilsdon, A., Breckpot, J., Turki, S. H. A., Thienpont, B., McRae, J., Fitzgerald, T. W., Singh, T., Swaminathan, G. J., Prigmore, E., Rajan, D., Abdul-Khaliq, H., Banka, S., Bauer, U. M. M., Bentham, J., Berger, F., Bhattacharya, S., Bu'Lock, F., the Deciphering Developmental, D. S. (2016). Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. *Nature Genetics, 48*(9), 1060-1065. 10.1038/ng.3627

Sim, N., Kumar, P., Hu, J., Henikoff, S., Schneider, G., & Ng, P. C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Research, 40*(Web Server issue), W452-W457. 10.1093/nar/gks539

Simán, C. M., & Eriksson, U. J. (1997). Vitamin E decreases the occurrence of malformations in the offspring of diabetic rats. *Diabetes, 46*(6), 1054-1061. 10.2337/diab.46.6.1054

Simán, C. M., Gittenberger-De Groot, A. C., Wisse, B., & Eriksson, U. J. (2000). Malformations in offspring of diabetic rats: morphometric analysis of neural crest-derived organs and effects of maternal vitamin E treatment. *Teratology, 61*(5), 355-367. 10.1002/(SICI)1096-9926(200005)61:5<355::AID-TERA7>3.0.CO;2-W

Smith, K. A., Joziasse, I. C., Chocron, S., van Dinther, M., Guryev, V., Verhoeven, M. C., Rehmann, H., van der Smagt, J. J., Doevendans, P. A., Cuppen, E., Mulder, B. J., Ten Dijke, P., & Bakkers, J. (2009). Dominant-negative ALK2 allele associates with congenital heart defects. *Circulation, 119*(24), 3062-3069. 10.1161/CIRCULATIONAHA.108.843714

Smithells, R. W., & Newman, C. G. (1992). Recognition of thalidomide defects. *Journal of Medical Genetics, 29*(10), 716-723. 10.1136/jmg.29.10.716

Smits, B. M. G., van Zutphen, B. F. M., Plasterk, R. H. A., & Cuppen, E. (2004). Genetic variation in coding regions between and within commonly used inbred rat strains. *Genome Research, 14*(7), 1285-1290. 10.1101/gr.2155004

Soemedi, R., Wilson, I. J., Bentham, J., Darlay, R., Töpf, A., Zelenika, D., Cosgrove, C., Setchfield, K., Thornborough, C., Granados-Riveron, J., Blue, G. M., Breckpot, J., Hellens, S., Zwolinkski, S., Glen, E., Mamasoula, C., Rahman, T. J., Hall, D., Rauch, A., Keavney, B. D. (2012). Contribution of Global Rare Copy-Number Variants to the Risk of Sporadic Congenital Heart Disease. *The American Journal of Human Genetics, 91*(3), 489-501. 10.1016/j.ajhg.2012.08.003

Srivastava, D. (2001). Genetic assembly of the heart: implications for congenital heart disease. *Annual Review of Physiology, 63*, 451-469. 10.1146/annurev.physiol.63.1.451

Steinberger, E. K., Ferencz, C., & Loffredo, C. A. (2002). Infants with single ventricle: a population-based epidemiological study. *Teratology, 65*(3), 106-115. 10.1002/tera.10017

Steinhaus, R., Proft, S., Schuelke, M., Cooper, D. N., Schwarz, J. M., & Seelow, D. (2021). MutationTaster2021. *Nucleic Acids Research, 49*(W1), W446-W451. 10.1093/nar/gkab266

Stenson, P. D., Mort, M., Ball, E. V., Chapman, M., Evans, K., Azevedo, L., Hayden, M., Heywood, S., Millar, D. S., Phillips, A. D., & Cooper, D. N. (2020). The Human Gene Mutation Database (HGMD®): optimizing its use in a clinical diagnostic or research setting. *Human Genetics, 139*(10), 1197-1207. 10.1007/s00439-020-02199-3

Stenson, P. D., Mort, M., Ball, E. V., Shaw, K., Phillips, A., & Cooper, D. N. (2014). The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Human Genetics, 133*(1), 1-9. 10.1007/s00439-013-1358-4

Sun, R., Liu, M., Lu, L., Zheng, Y., & Zhang, P. (2015). Congenital Heart Disease: Causes, Diagnosis, Symptoms, and Treatments. *Cell Biochemistry and Biophysics, 72*(3), 857-860. 10.1007/s12013-015-0551-6

Sylva, M., van den Hoff, M. J. B., & Moorman, A. F. M. (2014). Development of the human heart. *American Journal of Medical Genetics. Part A, 164A*(6), 1347-1371. 10.1002/ajmg.a.35896

Tabaska, J. E., & Zhang, M. Q. (1999). Detection of polyadenylation signals in human DNA sequences. *Gene, 231*(1-2), 77-86. 10.1016/s0378-1119(99)00104-3

Takizawa, T., Ohashi, K., & Nakanishi, Y. (1996). Possible involvement of double-stranded RNA-activated protein kinase in cell death by influenza virus infection. *Journal of Virology, 70*(11), 8128-8132. 10.1128/JVI.70.11.8128-8132.1996

Tennant, P. W., Pearce, M. S., Bythell, M., & Rankin, J. (2010). 20-year survival of children born with congenital anomalies: a population-based study. *The Lancet, 375*(9715), 649-656. 10.1016/S0140-6736(09)61922-X

Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., Kang, H. M., Jordan, D., Leal, S. M., Gabriel, S., Rieder, M. J., Abecasis, G., Altshuler, D., Nickerson, D. A., Boerwinkle, E., Akey, J. M. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science (New York, N.Y.), 337*(6090), 64-69. 10.1126/science.1219240

The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators. (1989). *American Journal of Epidemiology, 129*(4), 687-702.

The ENCODE (ENCyclopedia Of DNA Elements) Project. (2004). *Science (New York, N.Y.), 306*(5696), 636-640. 10.1126/science.1105136

Thermo Fischer Scientific. (2009). *NanoDrop 2000/2000c Spectrophotometer V1.0 User Manual*. Thermo Fisher Scientific.

Thienpont, B., Mertens, L., de Ravel, T., Eyskens, B., Boshoff, D., Maas, N., Fryns, J., Gewillig, M., Vermeesch, J. R., & Devriendt, K. (2007). Submicroscopic chromosomal imbalances detected by array-CGH are a frequent cause of congenital heart defects in selected patients. *European Heart Journal, 28*(22), 2778-2784. 10.1093/eurheartj/ehl560

Thiltgen, G., & Goldstein, R. A. (2012). Assessing predictors of changes in protein stability upon mutation using self-consistency. *PloS One, 7*(10), e46084. 10.1371/journal.pone.0046084

Thomas, J. W., Touchman, J. W., Blakesley, R. W., Bouffard, G. G., Beckstrom-Sternberg, S. M., Margulies, E. H., Blanchette, M., Siepel, A. C., Thomas, P. J., McDowell, J. C., Maskeri, B., Hansen, N. F., Schwartz, M. S., Weber, R. J., Kent, W. J., Karolchik, D., Bruen, T. C., Bevan, R., Cutler, D. J., . . . Green, E. D. (2003). Comparative analyses of multi-species sequences from targeted genomic regions. *Nature, 424*(6950), 788-793. 10.1038/nature01858

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., & Narechania, A. (2003). PANTHER: a library of protein families and subfamilies indexed by function. *Genome Research, 13*(9), 2129-2141. 10.1101/gr.772403

Tikkanen, J., & Heinonen, O. P. (1991). Maternal hyperthermia during pregnancy and cardiovascular malformations in the offspring. *European Journal of Epidemiology, 7*(6), 628-635. 10.1007/BF00218673

Till, B. J., Cooper, J., Tai, T. H., Colowit, P., Greene, E. A., Henikoff, S., & Comai, L. (2007). Discovery of chemically induced mutations in rice by TILLING. *BMC Plant Biology, 7*, 19. 10.1186/1471-2229-7-19

Till, B. J., Reynolds, S. H., Weil, C., Springer, N., Burtner, C., Young, K., Bowers, E., Codomo, C. A., Enns, L. C., Odden, A. R., Greene, E. A., Comai, L., & Henikoff, S. (2004). Discovery of induced point mutations in maize genes by TILLING. *BMC Plant Biology, 4*, 12. 10.1186/1471-2229-4-12

Trembath, R. C. (2001). Mutations in the TGF-beta type 1 receptor, ALK1, in combined primary pulmonary hypertension and hereditary haemorrhagic telangiectasia, implies pathway specificity. *The Journal of Heart and Lung Transplantation: The Official Publication of the International Society for Heart Transplantation, 20*(2), 175. 10.1016/s1053-2498(00)00352-1

Trunzo, R., Santacroce, R., D'Andrea, G., Longo, V., De Girolamo, G., Dimatteo, C., Leccese, A., Lillo, V., Papadia, F., & Margaglione, M. (2013). Mutation analysis in hyperphenylalaninemia patients from South Italy. *Clinical Biochemistry, 46*(18), 1896-1898. 10.1016/j.clinbiochem.2013.06.009

Tsutsumi, R., Fukata, Y., Noritake, J., Iwanaga, T., Perez, F., & Fukata, M. (2009). Identification of G protein alpha subunit-palmitoylating enzyme. *Molecular and Cellular Biology, 29*(2), 435-447. 10.1128/MCB.01144-08

Vallaster, M., Vallaster, C. D., & Wu, S. M. (2012). Epigenetic mechanisms in cardiac development and disease. *Acta Biochimica Et Biophysica Sinica, 44*(1), 92-102. 10.1093/abbs/gmr090

van der Bom, T., Zomer, A. C., Zwinderman, A. H., Meijboom, F. J., Bouma, B. J., & Mulder, B. J. M. (2011). The changing epidemiology of congenital heart disease. *Nature Reviews Cardiology, 8*(1), 50-60. 10.1038/nrcardio.2010.166

van der Linde, D., Konings, E. E. M., Slager, M. A., Witsenburg, M., Helbing, W. A., Takkenberg, J. J. M., & Roos-Hesselink, J. W. (2011). Birth Prevalence of Congenital Heart Disease Worldwide: A Systematic Review and Meta-Analysis. *Journal of the American College of Cardiology, 58*(21), 2241-2247. 10.1016/j.jacc.2011.08.025

Van Durme, J., Delgado, J., Stricher, F., Serrano, L., Schymkowitz, J., & Rousseau, F. (2011). A graphical interface for the FoldX forcefield. *Bioinformatics (Oxford, England), 27*(12), 1711-1712. 10.1093/bioinformatics/btr254

Venselaar, H., te Beek, T. A., Kuipers, R. K., Hekkelman, M. L., & Vriend, G. (2010). Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. *BMC Bioinformatics, 11*, 548. 10.1186/1471-2105-11-548

Verheugt, C. L., Uiterwaal, C. S. P. M., van der Velde, E. T., Meijboom, F. J., Pieper, P. G., van Dijk, A. P. J., Vliegen, H. W., Grobbee, D. E., & Mulder, B. J. M. (2010). Mortality in adult congenital heart disease. *European Heart Journal, 31*(10), 1220-1229. 10.1093/eurheartj/ehq032

Vis, J. C., Duffels, M. G. J., Winter, M. M., Weijerman, M. E., Cobben, J. M., Huisman, S. A., & Mulder, B. J. M. (2009). Down syndrome: a cardiovascular perspective. *Journal of Intellectual Disability Research: JIDR, 53*(5), 419-425. 10.1111/j.1365-2788.2009.01158.x

Wang, H., Ji, R., Meng, J., Cui, Q., Zou, W., Li, L., Wang, G., Sun, L., Li, Z., Huo, L., Fan, Y., & Penny, D. J. (2014). Functional Changes in Pulmonary Arterial Endothelial Cells Associated with BMPR2 Mutations. *PLoS ONE, 9*(9), e106703. 10.1371/journal.pone.0106703

Wang, Q., Pierce-Hoffman, E., Cummings, B. B., Alföldi, J., Francioli, L. C., Gauthier, L. D., Hill, A. J., O'Donnell-Luria, A. H., Armean, I. M., Banks, E., Bergelson, L., Cibulskis, K., Collins, R. L., Connolly, K. M., Covarrubias, M., Daly, M. J., Donnelly, S., Farjoun, Y., Ferriera, S., . . . Genome Aggregation, D. C. (2020). Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes. *Nature Communications, 11*(1), 2539. 10.1038/s41467-019-12438-5

Wang, X., Li, P., Chen, S., Xi, L., Guo, Y., Guo, A., & Sun, K. (2014). Influence of genes and the environment in familial congenital heart defects. *Molecular Medicine Reports, 9*(2), 695-700. 10.3892/mmr.2013.1847

Wang, X., Lian, T., Jiang, X., Liu, S., Li, S., Jiang, R., Wu, W., Ye, J., Cheng, C., Du, Y., Xu, X., Wu, Y., Peng, F., Sun, K., Mao, Y., Yu, H., Liang, C., Shyy, J. Y. -., Zhang, S., Jing, Z. (2019). Germline BMP9 mutation causes idiopathic pulmonary arterial hypertension. *The European Respiratory Journal, 53*(3), 1801609. 10.1183/13993003.01609-2018

Ward, L. D., & Kellis, M. (2012). Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science (New York, N.Y.), 337*(6102), 1675-1678. 10.1126/science.1225057

Warnes, C. A., Liberthson, R., Danielson, G. K., Dore, A., Harris, L., Hoffman, J. I., Somerville, J., Williams, R. G., & Webb, G. D. (2001). Task force 1: the changing

profile of congenital heart disease in adult life. *Journal of the American College of Cardiology, 37*(5), 1170-1175. 10.1016/s0735-1097(01)01272-4

Wasik, K., Berisa, T., Pickrell, J. K., Li, J. H., Fraser, D. J., King, K., & Cox, C. (2021). Comparing low-pass sequencing and genotyping for trait mapping in pharmacogenetics. *BMC Genomics, 22*(1), 197. 10.1186/s12864-021-07508-2

Watanabe, M., Choudhry, A., Berlan, M., Singal, A., Siwik, E., Mohr, S., & Fisher, S. A. (1998). Developmental remodeling and shortening of the cardiac outflow tract involves myocyte programmed cell death. *Development (Cambridge, England), 125*(19), 3809-3820.

Webb, G., & Gatzoulis, M. A. (2006). Atrial Septal Defects in the Adult. *Circulation, 114*(15), 1645-1653. 10.1161/CIRCULATIONAHA.105.592055

Weiss, A., & Attisano, L. (2013). The TGFbeta superfamily signaling pathway. *Wiley Interdisciplinary Reviews. Developmental Biology, 2*(1), 47-63. 10.1002/wdev.86

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., & Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research, 42*(Database issue), 1001. 10.1093/nar/gkt1229

Werler, M. M., Hayes, C., Louik, C., Shapiro, S., & Mitchell, A. A. (1999). Multivitamin supplementation and risk of birth defects. *American Journal of Epidemiology, 150*(7), 675-682. 10.1093/oxfordjournals.aje.a010070

Whiffin, N., Armean, I. M., Kleinman, A., Marshall, J. L., Minikel, E. V., Goodrich, J. K., Quaife, N. M., Cole, J. B., Wang, Q., Karczewski, K. J., Cummings, B. B.,

Francioli, L., Laricchia, K., Guan, A., Alipanahi, B., Morrison, P., Baptista, M. A. S., Merchant, K. M., Armean, I. M., . . . 23andMe, R. T. (2020). The effect of LRRK2 loss-of-function variants in humans. *Nature Medicine, 26*(6), 869-877. 10.1038/s41591-020-0893-5

Wiel, L., Baakman, C., Gilissen, D., Veltman, J. A., Vriend, G., & Gilissen, C. (2019). MetaDome: Pathogenicity analysis of genetic variants through aggregation of homologous human protein domains. *Human Mutation, 40*(8), 1030-1038. 10.1002/humu.23798

Wiel, L., Venselaar, H., Veltman, J. A., Vriend, G., & Gilissen, C. (2017). Aggregation of population-based genetic variation over protein domain homologues and its potential use in genetic diagnostics. *Human Mutation, 38*(11), 1454-1463. 10.1002/humu.23313

Wiley, D. M., Kim, J., Hao, J., Hong, C. C., Bautch, V. L., & Jin, S. (2011). Distinct Signaling Pathways Regulate Sprouting Angiogenesis from the Dorsal Aorta and Axial Vein. *Nature Cell Biology, 13*(6), 686-692. 10.1038/ncb2232

Williams, C. J., Headd, J. J., Moriarty, N. W., Prisant, M. G., Videau, L. L., Deis, L. N., Verma, V., Keedy, D. A., Hintze, B. J., Chen, V. B., Jain, S., Lewis, S. M., Arendall, W. B., Snoeyink, J., Adams, P. D., Lovell, S. C., Richardson, J. S., & Richardson, D. C. (2018). MolProbity: More and better reference data for improved all-atom structure validation. *Protein Science: A Publication of the Protein Society, 27*(1), 293-315. 10.1002/pro.3330

Williams, K., Carson, J., & Lo, C. (2019). Genetics of Congenital Heart Disease. *Biomolecules, 9*(12)10.3390/biom9120879

Wilson, P. D., Loffredo, C. A., Correa-Villaseñor, A., & Ferencz, C. (1998). Attributable fraction for cardiac malformations. *American Journal of Epidemiology, 148*(5), 414-423. 10.1093/oxfordjournals.aje.a009666

Woods, S. E., & Raju, U. (2001). Maternal smoking and the risk of congenital birth defects: a cohort study. *The Journal of the American Board of Family Practice, 14*(5), 330-334.

Wrana, J. L. (2013). Signaling by the TGF superfamily. *Cold Spring Harbor Perspectives in Biology, 5*(10), a011197. 10.1101/cshperspect.a011197

Wren, C., Birrell, G., & Hawthorne, G. (2003). Cardiovascular malformations in infants of diabetic mothers. *Heart, 89*(10), 1217-1220. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1767924/

Wu, N., Zhao, Y., Yin, Y., Zhang, Y., & Luo, J. (2010). Identification and analysis of type II TGF-β receptors in BMP-9-induced osteogenic differentiation of C3H10T1/2 mesenchymal stem cells. *Acta Biochimica Et Biophysica Sinica, 42*(10), 699-708. 10.1093/abbs/gmq075

Wu, Y., Jin, X., Zhang, Y., Zheng, J., & Yang, R. (2021). Genetic and epigenetic mechanisms in the development of congenital heart diseases. *World Journal of Pediatric Surgery, 4*(2), e000196. 10.1136/wjps-2020-000196

Yamagishi, H., Maeda, J., Uchida, K., Tsuchihashi, T., Nakazawa, M., Aramaki, M., Kodo, K., & Yamagishi, C. (2009). Molecular embryology for an understanding of congenital heart diseases. *Anatomical Science International, 84*(3), 88-94. 10.1007/s12565-009-0023-4

Yang, B., Zhou, W., Jiao, J., Nielsen, J. B., Mathis, M. R., Heydarpour, M., Lettre, G., Folkersen, L., Prakash, S., Schurmann, C., Fritsche, L., Farnum, G. A., Lin, M., Othman, M., Hornsby, W., Driscoll, A., Levasseur, A., Thomas, M., Farhat, L., Willer, C. J. (2017). Protein-altering and regulatory genetic variants near GATA4 implicated in bicuspid aortic valve. *Nature Communications, 8*(1), 1-10. 10.1038/ncomms15481

Yang, C., Yang, L., Wan, M., & Cao, X. (2010). Generation of a mouse model with expression of bone morphogenetic protein type II receptor lacking the cytoplasmic domain in osteoblasts. *Annals of the New York Academy of Sciences, 1192*, 286-291. 10.1111/j.1749-6632.2009.05248.x

Yang, F., Zhou, L., Wang, Q., You, X., Li, Y., Zhao, Y., Han, X., Chang, Z., He, X., Cheng, C., Wu, C., Wang, W., Hu, F., Zhao, T., Li, Y., Zhao, M., Zheng, G., Dong, J., Fan, C., Cao, H. (2014). NEXN inhibits GATA4 and leads to atrial septal defects in mice and humans. *Cardiovascular Research, 103*(2), 228-237. 10.1093/cvr/cvu134

Yang, H., Zeng, Q., Ma, Y., Liu, B., Chen, Q., Li, W., Xiong, C., & Zhou, Z. (2018). Genetic analyses in a cohort of 191 pulmonary arterial hypertension patients. *Respiratory Research, 19*(1), 87. 10.1186/s12931-018-0789-9

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., & Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction. *Nature Methods, 12*(1), 7-8. 10.1038/nmeth.3213

Yates, C. L., Monaghan, K. G., Copenheaver, D., Retterer, K., Scuffins, J., Kucera, C. R., Friedman, B., Richard, G., & Juusola, J. (2017). Whole-exome sequencing on deceased fetuses with ultrasound anomalies: expanding our knowledge of genetic

disease during fetal development. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, *19*(10), 1171-1178. 10.1038/gim.2017.31

Yates, C. M., Filippis, I., Kelley, L. A., & Sternberg, M. J. E. (2014). SuSPect: Enhanced Prediction of Single Amino Acid Variant (SAV) Phenotype Using Network Features. *Journal of Molecular Biology*, *426*(14), 2692-2701. 10.1016/j.jmb.2014.04.026

Yeo, G. (2004). CB Burge Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals J. of Comp. *Bio*, *11*(2-3), 377-394.

Yeo, G., & Burge, C. B. (2003). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. Proceedings of the Annual International Conference on Computational Molecular Biology, RECOMB, 322–331. https://doi.org/10.1145/640075.640118.

Yılmaz, M., Ozic, C., & Gok, İ. (2012). *Principles of Nucleic Acid Separation by Agarose Gel Electrophoresis*. IntechOpen. 10.5772/38654

Yim, D., Dragulescu, A., Ide, H., Seed, M., Grosse-Wortmann, L., van Arsdell, G., & Yoo, S. (2018). Essential Modifiers of Double Outlet Right Ventricle: Revisit With Endocardial Surface Images and 3-Dimensional Print Models. *Circulation. Cardiovascular Imaging*, *11*(3), e006891. 10.1161/CIRCIMAGING.117.006891

Ylinen, K., Aula, P., Stenman, U. H., Kesäniemi-Kuokkanen, T., & Teramo, K. (1984). Risk of minor and major fetal malformations in diabetics with high haemoglobin A1c values in early pregnancy. *British Medical Journal (Clinical Research Ed.)*, *289*(6441), 345-346. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1442358/

Yue, P., Li, Z., & Moult, J. (2005). Loss of protein structure stability as a major causative factor in monogenic disease. *Journal of Molecular Biology, 353*(2), 459-473. 10.1016/j.jmb.2005.08.020

Yue, P., Melamud, E., & Moult, J. (2006). SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics, 7*, 166. 10.1186/1471-2105-7-166

Zaidi, S., & Brueckner, M. (2017). Genetics and Genomics of Congenital Heart Disease. *Circulation Research, 120*(6), 923-940. 10.1161/CIRCRESAHA.116.309140

Zhang, H., & Bradley, A. (1996). Mice deficient for BMP2 are nonviable and have defects in amnion/chorion and cardiac development. *Development (Cambridge, England), 122*(10), 2977-2986. 10.1242/dev.122.10.2977

Zhang, T., Wu, Q., Liu, Y., Lv, J., Sun, H., Chang, Q., Liu, C., & Zhao, Y. (2021). Environmental Risk Factors and Congenital Heart Disease: An Umbrella Review of 165 Systematic Reviews and Meta-Analyses With More Than 120 Million Participants. *Frontiers in Cardiovascular Medicine, 8*, 640729. 10.3389/fcvm.2021.640729

Zierler, S., & Rothman, K. J. (1985). Congenital heart disease in relation to maternal use of Bendectin and other drugs in early pregnancy. *The New England Journal of Medicine, 313*(6), 347-352. 10.1056/NEJM198508083130603

Appendix A – Gene Panel

AAGAB	*ARTN*	*CEP290*	*DNMT1*	*FGFR2*	*GSTP1*
ABCB1	*ASH2L*	*CERS1*	*DNMT3A*	*FIGN*	*GSTT1*
ABCC8	*ATP2A2*	*CFC1*	*DNMT3B*	*FKRP*	*GTF2H3*
ABL1	*ATP2B2*	*CFC1B*	*DNTT*	*FLNA*	*HAND1*
ABO	*AVP*	*CHD1L*	*DRC1*	*FLT4*	*HAND2*
ACE	*AVSD1*	*CHD5*	*DSCAM*	*FOLR1*	*HAS2*
ACR	*AXIN2*	*CHD7*	*DSP*	*FOXC1*	*HAVCR1*
ACTA2	*B3GAT3*	*CHDH*	*DYSF*	*FOXC2*	*HBG1*
ACTB	*BANF1*	*CHRM3*	*EBP*	*FOXF1*	*HCN4*
ACTC1	*BAZ1B*	*CIT*	*ECE1*	*FOXH1*	*HDAC3*
ACVR1	*BBS1*	*CITED2*	*EDN1*	*FOXP1*	*HEY2*
ACVR2B	*BCOR*	*COG7*	*EDNRA*	*FRAS1*	*HHEX*
ADA2	*BHMT*	*COL3A1*	*EGF*	*FSD1*	*HIC2*
ADAM9	*BICC1*	*COL4A1*	*EGFR*	*FSD1L*	*HIF1A*
ADAMTS13	*BMP2*	*COL4A2*	*EGR1*	*FXN*	*HIRA*
ADAP2	*BMP4*	*COL6A1*	*EIF2AK3*	*FXR1*	*HNRNPA1*
ADAR	*BMPR1A*	*COL6A2*	*EIF4E*	*G6PC3*	*HOTAIR*
ADARB1	*BMPR2*	*CORIN*	*EMD*	*GAB1*	*HOXA1*
AFF4	*BRAF*	*CPA1*	*ENO1*	*GALNT1*	*HOXA13*
AGRP	*BRD4*	*CPB1*	*ENO2*	*GATA3*	*HOXA3*
AHDC1	*BTF3P11*	*CPLANE1*	*EP300*	*GATA4*	*HPGDS*
AHR	*C1orf127*	*CPLANE2*	*EPAS1*	*GATA5*	*HTC2*
AHSA1	*C2orf74*	*CPS1*	*EPHX1*	*GATA6*	*HYAL2*
AIMP2	*CACNA1C*	*CREBBP*	*EPO*	*GCK*	*HYMAI*
AIRE	*CAD*	*CRELD1*	*EPRS1*	*GDF1*	*IDUA*
AKAP12	*CASZ1*	*CRIM1*	*ETS1*	*GDF15*	*IFNAR1*
AKT3	*CAV3*	*CRK*	*ETS2*	*GET1*	*IFT74*
ALB	*CBFB*	*CRKL*	*EVC*	*GFAP*	*IGF1*
ALDH1A2	*CBS*	*CRP*	*EYA1*	*GHR*	*IGF1R*
ALDH2	*CBSL*	*CST3*	*F3*	*GJA1*	*IGFBP7*
ALG9	*CC2D2A*	*CXCL12*	*FABP3*	*GJA5*	*IHH*
ANKRD1	*CCDC114*	*CXCR4*	*FAM149B1*	*GLB1*	*IL10*
ANKRD11	*CCDC151*	*DAAM1*	*FANCA*	*GLI1*	*IL11*
ANKS6	*CCDC39*	*DAND5*	*FANCC*	*GLI3*	*IL1RN*
AOS	*CCN1*	*DCHS1*	*FANCD2*	*GNA11*	*IL6*
AP1B1	*CCNH*	*DGCR*	*FANCE*	*GNAQ*	*IMPACT*
APLN	*CCR6*	*DGCR2*	*FASLG*	*GNG5*	*INO80*
APLNR	*CD276*	*DGCR8*	*FAT4*	*GNMT*	*INS*
APOA1	*CDH10*	*DICER1*	*FBLN7*	*GP1BB*	*IRF8*
APOE	*CDH5*	*DLC1*	*FBN2*	*GPR182*	*IRX4*
ARGLU1	*CDK13*	*DNAAF3*	*FEN1*	*GPT*	*IRX5*
ARID1A	*CDKN2A*	*DNAH11*	*FGF10*	*GRAP2*	*ISL1*
ARMC4	*CECR*	*DNAH5*	*FGF19*	*GRIN2A*	*JAG1*
ARSA	*CELF2*	*DNAH8*	*FGF23*	*GRK2*	*JAM3*
ARSD	*CENPJ*	*DNAI1*	*FGFR1*	*GRK5*	*JARID2*

KANSL1	*MIR138-1*	*NKX2-5*	*PITX2*	*RFX3*	*SLC29A3*
KATNB1	*MIR143*	*NNMT*	*PKD1*	*RGS6*	*SLC2A14*
KCNE5	*MIR145*	*NODAL*	*PKD1L1*	*RIPPLY3*	*SLC2A3*
KCNJ11	*MIR146B*	*NONO*	*PKD2*	*RIT1*	*SLC50A1*
KCNJ6	*MIR184*	*NOS3*	*PKHD1*	*RLF*	*SLC6A4*
KCTD10	*MIR21*	*NOTCH1*	*PLAGL1*	*RN7SL263P*	*SLC7A7*
KIAA0753	*MIR27A*	*NPHP3*	*PLF*	*RNF19A*	*SLC8A1*
KIF7	*MIR29C*	*NPHP4*	*PLN*	*RNF41*	*SLIT2*
KLF13	*MIR320A*	*NPPA*	*PMP22*	*ROCK1*	*SLIT3*
KLF4	*MIR328*	*NPPB*	*PNN*	*RPGRIP1L*	*SLN*
KLHL24	*MIR34A*	*NR2F1*	*POC1B*	*RRDX*	*SMAD1*
KLHL3	*MIR34B*	*NR2F2*	*POGZ*	*RTN4RL1*	*SMAD2*
KMT2D	*MIR499A*	*NREP*	*POLDIP2*	*RUNX1T1*	*SMAD3*
LCN2	*MIR545*	*NTM*	*POTEF*	*RXRA*	*SMAD4*
LGALS3	*MIR592*	*NTRK3*	*POU5F1*	*RYR2*	*SMAD7*
LMBR1	*MIR873*	*NUP98*	*POU5F1P3*	*S100A4*	*SMARCA1*
LMNA	*MKKS*	*OFD1*	*POU5F1P4*	*S100B*	*SMARCA4*
LOXL2	*MME*	*OPA1*	*PPARGC1A*	*SAA1*	*SMARCB1*
LRP1B	*MMP21*	*OPCML*	*PPM1K*	*SAI1*	*SMARCD3*
LRP2	*MRPS22*	*OPN1LW*	*PPP1CB*	*SALL4*	*SMARCE1*
LRPAP1	*MSX1*	*OTUD6B*	*PPP1R1B*	*SAP130*	*SMG9*
LRRC59	*MSX2*	*PAG1*	*PRDM6*	*SAR1B*	*SMN2*
MALAT1	*MTHFD1*	*PAH*	*PROX1*	*SCN1A*	*SMUG1*
MAML3	*MTHFR*	*PAPPA*	*PTGIR*	*SCN5A*	*SMYD4*
MAP3K7	*MTHFS*	*PART1*	*PTGS2*	*SELP*	*SNX8*
MAPK1	*MTR*	*PBRM1*	*PTH*	*SEM1*	*SOCS3*
MAPK14	*MTRR*	*PBX1*	*PTPN1*	*SEMA3D*	*SOD1*
MAPK3	*MVP*	*PBX3*	*PTPN11*	*SENP2*	*SOD2*
MARCHF3	*MYH6*	*PCBP4*	*PTX3*	*SERPINA5*	*SOS1*
MED12	*MYH7*	*PCSK5*	*PUF60*	*SERPINF2*	*SOX11*
MED13	*MYL3*	*PDE2A*	*QRSL1*	*SETBP1*	*SOX12*
MED13L	*MYL4*	*PDE6D*	*RAC1*	*SETD2*	*SOX17*
MED23	*MYLK3*	*PDLIM1*	*RAD51C*	*SETD5*	*SOX4*
MED25	*MYOCD*	*PDSS1*	*RAF1*	*SFTPA1*	*SOX7*
MEF2C	*NAA15*	*PDX1*	*RAI1*	*SFTPA2*	*SOX9*
MEGF8	*NANOGP1*	*PEX2*	*RAPGEF5*	*SFTPB*	*SPP1*
MEIS2	*NAT2*	*PEX5*	*RBM20*	*SH3BGR*	*SRPX*
MESP1	*NCOA6*	*PGF*	*RBM24*	*SH3PXD2B*	*SSPN*
MGRN1	*NEK8*	*PHC1*	*RCAN1*	*SHFM5*	*STAG2*
MGST1	*NF1*	*PHOX2B*	*REC8*	*SHH*	*STAT3*
MIB1	*NFATC1*	*PIGV*	*RECQL4*	*SHMT1*	*STRA6*
MIB2	*NFKB1*	*PIK3CA*	*REM1*	*SIRT1*	*STX18*
MID1	*NHS*	*PIK3CB*	*REN*	*SLC24A4*	*STX18-AS1*
MIR10A	*NIPBL*	*PIK3CD*	*RERE*	*SLC25A24*	*SUMO1*
MIR10B	*NKX2-1*	*PIK3CG*	*RFC1*	*SLC26A3*	*SUV39H1*

TAB2	*TRRAP*
TAC1	*TSHR*
TAF1	*TWIST1*
TAGLN	*UBE2A*
TAMM41	*UFD1*
TBC1D32	*UROD*
TBC1D9	*USP9X*
TBCC	*VEGFA*
TBX1	*VWF*
TBX18	*WDR62*
TBX2	*WNT11*
TBX20	*ZC3H12D*
TBX3	*ZDHHC24*
TBX5	*ZEB2*
TCF21	*ZFP57*
TCOF1	*ZFPM2*
TCTN3	*ZHX2*
TDGF1	*ZIC3*
TEF	*ZNF778*
TEK	*ZRS*
TFAP2B	
TGFB1	
TGFB2	
TGFBR1	
TGFBR2	
THAS	
TIMP2	
TLR4	
TMEM135	
TMEM216	
TMEM67	
TMEM87B	
TMEM94	
TNF	
TNFRSF10C	
TNFRSF10D	
TNFRSF11A	
TNFRSF11B	
TNFSF11	
TNNI3	
TNXB	
TPM1	
TRAF7	
TRDMT1	
TRPV1	

Table 1 – Genetic panel for the phenotype of Congenital Heart Disease

HOA	**PAH**	**Others**
BGLAP	*BMPR2*	*CHD4*
CBR1	*GDF15*	*COL1A1*
COX2	*PAH*	*COL5A2*
CTNNB1	*SMAD9*	*NR1D2*
DKK1	*TBX4*	*NSD1*
DPEP1	*SOX17*	*RBPJ*
FN1	*BMPR2*	*RFX3*
GAST		*SMAD6*
HPGD		
IL6		
MTCO2P12		
PTGS2		
SERPINE1		
SLCO2A1		
TNF		

Table 2 – Genetic panel for the phenotypes of Hypertrophic osteoartrophathy (HOA), pulmonary arterial hypertension (PAH) and other genes found through literature relevant to the proband's phenotypes.

Appendix B – In Silico modelling BMPR2 p.Arg491Trp.

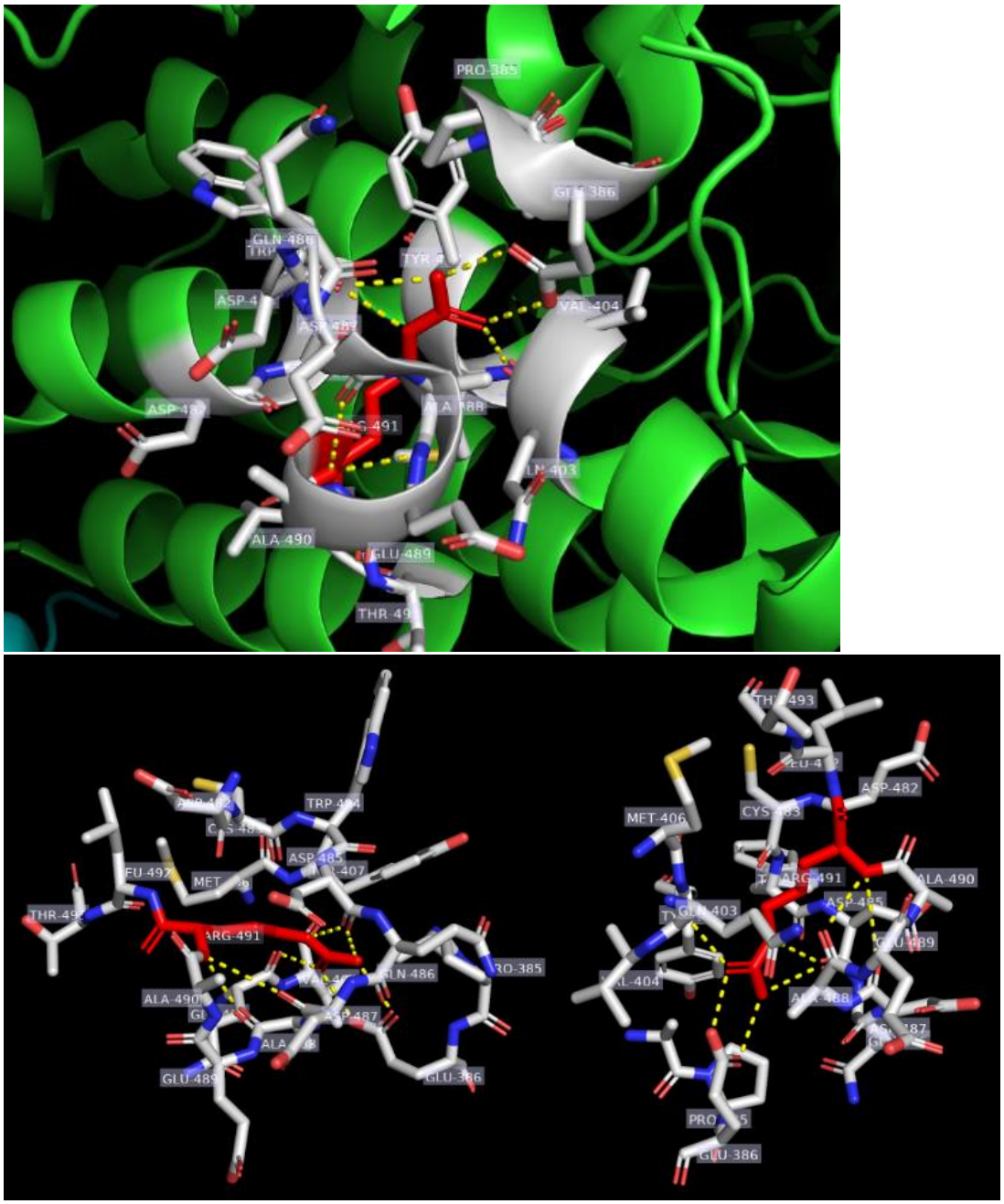


Figure 1. Wildtype BMPR2 protein. Figure A shows the residue of interest at position 491 along with its surrounding residues and the polar interactions to the surrounding residues. Arginine at position 491 of the protein shown in red. The green cartoon figure shows the remaining protein surrounding the 491 residue. Figure B and C show a clearer representation of R491 and its surrounding residues within 5Å along with any interactions. The polar interactions are presented as yellow dotted lines between ARG491 and GLU386, ASP487, ASP485 and GLN403. The ARG491-GLU386, ARG491-GLN403 and ARG491-ASP485 are hydrogen bonds, the ARG491-ASP487 and ARG491-GLU386 are salt bridge formations

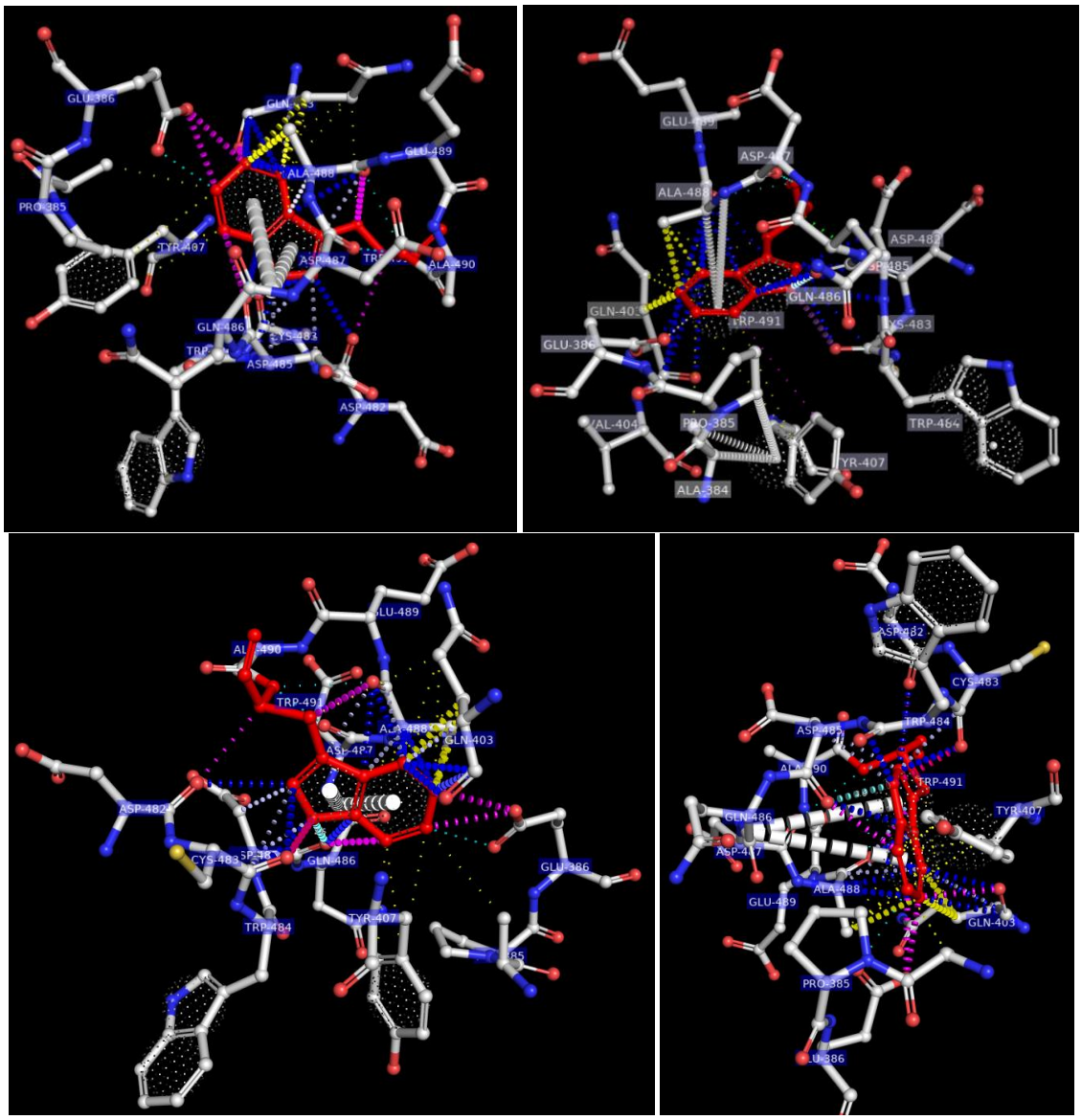


Figure 2 BMPR2 protein 491 having the variation of Tryptophan at position 491. Tryptophan shown in red along with the surrounding amino acids and their interactions in a stick representation. The amino acids having interactions with the amino acid of interest (red tryptophan at position 491) are presented in a grey spectrum according to element, nitrogen in red and oxygens in blue. The grey dotted spheres represent the electronic configuration of the centre of the benzene rings within TRP-484, TYR-407 and TRP-491. The dashed yellow line represents hydrophobic van der wall interactions between the mutated TYR-491 and the surrounding residues GLN-403 and ALA-488 along with a few weaker hydrophobic interactions between TRP-491 and PRO-385 and TYR-407. The smaller light blue dashes represent weak hydrogen interactions between TRP-491, GLU-386, ALA-488 and ASP-487. The large dark blue dashed lines represent van der wall clashes between residues; TRP-491, GLN-403, ALA-488, CYS-483, ASP482 and ASP-485. The pink dashed lines represent the polar van der wall clashes between the residues TRP-491 and CYS-483, ASP-485, ALA-488, GLN-403, GLU-386. The large grey dashed line represents electron donation from the amide ring to the oxygen on ASP-487.

Appendix C – URECA Ethical Approval

L-Università ta' Malta

Faculty of Medicine & Surgery
University of Malta
Msida MSD 2080, Malta

Tel: +356 2340 1879/1891/1167
umms@um.edu.mt

www.um.edu.mt/ms

Ref No: MED-2022-00328

23 November 2023

Ms Giulia Aquilina
7, Corbiere,
Dahlet ic-cipress
Attard, ATD2810

With reference to your application submitted to the Faculty Research Ethics Committee in connection with your research entitled:

Genetic Characterisation of Selected probands/kindreds with congenital heart disease

The Faculty Research Ethics Committee is granting ethical approval for the above-mentioned application.

Professor Anthony Serracino Inglott
Chair
Faculty Research Ethics Committee

Appendix D – Sequencing quality metrics

Sample	Sequence read	Deduplicated (%)	Mapping (%)	Unique (%)	On target (%)
Proband	57,033,974	33,184,668(76.47)	32,898,025(99.14)	31,123,885(93.79)	28,166,931(84.88)
Mother	49,196,646	28,989,287(78.85)	28,771,863(99.25)	27,254,459(94.02)	24,877,235(85.82)
Father	67,411,130	38,921,201(76.11)	38,541,318(99.02)	36,483,824(93.74)	29,866,886(76.74)

Table 1. Raw sequencing metrics. Sequence read - unfiltered sequence. Deduplicated (%): Discarded clean reads following PCR duplicate. Mapping (%): De-duplicated reads followed by mapping onto the reference genome. Unique (%): Reads with the same starting position on each end. On-target (%): Mapped de-duplicated reads (on-target region).

Sample	Raw depth	On target depth (SD)	Coverage 5x %	Coverage 20x %	Coverage 50x %
Proband	126.52	49.04 (40.41)	99.36	92.08	49.39
Mother	139.66	54.19 (46.1)	99.46	92.23	48.68
Father	130.73	51.93 (42.03)	99.19	90.48	45.54

Table 2. Raw sequencing metrics. Coverage 5X %: The rate of cumulative mapping depth exceeding 5X compared with the reference genome. Coverage 20X %: The rate of cumulative mapping depth exceeding 20X compared with the reference genome. Coverage 50X %: The rate of cumulative mapping depth exceeding 50X compared with the reference genome.

Sample	Ts	TV	Ts/Tv	Hetero variants	Homo variants	Hetero/homo
Proband	52428	21568	2.431	43257	19179	2.255
Mother	52678	21533	2.446	37918	21944	1.728
Father	52638	21810	2.413	38681	21840	1.771

Table 3. Raw sequencing metrics. TS (Transitions): Number of transitions, which are point mutations that changes a purine nucleotide to another. purine or a pyrimidine nucleotide to another pyrimidine. TV (Transversions): Number of transversions, which refer to the substitution of a purine for a pyrimidine or vice versa, in deoxyribonucleic acid (DNA). Ts/Tv ratio: The ratio of the number of transitions to the number of transversions for a pair of sequences. Hetero Variants: Number of Heterozygous Variants. Homo Variants: Number of Homozygous Variants. Hetero/Homo ratio: The ratio of the number of heterozygous variants to the number of homozygous variants for a pair of Sequences.

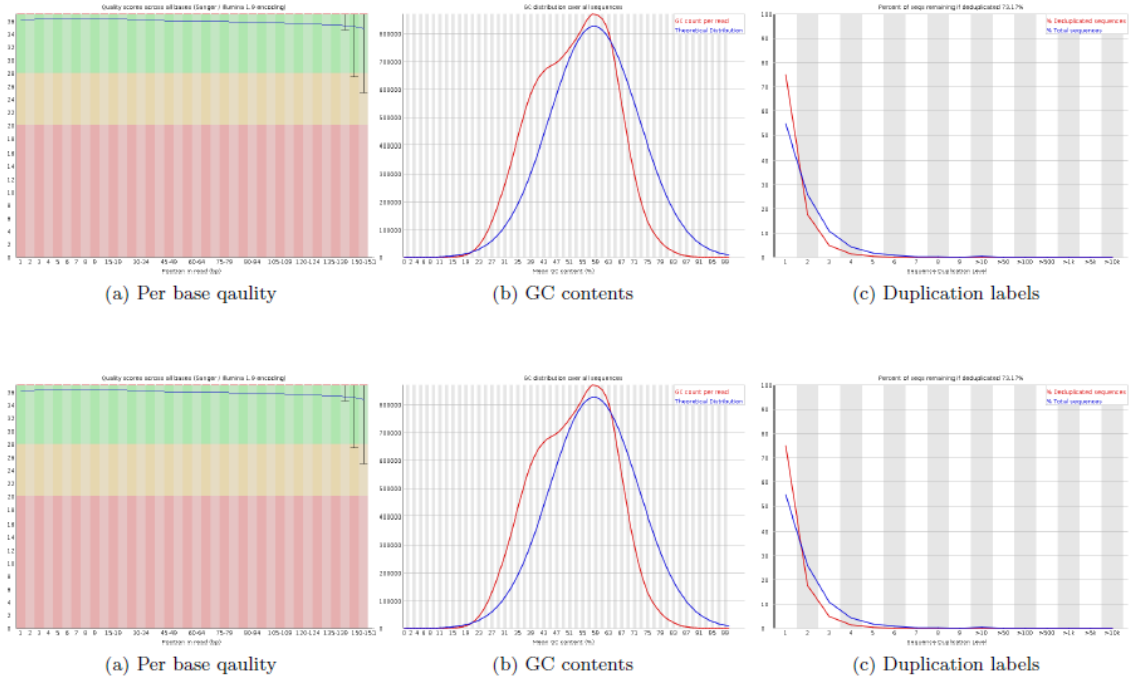


Figure 1. FastQC results generated for the proband using FastQC software. Top and bottom image for Fastq files read 1 and read2 (left read and right read) respectively. (a) Per base qual: Quality values across all bases at each position. (b) GC contents: GC content of each base position in a sample. (c) Duplication level: Proportion of sequence duplication level.

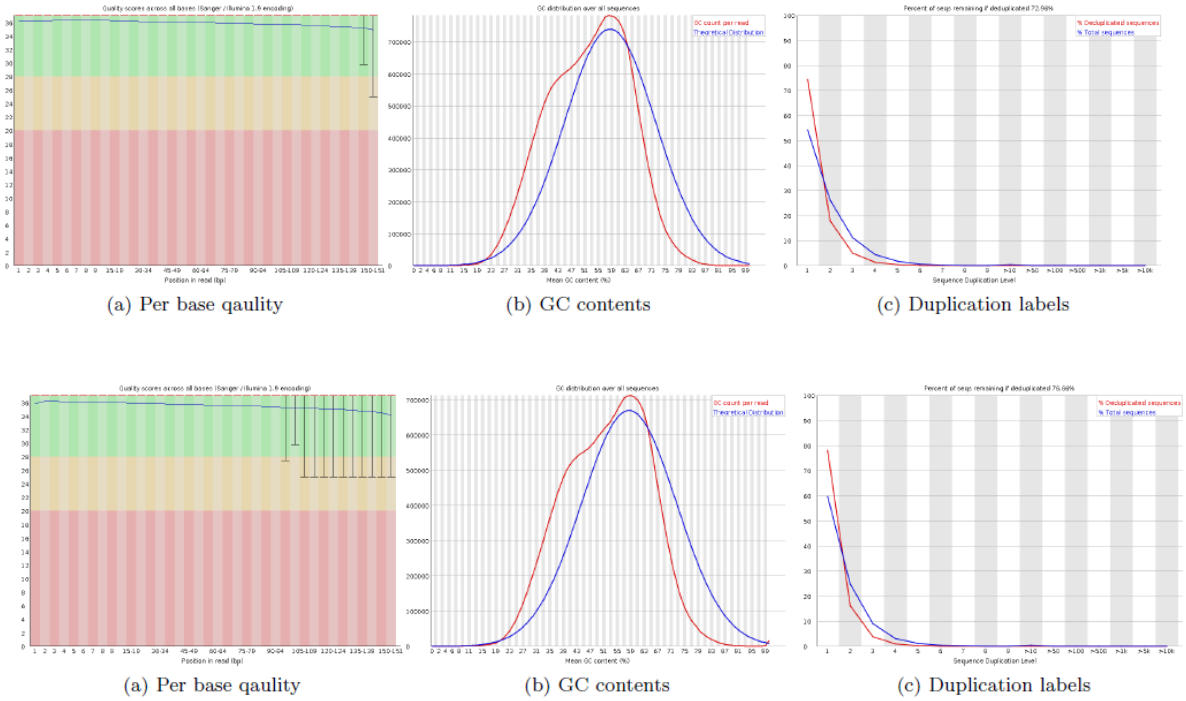
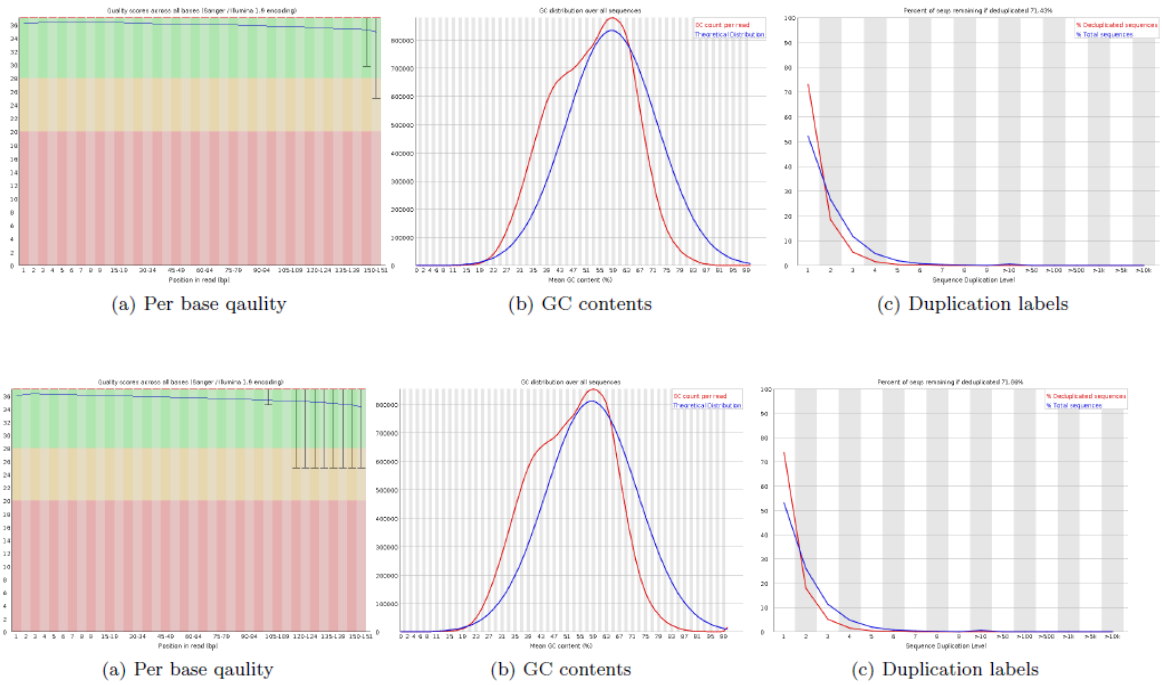


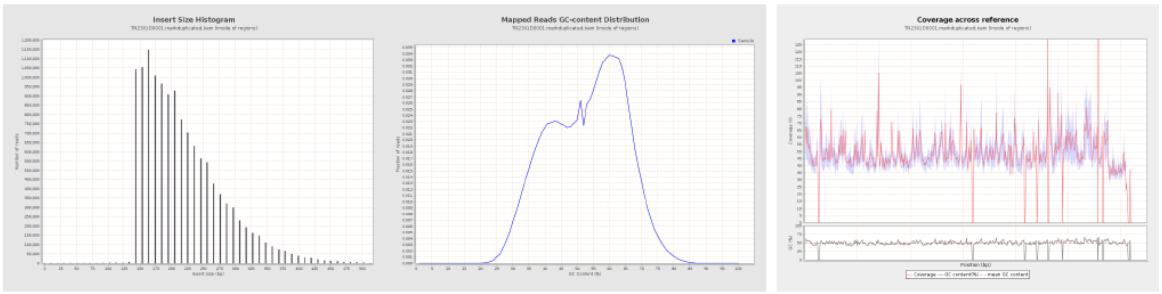
Figure 2. FastQC results generated for the mother using FastQC software. Top and bottom image for Fastq files read 1 and read2 (left read and right read) respectively. (a) Per base qual: Quality values across all bases at each position. (b) GC contents: GC content of each base position in a sample. (c) Duplication level: Proportion of sequence duplication level.



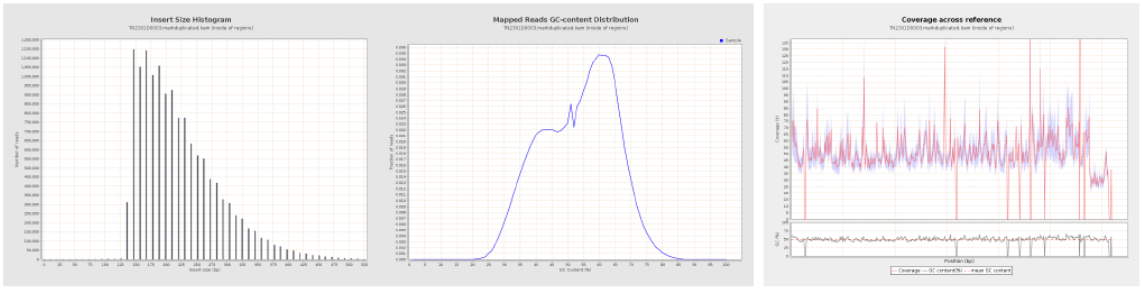
(a) Per base qaulity (b) GC contents (c) Duplication labels

(a) Per base qaulity (b) GC contents (c) Duplication labels

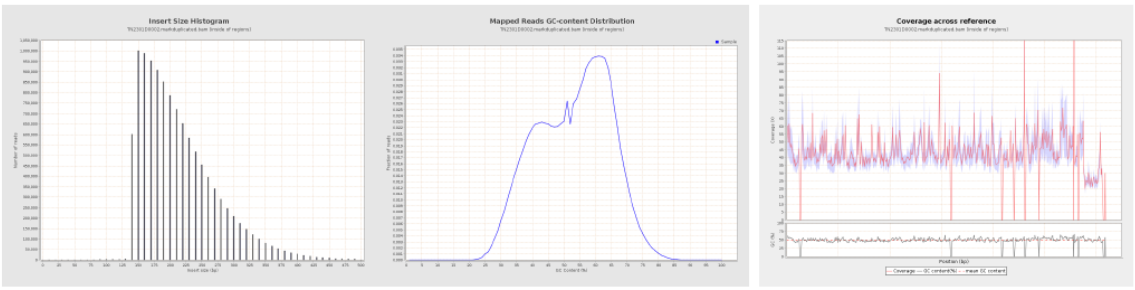
Figure 2. FastQC results generated for the father using FastQC software. Top and bottom image for Fastq files read 1 and read2 (left read and right read) respectively. (a) Per base qual: Quality values across all bases at each position. (b) GC contents: GC content of each base position in a sample. (c) Duplication level: Proportion of sequence duplication level.



(a) Insert size (b) GC contents (c) Reference coverage



(a) Insert size (b) GC contents (c) Reference coverage



(a) Insert size (b) GC contents (c) Reference coverage

Figure 4. QC results generated using Qualimap software for the Proband, mother and father respectively (top to bottom). Insert Size: Histogram of insert size distribution of reads mapped on reference genome. X- axis shows insert size distribution in bp, y -axis shows number of reads. GC contents: GC content of each base position in a sample. This graph shows the distribution of GC content per mapped read. X-axis shows GC content (%) and y-axis shows fraction of reads. Reference coverage: Coverage per position(bp) of reference genome. The upper figure provides the coverage distribution (red line) and coverage deviation across the reference sequence. The coverage is measured in X. The lower figure shows GC content across reference (black line) together with its average value (red dotted line).