Investigation of the molecular basis of inherited developmental conditions in high risk population isolates.

Submitted by

Hannah Faye Jones

To the University of Exeter as a thesis for the degree of Doctor of Philosophy in Medical Studies, April, 2019.

Investigation of the molecular basis of inherited developmental conditions in high risk population isolates.

Submitted by

Hannah Faye Jones

To the University of Exeter as a thesis for the degree of Doctor of Philosophy in Medical Studies, April, 2019.

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

(Signature) ……

ACKNOWLEDGMENTS

First and foremost I would like to thank the families who took part in these studies and the Amish community for their support of the wider Windows of Hope project.

My sincerest thanks goes to my primary supervisor Professor Andrew Crosby for allowing me the opportunity to undertake my PhD studies and for the invaluable academic support, guidance and advice provided over the last four years.

I would also like to extend my sincerest thanks to my secondary supervisor Dr Emma Baple for her vital input particularly for providing her outstanding expertise on all clinical aspects of my projects.

My sincerest gratitude goes to Dr Barry Chioza, my third supervisor, for the endless personal support and encouragement he has shown me over the course of my studies. I will forever be indebted to him for everything he has taught me and can honestly say I would not have got this far without his continued patience, reassurance and confidence in my ability.

I would like to thank all members of the Crosby group, both past and present, for their help and support. In particular Dr Martina Muggenthaler, Dr Gaurav Harlalka, Ilaria D'Atri, Olivia Rickman and Dr Serene Lin for their help and advice but most importantly their friendship and encouragement which has made the undertaking of my studies so enjoyable.

During my PhD I have been fortunate enough to collaborate with many other scientists and clinicians, all of whom I would like to thank for their contributions. I am particularly grateful to; Dr Zineb Ammous, a Clinical Geneticist from the

Community Health Clinic, Matthew Wakeling, a Research Fellow at the University of Exeter and Dr Ryan Ames, a Postdoctoral Research Fellow at the University of Exeter.

I would also like to extend my special thanks to Dr John Chilton and Holly Hardy for welcoming me into their lab and taking the time to teach me so many invaluable techniques. Their enthusiasm and support are central to many of the wonderful experiences this PhD brought me.

I will also like to thank Professor Lorna Harries and Team RNA, in particular Dr Nicky Jeffery and Ben Lee for letting me seek sanctuary in their office during my write up and for their support, advice and friendship over the past few months. I am so grateful to you all; it has meant more than you know!

On a personal level I would like to thank my friends and family for their continued love, support and understanding during one of the best, yet most challenging, periods of my life to date. I would like to specifically thank Sam Faro, Laura Love, Kate Brimblecombe, Vy Catley, Jubair Miah, Lauren Kemp, Bejoy Pal, Wendy Burnman and Tina and Andy Butterworth who have gone above and beyond to help me through this amazing journey. Without you in my corner none of this would have been possible and for that I will be eternally grateful.

ABSTRACT

The Amish communities of Ohio (USA) are a distinct group of endogamous, rural-living Anabaptist Christians. An ancestral bottleneck, caused by migratory events in the 17th century and subsequent rapid population expansion, has led to the enrichment of a number of inherited conditions within these communities. This provides significantly enhanced power to identify genes responsible for rare monogenic disorders, as well traits with more complex inheritance patterns. The studies detailed in this thesis aims to provide diagnoses to individuals and their families for the underlying genetic causes responsible for the difficulties they experience and contributes to a long-running, non-profit community clinical-genetic research programme called the Windows of Hope (WoH).

Forming part of a wider Amish Hearing Loss Program the studies described in chapter three document the discovery of the genetic causes of hearing loss for eight Amish families. Through a combination of targeted gene sequencing, genome-wide SNP mapping and exome sequencing this study identified a variant in the Gap junction beta-2 (*GJB2*) gene, not previously reported in the Amish, as the cause of non-syndromic hearing loss in six families. Additionally, one family initially thought to be affected by a neurodevelopment disorder which included syndromic hearing loss, was found to possess two distinct genetic disorders; a 16p11.2 microdeletion, responsible for the developmental delay, and a homozygous *GJB2* variant, responsible for the hearing loss. Finally, this chapter proposes two novel hearing loss genes and details the functional work undertaken to assess the pathogenicity of one of these genes (*SLC15A5*). This work provided important diagnoses for many families and acquired significant information regarding the spectrum and frequency of hearing loss-associated gene variants across distinct Amish communities.

Chapter four details work undertaken to define the clinical phenotype and molecular basis of a novel complex autosomal recessive neurological disorder. Work undertaken by one of our collaborators, Dr Zineb Ammous, was instrumental in precisely defining the clinical phenotype of this disorder. A combination of genome-wide SNP mapping and exome sequence identified a sequence variant in Smad Nuclear Interacting Protein 1 (*SNIP1)*, which encodes

an evolutionary-conserved transcriptional regulator, as the likely underlying genetic cause. Due to its role as a transcription regulator whole transcriptome sequencing was undertaken to determine the impact of this gene mutation. This work provided important information regarding the specific biological role of *SNIP1* and identified gene expression pathways of direct relevance to the clinical phenotype, highlighting therapeutic approaches likely to benefit affected individuals. Additionally, this study determined that *SNIP1*-associated syndrome is one of the most common conditions across many Amish communities.

In recent years the WoH Project has accumulated extensive single nucleotide polymorphisms (SNP) and exome sequencing datasets from patients and individuals from the Amish community. Chapter five outlines a pilot, proof-of-principle study undertaken to explore this data with the aim characterising the architecture of the Amish genome. The interrogation of 26 exomes identified the presence of 12 pathogenic variants known to cause autosomal recessive (AR) diseases that have not yet been reported in the Amish but are likely to be present. Additionally, a PLEXseq sequencing approach was implemented to determine the prevalence of 165 pathogenic variants in 171 unaffected Amish individuals. The findings indicated diverse carrier frequencies within the different Amish communities and contributed to the consolidation of two genes responsible for ultra-rare inherited AR diseases (*CEP55*, *MNS1*). By developing approaches to improve knowledge of the specific causes of inherited diseases in the community, this work has laid the foundation for the development of a new genetic-based approach to diagnostic testing in the community.

This thesis, and the wider programme of work of Windows of Hope, occupies a privileged positioned at the interface between scientific research and clinical care. The findings described here have made a significant contribution to our understanding of the pathomolecular cause of a number of rare inherited disorders by increasing our knowledge of the nature and spectrum of inherited disease within the Amish laying the foundations to aid the future discovery of new disease genes and improving clinical outcomes by enabling focussed clinical diagnostic and management strategies to be implemented.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Disorder/ Phenotype	**Gene**	**Variant [GRCh38)**	**PLEXseq Data**			
			Region	AF	No.of Hets	No. of Homs
Hydranencephaly with renal aplasia-dysplasia	CEP55	c.514dup; p.Ile172Asnfs (NM_001127182) chr 10:g.93507042dup	Indiana	-	0	0
			Ohio Holmes	0.0221	3	0
			Ohio Geauga	0.0273	3	0
			Wisconsin	0.0200	1	0
			Total	0.0205	7	0
Situs inversus (SI) and male infertility	MNS1	c.407_410del;p.Glu136Glyfs*16 (NM_018365.2) chr15:g.56446887_56446890Del	Indiana	n/a	n/a	n/a
			Ohio Holmes	n/a	n/a	n/a
			Ohio Geauga	n/a	n/a	n/a
			Wisconsin	n/a	n/a	n/a
			Total	n/a	n/a	n/a
Psychomotor retardation, epilepsy, and craniofacial dysmorphism	SNIP1	c.1097A>G; p.Glu366Gly (NM_024700.3) chr1:g.37537842T>C	Indiana	0.0870	4	0
			Ohio Holmes	0.0441	6	0
			Ohio Geauga	0.0182	2	0
			Wisconsin	-	0	0
			Total	**0.0351**	**12**	**0**

CHAPTER 1

INTRODUCTION

1 Introduction

1.1 The Amish

1.1.1 A brief history of the Amish

The Amish are a distinct group of rural-living Anabaptist Christians whose heritage dates back to 16th century Europe. Setting out to revive the Anabaptist movement one leader, Jakob Amman, suggested the implementation of stricter practices including; more regular communions, forbidding the trimming of beards, the wearing of fashionable clothes and advocating the shunning of excommunicated members. The practice of shunning or “Meidung” was not intended to be a punishment but instead used to show to an individual that they needed to repent their sin. Shunning could be demonstrated in a number of ways including; avoiding the individual, refusal of goods from offenders or refusing to share a meal with them. In 1693 these proposals caused a schism within a group of Swiss Anabaptists in Alsace; those members choosing to follow Amman became known as the Amish.

Since the formation of the Anabaptist church, during the Protestant Reformation in 16th century Europe, the religious convictions of these individuals has created a social and cultural divide. Their rejection of infant baptism and belief in adult baptism, was one such division. The name Anabaptist itself means “rebaptisers” due to their practice of baptising adults who had previously been baptised as infants in a Catholic or Protestant church. The rapid spread of the Anabaptist movement and their request for a voluntary church separate from the state angered other religious leaders and civil officials. This unfortunately led to the

persecution of many Anabaptist followers and culminated in the social, cultural, and subsequently genetic, isolation of Anabaptists from the rest of the Europe.

In 1737, as a consequence of continued persecution, twenty-one Amish families migrated to the USA, originally settling in the Mid-west, particularly Pennsylvania, Ohio and Indiana. Between 1815 and 1860, there was a second wave of immigration resulting in approximately 3000 Amish individuals from Europe settling in the US.

1.1.2 Population bottleneck and the founder effect in the Amish

These migrations created an ancestral bottleneck (Figure 1.1Figure 1.1) which further reduced the genetic diversity of the already modest gene pool that has established the now large Amish population, estimated to be 308,000.

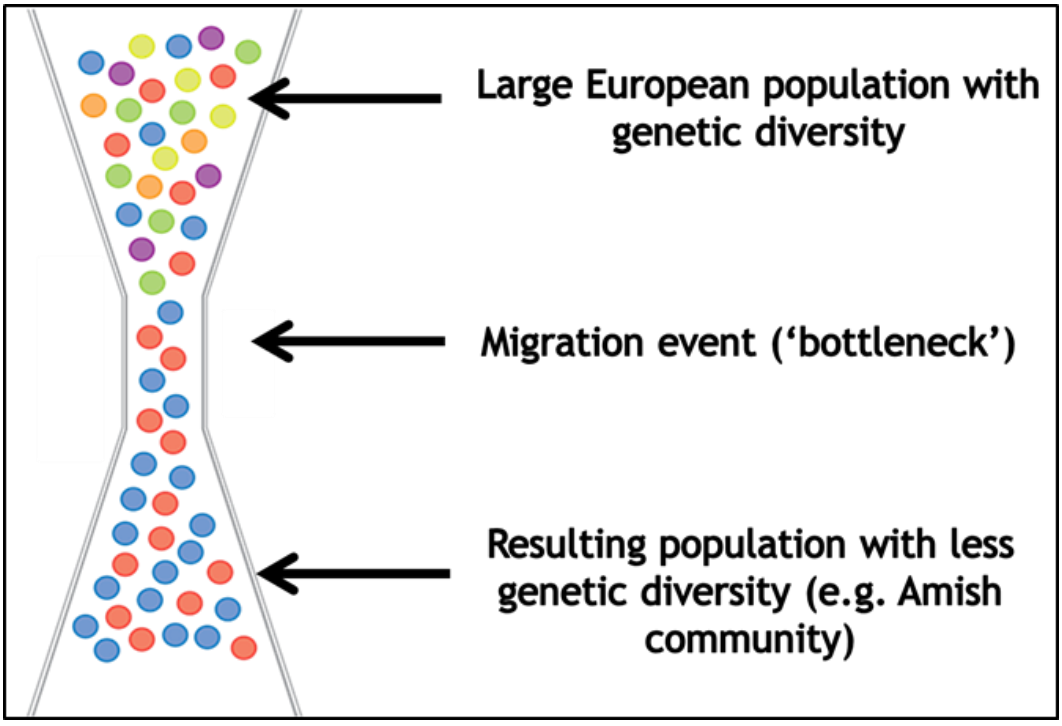


Figure 1.1: Ancestral bottleneck leading to the founder effect. Image hand-drawn.

As a result of a limited number of founder individuals the frequency of autosomal recessive (AR) alleles, present in these founders, may increase within that population meaning any two individuals, selected at random, from the community have an increased chance of being a carrier for a particular variant [1].

As with any community arising in this way, including other genealogically-related Anabaptist communities, such as the Mennonites and Hutterites, a number of causal genetic variants have become enriched within the population. This has unfortunately led to a higher incidence of particular genetic diseases, compared to the general population, caused by a phenomena known as the “founder effect”. Dwarfism (Ellis–van Creveld syndrome) [2], Angelman syndrome [3] and various metabolic disorders, including maple syrup disease [4] and phenylketonuria [5] are examples of genetic diseases that have become enriched within the Amish population.

Despite no longer being persecuted for their beliefs the Amish are still considered to be isolated, both culturally and geographically. The Amish are relatively immobile as a result of religious constraints on transportation (Figure 1.2). Very few members join their communities and the Amish advocate endogamous marriages, the practice of marrying within a specific social group rejecting those from others. This has resulted in little gene inflow into the population meaning that the current day gene pool is essentially the same as the original founders.

Figure 1.2: Amish horse and buggy transportation. Personal photograph provided by Professor Andrew Crosby, taken in Ohio, June 2014.

The Amish Population has increased dramatically since the initial migrations; expanding by 18% between 2011 and 2016 [6]. It is estimated that the population doubles approximately every 18-20 years. The driving force behind this growth can be attributed to the large nuclear families, frequently having more than five children, and the high retention rate of around 85% within the Amish faith. As a result of this growth the number of settlements (geographical communities) continues to increase with 138 new settlements being established from 2009 to 2018. This included six new settlements in Canada and surprisingly two settlements in South America located in Argentina and Bolivia [6]. The establishment of new settlements is not solely to accommodate the increasing population. New settlements can arise for a variety of reasons including; the availability of affordable farmland or non-farm work in specialised occupations, to achieve rural isolation that supports the Amish lifestyle in terms of social and

physical environments, to move closer to family or similar Amish church groups (demes) or to resolve church or leadership conflicts. Conflicts can arise due to differing opinions on the practices within a particular group or affiliation (collection of church districts that have similar practices).

As with any religious following there is diversity between the different groups within the Amish faith. One of the greatest sources of this variety is the differing views regarding the acceptance of technology. It is a common misconception that all Amish reject all forms of modern technology. Some affiliations choose to use battery powered lights and sewing machines where the more conservative decide to use kerosene lanterns and only permit the use of treadle (foot-powered) sewing machines. A more accurate description of the use of technology within the Amish is that is it used selectively in a manner that preserves their traditional way of life and prevents the introduction of foreign values into their communities. For these reasons many Amish groups reject technology that enables access to the mass media including the use of televisions, radios and the internet.

There are four groups that carry the Amish name: Beachy Amish, Amish Mennonites, New Order Amish and Old Order Amish. Although practices between these groups display large amounts variation they can largely be separated into two distinct classes. The Beachy and Mennonite Amish own automobiles and use public utility electricity whereas the New and Old Order Amish use horse-and-buggy transportation and do not use public utility electricity. All families included within this study are from New or Older Order Amish communities.

1.1.3 Amish demes

One aspect of this study involved undertaking the first large scale investigation of carrier frequencies within different Amish demes for the most commonly occurring pathogenic variants seen in the Amish. These studies have the potential to expedite diagnostic testing for families by helping identify the most likely causative genetic variants based on the community in which the affected individual belongs.

Due to the circumstances under which new demes are typically formed (discussed above) each deme included in this study occupied a different geographical location. We hypothesised the allele frequencies observed in each region would reflect the ancestral histories and migration patterns of each deme. For this reason, the allele frequency data was grouped depending on the region in which an individual originated. This gave rise to four cohorts within our allele frequency data; Indiana, Ohio Holmes County, Ohio Geauga County and Wisconsin.

1.2 Other population isolates

Population isolates can arise as a result of a founder effect, where a new population is established by a (relatively) small number of individuals from a larger population, or from the extreme reduction in size of a population resulting in a genetic bottleneck [7]. The Amish are not the only founder population that display an enrichment of particular diseases. Other founder populations include the Finns, the Samoans, the Orkney islanders and the Ashkenazi Jews.

1.2.1 The Finnish population

The Finnish population has been a target of extensive genetic studies since the 1950s [8]. In 1973 a “landmark paper” describing 10 'Finnish' disorders was published and coined the term 'Finnish Disease Heritage' (FDH) to describe a group of rare hereditary diseases that are overrepresented in Finland [8, 9]

The unique Finnish genetic architecture is explained by repeated population bottlenecks that have subsequently given rise to the large current day Finnish population of 5.5 million [10] from a small founder population. The initial founder effect arose as a result of two waves of colonisation that took place between 4000-2000BP (years before present) in southern and western Finland. These areas remained populated at a low density until internal emigration occurred in the 15-16th century when small family groups, from the original colonisations of southern Finland, moved into the northern and eastern areas [10, 11]. These movements formed sub-isolate populations, isolated through distance, that are believed to be behind the presence of more than 35 recessive monogenic illnesses including Finnish nephropathy (*NPHS1)* and cartilage hair hypoplasia (*RMRP*) [9], that are more commonly seen in the Finnish population, particularly in eastern Finland [11]. These movements are also assumed to explain the

exceptionally low prevalence of other diseases such as cystic fibrosis, with an incidence a 10th of that in other parts of Europe, and phenylketonuria, which was reported at a carrier frequency of 1:180 [8].

Like the Amish, the Finnish population is well suited for gene mapping studies due to its reduced diversity and increased homogeneity. However, studies have reported substantial differences in genetic composition between different parts of Finland [11].

1.2.2 The Samoan population

It has been suggested that the Samoan populations of the Western Pacific are one of the best examples of an isolated population given the archaeological, cultural and linguistic evidence supporting the long-distance migration undertaken by the ancestors of today's islanders [12, 13]. Samoan settlement is believed to have occurred 5000-4000BP after rapid migration from Southern China [12]. After ~1000years (3000BP) island culture was reportedly thriving and supporting an estimated population of 100,000 to 300,000 people [12, 14]. After European contact in the 1700s Samoa suffered significant population declined, attributed to the introduction of disease [12]. By 1900 the population was estimated to be as low as 30,000 people. After numerous epidemics throughout the 18th century the population recovered to an estimated 69,000 by the 1940s [12]. These historical events are likely to have influenced the Samoan genome.

Over the last 35 years there has been a documented rise in the prevalence of a number of non-communicable diseases, including Type 2 diabetes mellitus (TD2) and cardiovascular disease (CVD) in addition to high levels of adiposity (obesity) [14, 15]. Although obesity is a complex phenotype with genetic and environmental factors impacting its presentation the relative isolation, large family sizes and

recent exposure to modernization of the Samoan population provide a unique opportunity to identify novel genetic contributors to obesity [14, 16].

In the last decade a number of susceptibility loci for obesity have been identified within the Samoan population [14]. More recently one study has identified a variant (NM_153607.2:c.1370G>A, p.Arg457Gln) in the *CREBRF* gene associated with an extreme increase in body mass index (BMI), very rare in other populations but common in Samoans, that selectively decreases energy use and increases fat storage in adipocyte cell studies [16, 17]. Interestingly there is evidence that this variant has been positively selected in Samoan genomes, supporting the “thrifty” variant hypothesis proposed by James Neel in 1962, suggesting that this variant may historically have been an asset to the population during historical periods of “feast-or-famine” [18].

1.2.3 Orkney Island population

Located in an isolated position off the northern coast of Scotland, geographic distance has acted as a barrier to migration resulting in minimal movement in to and out of the Orkney Islands [19]. This has impacted the genetic structure of the population and led to individuals displaying higher levels of genetic similarity than would be expected in a non-isolated population [20]. This relatively low genetic variability, in addition to the comparatively low environmental variability, means this population is well suited to identify risk factors of diseases found within the community at a higher than expected prevalence [19].

Over 30 years ago high rates of multiple sclerosis (MS) were reported in the Orkney Islands and northern Scotland. MS, a complex inflammatory autoimmune disorder, is the most common disabling neurological disorder in young adults [21] with strong evidence suggesting both genetic and environmental risk factors

effect disease susceptibility [19]. Recent epidemiological studies have shown that the prevalence of this condition on the island has continued to increase with Orkney now reporting the highest prevalence rate of MS in world [22, 23]. As a result the communities of the Orkneys have been involved in a number of studies investigating the role of both genetic [19] and environmental [21] factors on the presentation of the disorder.

1.3 Genetic studies in the Amish

In 1962 the first genetic studies of the Amish were undertaken by Victor McKusick, a prominent figure in the formation of the medical genetics field, investigating dwarfism within the community. These investigations were based on ideas he had developed from reading an article written by David Krusen, a local family doctor, and a manuscript entitled “Amish Society” submitted by John Hostetler. The article mentioned achondroplasia was extremely common in the Amish and the manuscript highlighted the many characteristics of the Amish community that are advantageous to the study of genetic traits (Table *1.1*) [24].

Table 1.1: Advantages of Studying Genetics in the Amish. Taken [24]

Advantages of Studying Genetics in the Amish
The Older Order Amish are a self-defined population
It is a closed population; gene flow is almost exclusively centrifugal
The Western European origins of the population are well known
Extensive genealogic records
The standard of living is high
The standards of medical care are relatively high
An evident interest in illness (and its cause)
There is a high coefficient of inbreeding due to the relatively small number of founder couples
The illegitimacy rate is apparently low
The Amish are interested and knowledgeable about the health of their relatives. They seek out information on rare disorders shared by other Amish families
Socio-economic and occupational circumstances are notably uniform
Because of constraints on transportation, the Amish are relatively immobile
Most Amish families are large, with an average of seven to nine children
Children with birth defects or genetic disorders are usually kept at home rather than institutionalized
The existence of several Amish isolates makes comparisons of sub-populations possible

Taken together these factors greatly facilitate the discovery of genes responsible for inherited disease, which might otherwise have been impossible in studies of

other populations due to the genetic and environmental complexities of a condition.

In 1962 McKusick was approached by a young doctor, Dr Harold Cross, who was raised in an Amish community in Indiana and wished to undertake a PhD. The resulting studies focused on neurological conditions in the Ohio Amish and produced publications describing a number of disorders including Troyer syndrome [25] and Mast syndrome [26] two novel forms of complex hereditary spastic paraplegia. However, at the time, it was not possible to identify the causative genes for these novel conditions.

The work of McKusick in the early 1960's led to the development and publication of the Mendelian inheritance in Man (MIM). First published in print in 1966 as a "comprehensive knowledge base of human genes and genetic disorders" [27] it contained a trilogy of catalogues detailing autosomal dominant (AD), autosomal recessive (AR) and X-linked phenotypes. Twelve print editions of this compendium were released between 1966 and 1998. These print versions have been superseded by an online version, the Online Mendelian Inheritance in Man (OMIM) which was first made available in 1987. Since 1995, and its distribution by the National Centre for Biotechnology Information (NCBI), it has been updated daily and become a critical, frequently used resource for anyone involved in the field of medical genetics [28].

Professor Andrew Crosby, who has a particular interest in neurological conditions, contacted Dr Cross in 2000 to suggest a collaborative research project with the aim of using genetic technologies, developed since the original

study, to identifying the underlying genetic causes of these conditions. These studies successfully identified the causative genes for both Troyer syndrome (*SPG20*) in 2002 [29] and Mast syndrome (*SPG21*) in 2003 [30].

As a result of these successful investigations and realising the potential impact to the Amish community Crosby and Cross established the Windows of Hope Project (WoH) [31] which conducted the largest survey of inherited conditions amongst the Amish communities. In partnership with the non-profit Amish-led Windows of Hope Genetic Information Centre (WHGIC) the project has made a significant contribution to determining the underlying molecular causes and clinical manifestations of inherited diseases amongst the Ohio Amish community.

1.3.1 Windows of Hope Project

The Windows of Hope Project is a long-running, non-profit community genetic research program which has undertaken the largest survey of inherited conditions within Anabaptist communities. By working closely with the Anabaptist communities and local healthcare and special educational needs providers in both Ohio and Wisconsin the group has made significant progress in determining the molecular causes and clinical manifestations of inherited diseases. To date it has assisted in the identification and description of the genetic cause of over 30 inherited disorders, including 16 novel conditions being described as a direct result of their work [31]. Almost all of these disease genes, initially identified in the Amish, have subsequently been found worldwide in other populations causing similar diseases. Highlighting the global significance of studying inherited conditions in genetic isolates such as the Amish.

The high number of undiagnosed childhood development disorders among the Holmes County Amish population places a significant social and financial burden

on the community. Whilst community schemes are available to help with medical expenses, the rising costs of medical care means many Amish families are not supported so do not undergo important clinical investigations or receive treatment.

The primary aim of the WoH project is to utilise modern genomic technologies, combined with clinical expertise, to advance the understanding of inherited disease within Anabaptist communities, improve healthcare outcomes and provide substantial cost savings for families affected by these conditions. Genomic findings from research studies enable these improved outcomes by empowering molecular diagnostic services, enabling focused disease-specific clinical management strategies and the development of new treatment strategies. The WoH undertakes studies of a wide range of inherited conditions including neurological, cardiovascular and developmental disorders. Clinical and diagnostic laboratory collaborations, supported by the WoH, have assisted families in receiving much-needed diagnoses for previously unrecognized conditions.

The WoH Project now has a mature infrastructure including an Amish-led regional centre, WHGIC located in Ohio, the second largest Amish settlement in the US. The centre is a source of information and support for local families which, through collaborations with clinical partners, has developed an extensive community-appropriate educational programme including family information days and disease-specific leaflets, and practical educational symposia. In addition to this the WoH has established an extensive searchable clinical online database to provide information to medical practitioners about all the conditions currently known to be present in the Plain communities [31].

1.4 Genomic technologies

The advent and widespread availability of next-generation sequencing (NGS) has decreased the cost and associated timelines of genetic and biological research [32]. A significant number of studies undertaken by the WoH project within the Amish have utilised various high-throughput NGS techniques including autozygosity mapping, in combination with whole exome sequencing (WES) to identify causative pathogenic variants.

Outlined in this project is the first time the WoH project has employed the use of the PLEXseq process, to determine the allele frequencies of the most commonly occurring pathogenic variants seen within the various Amish communities.

1.4.1 Autozygosity mapping

Rare recessive mutations are predicted to occur in every population but are unlikely to achieve homozygosity in the general population. However, unions between closely related (consanguineous) or distantly related (endogamous) individuals dramatically increases the probability of resultant offspring being homozygous at any given genetic locus [33]. Endogamous and consanguineous unions often result in autozygosity, a special form of homozygosity, where two copies of a section of DNA, shared by two or more people, are identical by descent (IBD) as they have been inherited from a common ancestor, without any intervening recombination (Figure 1.3) [34].

The use of ‘autozygosity mapping’ to define the chromosomal location of a disease locus was first proposed in 1987 using consanguineous families based on the principle that affected individuals will be IBD for the disease causing variant and surrounding haplotype [35]. Over the last 30 years this approach has

become a powerful method for identifying recessively inherited disease genes within genetically isolated populations, such as the Amish.

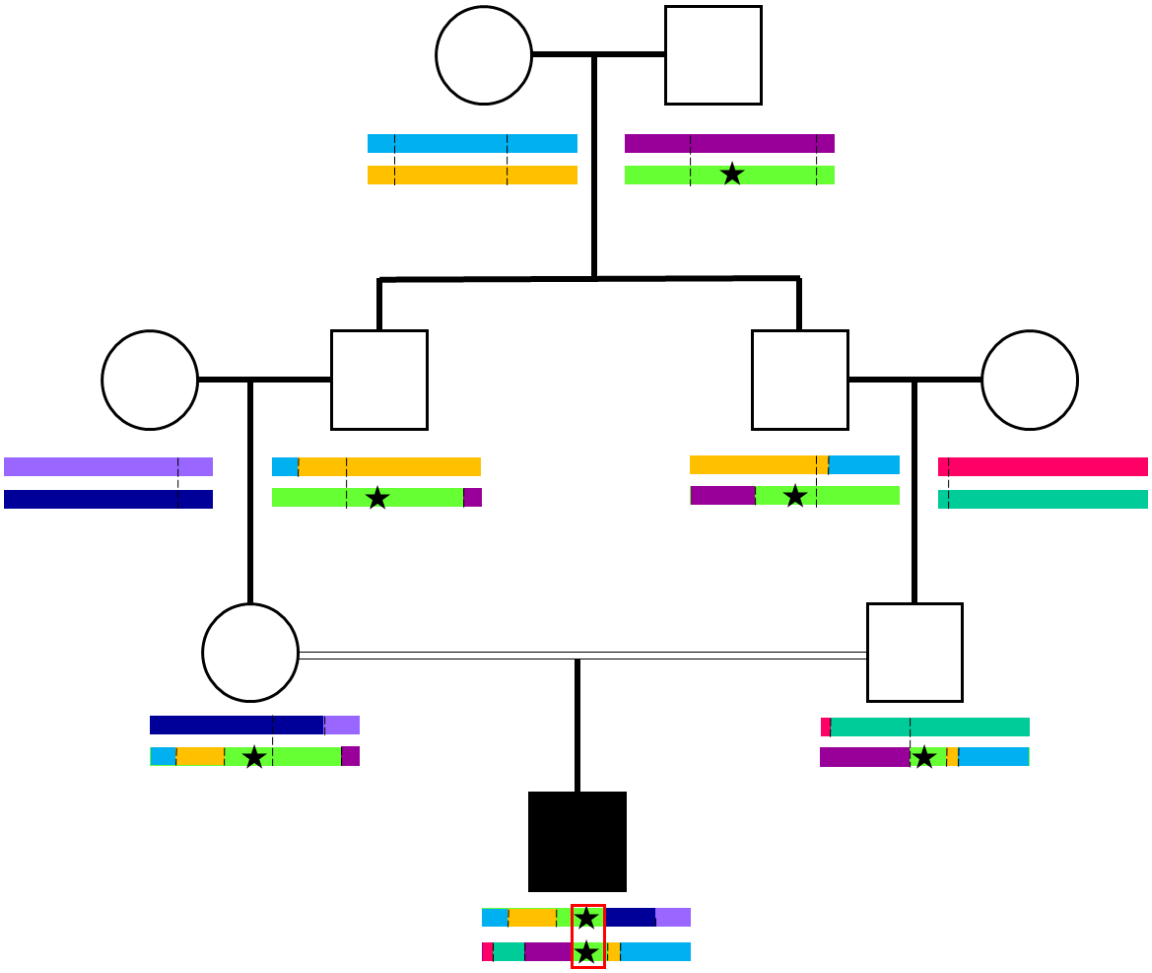


Figure 1.3: Simplified schematic presentation of the concept of autozygosity. An ancestral haplotype (red box) containing a variant of interest (★) is transferred through the generations. In each generation different haplotypes enter the pedigree represented with a different coloured bar. Recombination events (shown by dotted lines) in each generation shorten the size of the haplotype. (Modified from [36] and [34]).

Advances in the experimental techniques used to generate and analyse data has dramatically increased the speed of detecting autozygous regions [37]. Early methods used highly polymorphic microsatellite, or small tandem repeat (STR), markers to identify autozygous regions. Despite these markers being more powerful at detecting homozygous chromosomal segments than single-

nucleotide polymorphisms (SNPs) the increased time, cost and inability to identify smaller autozygous sections, due to the increased distance (10–12cM) between markers, led to them being superseded by SNP microarrays [38].

SNP microarrays are silicon chips (SNP-chips), originally designed by Affymetrix and Illumina that detect SNPs across the whole genome in a single hybridisation reaction making them much more time and cost efficient than microsatellite analysis. Although each SNP, as mentioned previously, is far less powerful at detecting a homozygous region than a microsatellite marker, they offer a number of other benefits that have encouraged their increased and extensive use.

SNP-chips offer greater coverage of the genome, due to their increased number. For example an early chip containing 10,913 SNPs is reported to be equivalent to a 3–4cM microsatellite marker map [39], enabling the identification of smaller autozygosity regions. The increased number of SNPs also enables the detection of heterozygous regions more effectively than with microsatellite markers. A single average microsatellite marker was projected to have a 70% chance of detecting heterozygosity. A genomic region of the same size on a SNP-chip containing approximately 30 SNPs would have a 99% chance of detecting a heterozygous region [39].

More recently it was suggested that exome sequencing could be used to concurrently define autozygous regions and identify possible causative variants. However, initial investigations found poor coverage, compared to SNP-chip genotyping, due to the uneven distribution of coding regions across the genome [40] and suggested that shorter autozygous regions could be missed should they be located within in gene-poor regions [37].

Today high-density and high-resolution SNP-chips, available from Affymetrix and Illumina both containing more than 1 million genetic markers (Affymetrix SNP array 6.0 and Illumina 1M respectively), enable the detection of even the smallest, (54kb) structural changes [41, 42]. Copy number variation (CNV) is a type of genetic variation that is widely found in mammalian genomes and includes genomic deletion and duplication as well as complex rearrangements that range from 100 base pairs to several mega base pairs in size [43]. CNVs have been shown to have a significant impact on complex human diseases, such as autism [44] and cancer [45], due to the fact that they can disrupt gene structure and affect gene regulation. However, not all CNVs are linked to adverse phenotypes. To date, approximately 552,586 CNVs are included in the Database of Genomic Variants [http://dgvbeta.tcag.ca/dgv/app/home] [46]. A recent study mapping CNVs, not associated with disease, found around 100 genes that can be homozygously deleted without producing an adverse disease phenotype and estimated that up to 9.5% of the human genome contributes to CNV [47].

These high-density SNP genotyping array platforms target biallelic SNPs. For each SNP, an array platform includes two types of hybridisation probes specific to two types of known alleles, usually coded as A and B, and the SNP genotype can be determined by the ratios of the hybridisation intensities for A and B probes. CNVs such as duplications and deletions increase or decrease the total measured intensities. Large CNVs, that span multiple SNPs, have intensity ratio patterns distinct from normal genomic regions (Figure 1.4).

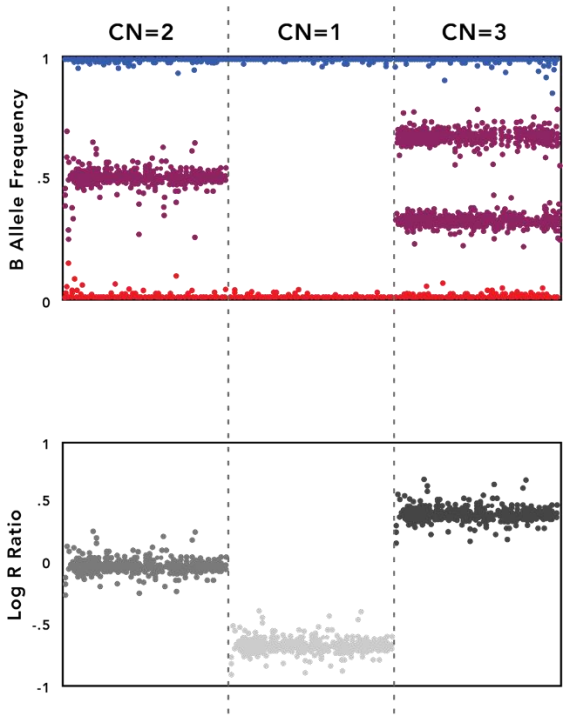


Figure 1.4 CNV Analysis of Illumina Bead Chip Data. B allele frequency and log_2 R (normalised signal intensities) ratio are plotted over the entire genome for all SNPs. The plot exhibit diagnostic signature profiles of copy number. CN=2, 1; & 3 shown (adapted from [48]).

The number of SNP-chips now readily available allows users to select the array most suited to their research needs. The Illumina HumanCytoSNP-12 Bead Chip array was used in all studies outlined in this project. This platform permits the processing of up to 12 samples in parallel which increases sample throughput, to a level appropriate to the number of samples required for our studies, and decreases experimental variability [49]. It offers a low cost per sample, compared to other methods, and only requires 200ng of DNA per sample which is easily achievable when extracting DNA from whole blood lymphocytes and possible, though more variable, from extracting DNA from buccal swabs. Additionally this SNP-chip incorporates ~300,000 SNPs, offering dense coverage of ~250 disease regions shown to be important for detecting cytogenetic abnormalities most relevant to human disease.

Whilst it is generally accepted that SNP-chips are the best way to generate the data required to calculate regions of homozygosity (ROH) within samples and produce autozygosity maps there is currently a lack of consensus regarding how ROHs should be defined [50].

Although there are several computational methods for identifying ROHs within a dataset they can all be classified as either genotype-counting or model-based.

Genotype-counting

Genotype-counting software, such as PLINK [51], GERMLINE [52] or cgaTOH [53], all search for long, consecutive runs of homozygous genotypes that occur within a set of predefined parameters including the maximum number of heterozygous calls within a given region and the number of allowable missing genotypes.

PLINK v.19, for example, uses a sliding window approach. This is where an algorithm scans each chromosome by moving a fixed sized window along the whole genome searching for consecutive homozygous SNPs [54]. To identify ROHs the location of each SNP is considered. This is achieved by calculating the proportion of completely homozygous windows that incorporate each SNP. If this proportion is higher than a defined threshold, a given SNP appears in more consecutive homozygous windows than expected, the SNP is considered to be in a ROH. The simplicity of this method permits large amounts of SNP data to be analysed efficiently [54].

Model-based

Model-based software, such as BEAGLE [55], FILTUS [56], BCFtools/RoH [57] and GARLIC [58], use probability to differentiate between autozygous and non-

autozygous regions using the allele frequency of SNPs and recombination rate estimated from the data. Each of these methods utilise different algorithms which determines the achievable sensitivity, specificity and false positive rate of the resulting analysis. However, all current model-based methods use statistical modelling in the form of hidden Markov models (HMMs) to account for background levels of linkage disequilibrium [54].

HMMs were first introduced to computational biology during the late 1980s and are a class of probabilistic models [59] that represent probability distributions over linear sequences [60] and offer a consistent mathematical basis for assigning position-specific residue scores [61]. The central idea behind a HMM is that it is a finite model describing the probability distribution across an infinite number of possible sequences [61]. Being described as the “Legos of computational sequence analysis”, HMMs are now central to a diverse range of analysis programmes including; multiple sequence alignment, gene finding and regulatory site identification [62].

In addition to the lack of consensus in how ROHs are delineated there is also a lack of standardisation regarding the sizing of ROHs which can be expressed using either a physical distance, in kilobases (kb) or megabases (Mb) or a genetic distance expressed in centimorgans (cM). A centimorgan is a measure of recombination frequency, with 1cM corresponding to a recombination frequency of 1% [63]. Whilst it has been suggested that the use of genetic distances to describe ROHs is preferable due its capacity to mitigate the effect of linkage disequilibrium (LD) [64], many studies still describe ROHs using a physical distance.

LD refers to the non-random association of alleles at two, or more, genetic loci [65]. Loci are said to be in LD when the frequency of association of their different alleles is higher or lower than what would be expected if the loci were independent and associated randomly. The degree of LD varies greatly across the human genome. Regions that have undergone little to no historical recombination, displaying high LD, are often referred to as “haplotype blocks” [66] and are commonly bordered by recombination hotspots [64]. As a result of these different regions it is not possible to accurately compare ROHs located in different genomic locations based solely on physical distances.

A population that has accumulated many recombinations at every position in the genome would display little LD as there would be no correlation between the inheritance of alleles at any particular loci [67]. The amount of LD between two alleles is related to the time of the mutation event, the genetic distance between alleles and the history of the population in which they are located. Relatively young (<2000 years) populations arising from a small founder population, such as the Amish, tend to display higher levels of LD [67]. This means the genomes of individuals from this population are more likely to contain longer ROH, or haplotype blocks, compared to the general population.

The ROHs in the studies outlined in this project where measured in physical distances and determined using an in-house genotype-counting method which allowed for regions to be confirmed manually, taking into consideration potential miscalled or missing genotypes. Regions >1Mb in size were preferentially interrogated, through cross-referencing with exome sequencing data.

1.4.2 Whole exome sequencing

Since the completion of the Human Genome Project (HGP) in 2001, at a cost of ~$3billion and in excess of 200 scientists working for over a decade [68], substantial improvements have been made in the approach to genome sequencing [32]. Current methodologies are far less laborious and offer significantly cheaper costs per sample when compared to BAC (bacterial artificial chromosome) based sequencing, used by the HGP, and first generation sequencing techniques such as chemical sequencing [69] and dideoxy chain terminator, “Sanger”, sequencing [70]. Second generation, or next generation sequencing (NGS) technologies implement massively parallel sequencing (MPS), the simultaneous sequencing of multiple variants within multiple samples, of short-read lengths of DNA (50–500bp) which are amplified then assembled, by alignment to a reference sequence, using a bioinformatic pipeline [71, 72]. The high-throughput nature of these short-read technologies has enabled the cost of sequencing, per megabase, to reduce at a rate exceeding that of Moore’s Law (which proposed technology reduces microprocessor costs by half every 18 months) [73]. Recently the lowest cost of whole genome (WGS) and whole exome (WES) sequencing was estimated to be $1906 per genome and $555 per exome, which has facilitated their increased application within both a research and clinical setting [74]. There are now several types of NGS tests available for use in a clinical setting; exome, genome, and panel NGS, which offer varying degrees of genome coverage [75].

Despite WGS investigating sequence changes such as; single-nucleotide variants (SNVs), insertions and deletions (indels), chromosomal rearrangements and copy-number variation (CNVs), across the whole genome, WES is reported

to be the more popular technique [76] as it covers the “more actionable areas of the genome” [74].

Protein coding genes only constitute ~1% of the human genome [76] but contain 85% of disease-causing mutations responsible for Mendelian disorders [77] which result from the mutation of a single genetic locus. As WES only targets 95% of the coding regions, or exons, of protein coding genes [77] across the genome it requires less sequencing space, allowing more samples to be analysed, and produces less, more interpretable data at a fraction of the cost of WGS [68, 76]. Even with the rapid and significant improvements in sequencing technologies identifying exonic variants that affect phenotypic expression through WGS is approximately four times more expensive than an exome sequencing approach [68, 74].

Since its first successful use to diagnose and inform subsequent treatment of an infant patient with a rare form of inflammatory bowel disease [68] WES has been instrumental in revolutionising our understanding of rare and common human diseases and supporting the implementation of health-improvement projects throughout the world [77, 78]. With the increasing interest and drive towards personalised medicine the development of efficient targeted sequencing strategies is likely to continue. Currently WES methodologies can be classified in two ways; solution or array based [68].

Array-based methods where the first used to enrich specific regions of the genome [79] and sequence a whole exome [80]. They involve the hybridisation of randomly sheared, adapter-ligated genomic DNA (target sequences) to synthetic oligonucleotides (probes) bound to a high-density microarray [81]. An additional array-based method, multiplex amplification, was also proposed that cleaved off and amplified, through a polymerase chain reaction (PCR), the

oligonucleotides synthesised on the microarray to perform a padlock and molecular-inversion reaction [82]. However, this method was initially reported to miss more than 80% of targeted exons, represent sequence targets unevenly and showed poor reproducibility between replicates [81]. Since these methods were introduced, increasingly powerful sequencing techniques requiring smaller amounts of template DNA and involving less manual work where in high demand [83].

A solution-based method, proposed by Gnirke *et al*. in 2009 overcame some of the short-comings of previous methods utilising a hybrid-selection method for enriching specific genomic regions. It combines the robust performance of oligonucleotide synthesis on an array with the favourable kinetics of RNA-driven hybridisation in solution [81]. The first commercially available sample preparation kit using this method was the SureSelect Human All Exon capture kit (Agilent) closely followed by the NimbleGen with the SeqCap EZ Exome capture system (Roche). A systematic comparison of these two platforms, using the same Illumina sequencing machine and bioinformatics pipeline to annotate the sequences, found the NimbleGen kit aligned more accurately to the target regions whilst the Aligent kit had less duplicated reads. Alignment of the Aligent kit to the human reference genome was equal to that of the NimbleGen kit with neither kit capturing all of the consensus coding sequence (CCDS) exons [83].

Available WES capture methods are constantly being improved and updated. Solution-based kits are currently the most commonly used with improvements to these platforms focusing on increasing the read depth, the number of aligned sequencing reads covering a specific genomic position [84], the coverage, the average raw or aligned read depth and the breadth of coverage, the percentage

of target bases that are sequenced a given number of times [85], increasing both sensitivity and specificity.

Two solution-based WES platforms where used in this study; the SureSelect Human All Exon V4 (Agilent) exome enrichment kit sequenced on an Illumina HiSeq2000 sequencer and the BGISEQ-500 sequencing system both obtaining mean read depths and breadth of coverage sufficient for accurate variant calling for clinical purposes.

Current technologies now generate accurate and reliable sequencing data covering the majority of the genome [75], removing the historical “sequencing bottleneck” [86], where the sequencing of genetic variants was the most time, labour and cost intensive aspect of providing a genetic diagnosis. This has allowed these technologies to be widely integrated into clinical settings but has posed several new challenges including [87];

- The setup, validation and implementation of appropriate bioinformatic analysis to accurately determine genotypes
- The standardisation of variant interpretation and classification
- The development of polices and guidelines to inform the identification and disclosure of secondary variants (incidental findings) not directly linked to the patient’s phenotype under investigation
- The storage, accessibility and dissemination of sequencing data

Quality control (QC) is an essential step in the analysis of sequencing data to ensure accurate genotyping when defining a variants pathogenicity. It can be difficult to determine a genotype due to errors introduced in the base-calling process, which can vary significantly across different sequencing platforms. In addition genotype calls are extremely dependent on the achieved read depth;

only genotypes with many reads can be reliably assigned [88]. A Phred, or Q, score (**Error! Reference source not found.**) is calculated using several predictors of possible errors and widely used by all major sequencing platforms to measure the probability that a base has been called incorrectly [88].

Table 1.2: Phred quality scores with associated Q score. Q scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
Q10	1 in 10	90%
Q20	1 in 100	99%
Q30	1 in 1000	99.9%
Q40	1 in 10,000	99.99%
Q50	1 in 100,000	99.999%

Once the reliability of the genotyping calls has been calculated the next important step of interpreting sequence variants is to determine their frequency in large population databases. The Genome Aggregation Database (gnomAD) is one such database which has complied the data of 125,748 exomes and 15,708 genomes from human sequencing studies [89]. Pathogenic variants are expected to occur at extremely low frequencies or not have been previously observed in the general population. It has become widely accepted that for a variant to be considered rare it must have a frequency of <1% in the general population [71]. However, it has been reported that genetic studies show a high degree of population bias with a greater representation, ~80%, of participants being of European decent which should be considered when interpreting variants from underrepresented populations [90, 91].

Good quality variants occurring at a low frequencies can then be analysed through the use of *in silico* prediction tools such as SIFT [92], PolyPhen-2

(Polymorphism Phenotyping v2) [93] or MutationTaster2 [94] to establish if the variant is likely to alter protein function through disruption of the amino acid sequence. Whilst it is possible to undertake these investigations in isolation, investigating all variants from an individual's exome sequencing data via a number of tools, would be far too laborious. It is far more common for the VCF files which is the standard bioinformatics format for storing gene sequence variations, containing WES and WGS data to be annotated using a software application that integrates several prediction tools. In these studies the Alamut Software Suite (v1.4.4) was used to assess variants so that a small number of candidate variants could be selected to undergo further interrogation. Alamut complies with ACMG/AMP variant interpretation guidelines [95] and as a result is used extensively within clinical settings.

1.4.3 The future of sequencing technologies

NGS technologies have dramatically reduced the cost of DNA sequencing increasing its accessibility [96] and making the use of gene panels, WES and WGS in clinical diagnostics possible [72]. Despite this, the short-read sequences used by these techniques are not without issue. The dependence on clonal amplification and creation of clusters of DNA molecules requires read lengths to be short (50-500bp), providing an opportunity for errors in base incorporation to occur thus increasing noise within samples [72]. Furthermore these short lengths then require extensive assembly which can cause difficulties for complex regions, particularly those containing a high number of repeated sequences by producing misalignments or misassemblies and impairing the phasing of variants [71, 72].

In 2011, Pacific Biosciences (PacBio) released the first commercially available third-generation sequencing (TGS) technology employing single molecule real-

time sequencing (SMRT) [97] which has proved to be useful in sequencing extended repetitive regions of the genome [71].

The key differences of TGS is that sequencing is done in real-time, unlike NGS where sequencing is paused after the incorporation of each base, and that it utilises long-read technologies which are reported to be revolutionising genomics research [98].

One such example is the development of nanopore sequencing introduced in 2014 by Oxford Nanopore Technologies (ONT) [99] which identifies nucleotides by measuring their electrical conductivity as they pass through the nanopore membrane. In addition to its innovative sequencing chemistry an attractive feature of this sequencer is its incredibly small size and USB port connectivity making it the first fully portable DNA sequencer [71].

Alongside the continued development and innovation in sequencing technologies there is growing interest in the increased incorporation of artificial intelligence (AI) platforms into clinical diagnostic practices. Being employed initially to cut costs associated with analysing the ever increasing volume of patient data by accelerating the annotation and prioritisation of sequence variants from WES. Whilst is it likely to be some time before AI technologies are commonplace in mainstream medicine a number of companies are already deploying aspects of AI technologies through clinician-friendly web-based interfaces that support the clinical prioritisation of variants [74].

1.5 Aims

The studies detailed in this thesis can be summarised by two distinct yet connected aims. The first aim was to identify the underlying molecular cause of developmental conditions within Amish communities which included; investigating congenital forms of hearing loss in multiple families to provide genetic diagnoses and defining the clinical phenotype and molecular basis of a novel complex autosomal recessive neurological disorder. The second aim was to conduct a pilot, proof-of-principle study to characterise the architecture of the Amish genome which included; determining the carrier frequencies of pathogenic and potentially pathogenic variants known to be present in the Amish communities in addition to identifying potentially pathogenic variants known to cause disease but yet to be reported in the Amish community. In order to meet these aims the following objectives where pursued:

- To undertake in-depth genetic studies, including autozygosity mapping, traditional and next generation sequencing technologies, to identify the underlying molecular cause of congenital hearing loss in two families with multiple affected individuals.
- To functionally characterise putative disease genes and elucidate the effect of the pathogenic variants identified.
- To investigate the allele frequencies of variants known to cause hearing loss in different Amish communities (demes).
- To interrogate exome sequencing datasets to identify coincidentally carried, potentially deleterious, autosomal recessive variants found in genes known to cause disease but yet to be seen in the Amish community.
- To utilise a PLEXseq sequencing approach to determine the prevalence of variants associated with disease and present in the Amish communities.

CHAPTER 2
MATERIALS & METHODS

2 Materials and Methods

2.1 Family recruitment and sample acquisition

2.1.1 Recruitment to the Amish Windows of Hope (WoH) project

All the WoH project studies were reviewed and approved by the Institutional Review Board of the Office for Responsible Conduct of Research, University of Arizona (Tuscon, Arizona, USA) (reference 10-0050-01) and by the University of Exeter Research Ethics Committee (reference 14/04/048).

Research was carried out in compliance with the Code of Practice for Human Tissue and Research (code E) provided by the Human Tissue Authority (HTA), which defines human tissue as relevant material consisting of, or containing cells, therefore includes blood and buccal samples. All blood and buccal samples, and subsequent DNA extractions, used in this project where used and stored in HTA-licensed premises with research carried out in accordance with the Human Tissue Act 2004.

Recruitment to the WoH project requires submission of the appropriately signed consent, clinical details and a blood or buccal sample. Signed consent is given, in accordance with the HTA's code of practice, only when individuals, or parents of individuals, feel they are sufficiently informed, about the purpose of the research, how their samples are to be stored and used and satisfied with the purpose of the research in which they are to be involved.

2.1.2 Phenotyping of affected individuals

Full medical and family histories for the purposes of the research study was obtained for all individuals. Individuals were examined with clinical phenotypes assessed and described by a member of our research group, Dr Emma Baple, Consultant Clinical Geneticist at the Royal Devon and Exeter Hospital. Phenotypic data for each family was then collated and reviewed for the purposes of the studies outlined in this thesis.

2.1.3 Data management

On receipt of blood or DNA samples each sample was assigned a sample ID. The tubes containing the samples were anonymised and labelled with the relevant sample ID. The clinical and molecular information was recorded alongside the samples ID in a password protected database.

Family pedigrees are constructed using the online Swiss Anabaptist Genealogical Association (SAGA) database (www.saga-omii.org).

2.2 Molecular DNA methods

All general-purpose chemicals were acquired from Fisher Scientific. All primers where supplied by IDT (Integrated DNA Technologies). Specialist kits, chemicals and consumables purchased from alternate sources are noted in the text where appropriate. Components of solutions made in-house are detailed in Table 2.1.

All plastic ware was acquired from StarLabs or Sarstedt. Kits for DNA extraction where purchased from Promega.

2.2.1 Buffers, Reagents and Stock Solutions

Table 2.1 Buffers, reagents and stock solutions required for molecular DNA techniques and their constituents.

Solution	Constituents
Agarose loading buffer	40% (w/v) ficoll 0.2% (w/v) xylene cyanol 0.1% (w/v) bromophenol blue
ExoSAP	For 1 millilitre: 50 U/ml Exonuclease I, 50U/ml shrimp alkaline phosphatase (both bought from New England BioLabs), ddH_2O to final volume
50X LAB	5.1% (w/v) lithium acetate dihydrate 3.1% (w/v) boric acid ddH_2O to final volume

2.2.2 DNA extraction from whole blood

DNA was extracted from whole blood lymphocytes using the ReliaPrep™ Blood gDNA Miniprep system (Promega) according to the manufacturer's instructions which is summarised below.

On arrival blood samples were stored at -20°C. Prior to extraction, blood samples were thawed completely and mixed thoroughly for 10 minutes at room

temperature on a rotisserie shaker. Filter tip pipette tips were used at all stages during the procedure to prevent contamination of samples and equipment.

For each sample, 20µl of Proteinase K (10U/µl) was dispensed into a 1.5ml microcentrifuge tube. 200µl of blood was added to the Proteinase K and mixed by repeat pipetting. 200µl cell lysis buffer was added to the tube. The tube contents were vortexed for 10 seconds then incubated at 56°C for 10 minutes. Following incubation, 250µl of binding buffer was added to the tube with the contents vortexed for 10 seconds. The lysate was checked to ensure that it was dark green in colour. The contents of the tube were added to a ReliaPrep™ binding column placed in a collection tube and centrifuged at 16,200xg (13,000rpm), max speed of microcentrifuge) for 1 minute. If the lysate was still visible at the top of the membrane following centrifugation, the column was spun for a further 1 minute. The column was moved to a fresh collection tube, and the flow through from the old one was discarded as hazardous waste. The column was then washed by adding 500µl of column wash solution to the column and centrifuging it at 16,200xg (13,000rpm) for 3 minutes. If any of the solution remained visible on the membrane, the column was spun for a further minute. The flow through was again discarded as hazardous waste. This wash step was repeated a further two times to make a total of three washes. The column was then transferred to a clean 1.5 microcentrifuge tube and 50µl of 70°C nuclease free water added to the column which was centrifuged at 16,200xg (13,000rpm) for 1 minute to elute the DNA. The binding column was discarded.

The DNA concentration and purity of the sample was measured using the NanoDrop 2000c UV-Vis Spectrophotometer (Thermo Scientific) by measuring absorption at 260nm (A260) in 1-2µl of undiluted sample. The NanoDrop software automatically uses a modified Beer-Lambert equation to calculate the

concentration (in ng/µl). DNA purity was assessed simultaneously by measuring absorption at 280 nm (A280). A ratio of A260 to A280 of ~1.8 indicates “pure” DNA. A secondary measure of absorbance at 230nm (A230) was also taken, values for a “pure” nucleic acid are often higher than the respective A260/A280 values being within the range of 1.8-2.2.

An aliquot of working stock of 10-30ng/µl was prepared by diluting the DNA with molecular grade water. Samples were then stored at -20^{o}C.

2.2.3 DNA extraction from buccal swabs

DNA was extracted from buccal swabs using the Xtreme DNA Kit (XME-5/50, Isohelix) according to the manufacturer’s instructions which is summarised below. The composition of the buffers and solutions used in this protocol is proprietary information.

Prior to extraction a hot block was preheated to 60^{o}C. The proteinase K was reconstituted by adding 550µl ddH_2O before first use (then stored at 4^{o}C after reconstitution) and 60ml of 98-100% ethanol was added into the WB solution before first use.

500µl LYS buffer was added to each sample which was then vortexed to ensure the solution covers the swab head. 20µl Proteinase K solution was added to each sample then mixed immediately by vortex. The tubes where then incubated at 60^{o}C for a minimum of 10minutes to lyse the sample. Following incubation, the liquid was transferred to a 5ml tube. 750µl CB buffer was then added to the samples and mixed by vortexing for 30 seconds. 1.25ml of ethanol was added to each sample then vortexed to mix.

100µl of EB buffer per sample, was preheated in a hot block, at 70ºC.

An Xtreme DNA column was placed into a collection tube, one per sample, with 750µl of the sample was carefully added to the column without touching the rim. Samples were then centrifuged at 16,200xg (13,000rpm) for 1 minute. The flow through was discarded as hazardous waste. This step was repeated until all of the samples had been loaded onto the columns.

The columns were then washed by adding 750µl of WB solution and centrifuged for 1 minute at 16,200xg (13,000rpm). The flow through was again discarded as hazardous waste. This was step was then repeated, again discarding the flow through.

Following the wash steps the columns were then placed into clean collection tubes and centrifuged at 16,200xg (13,000rpm) for 3 minutes to remove all traces of ethanol.

The columns were then placed into clean 1.5ml microcentrifuge tubes. 100µl of preheated EB buffer was then added to the centre of the membrane of each column. The columns were left to stand at room temperature for 3 minutes then centrifuged at 16,200xg (13,000rpm) for 1 minute to elute the DNA.

The DNA concentration and purity of the sample was measured using the NanoDrop 2000c UV-Vis Spectrophotometer (Thermo Scientific) as previously described with an aliquot of working stock of 10-30ng/µl prepared by diluting the DNA with molecular grade water and samples being stored at -20ºC.

2.2.4 Single nucleotide polymorphism (SNP) genotyping

SNP genotyping was carried out using Illumina CytoSNP-12v2.1 arrays following the Infinium® HD Assay Ultra manual protocol and assistance from Dr Barry Chioza, University of Exeter.

The assay requires 200ng of DNA per sample at a concentration of 50ng/µl with each chip holding 12 samples. The protocol is carried out over three days following the Infinium HD Assay Ultra Manual Workflow which is summarised below.

Day 1: DNA samples were denatured using a buffer containing 0.1N NaOH and then neutralised in preparation for amplification. Samples were incubated overnight at 37^{o}C to amplify.

Day 2: Amplified DNA samples were enzymatically fragmented using the Illumina FMS buffer which utilises end-point fragmentation (to avoid over-fragmentation). The DNA was then precipitated using 2-propanol and the Illumina solution PM1, then collected via a 20 minute centrifugation carried out at 4^{o}C. Following resuspension, using the Illumina solution RA1, the DNA was denatured at 95^{o}C for 20 minutes. The denatured samples were cooled then 12µl of each sample was loaded onto the BeadChip. This was then incubated in the Illumina Hybridisation Oven at 48^{o}C for a minimum of 16 hours (but no more than 24 hours).

Day 3: The BeadChips were prepared for the staining process. This involved washing away any un-hybridised and non-specifically hybridised DNA using the PB1 Illumina buffer. Following the wash step, labelled nucleotides were

dispensed onto the BeadChip through the Flow-Through Chambers to perform single-base extension of primers hybridised to the DNA. The BeadChips where then stained using the Illumina XStain HD BeadChip process then imaged on an Illumina iScan Reader.

The iScan Reader uses a laser to excite the fluorophores of the single-base extension product on the beads of the BeadChip. Light emissions from the fluorophores are recorded by the reader, taking high-resolution images of the BeadChip. The data from these images were analysed using the Illumina GenomeStudio Integrated Informatics Platform allowing for the genotype to be determined. Further analysis was then undertaking by exporting the data into Microsoft Excel and using a macro to highlight notable regions of homozygosity (>1Mb) and to compare genotyping across samples.

2.2.5 Whole-exome sequencing (WES)

DNA Whole-exome sequencing of individuals was performed using two different sequencing platforms, Otogenetics and BGI, summarised below;

Whole-exome Otogenetics sequencing platform Otogenetics Corporation using the SureSelect Human All Exon V4 (Agilent Technologies)

Patient DNA was sent to Otogenetics Corporation (Norcross, GA, USA) where whole exome sequencing of genomic DNA was performed on an Illumina HiSeq2000 using the Agilent SureSelect Human All ExonV4 (51Mb) enrichment kit and a paired-end (2 × 100) protocol at a mean coverage of 30X. The exome sequencing produced 31,783,299 mapped reads, corresponding to 93% of targeted sequences covered sufficiently for variant calling (>10X coverage, mean depth 45X).

Whole-exome BGI sequencing platform

WES was also performed by BGI Tech Solutions (Hong Kong) on the BGISEQ-500 sequencing system. A total of 1,403,229,858 clean reads were aligned to the human reference genome (GRCh37) using the Burrows-Wheeler Aligner (BWA). On average, 99.79% of the whole genome excluding gap regions had at least 99.40% had at least 4X coverage and 98.20% at least 10X coverage. Average sequencing depth across the genome was 45.73X.

Bioinformatics Pipeline

The FASTQ files obtained from Otogenetics and BGI were mapped to the reference genome using the Burrows-Wheeler Aligner BWA-MEM algorithm [100, 101]. This algorithm was used due to its improved performance, compared to

other BWA algorithms, being faster and more accurate than previous versions as well as providing higher quality queries. The sequence alignment map (SAM), was converted to binary SAM format (BAM) to produce a smaller file and to increase the processing speed.

Duplicate reads were marked by Picard (version 1.46). The BAM file was then realigned, to account for indels, with variants called using GATK-HaplotypeCaller and subsequently quality filtered based on; mapping quality (MQ40), read depth (QD2), strand bias (FS60), the average position of a variant in a read (RPRS-8) and SNP quality (MQRankSum-12.5). The variant call file (VCF) file was annotated using the Alamut Software (v1.4.4) Suite. Variants were quality control (QC) checked and filtered for rare, non-synonymous exonic or splice variants, with a population frequency of <0.005 in control databases (including the Genome Aggregation Database; gnomAD, the Exome Aggregation Consortium; ExAC, and the 1000 Genomes Project) (Interactive Biosoftware). Annotated vcf files were then interrogated depending on the condition(s) under investigation (Figure 2.1).

Support for analysis of exome sequencing data was provided by Matthew Wakeling based at the University of Exeter.

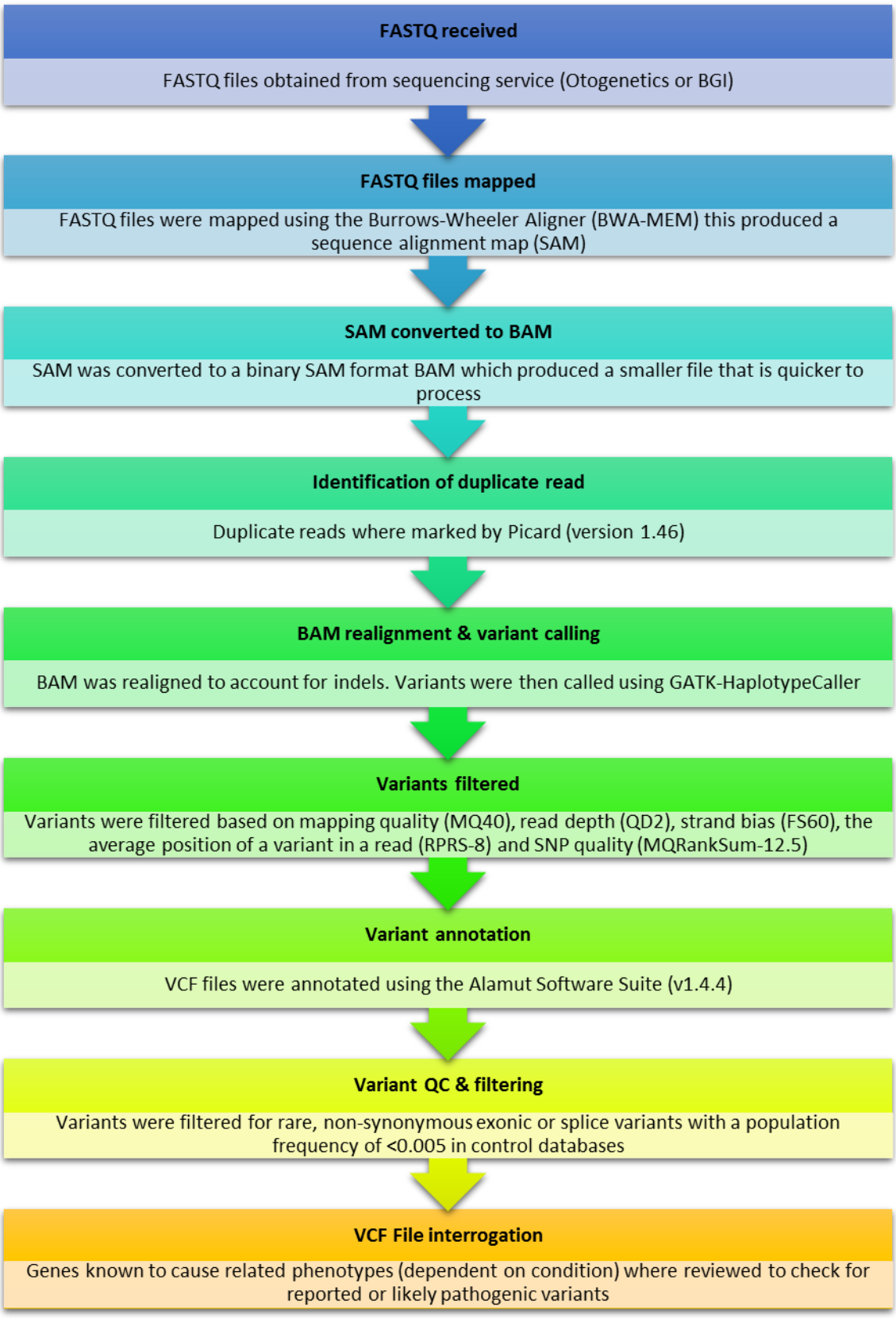


Figure 2.1: Summary of bioinformatics pipeline undertaken on exome sequencing.

2.2.6 PLEX-seq sequencing

The genotyping of 176 Amish individuals for 165 variants commonly seen in the Amish communities (see **Appendix A**) was performed using the PlexSeq process by Plexseq Diagnostics. This process, previously described by Kayima et al [102], uses a multiplexed approach to amplify regions surrounding each SNP. The primers contain an additional sequence at the 5' end to which universal barcoded Illumina primers were annealed during a secondary amplification reaction. All samples, including negative and positive controls, were uniquely barcoded and sequenced simultaneously using the first DNA-to-data sequencing platform, MiSeq (Illumina). Genotype calls for all SNPs in each sample where determined through analysis of sequence data, in the form of FASTQ files, using Plexcall software (PlexSeq Diagnostics).

2.2.7 Primer design

The Ensembl Genome Browser (December 2013 (GRCh38/hg38) assembly) was used to obtain the gene sequences (website http://www.ensembl.org/index.html). Primers used for PCR amplification were designed using Primer3 software version 0.4.0 (website http://frodo.wi.mit.edu/primer3/).

Primers were designed using the following criteria:

- Primer sizes were between 18 and 22 nucleotides.
- The difference in melting temperatures for the forward and reverse primers were no more than 1^{o}C and between 55-65^{o}C.
- Guanine-cytosine (GC) base content was kept between 40-60%.
- The primer sequences selected were specific and a 100% complimentary and unique to the region of interest to ensure only that region is amplified.

- Use of self-complimentary primers with inter or intra-primer efficiency extending more than 3 bases were avoided to reduce the formation of primer dimers and prevent the formation of secondary structures.

In silico PCR and BLAST analysis were performed using the UCSC Genome bioinformatics website to ensure primers where specific to the region of interest and to confirm the corresponding primer sequence (available in **Appendix B**) contained no common SNPs (>1%).

2.2.8 Resuspension of lyophilised primers

Primers, designed as described in section 2.2.7, arrive lyophilised. They are resuspended in molecular grade water to a concentration of 100µM to produce a master stock, which is stored at -20^{o}C. Before use in PCR reactions a 10µM working stock is produced by diluting the master stock 1 in 10 with molecular grade water.

2.2.9 Optimisation of primer conditions

To determine the optimal annealing temperature for a primer pair a PCR reaction was carried out using a temperature gradient of 52-64^{o}C across the PCR block of an Eppendorf Mastercycler thermocycler.

This involved setting up 12 reactions for each primer pair; each reaction had a different annealing temperature which increased incrementally across the PCR block from 52°C to 64°C by approximately 1°C. For these reactions a control DNA of high concentration and quality which had amplified well in a previous reaction was used.

If the PCR reaction produced weak or no product across the range of temperatures the contents of the mix and amplification conditions were altered. This might involve increasing the primer concentration or if the GC content was high, above 60%, a second gradient using 10% dimethyl sulfoxide (DMSO, Fisher Scientific) was performed. DMSO is an organic sulphur compound which binds to cytosine residues on DNA, this changes its conformation making it more liable to heat denaturation. Fortunately the primers used in the studies included in this thesis (**Appendix B)** did not require the conditions to be altered, standard 10µm concentrations of each primer were used without the addition of DMSO.

2.2.10 LabTAQ Polymerase Chain Reaction (PCR)

PCR is an *in vitro* laboratory technique used to selectively amplify DNA sequences. The process involves subjecting a small quantity of DNA to repeated temperature cycles permitting the exponential amplification of specific sequences of DNA, located between the forward and reverse primers, by up to 10^9 times.

A master reaction mix was made for each primer pair, which includes all of the reaction components, with the exception of the sample DNA. The master mix volume is dependent on the number of reactions required. This is dictated by the number of samples (patients) plus positive and negative controls. The PCR reaction for each sample was either 10 or 20µl depending on further downstream analysis required (for example Sanger sequencing or sequencing by restriction digestion) (Table 2.2).

Table 2.2: Components of a 10µl and 20µl LabTAQ PCR reaction

Component	Volume (µl) (10µl reaction)	Volume (µl) (20µl reaction)
ddH_2O	6.85	13.7
5x labTAQ reaction buffer	2	4
10µM forward primer	0.3	0.6
10µM reverse primer	0.3	0.6
labTAQ enzyme	0.05	0.1
DNA	0.5	1
Total volume	**10µl**	**20µl**

The master reaction mix was aliquoted into reaction tubes (individually, on a strip or on a plate, depending on the number of samples). The sample (patient) DNA was then added to the master reaction mix. For the negative control, ddH_2O was added in place of DNA to ensure that the desired DNA template was being amplified, and not DNA from a contaminant in one of the reaction constituents.

The PCR mix was then placed in an Eppendorf 96-well Mastercycler thermocycler and exposed to repeated heating cooling in order to separate the strands the template DNA (denaturation), allow the primers to bind to their complementary sequence (annealing) and permit the *Taq* enzyme to replicate the DNA strands within the region of interest through the addition of dinucleotides (elongation).

To reduce the amplification of non-specific products, and therefore improve specificity of primer binding, a touchdown (TD) PCR protocol was implemented. This involves using an initial annealing temperature 4^{o}C higher than the optimum

annealing temperature (T_m), then incrementally lowering the annealing temperature by 2^{o}C every two cycles until the desired T_m is reached (Table 2.3).

Table 2.3: LabTAQ PCR reaction thermocycler program

NUMBER OF CYCLES	TEMPERATURE (^{o}C)	TIME (S)
1	95	120
2	95 T_m +4 72	15 15 15
2	95 T_m +2 72	15 15 15
35	95 T_m 72	15 15 15
1	72	120

2.2.11 Agarose gel electrophoresis

To determine if the amplification of the DNA was successful, and adequate for sequencing, the resulting PCR products underwent agarose gel electrophoresis. Agarose gel electrophoresis is a technique used to separate DNA (PCR products), according to their size, using an electric current. The agarose gel forms a matrix through which the negatively charged DNA travels when an electric current is applied across the gel. The smaller DNA molecules migrate faster towards the positive electrode so therefore travel further down the gel than the larger molecules in a given time.

For resolving smaller DNA fragments, such as PCR products which are typically 500bp a 1% agarose gel was made by mixing 1g of agarose powder (Sigma-

Aldrich) with 100ml 1X LAB (Table 2.1) and heating the mixture in a microwave for 2-3 minutes. After checking the powder had completely dissolved, 2µl of 10mg/ml ethidium bromide solution was added to the gel and swirled gently to mix evenly. Ethidium bromide (EtBr) is a DNA intercalating agent that fluoresces brightly when exposed to ultraviolet (UV) light.

The gel was then left to cool while the casting tray was set up. This involved securing a rubber gasket to each end of the gel tray and placing a 28-toothed comb at the top of the gel (the number of rows of combs is dependent on the number of samples being run). The molten gel was then carefully poured into the casting tray and left to set for 10-15 minutes. Once set, the rubber gaskets and comb(s) were removed from the gel creating the wells. The gel was then placed into an electrophoresis tank and submerged in 1X LAB.

The first well of each row (if using multiple combs) was then loaded with 2µl DNA ladder (Gene Ruler 100bp DNA Ladder, Thermo Scientific) to allow the size of PCR product to be estimated. 5µl of each PCR product was mixed with 2µl agarose loading buffer, and loaded into one of the wells alongside the DNA ladder.

A power pack was used to apply a 130V across the gel for 20 minutes. The gel was removed from the gel tank and the gel plate then placed on the illuminator (UV light box with a camera). This causes the EtBr to fluoresce under the UV light visualising the PCR products in which it is intercalated.

The PCR reaction can be deemed as successful if a band, of the correct size, is seen in the lanes containing sample DNA with no band being visible in the negative control lane. If a band (therefore DNA/PCR product) is visible in the lane containing the negative control it shows the presence of contamination. As the

source of contamination is not clear it must be assumed that all samples have been contaminated and therefore the reaction must be repeated.

For larger DNA fragments (>500bp), PCR products a lower concentration agarose gel (0.8%) was used; prepared as above with less (0.8g) agarose powder dissolved in the same volume (100ml) of 1xLAB.

2.2.12 PCR product purification

Prior to sequencing a PCR product it is important to remove any unincorporated primers and dNTPs. This was achieved by undertaking an ExoSAP reaction containing exonuclease-1 (Exo) and shrimp alkaline phosphatase (SAP). Exonuclease-1 is an enzyme capable of degrading the single stranded DNA of the unincorporated primers in a 3'-5' direction. This step produces dNTPs which are subsequently removed by the shrimp alkaline phosphatase.

For this reaction 2µl of ExoSAP was added to 5µl of each of the PCR products. This mixture was then incubated at 37^{o}C for 30 minutes, the optimum temperature for enzyme activity, and then at 85^{o}C in order to inactivate the enzymes by denaturation.

2.2.13 Sequencing reaction

Purified PCR products underwent a sequencing reaction using the BigDye Terminator Cycle Sequencing Kit v3.1 (ABI, Applied Biosystems) which uses a classic chain termination PCR method to incorporate ddNTPs labelled with fluorescent dyes into the resultant PCR products. These dyes emit light at different wavelengths which are read by the sequencing machine. A master reaction mixture, containing the BigDye Terminator, BigDye Terminator buffer,

primer and ddH_2O, was made for each region of interest. 7µl of this mixture was aliquoted into appropriately labelled 0.2ml tubes and mixed with 3µl of the appropriate PCR product. For each sample two separate 10µl sequencing reactions were carried out for the forward and reverse primers (

Table 2.4).

Table 2.4: Sequencing reaction components and quantities

Component	Volume (µl)
BigDye Terminator (Applied Biosystems)	0.5
BigDye Terminator Buffer (Applied Biosystems)	1.7
Primer (Forward or reverse)	0.5
ddH_2O	4.3
Cleaned PCR product	3.0
Total Volume	**10**

The reaction mixtures where placed into an Eppendorf 96-well Mastercycler, thermal cycler machine and run through following programme (Table 2.5) for 25 cycles:

Table 2.5: Sequencing Reaction Thermocycler Program

Process	***Temperature (°C)***	***Time***
Denaturation	96	30 seconds
Annealing	50	15 seconds
Elongation	60	4 minutes

2.2.14 Sequencing reaction purification

In order to prepare the sequencing reaction products for automated DNA sequencing, they were first be purified. Purification of was carried out using the BigDye® XTerminator™ Purification Kit (Applied Biosystems) according to the manufacturer's instructions.

Unincorporated BigDye, salts and other charged molecules from sequencing reactions may interfere with base calling and electrokinetic sample injection during DNA sequencing. The BigDye® XTerminator™ Purification Kit cleans samples by utilising two reagents; XTerminator™ Solution which scavenges unincorporated dye terminators along with other charged molecules and SAM™ Solution that enhances the performance of the XTerminator™ Solution and stabilises the post-purification reactions.

Sequencing reaction products were loaded into a 96-well plate with 5µl XTerminator™ Solution aliquoted into each well (containing the sequencing reaction products) and vortexed briefly. 30µl of SAM™ Solution was then added to each well. The plate was sealed using clear adhesive film (Thermo Scientific) and vortexed for 30 minutes then briefly centrifuged. The plate was then placed into a 16-capillary 3130xl Applied Biosystems® Automated DNA Sequencer for sequencing on a 36cm array POP7 polymer programme setting.

2.2.15 Genotyping by restriction digest

A restriction digest can be used to detect variation in a DNA sequence providing a genotype. This method can only be used for DNA polymorphisms that create or destroy the restriction site of a restriction enzyme, commonly known as a restriction fragment length polymorphism (RFLP).

RFLP analysis can be a cheap and effective way to genotype a large number of samples. As the polymorphism under investigation will interrupt the palindromic recognition sequence of a restriction enzyme, DNA fragments of differing lengths are produced depending on the presence or absence the variant. After carrying out the restriction digest reaction, at the optimum conditions, agarose gel electrophoresis can be used to size the fragments and then determine the genotype of individuals.

A restriction digest reaction mixture contains PCR product of the area of interest, the appropriate 10X restriction digest buffer, the appropriate restriction endonuclease and ddH_2O.

To determine the most appropriate restriction enzyme to use the sequence, containing the variant and the wild type sequence, was entered into NEBcutter V2.0 online software (http://tools.neb.com/NEBcutter2/).

A restriction digest master mix was made with all the constituents, except for the PCR product with 12.5µl aliquoted into appropriately labelled 0.2ml microcentrifuge tubes. 2.5µl PCR product was then added to each tube, and the tubes were incubated at 37˚C overnight (8+ hours). In addition to sample DNA, a negative control (water replacing DNA) and a known homozygote and heterozygote for the variant were included to provide control bands to confirm digestion and aid the interpretation of results. Following incubation samples were

loaded, along with a loading buffer, onto a 3% agarose gel (3g Agarose powder, 100ml 1XLAB) and were electrophoresed for 60mins at 100V. The resultant bands were visualised in an illuminator to identify if the specific variant under investigation was also present any DNA samples.

2.3 Molecular cloning techniques

The pCMV6-entry-SLC15A5 clone was transformed into *E.coli* DH5-alpha bacteria through the incubation of the bacteria with the clone and applying a heat shock. Liquid bacterial culture was inoculated to amplify the plasmid DNA which was then extracted through mini and midi preps. Restriction digests of the plasmid DNA where undertaken to check the plasmid DNA was what it was expected to be which can be determined by the fragment pattern seen after digestion.

Initial immunocytochemistry experiments were undertaken using the pCMB6-entry vector but due to low levels of expression a construct using a pCAGGs plasmid was produced, to increase mammalian expression, this included a YFP tag in place of the Flag (DDK) tag, to avoid potential issues with SLC15A5 antibody binding. The transfection process was repeated with the new pCAGGs-SLC15A5-YFP plasmid DNA.

The methods used for these experiments are detailed below.

2.3.1 Buffers, reagents and stock materials

All general-purpose chemicals, with exception of alcohols (purchased from Fisher Scientific) were acquired from Sigma-Aldrich. All solutions for cell culture were supplied by Lonza, except and penicillin/streptomycin from PAA laboratories. Components of solutions are detailed in (Table 2.6).

Table 2.6: Solutions for molecular cloning techniques

Solution	Constituents
10% APS	10% (w/v) APS
10% SDS	10% (w/v) SDS
100X SOC	2M glucose, 1M $MgCl_2$, 250mM KCl, sterilised by push filtration (0.22µm filter)
2X PFA	8% (w/v) paraformaldehyde, in PBS, pH adjusted to 7.4
3X Laemmli sample buffer	10% (v/v) glycerol, 2% (w/v) SDS, 5% (v/v) β-2 mercaptoethanol, 0.002% (w/v) bromophenol blue, 0.125M Tris-Cl (pH 6.8)
Ampicillin	100 U./ml
Destain	40% (v/v) MeOH, 10% (v/v) acetic acid, in dd.H20
DMEM 10 % serum (HEK)	10% (v/v) heat-inactivated FBS, 100 U./ml penicillin, 100ug/ml streptomycin, in DMEM
6X DNA loading buffer	30% (v/v) glycerol, 0.25% (w/v) bromophenol blue, 0.25% (w/v) xylene cyanol
HE lysis buffer	25 mM HEPES, 5 mM EDTA, 1 mM MgCl2, 10% (v/v) Glycerol, 1% (v/v) Triton-X100, 100 µM PMSF
Kanamycin	50 µg/ml
LB agar	15 g/L agar, 10 g/L tryptone, 10 g/L NaCl, 5 g/L yeast extract
Luria Bertani Broth	10 g/L tryptone, 10 g/L NaCl, 5 g/L yeast extract
Lysine block	5% (v/v) horse serum, 5% (v/v) goat serum, 50 mM poly-D-lysine, 0.2% (v/v) Triton X-100
Running buffer	25 mM Tris-base, 192 mM glycine, 0.1% (w/v) SDS
TAE	40 mM Tris-base (pH 7.6), 20 mM acetic acid, 1 mM EDTA
TBS	20mM Tris-Cl, 150mM NaCl
TBS-T	20mM Tris-Cl, 150mM NaCl, 0.1% (v/v) Tween-20
Transfer buffer	25mM Tris-base, 192mM glycine, 0.1% (w/v) SDS, 20% (v/v) MeOH

Plastic ware for tissue culture was acquired from Greiner Bio-one, with other general laboratory consumables purchased from Alpha-Labs and Fisher Scientific. Kits for deoxyribonucleic acid (DNA) amplification and gel extraction were purchased from Qiagen, restriction enzymes were supplied by Promega and New England Biolabs.

Specialist kits, antibodies, chemicals and consumables bought from alternate sources are noted in the text where appropriate.

2.3.2 DNA Plasmid preparation

Constructs

The *SLC15A5* plasmid for expression studies was purchased from Origene. *SLC15A5* fusion genes were subcloned by PCR (primer details outlined in Table 2.7) from its original cytomegalovirus (CMV) vector into a pCAGGS backbone (gift of Dr John Chilton) to increase expression efficiency [103]. Full length *SLC15A5* was excised from the pCMV6-Entry vector by *HindIII* overnight digestion and inserted into the corresponding sites of pCAGGS-mycFLAG. Maps for key expression vectors and cloning schemes can be found in **Appendix C**. Constructs were checked at all steps by restriction enzyme digest using unique sites within the inserts and analysis by agarose gel electrophoresis.

Primer design

Table 2.7 describes all primers that were designed, following the general rules of primer design [104], to have flanking sequences containing convenient sites for restriction digest and to omit the proteins native stop codon.

Table 2.7: Details of custom designed primers

Description	Sequence	Complement		Length (bp)	Restriction digest site
		Forward	Reverse		
CMV6BamF	GAC TGG ATC CGG TAC CGA GGA G	✓		22	BamH1
FLAGSalR	ATA TGT CGA CTT AAA CCT TAT CGT CGT CAT C		✓	31	Sal1
SLC15BamF	ATA TGG ATC CAT GTC TGT TAC AGG CTT TAC C	✓		31	BamH1

SLC15SalR	AAT TGT CGA CTC ATA GGG CTG TCT CCC AAA GAT C	✓	34	Sal1

2.3.3 Mini preps: Inoculating a liquid bacterial culture and recovering plasmid DNA from bacterial culture

Small amounts of plasmid DNA, usually around 10µg per ml of bacterial culture, were extracted using the QIAprep Spin Miniprep Kit. These plasmids were used for restriction mapping, ligations and transient transfections to confirm fluorescence for fluorescent protein-tagged plasmids.

A 5ml culture of bacteria was grown overnight in LB broth (Sigma) with 5µl of kanamycin (30mg/ml) selective antibiotic. This produced a LB broth with a final kanamycin concentration of 0.03.

A 1.5ml aliquot of overnight bacterial culture was removed and centrifuged at 16,200xg (13,000rpm) for 1 minute at 20^{o}C. The supernatant was discarded and the pellet resuspended in 250µl of Buffer P1 (a resuspension buffer; 50mM Tris.HCl, pH8.0; 10mM EDTA; 100µg/ml RNAse A, without LyseBlue reagent). 250µl of Buffer P2 (a lysis buffer; 200mM NaOH; 1% w/v SDS) was added to the suspension, and mixed via 4-6 inversions, to rupture the bacteria by alkaline lysis. Cellular debris was precipitated by the addition of 350µl Buffer N3 (a proprietary neutralization Buffer) and mixed by further 4-6 inversion until the flocculent precipitate was evenly dispersed. The mixture was then centrifuged at 16,200xg (13,000rpm) for 10 minutes. The supernatant was removed then added to a QIAprep spin column and centrifuged at 16,200xg (13,000rpm) for 1 minute. The column was washed by addition of 500µl Buffer PB (a proprietary binding buffer) which was then centrifuged at 16,200xg (13,000rpm) for 1 minute. A second wash was undertaken by the addition of 750µl Buffer PE (a low salt, high ethanol

proprietary solution) to the column which was then centrifuged at 16,200xg (13,000rpm) for 1 minute. The flow through was discarded and the column centrifuged for a further 1 minute at 16,200xg (13,000rpm) (to remove residual ethanol). The column was then transferred to a fresh microcentrifuge tube. 55µl of molecular biology grade water (Sigma-Aldrich) was added to the column which was then incubated for 1 minute before a final spin at 16,200xg (13,000rpm) for 1 minute to elute the DNA. Molecular biology grade water is 0.1mm filtered, has been analysed for the absence of nucleases and proteases and has undergone bioburden analysis.

2.3.4 Midi preps: Inoculating a liquid bacterial culture and extracting plasmid DNA from bacterial culture

Midipreps were used in the purification of up to 300µg of plasmid DNA using HiSpeed Midi Prep Kit (Qiagen) which was then used for expression and functional studies.

A 5ml starter culture was grown overnight in LB broth with kanamycin a selective antibiotic. This was diluted 1:1000 in 50ml of LB broth with the kanamycin selective antibiotic and grown overnight. The bacteria were harvested by centrifugation at 1,900xg (4,500rpm) for 15 minutes at 4^{o}C then resuspended in 6ml of Buffer P1 (without LyseBlue added). The bacteria were lysed by the addition of 6ml of Buffer P2, mixed thoroughly by 4-6 inversions and incubated at room temperature for 5 minutes. Cellular debris was precipitated by the addition of 6ml of Buffer P3 (a neutralization Buffer; 3M potassium acetate, pH5.5) and 4-6 inversions. This was then added to a QIAfilter Midi Cartridge and incubated at room temperature for 10 minutes. During this incubation, an anion-exchange

resin column (Qiagen HiSpeed Midi Tip) was prepared by the addition of 4ml Buffer QBT (an equilibration Buffer; 750mM NaCl; 50mM MOPS, pH 7.0; 15% v/v isopropanol; 0.15% v/v Triton X-100) which was allowed to empty under gravity. The cell lysate was push filtered in to the column and allowed to move through under gravity. The column was washed by the addition of 20ml Buffer QC (a wash buffer; 1M NaCl; 50mM MOPS, pH7.0; 15% v/v isopropanol). 5ml Buffer QF (an elution buffer; 1.25 M NaCl; 50 mM Tris.HCl, pH 8.5; 15% v/v isopropanol) was added to the column to elute the DNA and precipitated by the addition of 0.7 volumes of isopropanol, inverted 4-6 times and incubated at room temperature for 5 minutes. Precipitated DNA was bound to a QIAprecipitator module, washed with 2ml 70% v/v ethanol and dried by pushing air through the module. DNA was recovered by the addition of 350µl of Buffer TE (a resuspension and storage buffer; 10 mM Tris.HCl, pH 8.0; 1 mM EDTA).

The concentration of DNA recovered was quantified using a NanoDrop 2000c UV-Vis Spectrophotometer (Thermo Scientific) as previously described.

2.3.5 Pfu PCR of extracted bacterial DNA

DNA, containing the gene of interest, was amplified via PCR using *Pfu* DNA polymerase (Promega). The highly thermostable DNA polymerase (from the hyperthermophilic archaeum *Pyrococcus furiosus*) was used in placed of labTAQ DNA polymerase due to its 3'→5' exonuclease (proofreading) activity, which enables the *Pfu* polymerase to correct nucleotide incorporation errors.

The reaction mixture (outlined in Table 2.8) was combined in a sterile, nuclease-free microcentrifuge tube on ice. It is critical to add *Pfu* DNA Polymerase after the addition of dNTPs or the proofreading activity of the polymerase may degrade the primers which may result in nonspecific amplification and reduced product yield.

Table 2.8: Components Pfu PCR reaction mixture

Component		Final concentration
Pfu DNA Polymerase 10X Buffer with $MgSO_4$	5µl	1X
dNTP mix, 10mM each	1µl	200µM (each)
Upstream primer	25 pmol	0.1–1.0µM
Downstream primer	25 pmol	0.1–1.0µM
DNA template	Variable	<0.5µg/50µl
Pfu DNA Polymerase (2–3u/µl)	Variable	1.25u/50µl
Nuclease-Free Water (to final volume of 50µl)	Variable	

The mixture was mixed gently then placed into the SimpliAmp Thermal Cycler (Applied Biosystems by Life Technologies) and heated to 95°C for an initial 2 minute denaturing step. This was followed by 35 cycles of; 30 second DNA strand melting, 30 second annealing at 50°C, 2 minute extension at 72°C (which required 1 minute per kilobase (Kb) of final PCR product). Finally a 5 minute extension

step at 72ºC was carried out before samples were cooled to 4ºC until the tube was collected. A 5µl aliquot of PCR product was mixed with 6X Agarose loading buffer (Table 2.1) and loaded onto a 0.8% w/v agarose gel to undergo electrophoresis and a verify the amplicon was of the predicted size.

2.3.6 PCR product purification

QIAquick PCR Purification Kit was used to remove any impurities such as salts, unincorporated nucleotides, agarose, or dyes and unincorporated nucleotides which can affect subsequent processing.

Five volumes of buffer PB to one volume of PCR product was added to 10µl was added to 3M sodium acetate (pH5) and mixed by slowly pipetting up and down. The sample was when added to a QIAquick column which was placed inside a collection tube. The sample was centrifuged at 16,200xg (13,000rpm) for 13 minutes. The flow through was discarded and the column placed back into the collection tube. 750µl buffer PE was added to the column then centrifuged at 16,200xg (13,000rpm) for 1 minute. The flow through was again discarded. This time the column was placed into a clean 1.5ml centrifuge tube with 50µl of molecular grade water added to the centre of the QIAquick membrane. The tube was then centrifuged for a further minute at 16,200xg (13,000rpm).

Purified PCR products underwent overnight restriction digest with an appropriate restriction enzyme at 37ºC. The digested product was combined with 6X DNA loading buffer and loaded onto a 0.8% w/v agarose gel which was run for 40 minutes at 80V to undergo electrophoresis. Afterwards the gel was stained in GelRed® nucleic acid gel stain for 20 minutes on the rocker at room temperature.

2.3.7 Gel purification of DNA

QIAquick Gel Extraction Kit was used to clean the DNA fragments from enzymatic reactions and remove unwanted impurities such as salts, agarose, or dyes which can affect subsequent processing.

After suitable restriction digestion, DNA was electrophoresed at 100V on a 0.8% w/v agarose gel in TAE buffer containing 1:20,000 Sybr Safe (Life Technologies) until DNA bands could be resolved and the appropriate fragment excised and placed into a microcentrifuge tube. The DNA was purified using a QIAquick Gel Extraction Kit.

The volume of gel was estimated by weight, with 100 mg ≈ 100 µl. Three volumes of buffer QG were added and incubated at 50^{o}C for 10 minutes, vortexing every 3 minutes. If the colour of the mixture is orange or violet, then 10µl of 3M sodium acetate was added to the dissolved gel solution to ensure the correct pH as indicated by a yellow colour. The solution was then added to a QIAquick column and spun at 16,200xg (13,000rpm) for 1 minute. The column was washed with 750µl buffer PE (a low salt, high ethanol proprietary solution) and the column centrifuged at 16,200xg (13,000rpm) for 1 minute. The flow through was discarded and the column was spun for a further minute at 16,200xg (13,000rpm) to remove residual ethanol. The column was transferred to a fresh microcentrifuge tube and 30µl of molecular biology grade water was added to the column and incubated for 1 minute before a final spin at 16,200xg (13,000rpm) for 1 minute to elute the DNA.

2.3.8 Ligation

The reaction mixture (Table 2.9) was made up to a total of 10µl in molecular biology grade water then incubated overnight at 14^{o}C.

Table 2.9: Ligation reaction mixture

Component	Final Concentration
10X T4 ligase buffer* (Promega)	1 µl
Restriction digested backbone	~80ng
Restriction digested insert	~240ng
T4 DNA ligase (Promega)	2U

*(300mM Tris.HCl, pH 7.8; 100 mM MgCl2; 100 mM DTT and 10 mM ATP)

The amount of cDNA was estimated by comparison to known amounts of DNA in standard size marker ladders in an agarose gel. 1µl of ligation product was used to transform bacteria.

2.3.9 Bacterial transformation

All bacterial transformations were carried out using standard sterile practice in a designated category 2 laminar flow containment hood. A 40µl aliquot of competent DH5α *Escherichia coli* (*E.coli*, NEB) was thawed on ice. 1µl of plasmid DNA was then added and mixed to the *E.coli* and mixed by gentle tapping. The bacteria were left on ice for a further 15 minutes.

A heat shock method was used to allow the bacterial cells to take up the plasmid DNA. During heat shock transformation a sudden increase in temperature creates pores in bacterial plasma membranes allowing for plasmid DNA to enter the cell. This method involved cells being incubated at 42^{o}C for 45 seconds, then returned to the ice for a further 2 minutes.

The bacteria were then added to 1ml of SOC (LB containing 20mM glucose, 10mM $MgCl_2$ and 2.5mM KCl) and placed in a shaker at 37^{o}C for 1 hour, to allow expression of antibiotic resistance proteins. The bacteria cells were collected by centrifugation at 1,500xg (4,000rpm) for 2 minutes at 4^{o}C. 800µl of the supernatant was removed, the bacteria were gently resuspended in the remaining liquid which was then plated out on an agar culture plate containing the ampicillin selection antibiotic.

2.3.10 Human Embryonic Kidney (HEK) 293 cell culture

HEK 293 cells were originally transformed in 1977 by Frank Graham, a post-doc in Alex Van der Eb's laboratory. They are named after the cell type, human embryonic kidney, and the fact that this transformation was Graham's 293rd experiment. The cells were transformed through exposure to sheared fragments of human adenovirus type 5 (Ad5) DNA which lead to the incorporation of Ad5 into chromosome 19 of the kidney cell genome [105].

Cell lines used were cultured in Dulbecco's Modified Eagle's Medium (DMEM), containing 100U/ml penicillin and 100µg/ml streptomycin and 10 % (v/v) foetal calf serum. HEK 293 cells were incubated at 37^{o}C, 5% CO_2, and were routinely passaged at 80-90 % confluency.

The plasmid DNA was then used to transfect HEK 293 cells. After cells had been grown to 80-90 % confluency they were counted using a Neubauer chamber then seeded onto cover slips at a concentration of 30x10^{4}. Overnight lipofectamine transfections were carried out with cells then fixed to the cover slips (4% PFA at room temperature for 15 minutes). Immunocytochemistry was then carried to

stain the cells ready for visualisation under the confocal microscope. Details of the methods used are outlined below;

2.3.11 Transient transfection

Lipofectamine LTX (Invitrogen) transfections where undertaken following the manufacturer's protocol. Cells were grown until 40-80% confluency. An appropriate amount of sterile DNA (2.5µg per 35mm dish) was diluted in Optimem (Gibco), to a suitable volume for the size of the culture dish being used. If multiple constructs were to be expressed, plasmids were combined prior to addition to enhance co-expression efficiency. The specified amount (3.75-8.75µl) of Lipofectamine LTX reagent required was added and the contents mixed by gently tapping the microcentrifuge tube. The transfection mixture was incubated for a minimum of 30 minutes at room temperature (Table 2.10).

Table 2.10: Transfection mixture composition

Component	Volume (µl) per 35mm dish
Plain DMEM	493
DNA (1µg/µl)	2.5
Lipofectamine LTX	4.5
Total	**500**

Following incubation, the transfection mixture was added to the cell culture medium and the cells returned to the incubator at 37^{o}C and 5% CO_2 overnight.

2.3.12 Immunocytochemistry

Cells were seeded on to coverslips at a density of approximately 1 x 10^4 cells per 13mm diameter coverslip, and cultured for a minimum of two days prior to staining. All washes were carried out at room temperature. Cells were fixed by adding matching volume of 4% w/v paraformaldehyde (PFA), pre-warmed to 37ºC, to the culture medium. This was then removed and replaced with fresh PFA (pre-warmed) with cells fixed for 15 minutes at 37ºC.

After fixing, the coverslips underwent three 5 minute washes with PBS. Non-specific binding was pre-blocked by incubation at room temperature in lysine block for 1 hour. The primary antibodies (Table 2.11) was diluted in lysine block with 50µl applied to each coverslip to cover it completely. This was then left for 1 hour at room temperature.

Table 2.11: Details of primary antibodies used for immunocytochemistry (ICC) and Western blot (WB) analysis

Antigen	Host	Dilution	Supplier	Application
SLC15A5	Rabbit	1:100	Atlas Antibodies	ICC & WB
Myc	Mouse	1:100	Abcam	ICC & WB
FLAG	Mouse	1:100	Abcam	ICC & WB

The coverslips were then underwent three further 5 minute washes with PBS. The appropriate fluorescently conjugated secondary antibodies (Table 2.12) were diluted in lysine block and applied as with the primary antibodies and incubated for a minimum of 1 hour at room temperature.

Table 2.12: Details of secondary antibodies used for immunocytochemistry (ICC) and Western blot (WB) analysis

Target	Host	Conjugate	Dilution	Supplier	Application
Rabbit IgG	Goat	AlexaFluor 568	1:400	Invitrogen	ICC
Mouse IgG	Goat	AlexaFluor 488	1:400	Invitrogen	ICC
Mouse IgG	Goat	AlexaFluor 633	1:400	Invitrogen	ICC
Rabbit IgG	Goat	Horseradish Peroxidase	1:5000	Sigma	WB
Mouse IgG	Goat	Horseradish Peroxidase	1:5000	Sigma	WB

Coverslips were then washed three more times with PBS for 5 minutes. When appropriate the final wash was replaced with 4, 6- diamidino-2-phenylindole (DAPI, Invitrogen) at 1:1000 in PBS to stain nuclei. After removing excess PBS, by blotting, coverslips were mounted on Superfrost slides (VWR) in Fluorsave (Calbiochem) and left to cure for 24 hours, in the dark, at room temperature. Images were captured using a Leica SP8 confocal microscope.

2.4 Western blotting

To detect and analyse endogenous proteins and expression of fusion protein products from exogenous DNA transiently transfected in cell lines as previously described, the proteins were first extracted then subjected to Western blot analysis.

2.4.1 Preparation of cell lysates

Cells were typically seeded in 6-well plates one day prior to experimentation. Where appropriate, cells were transfected and allowed to express the exogenous constructs for 18-24 hours. Growth media was removed from culture vessel prior to being placed on ice. Cells were washed with chilled PBS then 150µl of HE lysis buffer was added. The dish was swirled to ensure coverage then left on ice for 10 minutes. The bottom of the culture vessel was scraped, the lysates collected in 1.5ml microcentrifuge tube and then spun at 16,200xg (13,000rpm) for 10 minutes at 4^{o}C. The supernatant was carefully transferred to a fresh microcentrifuge tube, without disturbing the pellet of non-solubilised cellular material. The supernatants were retained for further analysis.

2.4.2 Protein quantification

A Pierce bicinchoninic acid (BCA) Protein Assay Kit (Thermo Scientific) was used, according to the manufacturer's protocol to estimate the amount of protein present in cell lysates.

Working reagent was prepared by combining 50 parts of BCA reagent A with 1 part of BCA reagent B. Bovine serum albumin protein standards (0.2-1.2 µg/ml) were prepared in a 96-well plate by diluting 2mg/ml stock with the same diluent as the samples. As HEPES-EDTA (HE) lysis buffer contains more EGTA than the

microplate assay can accommodate, it was necessary to first dilute the lysis buffer 1:10 with PBS before preparing the standards. The samples were also diluted 1:10 in PBS and 10µl of both samples and standards were loaded into wells in triplicate.

Following an 8 minute incubation at room temperature on an orbital shaker (LSE Low Speed Orbital Shaker, Corning), absorbance was measured at 562nm using a PHERAstar FS microplate reader (BMG Labtech). Concentrations of each sample were calculated against the standard curve. Each sample was then diluted in appropriate lysis buffer and 1x Laemmli sample buffer to give 1mg/ml. Samples were then stored at -20^{o}C until required.

2.4.3 SDS-PAGE separation of proteins

Sodium dodecyl sulphate (SDS) polyacrylamide gel electrophoresis (SDS-PAGE) was used to separate proteins according to their molecular weight.

SDS is an anionic detergent with a net negative charge across a wide pH range that denatures proteins, eliminating the influence of the structure and charge of a polypeptide. During electrophoresis proteins are separated based solely on chain length when a voltage is applied and they migrate through the acrylamide gel matrix towards the positively charged electrode.

The resolution of the resolving gel required is determined by the length of the polypeptide (Table 2.13). A 10% resolving gel was used to investigate SLC15A5 due to its size of 65.2kDa (65,263Da).

Table 2.13: SDS-PAGE gel percentage recommendation based on protein size.

Protein size (kDa)	Gel Percentage (%)
4-40	20
12-45	15
10-70	12.5
15-100	10
25-200	8

The 10% SDS-PAGE gels were prepared according to the recipe outlined in Table 2.14 using the Mini-PROTEAN™ Tetra Handcast System (BioRad). Ammonium persulphate (APS) and N,N,N',N'-tetramethylethane-1,2-diamine (TEMED) were added last to initiate polymerisation (setting) of the acrylamide.

Table 2.14: SDS-PAGE gel composition.

Component	Stacking Gel (4%)	Resolving Gel (10%)
40% Acrylamide/Bis	0.7	6.3
1M Tris-HCl pH 6.8	0.88	-
1.5M Tris-HCl pH 8.8	-	6.25
10% SDS	0.07	0.25
10% APS	0.07	0.125
ddH_2O	5.28	12.1
TEMED (µl)	2.5µl	2.5µl
Total volume*	**7**	**25**

**Volumes (in ml, unless otherwise stated) are for 40% Acrylamide/Bis solution 37.5:1 ratio and makes 2 gels.*

Each well was loaded with 25µg of protein in Laemmli sample buffer and 5µl BLUeye Prestained Protein Ladder (Gene Flow) was loaded at either end of the gel. The loaded gel was then immediately run at 200 mA on ice for 90 minutes in running buffer (detailed in Table 2.6).

2.4.4 Membrane transfer

Following the separation of proteins by SDS-PAGE, gels were incubated in transfer buffer for 10 minutes to reach stability. This step was required to allow expansion of the gel as the polyacrylamide takes on water. Gels that do not undergo this equilibration step can swell during transfer resulting in poor protein resolution.

Proteins were then transferred to Immobulon-P polyvinylidene difluoride (PVDF) transfer membrane (Merck) by sandwiching the two together between chromatography-grade blotting filter paper (GE Healthcare) and applying a constant current of 200mA over the stack to enable efficient migration of the proteins towards the anode and binding to the PVDF membrane.

2.4.5 Immunodetection and visualisation

After transfer, PVDF membranes were blocked with 5% (w/v) skimmed milk in tris-buffered saline with 0.1% Tween-20 (TBS-T) for 1 hour at room temperature to prevent non-specific binding of the detection antibodies. Blocked membranes were then incubated in primary antibody solution (primary antibodies listed in Table 2.11).

Membranes were then removed from the primary antibody and washed in TBS-T three times for 5 minutes. Incubation for 1 hour at room temperature with appropriate secondary antibody (detailed in Table 2.12) conjugated to horse-

radish peroxidase (HRP) followed with another set of washes. Proteins were visualised using enhanced chemiluminescence (ECL) detection reagent. This method involves the breakdown of the ECL detection reagent by the HRP conjugated to the secondary antibody.

2.5 Gene expression methods

2.5.1 RNA extraction

RNA was extracted from whole blood collected in PAXgene Blood RNA Tubes (BRT) and purified for sequencing using the Qiagen PAXgene blood RNA kit according to the manufacturer's instructions.

The RNA concentration of samples was measured using the NanoDrop 2000c UV-Vis Spectrophotometer (Thermo Scientific), as previously described, with the quality of the RNA measure by the Agilent 2200 TapeStation System (Agilent Technologies).

Assessing RNA quality is a critical step as the integrity of the RNA determines the success of downstream experiments including cDNA library construction and qPCR. The Agilent 2200 TapeStation System compares relative ratios of signals to produce an RNA integrity number (RIN) equivalent (RINe) score. This RINe uses the 1-10 scale as RIN where a score of 10 describes the highest quality RNA and 1 indicates the RNA is completely degraded.

2.5.2 Whole transcriptome sequencing

Experiments outlined below were undertaken by the Exeter Sequencing Service (Exeter Sequencing Service and Computational core facilities at the University of Exeter. Medical Research Council Clinical Infrastructure award (MR/M008924/1). Wellcome Trust Institutional Strategic Support Fund (WT097835MF), Wellcome Trust Multi User Equipment Award (WT101650MA) and BBSRC LOLA award (BB/K003240/1).

Reverse transcription is the process by which single stranded complementary DNA (cDNA) is synthesized using a reverse transcriptase enzyme and an RNA

template. A TruSeq stranded total RNA sample preparation with RiboZero Globin kit (Illumina) was used to prepare whole-transcriptome sequencing libraries from blood-derived RNA by depleting rRNA, via a bead-based method, and synthesising cDNA. The quality of the cDNA was checked using the High Sensitivity D1000 ScreenTape system (HS D1000) (Aligent).

Accurate quantification of the library was undertaken using the NEBNext® Library Quant Kit (Illumina) in conjunction with the StepOnePlus™ Real-Time PCR System prior to undergoing NGS sequencing. This quality control (QC) step is vital to obtain maximum, high quality NGS sequencing data.

The NGS data was then analysed by Dr Ryan Ames (University of Exeter) to infer transcriptome wide expression levels.

CHAPTER 3

INVESTIGATING INHERITED FORMS OF HEARING LOSS IN THE AMISH COMMUNITY

3 Investigating inherited forms of hearing loss in the Amish community

3.1 Hearing

3.1.1 The mammalian ear

Consisting of the three basic parts; the outer ear, the middle ear and the inner ear [106], the mammalian ear is an intricate physiological apparatus that collects a wide spectrum of sounds of various frequencies and intensities from the environment. By transferring this information to the brain, as an electrical signal, it facilitates the interpretation of sound enabling them to not only be identified but their relative distance and direction to be determined [107].

The outer ear consists of the visible auricle and the ear canal (Figure 3.1a). Its primary function is to collect sound waves from the environment, funnel them into the ear canal and onto the tympanic membrane (ear drum), a thin, circular layer of tissue. Sound waves cause this thin membrane to vibrate, passing the sound waves into the middle ear.

The middle ear contains three small bones, or ossicles, called the malleus (hammer), the incus (anvil) and the stapes (stirrup). The malleus, being attached to the inside surface of the tympanic membrane, vibrates when sound waves hit the membrane. This is turn causes the subsequent bones to vibrate, passing the waves through the middle ear. Upon vibration the stapes makes contact with the oval window, a membrane covered opening leading into the cochlear, that divides the middle and inner ear.

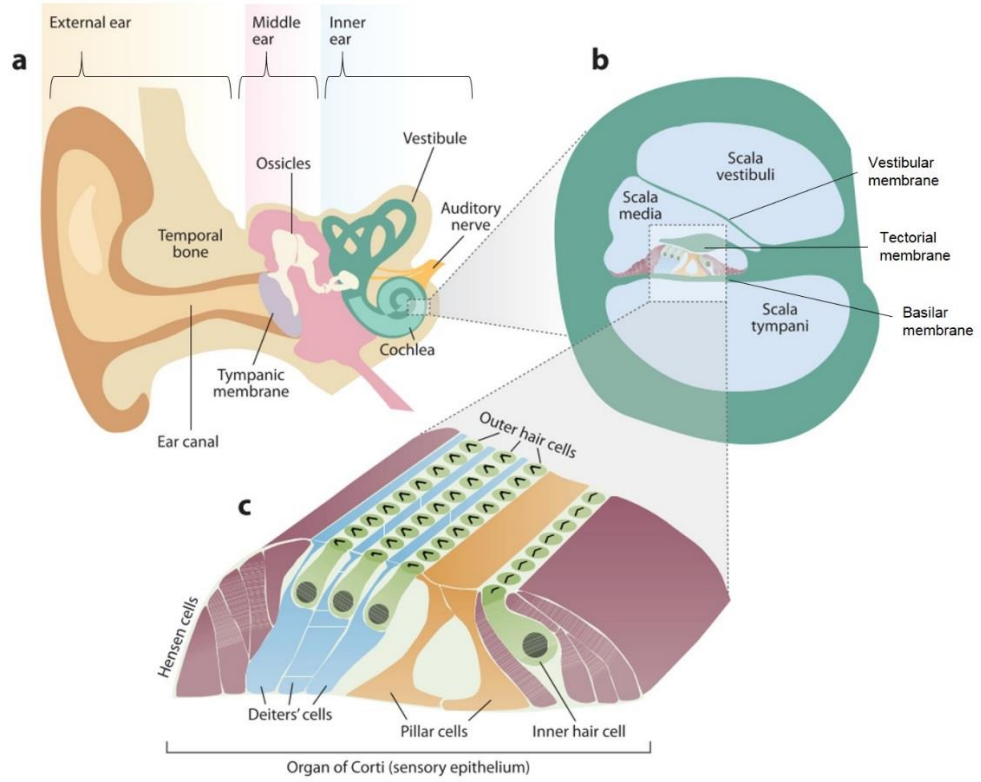


Figure 3.1: Schematic illustration of the human ear. (a) The ear consists of the outer, middle and inner ear. (b) A section through the cochlear. (c) The organ of Corti. Image adapted from [107].

The inner ear occupies the cavity in the temporal bone and houses the sensory organs for hearing and balance. This location serves as an acoustic chamber enabling the capture of low intensity sounds [107].

The cochlear is a snail-shaped organ consisting of a spiral canal that turns 2.5 times around an axis, the modiolus, ending in the helicotrema [108]. This canal is divided by two membranes the Reissner's, or vestibular, membrane and the basilar membrane which forms three fluid-filled ducts; the scala vestibular and the scala tympani, filled with perilymph, and the scala media (Figure 3.1b). The scala media, also called the cochlear duct, is an endolymph-filled cavity that houses the organ of Corti, commonly referred to as the organ of hearing, which sits on the basilar membrane (Figure 3.1c).

Mechanical vibration of the oval window, initiated by the stapes, causes the endolymph to move through the cochlear duct and vibrate the basilar membrane against the tectorial membrane [107]. Embedded within the basilar membrane are specialised cells called hair cells which, through a process of mechanoelectrical transduction (MET), are responsible for transferring this mechanical stimuli into an electrochemical signal which is the basis of hearing.

3.1.2 Mechanoelectrical transduction in the mammalian ear

There are two types of hair cell; inner hair cells (IHC) and outer hair cells (OHC) that are arranged in a highly order pattern [109] consisting of one row of IHCs and three rows of OHCs supported by various other non-sensory cells [110] forming the organ of Corti (Figure 3.1c). The IHCs are the main sensory cells which convert the sound-induced motion of the endolymph into an electrochemical impulse that is transmitted to the auditory cortex of the brain, via the auditory nerve, capturing information regarding the frequency, intensity and timing of sounds. The role of the OHCs is to act as amplifiers modifying the sensitivity and selectively to sound [111].

Both hair cell types have a hair bundle located on its apical surface, consisting of between 20-300 actin-rich projections called stereocilia and a single microtubule-based kinocilium, [112]. The hair bundles of the IHCs extend into the scala media whereas the stereocilia of the OHC bundles are connected to the tectorial membrane [111].

The stereocilia undergo a precise assembly process that gives the hair bundle an asymmetric, staircase appearance [113] (Figure 3.2a). Each stereocilia within in a hair bundle is connected to an adjacent stereocilia via a cadherin tip link (Figure

3.2bi) which connects the tops of the small and middle row stereocilia with the sides of the taller neighbouring stereocilia [110]. These connections are considered to be an essential part of the MET machinery [114, 115].

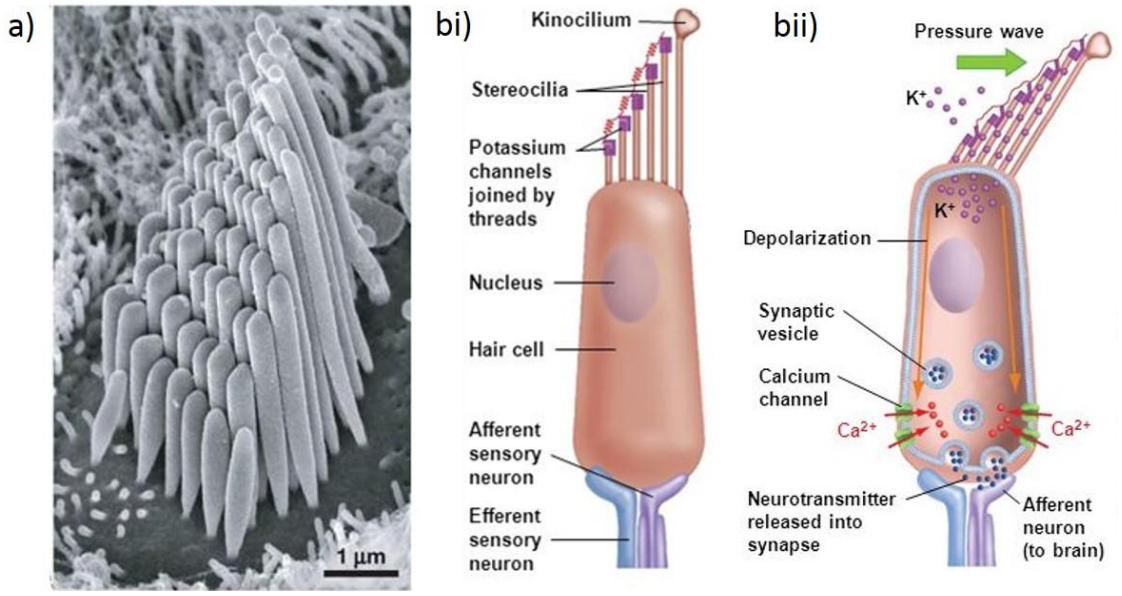


Figure 3.2: (a) Scanning electron microscopy showing the organisation of the IHCs viewed from the top of the organ of Corti. Schematic representation of an IHC with relaxed tip links and closed MET channels (bi) and with tip links under tension and MET channels open as a result of mechanical deflection by a sound wave (bii). Adapted from Benjamin Cummings, Pearson 2008.

Small (1-100nm) mechanical deflections of the hair bundle in the direction of the taller stereocilia [110, 116] increases tension in a gating spring [114] that leads to the opening of MET ion channels located atop the stereocilia (Figure 3.2bii). Although much is known about the MET channels of the HCs, the molecular identity of the gating spring and the MET channel protein and how these channels are activated is yet to be confirmed [114, 116]. However, current literature suggests the MET channel may belong to the transient receptor potential (TRP) channels [117].

The opening of these channels initiates an influx of sodium (Na^+), potassium (K^+) and calcium (Ca^{2+}) ions into the hair cells. This depolarises the basolateral membrane of the hair cells and triggers the calcium dependent exocytosis of the neurotransmitter glutamate [111] via synaptic ribbons. These ribbons tether synaptic vesicles permitting synchronous auditory signalling [118]. The release of glutamate into the synaptic cleft excites adjacent afferent auditory neurones signalling the auditory centres in the brain [111].

3.1.3 Ion Homeostasis in the mammalian ear

Ion homeostasis is the maintenance of highly asymmetric concentrations of the major inorganic ions [119]. MET is heavily reliant on ion homeostasis to maintain the specific ionic gradient between the perilymph and endolymph, unique to the inner ear [120]. The ion composition of the perilymph is similar to other extracellular fluids, such as cerebrospinal fluid and blood plasma, consisting of 5mM K^+ and 150mM Na^+. However, the endolymph, in addition to protein, Mg^+ and Ca^{2+}, comprises of high (150mM) K^+ and low (5mM) Na^+ which results in a highly positive endocochlear potential (EP) of +80mV compared to the perilymph [121, 122]. The function of IHCs is highly dependent upon this EP [123] which greatly enhances the flow of ions into the cells during MET permitting greater sensitivity to sound [121].

The non-sensory epithelium cells of the stria vascularis generate the high concentration of K^+ and in turn the highly positive EP with virtually no input of metabolic energy [120]. These cells also play a key role in the recycling of the K^+ (Figure 3.3).

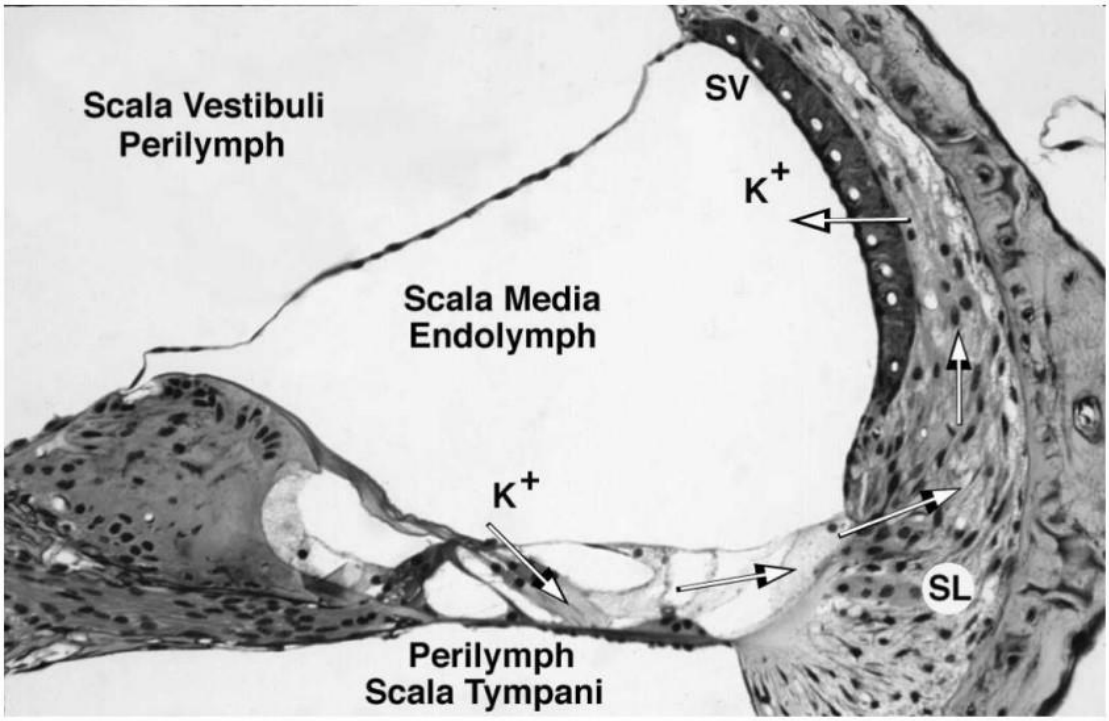


Figure 3.3: Potassium ion (K+) recycling in the inner ear. MET of the IHCs causes an influx of K+ into the hair cells. These ions are then secreted back into the endolymph by the stria vascularis (SV) via supporting cells and the spiral ligament (SL). Image taken from [114].

Several recycling pathways have been suggested comprising of a system of channels, transporters and gap junctions with many of the molecules involved identified via mutations in mice and humans that have led to hearing loss [120]. Any genetic mutation leading to an imbalance of K^+ in the endolymph and loss of EP can cause a number of hearing deficits including deafness [114, 117, 120, 121].

3.2 Hearing Loss

The World Health Organisation has reported that ~466 million people worldwide experience disabling hearing loss and predict this to rise to over 900 million by 2050 [124]. Hearing loss is the most prevalent sensory deficit disorder in developed societies [125] with congenital hearing loss affecting at least 1 in 500 new-borns [126, 127] and more than 50% of the population over the age of 80 suffering from presbyacusis (age-related hearing loss) [107].

3.2.1 Types of hearing loss

Hearing loss is an extremely heterogeneous condition [128, 129] and can be classified in a number of ways depending on the age of onset, type, severity, progression, frequency affected and laterality, if it affects one (unilateral) or both (bilateral) ears. It can also be grouped according to its aetiology (genetic or environmental factor) and whether the hearing loss occurs with (syndromic) or without (non-syndromic) other clinical features [128] (Table 3.1).

Table 3.1: Characteristics for classifying hearing loss [128-130].

Classification	Description of hearing loss
Age of Onset	Congenital (present at birth), prelingual (0-5years of age), or postlingual (>5 years of age)
Type	Sensorineural, conductive or mixed
Severity	Mild (20-40dB), moderate (41-55dB), moderately severe (56-70dB), severe (71-90dB), or profound (>90dB)
Progression	Progressive or stable
Tone frequency affected	Low (<1000Hz), middle (1000-2000Hz) or high (>2000Hz)
Laterality	Unilateral or bilateral
Aetiology	Environmental/acquired or genetic
Additional clinical features	Syndromic or non-syndromic
Inheritance	Recessive, dominant, X-linked or mitochondrial
Presence of vestibular disorder	Hearing loss with or without vestibular dysfunction

Sensorineural hearing loss (SNHL) occurs when the conversion of mechanical sound waves into an electrical signal is impaired due to malfunction of the inner ear [131] including the membranous labyrinth, the organ of Corti or the vestibulocochlear nerve [132]. SNHL most commonly has an underlying genetic cause. There are currently over 6000 causative variants in more than 110 genes linked to non-syndromic SNHL (NS-SNHL) [131] with additional genes being responsible for syndromic hearing loss. Conductive hearing loss is caused by abnormalities of the outer ear or the ossicles of the middle ear preventing the conduction of sound waves and is most commonly caused by an environmental factor [131]. Mixed hearing loss is when both the sensorineural and conductive parts of the ear are impaired.

Environmental factors such as exposure to ototoxic chemicals, for example aminoglycosides, exposure to excessive noise and neonatal insults including prematurity, jaundice or prenatal infection [125, 133] account for approximately 20% of prelingual hearing loss with genetic causes giving rise to the remaining 80% [134] (Figure 3.4).

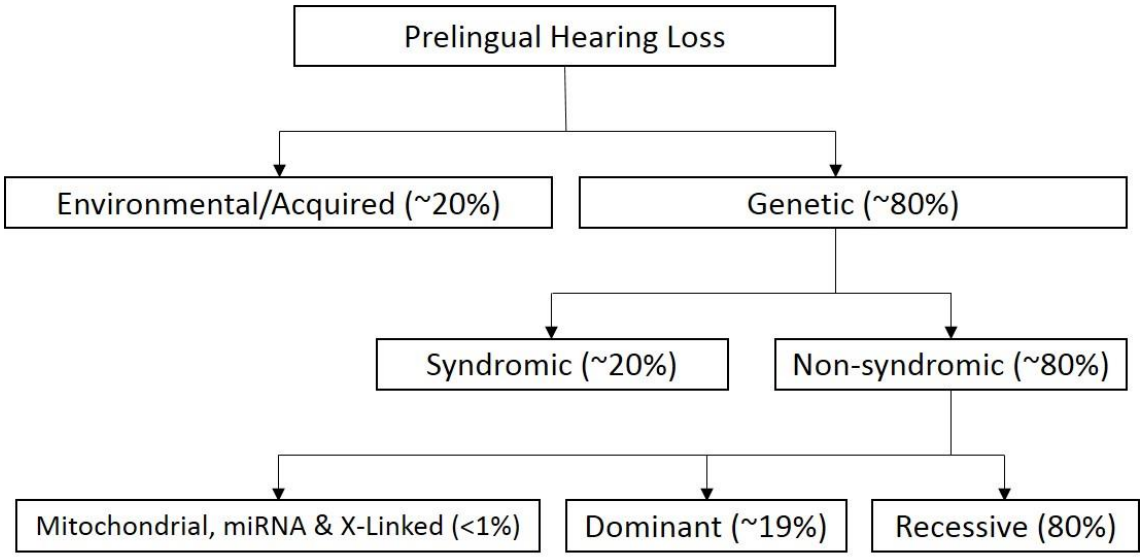


Figure 3.4: Causes of prelingual hearing loss in developed countries figure adapted from [131].

It is widely accepted that early identification and accurate diagnosis of the underlying genetic cause is crucial for selecting the most appropriate therapeutic option. This is of particular importance when assessing cochlear implant candidacy [111, 125] and when monitoring potential future health concerns of genetic syndromes allowing for the implementation of protective management strategies [132]. Discovering the underlying cause also allows families to receive more targeted genetic counselling and be provided with information on the prognosis for their child and the predicted chance of recurrence in future offspring [111].

3.2.2 Impact of hearing loss

Despite hearing loss being ranked as the fifth leading cause of years lived with disability by the Global Burden of Disease Study in 2103, it receives limited funding and public awareness [135]. Elucidating the underlying genetic cause and pathophysiology of congenital hearing loss presents numerous benefits to affected individuals, their families and to communities as a whole.

The social costs of untreated hearing loss is ~$1.1million dollars per individual [111], presenting a huge economic burden. Early intervention, in the form of Universal Newborn Hearing Screening (UNHS), is predicted to reduce these costs ~75% and most importantly improve the life opportunities of affected individuals [136]. The impact of hearing loss on an individual is profound effecting; speech development and language acquisition, cognitive and psychosocial development, independence and overall quality of life [135, 137, 138] resulting in substantial negative effects on educational achievement and work life opportunities [135, 138]. Interventions implemented before 6 months of age provide the greatest chance of developing age-appropriate speech and language skills, [137] making it imperative for the genetic cause to be identified as early as possible [125]. During the past 20 years UNHS has become a standard of care throughout the US and UK [139, 140] reducing the age of diagnosis, enabling the implementation of proactive intervention efforts [125, 137, 139, 141] and informing treatment options such as cochlear implantation candidacy [142].

3.2.3 Diagnosing hearing loss

It is widely accepted that hearing loss is an extremely heterogeneous condition [143]. Approximately 1% of all human genes are predicted to be involved in the hearing process and therefore could be responsible for the diverse phenotypes observed in hearing loss [137].

For a number of years this clinical and genetic heterogeneity slowed the discovery of new causative genes [125] and prevented comprehensive genetic testing and large scale population screening [133] thus making molecular diagnosis very difficult [128]. With more than 110 genes and in excess of 6000 mutations [131] reported to cause NS-SNHL traditional strategies involving single mutation screening using Sanger sequencing were not feasible due to cost and time constraints [131]. The advent of next generation sequencing (NGS) technologies, whole exome sequencing (WES) and whole genome sequencing (WGS), has increased the speed and proficiency of detecting hearing loss genes with 21 new genes identified between 2010 and 2015 [125].

In addition to expediting the discovery of causative genes these new technologies have also permitted the implementation of improved diagnostic testing increasing diagnostic rates by approximately 50% [128, 143]. This enables families to be counselled appropriately, with regard to the progression and prognosis of the condition, receive the most suitable treatment [111] and facilitates affected individuals in achieving optimal cognitive development [137].

In the UK the British Association of Paediatricians in Audiology (BAPA), whose aim is to maintain standards in audio-vestibular medicine [144], have published guidelines to standardise the investigations undertaken to diagnosis childhood hearing impairment after a failed UNHS. These guidelines (summarised in

Appendix D) utilise a gene panel approach to identify causative mutations for non-syndromic and syndromic hearing loss [145].

The new challenge for genetic investigations and genomic research no longer comes from the sequencing of variants, which has become the cheapest and simplest part, but instead resides in the accurate interpretation of identified variants which currently incurs the highest expense due to the time and expertise required from a number of highly qualified individuals [125].

3.2.4 Treatments

Hearing impairment is the only sensory defect that can be successfully treated, even when hearing loss is complete [137] due to the one of the most significant advances in modern medicine, the cochlear implant [146].

Cochlear implants

Cochlear implants (CIs) are a surgically placed electronic stimulus prostheses that can restore the missing, or impaired, function of the IHCs by taking on the role of transforming an acoustic signal into an electrical impulse capable of activating the auditory nerve (Figure 3.5) [141].

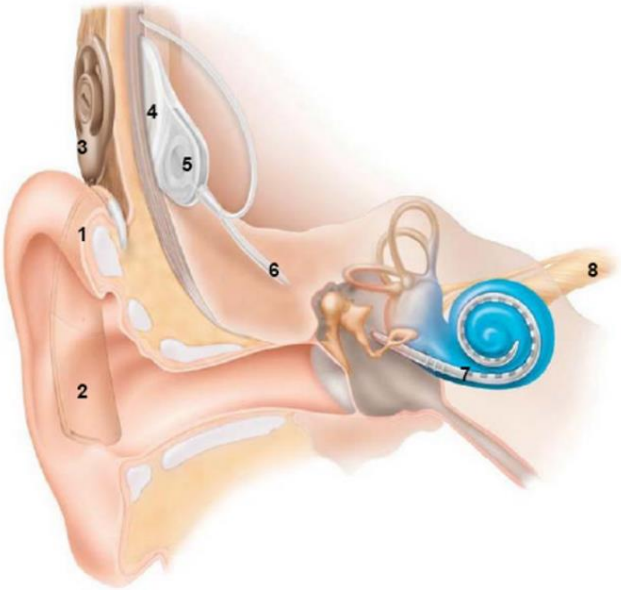


Figure 3.5: A typical modern cochlear implant system that converts sound to electric impulses delivered. Image shows the location of the; microphone (1), speech processor (2), transmitter (3), receiver (4), stimulator (5), electrodes (6&7) and the auditory nerve (8) [147].

However, to achieve the maximum therapeutic benefit of this treatment, and to minimise the effect on language development, it is necessary to detect and diagnose hearing loss as early as possible [141]. Studies have observed a negative correlation between outcomes with CIs and the duration of hearing loss. This indicates that those individuals that experience prolonged periods of auditory deprivation, for example those with congenital hearing loss who are implanted after the age of three years or individuals with postlingual hearing loss that experience a long period of severe hearing loss prior to implantation, are more likely to experience negative outcomes compared to those implanted earlier in life or within a year of developing substantial hearing loss for individuals affected by postlingual hearing loss [148].

Candidacy for CIs is comprehensively assessed via medical and audiological examination [146]. Conventionally children with severe to profound hearing loss with speech recognition <12-30% where considered CI candidates [149].

However, due to improvements in CI technology and surgical approaches, the criteria for CI implantation has expanded, with the FDA now approving children as young as one year of age and individuals with residual low-frequency hearing impairment [146].

Biological treatments

Although there is currently no biological treatment for hearing loss there is an apparent need for a treatment that restores auditory function without the need for a prosthesis or at the cost of any residual hearing [150, 151]. Within the auditory research community there is increasing interest into a variety of techniques that look to regenerate mammalian hair cells and restore their function [151].

For a number of years gene therapy, the treatment of human disease using genetic material, was the focus of many research groups (Reviewed in [151]). Recent advances in this field including improved methods of gene delivery and improvements in hair cell regeneration [150] have demonstrated exciting proof-of-principle studies in animal models which continues to suggest gene therapy as a possible treatment for some forms of hearing loss [152].

However, the clinical application of gene therapy, for hearing loss and other conditions, is currently limited by the perceived risk of side effects which are still under investigation [153]. With almost 2600 clinical trials, in 38 countries, having been completed, underway or approved [154] there is hope that the findings from these studies will aid the translation of gene therapy to human patients and the development of an effective treatment for individuals with hearing loss [152].

In addition to gene therapy, genome editing technologies are an exciting new area of research. Genome editing technologies modify the genome at a targeted locus to correct genetic variants, known to cause hearing loss (and other disorders), by restoring the wild-type sequence in native DNA through the use of programmable nucleases [151]. The CRISPR/Cas9 system, derived from prokaryotic immune systems, is the most recent and advanced programmable nuclease that is considered to be the most prevalent and easy-to-use system with multiple applications [153].

Currently there are considerable concerns about the use of this technology in a clinical setting due to the potential for off-target editing and the unknown resultant side effects [151]. Although, to date, there is no way to identify or prevent them given the increasing interest in this area it is likely these complications will be overcome in the future.

Public Health Awareness

In conjunction with developing therapies to treat hearing loss there a number of strategies that could be employed to help mitigate hearing loss in children. Improvements in public health awareness and the application of simple measures such as; avoiding the use of ototoxic drugs, immunisation and the early identification and intervention for acute and chronic ear conditions, have shown to reduce the contribution of environmental factors in the development of hearing loss [133, 135].

3.3 The genetics of hearing loss

3.3.1 Syndromic hearing loss

There are 600 conditions, cited in the London Medical Database V.1.0.31, linked to SNHL [128]. These account for approximately 20% of genetic causes of hearing loss. In a number of cases the hearing impairment may be the most obvious symptom, and therefore the first to be diagnosed, with other clinical features developing later [129]. Often a genetic diagnosis is the only way to determine if an individual will develop future complications, such as renal failure in Alport syndrome [132].

Perhaps one of most important syndromes to diagnose is the autosomal recessive Jervell and Lange-Nielsen syndrome where congenital, profound hearing loss occurs in conjunction with Long QT syndrome [131]. This syndrome is associated with a high rate of syncope, where individuals lose consciousness as a result of a sudden drop in blood pressure, and if left untreated, can lead to sudden death [128].

Some of the most common syndromes, grouped by inheritance pattern, are summarised in **Appendix E**.

Interestingly mutations in some genes associated with syndromic forms of hearing loss can also cause non-syndromic forms such as *CDH23* and *MYO7A* (Usher syndrome), *SLC26A4* (Pendred syndrome), *WFS1* (Wolfram syndrome) and *COL11A2* (Stickler syndrome).

3.3.2 Non-syndromic hearing loss

Determining the underlying genetic cause of an individual's hearing loss can exclude the presence of a hearing loss syndrome or confirm the presence of non-syndromic hearing loss. This information is useful for determining and implementing the most effective management strategy including potential treatments and for providing informed genetic counselling to families.

There are currently over 110 genes linked to non-syndromic hearing loss which can be referred to by the genes involved or by the genetic locus. Non-syndromic deafness loci are referred to by DFN (DeaFNess) followed by a letter, which classifies the mode of inheritance (DFNA autosomal dominant, DFNB autosomal recessive, DFNX X-linked) and a number representing the order of gene mapping/discovery [131].

GJB2

The most common genetic cause of congenital NS-SNHL are mutations in the gap junction beta 2 (*GJB2)* gene [130], which was first identified as a cause of hearing loss in 1997 [155] and is commonly referred to as DFNB1. For European populations, mutations in this gene account for approximately 50% of hearing loss cases [130]. The *GJB2* gene is located at 13q12.11 and encodes the gap junction protein connexin 26 (Cx26) which is expressed in the non-sensory epithelial cells of the inner ear [155].

Connexins are integral membrane proteins which are generally referred to by their predicted molecular weight. For example, connexin 26 is predicted to be ~26kDa [156] and exhibits the characteristic topological structure of a connexin protein containing; four transmembrane domains, two extracellular loops and one

intracellular loop, with both the N- and C- termini exposed to the cytoplasm (Figure 3.6) [157].

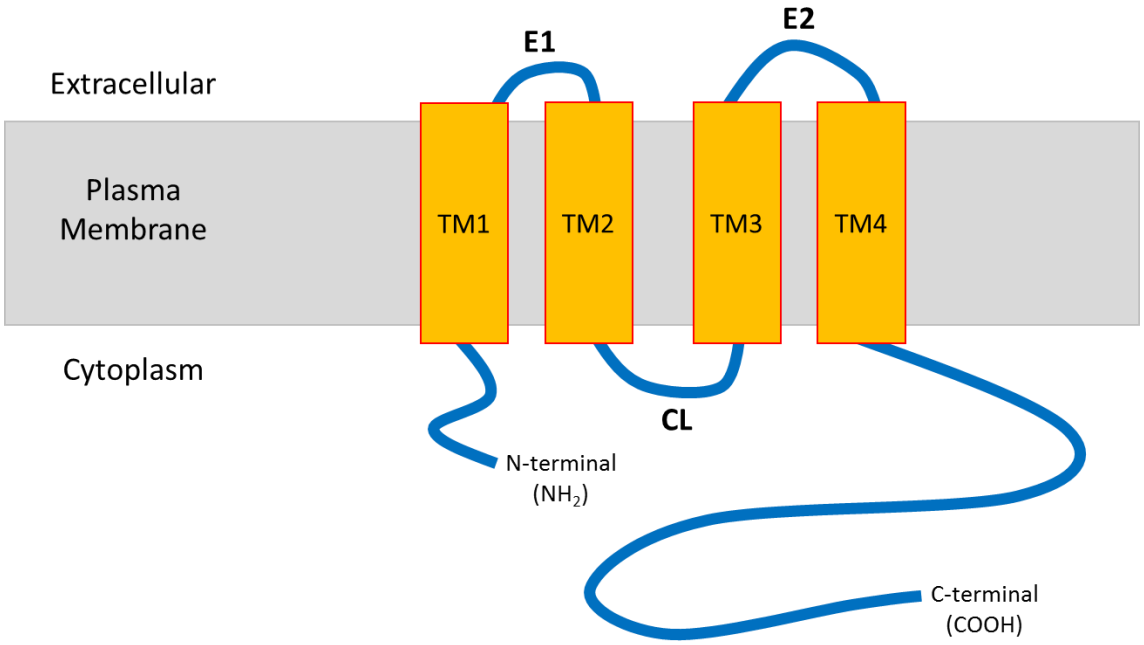


Figure 3.6: Topological structure of a typical connexin protein. Showing the four transmembrane (TM) domains, the two extracellular loops (E1 and E2) and the single cytoplasmic loop (CL).

Six connexin molecules join to form a pore-like transmembrane oligomer called a connexon, or hemichannel [107] (Figure 3.7). These gap junctions are intercellular channels that permit the passage of ions and small molecules (up to ~1.5kDa) [158]. It is widely accepted that these molecules play an important role in the recycling of K^+ during the process of ion homeostasis, which is essential for normal hearing [111].

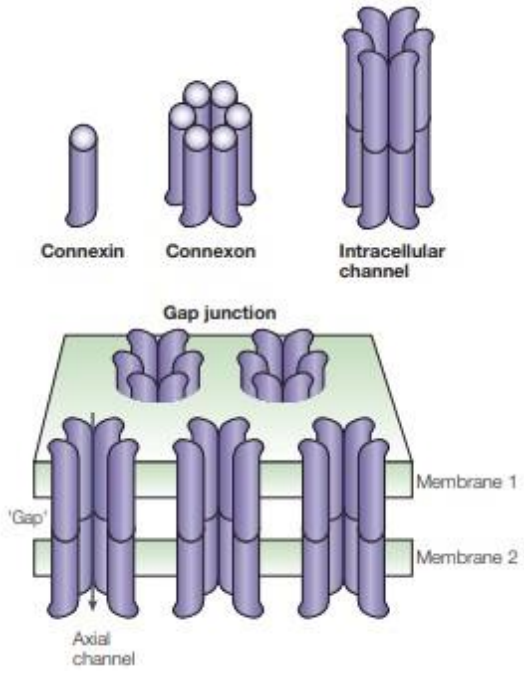


Figure 3.7: Connexins, connexons and gap junctions. Simplified diagram showing the assembly of a gap junction from the intercellular joining of two connexons, on adjacent membranes, each consisting of six single connexin molecules [156].

More than 100 mutations located within *GJB2* have been associated with hearing loss [158]. Most *GJB2* mutations display a recessive (DBFB1) inheritance pattern with a few being dominant (DNFA3) in nature [155, 159]. A wide spectrum of hearing loss phenotypes ranging from profound congenital deafness [158] to mild, progressive, late-onset hearing loss starting in childhood [160] are observed. This diversity in phenotype suggests that hearing loss arising from mutations in *GJB2* may involve several different underlying pathological mechanisms [158]. However, to date, there is no clear, demonstrable relationship between specific changes in Cx26 function and the observed phenotypes [158]. New evidence suggests that congenital hearing loss caused by mutations in *GJB2* are associated with cochlear development disorders, as opposed to EP generation [161] or K^+ recycling impairment [162]. This means that disruption of

K^+ recycling is not the principle mechanism causing Cx26 mutation-induced hearing loss as predicted [160, 161].

The most frequently occurring *GJB2* mutation, c.35delG (NM_004004.5: c.35delG), accounts for ~70% of cases worldwide [163, 164] with almost 4% of the white population in Southern Europe reported to be heterozygous [165, 166]. The high frequency of the c.35delG mutation could be due to a mutational hotspot resulting from its position in a $T(G)_6T$ sequence which, during DNA replication, may favour slippage and mispairing [167]. However, data from multiple other studies show strong evidence of linkage disequilibrium [168, 169] suggesting that c.35delG represents a common founder mutation and not a mutational hotspot. Other common *GJB2* mutations are observed in specific populations, including c.235delC in East Asia and Japan [170], c.167delT in Ashkenazi Jews [171], p.V37I in Southeast Asia [172], p.W24X in India [173], IVS1+1G>A in Mongolia [174] and del(GJB2-D13S175) in Russia [175].

Mutations in *GJB2* are often classified as either truncating or non-truncating depending on their possible effect on Cx26 [166]. Large systematic analysis of the *GJB2* genotype has determined that mutations in this gene do display genotype-phenotype correlation with the type of mutation significantly impacting the severity of hearing impairment [176]. Truncating, or inactivating mutations (nonsense, frameshift, indels) prevent synthesis of Cx26 and, in homozygotes, are associated with severe to profound SNHL [163]. Non-truncating mutations (missense, in-frame) only modify one or several amino acids meaning that the protein may retain its function [166]. Individuals with homozygous non-truncating

mutations, or compound heterozygotes with one non-truncating and one truncating mutation, typically express mild to moderate hearing loss [163].

Other genes responsible for NSHL

In addition to those seen in *GJB2,* mutations in other connexin genes can also lead to NSHL including, Cx30 (*GJB6*), Cx31 (*GJB3*) and Cx43 (*GJA1*). As with *GJB2, GJB6* is predominately expressed in the cochlea and is considered a major deafness gene [157]. Whilst mutations in both Cx26 and Cx30 can independently induce hearing loss, digenic Cx26 and Cx30 heterozygous mutations are more frequently observed and are the second most common cause of recessive hearing loss [162]. Interestingly mutations in *GJB6* have been shown to be responsible for both recessive (DFNB1) and dominant (DFNA3) forms of hearing loss. Connexin 30 is 30kDa protein, containing 261 amino acids, with the typical structure of a connexin (Figure 3.7). It is involved in gap junctional intercellular communication (GJIC) and is essential for maintaining homeostasis of the epidermis and inner ear [157].

Further to the connexion family, mutations in a wide array of other molecules have also been associated with inherited forms of hearing loss; information regarding these molecules may be accessed through extensive online databases including the Hereditary Hearing Loss Homepage OMIM, ClinVar and HGMD.

3.4 Causes of HL known in the Amish community

Genetic studies have previously defined a number of gene founder alterations as causes of inherited hearing loss in the Amish, including *PCNA* [177], *SLITRK6* [159, 178], *HARS* [179], *YARS [180]*, *KCNQ1* [181], *ST3GAL5* (GM3 synthase deficiency) [182], *LONP1 [183]*, *HYAL2* [184] and *COL1A2* [185, 186]. Details of these variants are summarised below (**Error! Reference source not found.** and **Error! Reference source not found.**) with a more detailed description of each variant available in **Appendix F**.

Table 3.2: Summary of genes associated with syndromic SNHL hearing loss identified in the Amish.

Gene	Variant	Hearing Loss	Other clinical features
PCNA	c.683G>T; p.Ser228Ile (NM_002592.2) chr20:g.5115472C>A	Prelingual onset, moderate to profound high frequency SNHL	Ocular/cutaneous telangiectasia, premature aging, photophobia/photosensitivity with predisposition to sun-induced malignancy, short stature, learning difficulties, neurodegeneration with cerebellar atrophy
SLITRK6	c.1240C>T; p.Q414X (NM_032229.2) chr13:g.85795269G>A	Bilateral, prelingual moderate to severe SNHL	Severe congenital myopia
HARS	c.1361A>C; p.Tyr454Ser (NM_002109.5) chr5:g.140674776T>G	Postlingual, severe progressive SNHL	Childhood progressive visual impairment, horizontal nystagmus, optic pallor, photosensitivity, bull's eye macula, retinitis pigmentosa, delay in gross motor development, lower limb brisk reflexes, ataxia, normal intellect
YARS	c.499C>A; p.Pro167Thr (NM_003680.3) chr1:g.32806493G>T	Bilateral SNHL	Severe nystagmus, visual impairment, developmental delay, pancreatic insufficiency, cholestatic liver disease, hypoglycaemia and subcortical white matter abnormalities
ST3GAL5	c.862C>T; p.R288X (NM_003896.3) chr2:g.85844542G>A	Variable onset & severity SNHL	Refractory infantile onset epilepsy, developmental delay, developmental regression, generalised irritability, feeding difficulties, dystonic arm movements, hypotonia, cortical blindness, optic nerve defects, skin pigment abnormalities
LONP1	c.2161C>G; p.Arg721Gly (NM_004793.3) chr19:g.5694546G>C	Mild to moderate SNHL	Developmental delay, facial dysmorphism, bilateral cataracts, dental anomalies, short stature, delayed epiphyseal ossification, metaphyseal hip dysplasia, vertebral abnormalities, hypotonia, scoliosis, laryngeal obstruction, swallowing difficulties, imperforate anus, omphalocele, rectovaginal fistula, cryptorchidism and tongue hemiatrophy

Table 3.3: Summary of genes associated with syndromic, conductive and mixed hearing loss identified in the Amish.

Gene	Variant	Hearing loss	Other clinical features
Conductive Hearing Loss			
HYAL2	c.443A>G; p.K148R (NM_003773.4) chr3:g.50320047T>C	Mild to moderate, unilateral/ bilateral and pre or postlingual (one individual with SNHL)	Cleft lip or palate (CLP), facial dysmorphism, congenital cardiac abnormalities, pectus excavatum, single palmar creases, 2,3 toe syndactyly, myopia, staphyloma and cataract
Mixed Hearing Loss			
COL1A2	c.2237G>T; p.Gly610Cys (NM_000089.3) chr7:g.94420590G>T	Progressive mixed hearing loss	Bone fractures with minimal or no trauma, bone deformity, short stature, dentinogenesis imperfecta and connective tissue abnormality

3.5 Results

As part of the Windows of Hope Project families with multiple individuals affected by hearing loss, with no genetic diagnosis, were recruited into the Amish Hearing Loss Programme. The aim of the program was to provide diagnoses for the underlying genetic cause of the hearing loss within these families. To date 19 families with between one and four siblings affected by SNHL have been identified and recruited into the program through the use of a community-appropriate newsletter and family support groups. Although *GJB2* gene mutations had not previously been described in the Amish, our initial studies included evaluation of this gene due to the high prevalence of *GJB2*-related hearing loss worldwide. The screening process for affected individuals was initially dependent on whether the hearing loss presented with or without additional clinical phenotypes (Figure 3.8).

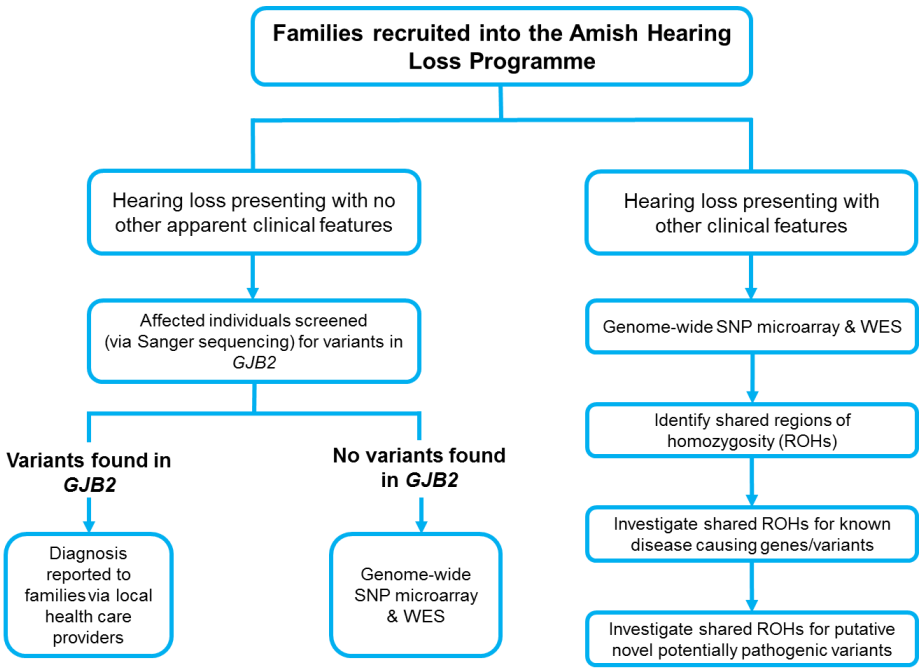


Figure 3.8: Initial screening strategy implemented for individuals recruited to the Amish Hearing Loss Programme.

Individuals presenting with no other clinical phenotypes (at the time of assessment) where considered to have a non-syndromic form of hearing loss so where screened, via Sanger sequencing, for variants within *GJB2*. If no variants were found further genetic investigations involving genome-wide SNP mapping in combination with whole exome sequencing was undertaken. If individuals were presenting with additional clinical features, and considered to have to syndromic form of hearing loss, screening for *GJB2* variants was omitted.

3.5.1 Molecular studies of a large Amish family with multiple individuals affected by a neurodevelopmental disorder and hearing loss

A family presented with six children affected by a neurodevelopmental disorder in form of intellectual disability (ID), four of whom also presented with pre-lingual SNHL (Figure 3.9).

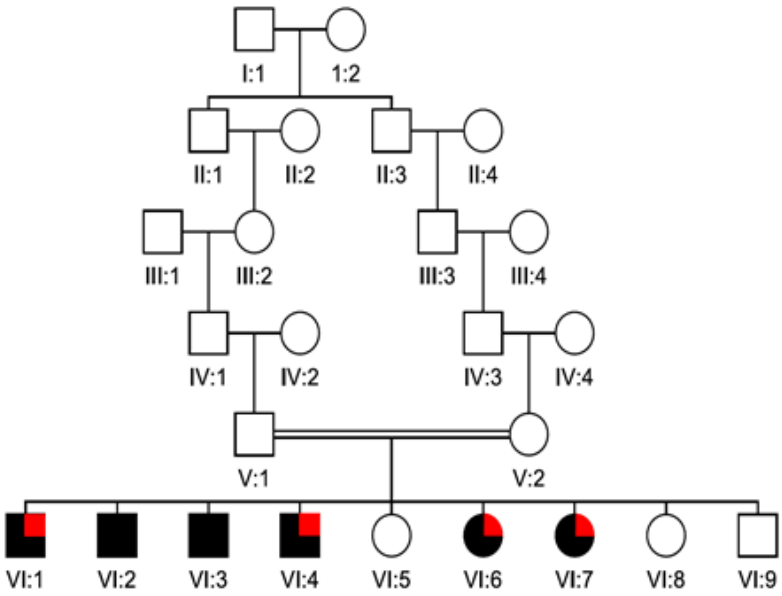
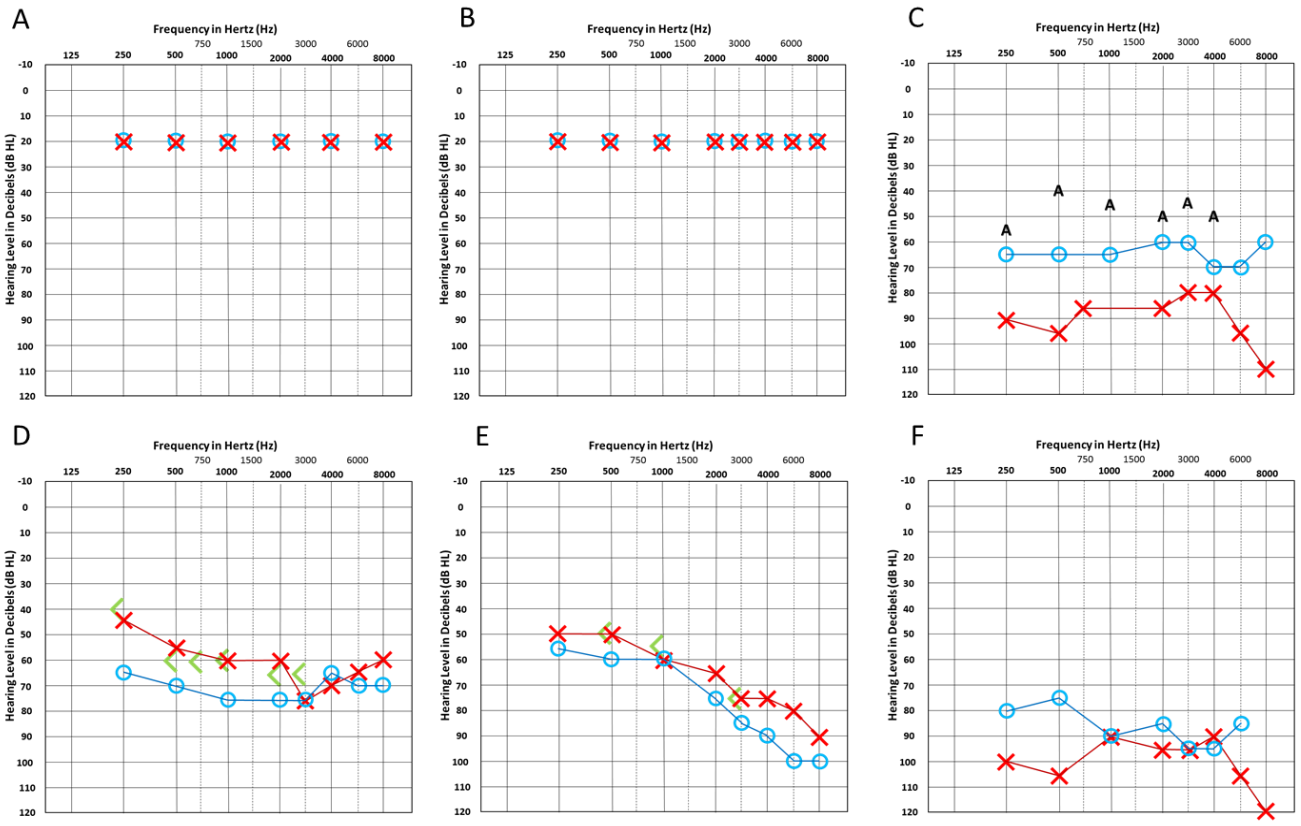


Figure 3.9: Simplified pedigree of the extended Amish family investigated for a syndromic form of hearing loss presenting with developmental delay. Development delay represented by black symbols, hearing loss denoted by red segments.

Hearing tests were carried out to confirm the individuals affected and to determine the extent of the hearing loss (Figure 3.10). Both air and bone conduction tests were carried out as part of these assessments. Although the audiograms do not display a characteristic audiometric configuration, so could not be used to aid the identification of an underlying genetic cause, they were able to confirm the type and severity of the hearing loss.



Ear	Air	Masked	Bone	Masked	Sound Field	Masked
R	O	△	<	[	S	**A**
L	✖	□	>	]	S	**A**

*Figure 3.10: Audiograms showing the two siblings unaffected by hearing loss (A and B) and the four affected individuals (C-F). Different symbols are used to plot the results of the different conduction tests for each ear. Air conduction tests in the right (O) and left (✖) ear are performed first with sounds being played through headphones. Masking can be used to prevent sound from the ear under test being detected by the other ear. This involves a noise being played into the left ear when the right ear is being tested (△) or the right ear when the left ear (□) is being tested. Bone conduction tests, used if the air conduction test identifies a hearing impairment, involve the use of an instrument that vibrates the bones of the skull and determines the function of the right (<) or left (>) cochlear. Again masking can be used to prevent the problem of "crossover" and ensures only the right ([) or left (]) ear is tested at one time. Sound field testing may be also be used, this is where a sound stimuli is played via a loud speaker, so is not ear specific, and can be conducted with (**A**) or without (S) masking.*

Two of the audiograms (Figure 3.10, D-E) confirm the type of hearing loss as sensorineural (SNHL) as the thresholds for air (O and ✖) and bone (<) conduction are similar. If normal thresholds (<25dB HL) were obtained in the bone conduction tests with the threshold for air conduction being poorer (>25dB HL) this would indicate the cochlear is still functional and a conductive form of hearing loss, where deformity of the outer or middle ear is preventing sound waves from entering the middle ear is responsible for the hearing impairment. If both thresholds where impaired, with the bone conduction thresholds being significantly better than those in air, this would suggest a mixed type of hearing loss [187].

The thresholds obtained in these tests show that the hearing loss experienced ranges from moderate (41-55dB HL) to profound (>90dB HL).

After confirming the degree of hearing loss in the four affected siblings a genome-wide SNP microarray, using the Illumina Human CytoSNP-12v2.1 330K array, was undertaken on DNA from all six siblings. Interrogation of the resultant genotypes identified a 0.56Mb microdeletion delimited by markers rs2549956 to rs35967690 (NC_000016.10: g.29622891-30188484) located on chromosome 16p11.2 common to all the affected family members (Figure 3.11).

Deletions and duplications in this region are one of the most frequently reported genetic cause of autism spectrum disorder (ASD) and other neurodevelopmental disorders [188]. The 16p11.2 microdeletion syndrome is characterised by speech articulation abnormalities, limb and trunk hypotonia with hyporeflexia, abnormal agility, sacral dimples, seizures/epilepsy, large head size/macrocephaly, lower brain herniation

(Chiari I/cerebellar tonsillar ectopia) and an increased tendency to be overweight [188]. The affected individuals in this family were not overweight and did not display a number of these characteristic phenotypes however, they did exhibit hyperactivity and social communication issues.

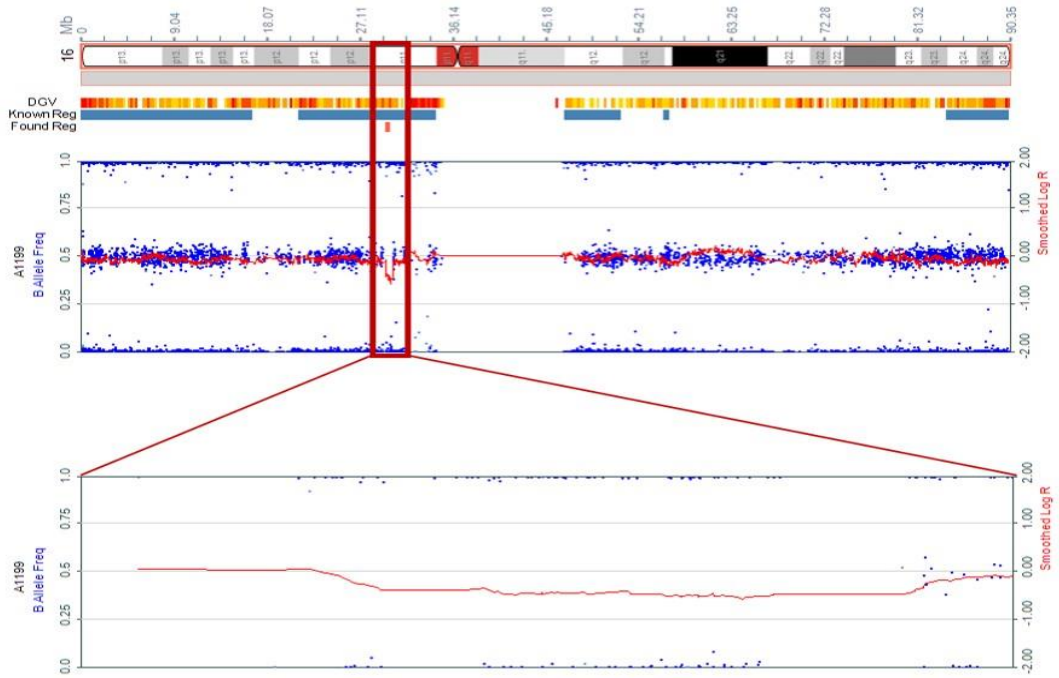


Figure 3.11: Output from KaryoStudio software (Illumina) showing ideogram of chromosome 16 and the presence of a hemizygous microdeletion at 16p11.2. Microdeletion is 0.56Mb spanning rs2549956 to rs35967690.

Whilst it is likely that this microdeletion is responsible for the neurodevelopmental delay seen in this family it is unlikely to be the cause of the inherited hearing loss. In order to identify other putative genetic causes of the hearing loss the CytoSNP microarray data was further investigated to identify regions of homozygosity specific to the four siblings affected by hearing loss. This lead to the identification of a 0.96Mb homozygous region on chromosome 13 (Figure 3.12, Aii) that encompassed the connexin 26 (*GJB2*) gene. Dideoxy sequencing analysis of the *GJB2*

gene identified a c.229T>C substitution (NM_004004.5:c.229T>C, p.W77R) (Figure 3.12, Ai) in exon 2. This is a well-established pathogenic variant that has been previously reported [155, 189] to cause autosomal recessive hearing loss so is the most likely cause of AR-SNHL observed in this family.

These findings confirm the presence of two distinct genetic disorders in the same family. Thus highlighting; the challenge of disentangling the genetic cause of complex phenotypes within the Amish community and the need to screen all individuals, recruited into the Amish Hearing Loss Programme, for variants within *GJB2*, even if the hearing loss presents with other clinical features.

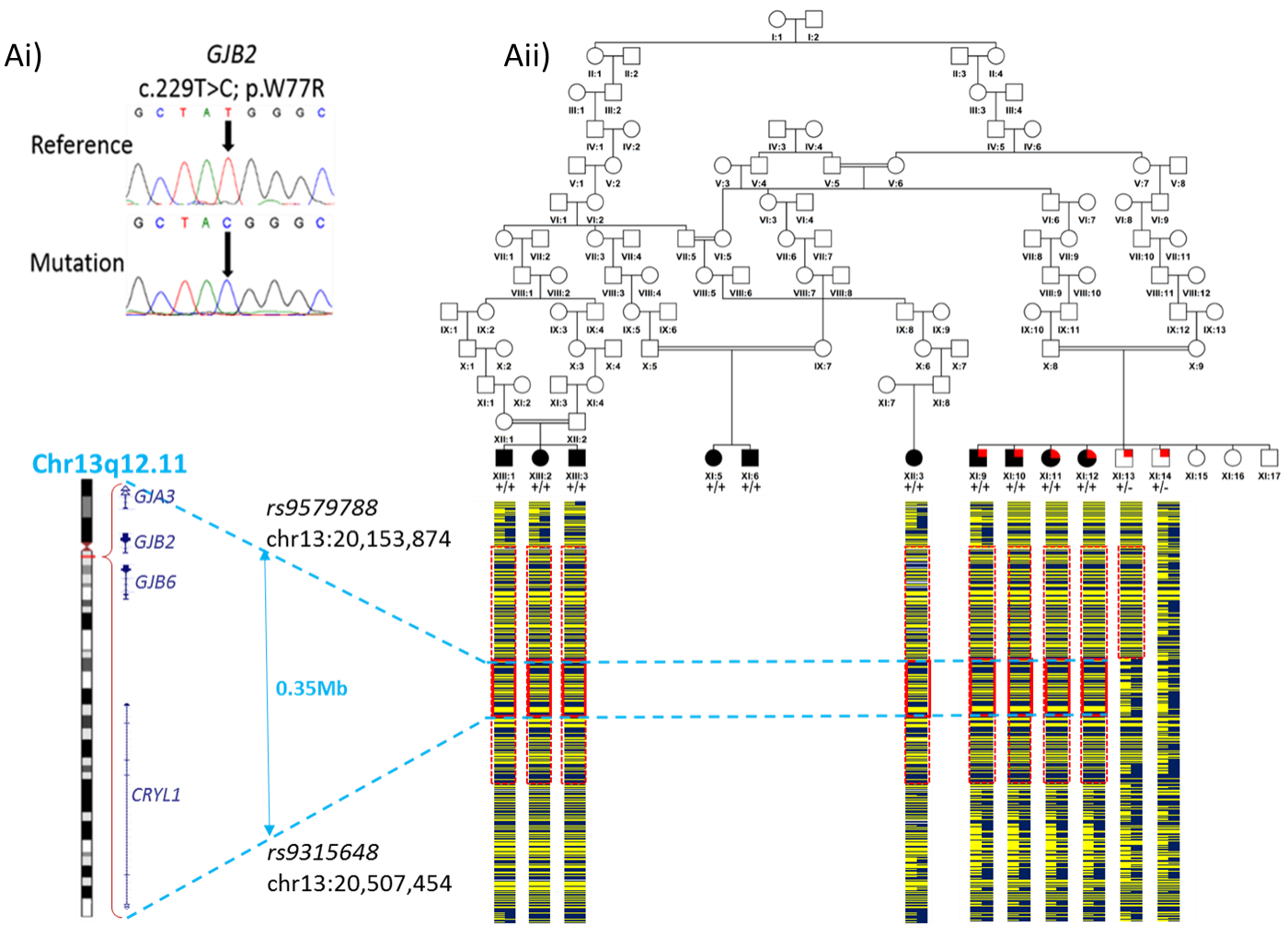


Figure 3.12: (Aii) Simplified pedigree of the extended Amish family investigated with pictorial representation of genotypes across a ~8Mb region of chromosome 13 encompassing the disease locus. Bold red box denotes the 0.96Mb region unique to the four affected siblings (and four other more distantly related individuals; XIII:1, XIIII:2, XIII:3 and XII:3) affected by hearing loss containing GJB2. Dashed red box highlights a homozygous region shared by all affected individuals and one unaffected sibling. Patient XI:13 appears to have a recombination event occur 33kb away from GJB2 preventing him from developing hearing loss, this also permitted the shared region containing GJB2 to be refined to 0.35Mb (Ai) Sequencing chromatograms showing the position of the GJB2 c.229T>C mutation.

3.5.2 GJB2 gene mutation is a common cause of NSHL in the Amish

Families recruited to the Amish Hearing Loss Programme presenting with non-syndromic hearing loss where initially screened for variants in *GJB2*, the most common genetic cause of congenital NS-SNHL. This initial screening process identified seven families with variants in *GJB2* (Table 3.4) including four families with affected individuals homozygous for NM_004004.5: c.229T>C; p.Trp77Arg variant and one family with an affected individual homozygous for the common NM_004004.5: c.35del; p.Gly12Valfs variant (Figure 3.13, Family 6,). In the remaining two families affected individuals were found to be compound heterozygotes for both c.35del and c.229T>C variants.

Table 3.4: Summary of families in which GJB2 variants were found during initial screening.

Variant(s) Identified	Genotype	Region	Number of affected	Family ID (Figure 3.13)
NM_004004.5:c.229T>C; p.Trp77Arg	Homozygous	Geauga	1	Family 1
		Indiana	3	Family 3
		Indiana	4	Family 4
		Indiana	2	Family 5
NM_004004.5:c.35del; p.Gly12Valfs	Heterozygous	Geauga	1	Family 6
NM_004004.5:c.229T>C; p.Trp77Arg	Compound heterozygous	Geauga	2	Family 7
& NM_004004.5:c.35del; p.Gly12Valfs		Geauga	2	Family 8

Figure 3.13 details all families in the Amish Hearing Loss Programme where a variant in *GJB2* has been found to be responsible for the observed hearing impairment, this includes the family with two distinct genetic disorders described in section 3.5.1 (Figure 3.13, Family 2).

There have been no previous studies published reporting specific *GJB2* variants and their frequencies in the Amish community. These findings suggest both the

c.35del and c.229T>C *GJB2* variants are likely to represent founder mutations. Interestingly the p.W77R variant has previously been reported in two interrelated families of Israeli-Arab origin [189], where both homozygosity and compound heterozygosity with c.35delG were observed.

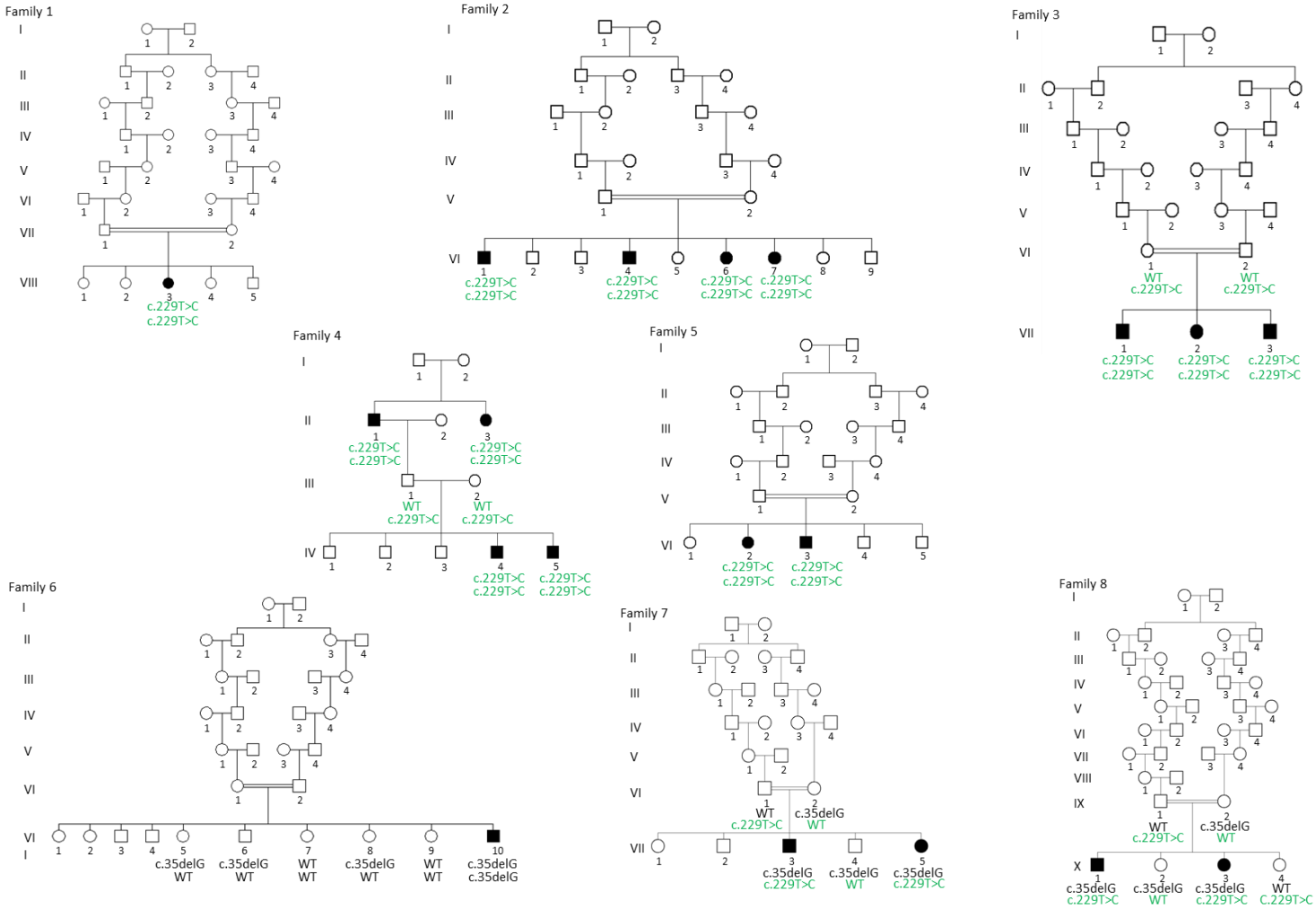


*Figure 3.13: Simplified pedigrees of eight Amish families with hearing loss caused by variants in GJB2. Hearing loss observed in families 1-5 is caused by the c.229T>C variant. Family 1 reside in Geauga, Family 2 are from Wisconsin and Families 3-5 are located in Indiana. Genotypes of sequenced individuals are shown below (in **green**). A wildtype allele for this variant is shown by WT. Hearing loss in family 6, from Geauga, is caused by a c.35del variant with the genotypes for sequenced individuals shown below (in black). A wildtype allele for this variant is shown by WT. Affected individuals in families 7 and 8, both located in Indiana, are compound heterozygous for both the c.229T>C (in **green**) and c.35del (in black) variants.*

3.5.3 Investigating SLC15A5 as a candidate molecule responsible for AR NS-SNHL

Genetic and Clinical data

A family presented with two siblings affected by non-syndromic hearing loss (Figure 3.14). The two siblings were initially screened for variants in *GJB2* but none were found. Due to the inheritance pattern observed in the family it was evident that the hearing loss experienced by the two affected individuals was a result of an autosomal recessive genetic abnormality. Hearing tests were conducted to determine the type and extent of the hearing loss (Figure 3.15).

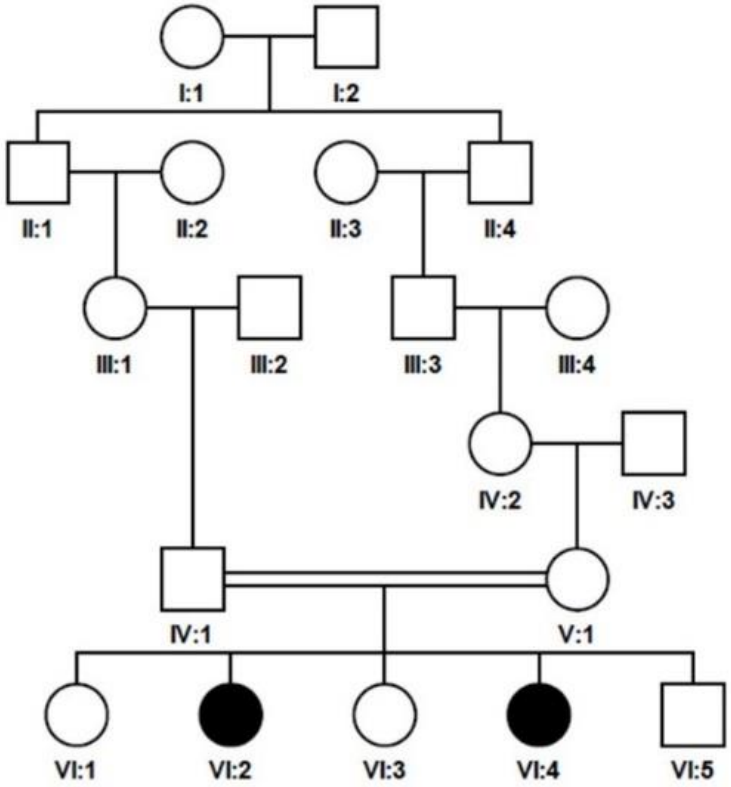


Figure 3.14: Simplified pedigree of the extended Amish family investigated for inherited hearing loss.

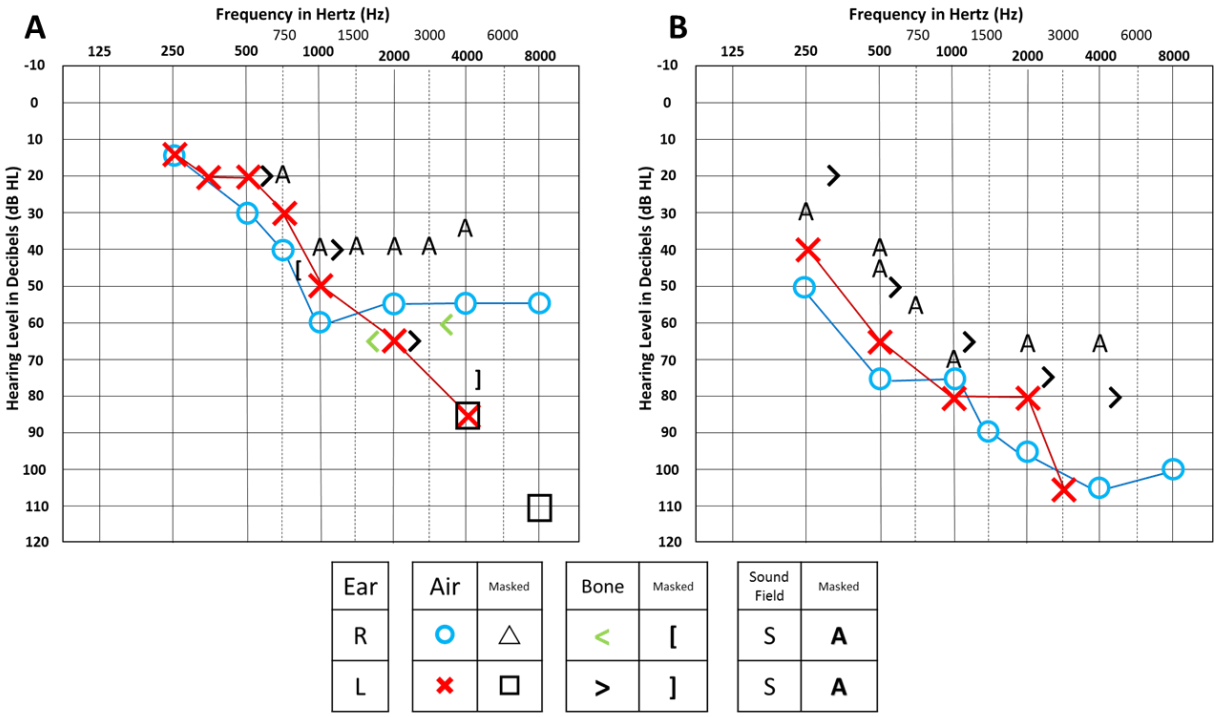


Figure 3.15: Audiograms showing the two siblings affected by hearing loss. Different symbols are used to plot the results of the different conduction tests for each ear. Air conduction tests in the right (□) and left (□) ear are performed first with sounds being played through headphones. Masking can be used to prevent sound from the ear under test being detected by the other ear. This involves a noise being played into the left ear when the right ear is being tested (□) or the right ear when the left ear (□) is being tested. Bone conduction tests, used if the air conduction test identifies a hearing impairment, involve the use of an instrument that vibrates the bones of the skull and determines the function of the right (<) or left (>) cochlear. Again masking can be used to prevent the problem of "crossover" and ensures only the right ([) or left (]) ear is tested at one time. Sound field testing may be used, this is where a sound stimuli is played via a loud speaker, so is not ear specific, and can be conducted with (A) or without (S) masking.

The audiograms (Figure 3.15Figure 3.10) confirm the type of hearing loss as sensorineural (SNHL) as the thresholds for air (O and ✖) and bone (<) conduction are similar. The thresholds obtained in these tests indicate that the hearing loss experienced by sibling A ranges from mild (20-40dB) to moderately severe (56-70dB) in their right ear and mild (20-40dB) to severe in their left ear (56-70dB). Sibling B appears to be experiencing a more severe hearing impairment with severity ranging from moderate (41-55dB) to profound (>90dB) in both ears.

To define the genetic abnormality responsible, genome-wide SNP mapping on affected family members was undertaken, using the Illumina CytoSNP-12 (330k) BeadChip platform in combination with whole exome sequencing, on a carrier parent (due to limitations in DNA availability from affected family members), using the Agilent Human exome V4 (51Mb) capture. The genome-wide SNP data identified a number of homozygous regions. Table 3.5 details the size and location of the four largest regions of homozygosity shared by the siblings.

Table 3.5: Regions of shared homozygosity between affected siblings.

SNPs Flanking Region	Genomic Positon of SNPs	Size of Region of Homozygosity
rs4763845;rs1472874	chr12:12728803-24575272	11.8Mb
rs4736695;rs11167068	chr8:134556752-142549778	8Mb
rs2824733;rs2830165	chr21:19689082-27676592	8Mb
rs408307; rs4876938	chr9:136892523-141213431	4.1Mb

Whole exome sequencing filtering for rare (frequency) and novel heterozygous variants in the carrier parent identified no candidate variants in known disease-associated genes. However, cross-referencing these datasets identified a single candidate nonsense variant (NM_001170798.1:c.865G>T; p.Glu289Ter) (Figure 3.16) located in a 8.2Mb homozygous block common to the affected individuals, in *SLC15A5*, a gene of unknown function not previously associated with inherited disease, as a candidate cause of this condition. The variant was validated using dideoxy sequencing, and found to cosegregate appropriately within the family.

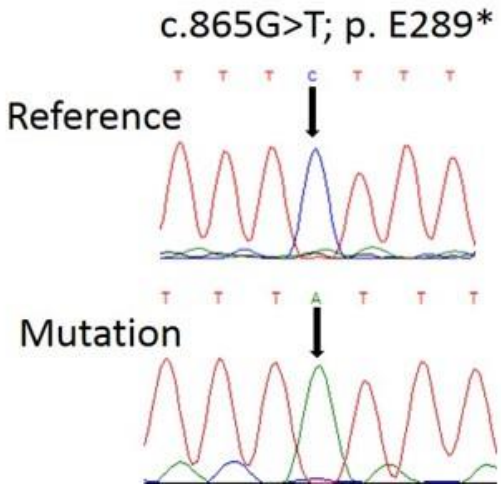


Figure 3.16: Sequencing chromatograms comparing the identified SLC15A5 mutation to the wildtype reference sequence.

In-silico investigations predicted the variant to be disease-causing as a result of nonsense mediated decay (NMD) of the protein product. Whilst the variant was listed in the Genome Aggregation Database (GnomAD) 16 times; in 14 individuals of European origin (MAF 0.0001989) and 2 of African descent (MAF 0.0001231), no homozygotes were reported. The variant was not reported in 1000 genomes database.

Genotyping studies of the variant was initially undertaken in 164 unaffected control Amish samples using dideoxy sequencing and identified 15 heterozygous individuals, corresponding to an allele frequency of 0.0457. Additional genotyping studies of the variant in 167 unaffected control Amish from different Amish communities (Table 3.6), using PlexSeq, a multiplexed amplicon sequencing approach (Illumina) [102], as described in chapter five, identified eight heterozygous individuals, corresponding to an allele frequency of 0.0227.

Table 3.6: Allele frequency data for SLC15A5 variant determined from an Amish population cohort of 167 unaffected individuals.

Gene	Variant	AF (by region)	AF	gnomAD Freq. (March 2019)*
SLC15A5	c.865G>T; p.Glu289Ter (NM_001170798.1) Chr12:g.g.16244690C>A	Indiana	-	0.0001915
		Ohio Holmes	0.0221	
		Ohio Geauga	0.0273	
		Wisconsin	0.0400	
		Total	0.0234	

**Allele frequency quoted from gnomAD refers to European (non-Finnish) AFs correct as of March 2019.*

Given this finding the *SLC15A5* variant remained a candidate cause of the condition, although no other individuals with SNHL were found to be homozygous for the *SLC15A5* variant.

Existing functional information on the SLC15A5 protein

Given the candidacy of *SLC15A5* as a cause of this condition, additional bioinformatic and molecular information was sought regarding the function of this molecule. The solute carrier family 15 member 5 (*SLC15A5*) gene is located on chromosome 12p12.3 comprising nine exons. *SLC15A5* predicted to encode a 579 amino acid proton oligopeptide cotransporter within the superfamily of solute carriers. The SLC15A5 polypeptide sequence is predicted to contain 11 helical transmembrane domains (Figure 3.17).

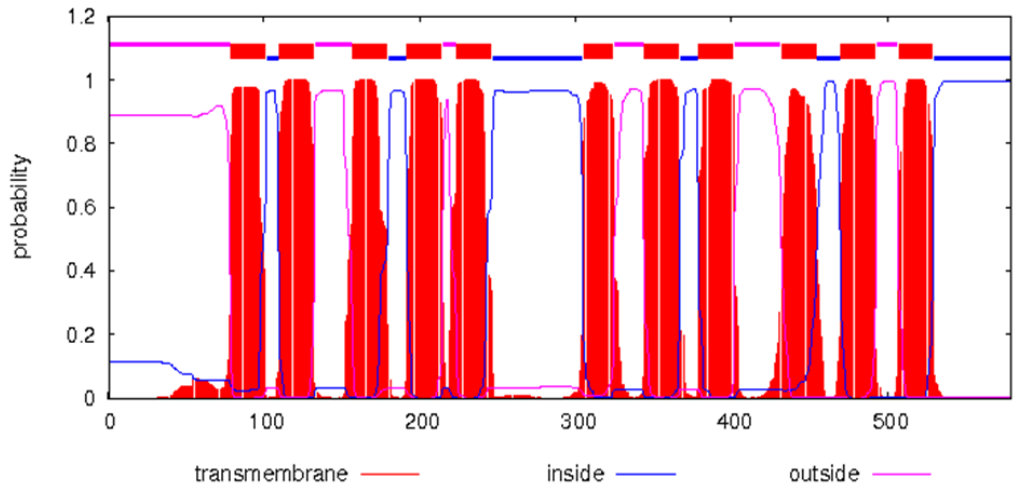


Figure 3.17: TMHMM Server v. 2.0 output. Showing the prediction of transmembrane helices in SLC15A5.

Very little is currently known about the specific function of SLC15A5 [190] as very few studies have been undertaken to characterise or determine its expression. Reviews detailing the expression of other family members, particularly SLC15A1 (PEPT1) and SLC15A2 (PEPT2), are available in the literature detailing their roles in;

- the absorption and conservation of dietary protein digestion products
- maintaining homeostasis of neuropeptides in the brain
- the absorption and disposition of a number of pharmacologically important compounds [190-193]

However, as the most distantly related member of the SLC15A proton-peptide exchange transporter family, sharing approximately only 50% amino acid identity with its other family members [194] it is hard to draw direct comparisons with regard to its expression or function.

Preliminary studies of *SLC15A5* gene product

Functional studies were carried out through a number of collaborations. Dr Morag Lewis, from The Wolfson Centre for Age-Related Diseases, King's College London, kindly undertook the Immunohistochemical investigations in the inner mouse ear and Dr Barbara Vona (University of Tübingen), undertook the RT-PCR studies. The aim of these studies was to ascertain the subcellular localisation of wildtype SLC15A5 to gain further insight into the molecular role of this molecule.

Immunocytochemistry experiments in HEK293 cells

A pCMV6-entry *SLC15A5* clone was transformed into *E.coli* DH5-alpha bacteria via heat shock. Initial immunocytochemistry experiments in HEK293 cells transfected with epitope tagged SLC15A5 showed SLC15A5 in a vesicular pattern (Figure 3.18). However, wide-field microscopy showed that the transfection efficiency of this construct was low. As approximately 70% fluorescence is required to ensure enough protein is produced to carry out a Western blot it was necessary to take steps to improve the uptake of the construct into cells. To do this is was necessary to determine what was causing the low transfection rate.

The first step was to rule out issues with the cell line being used for the transfections. To do this mouse fibroblast cells, NIH/3T3, where transfected with the pCMV6-*SLC15A5* construct. This cell line, named after the number of days the cells were originally allowed to grow (3) and the number of hundreds of thousands of cells (3) transferred (T) onto a new plate during each passage, are well known for being highly transfectable by DNA. Unfortunately, these transfections did not significantly improve the transfection efficiency indicating the cell type was not the cause of the low transfection rate.

To determine if aspects of the pCMV6-SLC15A5-myc-FLAG construct itself were affecting transfection a yellow fluorescent protein (YFP) tag was added to the C-terminal of the protein in place of the FLAG epitope and sub-cloned into a pCAGGs vector, creating pCAGGs-SLC15A5-myc-YFP. This construct did produce higher transfection levels and permitted some co-expression studies to be undertaken. When co-expressed with Rab7-RFP, a widely used marker of recycling endosomes [195], SLC15A5 showed extensive, but not complete, overlap (Figure 3.19). This may indicate that SLC15A5 is associated with lysosomes and/or endosomal trafficking and not the outer cell membrane.

However, due to persistently low expression of the SLC15A5 antibody it was proposed that constructs with the epitope tags at the N-terminus of the protein be created to rule out the possibility the large YFP epitope were preventing the antibody from binding properly and to determine if antibody binding to the C-terminus of the antibody was affecting the protein's native structure and/or cell localisation.

Advice regarding the binding efficiency of the commercially available polyclonal SLC15A5 antibody was sought from Dr Francesco Rao, Chief Scientific Officer at Dundee Cell Products. It was suggested the peptide sequence being targeted by the commercial antibody was relatively long (28 residues) so may be condensing to form an aggregate structure thus preventing antibody binding. It was also reported that our protein of interest (SLC15A5) shared some similarities to a proton-coupled transporter, nitrate transporter (NRT1.1).

Based on this information it was suggested further immunocytochemistry experiments be conducted using a custom monoclonal antibody targeting new epitope regions (Dr Francesco Rao, Dundee Cell Products/Dr John Chilton, personal communication).

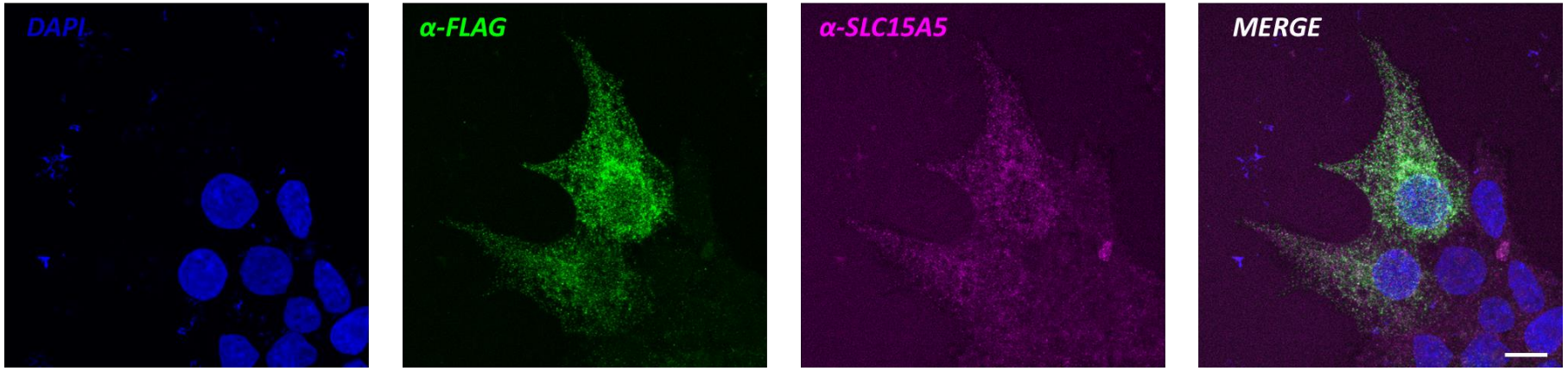


Figure 3.18: Distribution of tagged SLC15A5 in HEK293 cells. Cells were transfected with pCMV6-SLC15A5-myc-FLAG (origene) fixed in 4% PFA and immunolabelled with anti-FLAG (green) and anti-SLC15A5 (magenta) antibodies. Nuclei are counterstained with DAPI (blue). Scale bar = 7.5µm

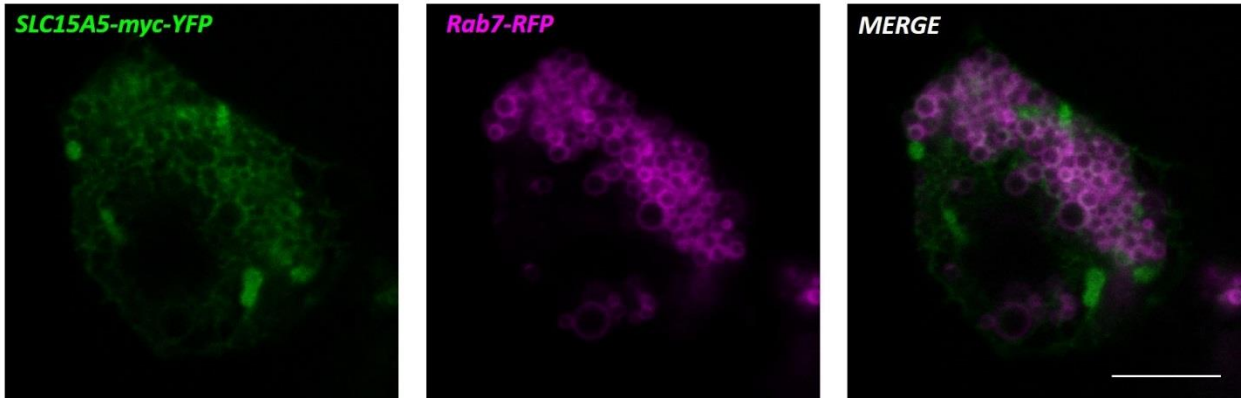


Figure 3.19: SLC15A5 partially associates with late endocytic structures in live cells. SLC15A5-myc-YFP (green) associates with Rab7-RFP (magenta). HEK 293 cells were transfected with SLC15A5-myc-YFP and viewed on the wide field microscope. Scale bar = 7.5µm

Western blot analysis

Western blot analyses were performed on lysed HEK 293 cells transfected with pCAGGs-SLC15A5-myc-YFP with membranes probed with rabbit polyclonal anti-SLC15A5 and anti-cmyc to detect the presence of SLC15A5 in transfected cells. Unfortunately, no SLC15A5 protein was detected (Figure 3.20a). The membrane was then post stained with Anti-GAPDH (Figure 3.20b), a protein integral for glycolysis and plays many roles in nuclear function, known to be expressed in HEK 293 cells. The result of this staining showed that protein was present on the membrane.

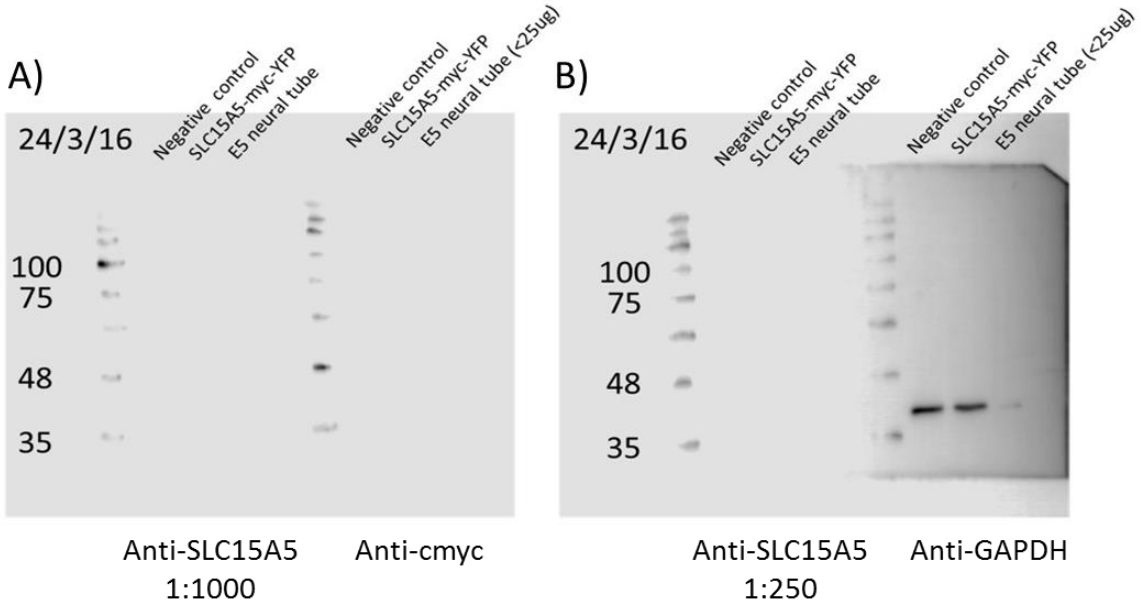


Figure 3.20: SLC25A5 Western blot

A lack of signal being detected by the SLC15A5 antibody could be a result of there not being enough SLC15A5 protein in the samples as a result of the low transfection efficiency or antibody binding issues, mentioned previously, which indicates a custom antibody for this protein or experiment type may be required.

Investigating the expression of *Slc15a5* in the mouse ear

Experiments to investigate the expression of murine *Slc15a5* in mouse tissue sections at various time points of embryonic development in the mouse using RT-PCR were undertaken (Figure 3.21). These experiments yielded promising results, showing cochlear expression.

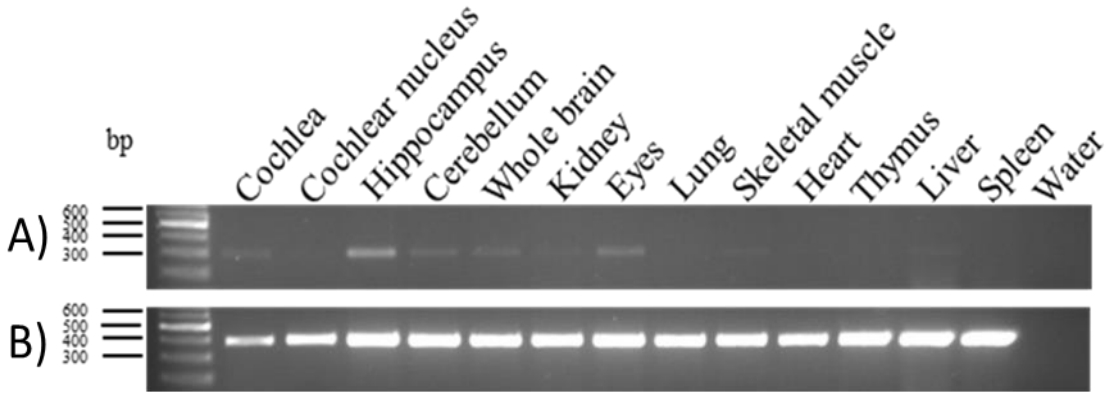


Figure 3.21: RT-PCR analysis of Slc15a5 in the P7 C57BL/6J mouse. A) RT-PCR amplification of Slc15a5 was obtained using primers in exons 5 and 6 (NM_177787). A 294 bp product is observed in the cochlea, hippocampus, cerebellum, whole brain, kidney, eyes, and liver. B) PCR amplification of ubiquitously expressed Hprt is shown below as a positive control. Amplification products were confirmed via Sanger sequencing. Image depicts results obtained from investigations carried out by Barbara Vona.

To investigate this further, studies were undertaken in collaboration with Professor Karen Steele's lab to determine the expression of Slc15a5 in paraffin sections of inner ear mouse tissue at embryonic day 16.5 (E16.5). Unfortunately however, no protein could be detected (**Appendix G**). As the peptide sequence targeted by the custom antibody appears to be located within a transmembrane (residues 542-570)

New genotyping data as part of a high-volume sequencing program excludes mutation of *SLC15A5* as a cause of NSHL.

Towards the latter stages of this study, new information became available regarding the frequency of the *SLC15A5* (NM_001170798.1:c.865G>T; p.Glu289Ter) variant in other Amish communities, through a recently developed data exchange collaboration with workers investigating the causes of inherited diseases in the Pennsylvanian Amish community. This study involved WES in >1000 Amish individuals from this community, including affected and unaffected individuals, in order to provide important gene variant annotation data. As part of this collaborative study, we recently received data which identified two individuals with no known NSHL to be homozygous for the c.865G>T; p.Glu289Ter variant. These findings likely exclude this *SLC15A5* variant as being causative of this condition.

3.5.4 Genetic studies define variant frequencies of causative gene mutations in distinct Amish communities

To learn more about the prevalence of the two *GJB2* variants, and other founder gene mutations linked to hearing loss in different Amish communities, genotyping studies were undertaken of each variant in 167 unaffected control Amish individuals from Ohio (Holmes County), Ohio (Geauga County), Indiana, and Wisconsin communities using PlexSeq sequencing. As expected, this defined remarkably divergent allele frequencies (AF) for each gene (Table 3.7), reflecting the distinct ancestral histories of each Amish community.

The most common genetic causes of hearing loss identified across all the Amish communities related to *ST3GAL5* (GM3 synthase deficiency) and *GJB2* gene variants. The severe neurodevelopmental disorder GM3 synthase deficiency is known to be common in both Ohio Amish communities, corroborated by the relatively high AF seen for this variant of 0.022% (Geauga County) and 0.055% (Holmes County). Corroborating our *GJB2* gene data (above), *GJB2* variants (c.229T>C; p.Trp77Arg and c.35del; p.Gly12Valfs) were also identified particularly in Amish families from Wisconsin and Ohio (Geauga County) in whom the combined AF for both *GJB2* variants is approximately 0.08% and 0.02% respectively. However, it was notable that the p.Trp77Arg *GJB2* variant, which is a common disease allele in white Caucasian families [196] was absent in controls from Ohio (Holmes County), indicative of a low incidence of NSHL due to this variant in this region.

The *HYAL2* founder variant (c.443A>G; p.Lys148Arg) is common in both Ohio communities with frequencies of 0.029% (Holmes County) and 0.018% (Geauga). Whilst hearing loss is not a cardinal feature the exclusion of any

known, or potentially deleterious, variants in all currently reported hearing loss genes, through the interrogation of WES from affected individuals, is highly suggestive that hearing impairment is a variable feature of this disorder (**Appendix F**).

The most common single gene variant detected overall was the *ST3GAL5* GM3 synthase deficiency founder variant which underlies a syndromic form of infantile epileptic encephalopathy with hearing loss present at birth [197].

Table 3.7: Summary of genes associated with hearing loss identified in the Amish community. Allele frequency was determined from an Amish population cohort of 167 unaffected individuals

Type	Gene	Variant	Hearing Loss	Allele Freq. (by Region)	AF	gnomAD Freq.
SNHL	*PCNA*	c.683G>T; p.Ser228Ile (NM_002592.2) chr20:g.5115472C>A	Prelingual onset, moderate to profound high frequency SNHL	Indiana Ohio Holmes Ohio Geauga Wisconsin TOTAL	- 0.015 - - 0.006	0.000007955
SNHL	*SLITRK6*	c.1240C>T; p.Gln414Ter (NM_032229.2) chr13:g.85795269G>A	Bilateral, prelingual moderate to severe SNHL	Indiana Ohio Holmes Ohio Geauga Wisconsin TOTAL	- 0.015 0.019 0.060 0.020	0.00001225
SNHL	*KCNQ1*	c.451_452delCT; p.Leu151Glyfs (NM_000218.2) chr11: g.2527992_2527993delCT	Congenital, bilateral, profound SNHL	Indiana Ohio Holmes Ohio Geauga Wisconsin TOTAL	- 0.007 0.018 - 0.008	0
SNHL	*ST3GAL5*	c.862C>T; p.Arg288Ter (NM_003896.3) chr2:g.85844542G>A	Variable onset and severity SNHL	Indiana Ohio Holmes Ohio Geauga Wisconsin TOTAL	- 0.022 0.055 0.02 0.028	0.00002527
CON.	*HYAL2*	c.443A>G; p.Lys148Arg (NM_003773.4) chr3:g.50320047T>C	Mild to moderate, unilateral/ bilateral and pre or postlingual (although one individual had SNHL)	Indiana Ohio Holmes Ohio Geauga Wisconsin TOTAL	0.022 0.029 0.018 - 0.020	0
MIXED	*COL1A2*	c.2237G>T; p.Gly610Cys (NM_000089.3) chr7:g.94420590G>T	Progressive mixed hearing loss	Indiana Ohio Holmes Ohio Geauga Wisconsin TOTAL	- - - - -	0
NS-SNHL	*GJB2*	c.35delG; p.Gly12Valfs (NM_004004.5) chr13:g.20189547delC		Indiana Ohio Holmes Ohio Geauga Wisconsin TOTAL	- 0.015 0.009 0.020 0.011	0.006258 (10 homs)
NS-SNHL	*GJB2*	c.229T>C; p.Trp77Arg (NM_004004.5) chr13:g.20189353A>G		Indiana Ohio Holmes Ohio Geauga Wisconsin TOTAL	- - 0.018 0.040 0.011	0.00004062

3.5.5 Identification of novel hearing loss gene

As the above data was being compiled for publication, ongoing studies in two interlinking Amish families comprising five affected individuals with autosomal recessive syndromic SNHL and neurodevelopmental delay, identified a gene variant as a putative candidate (Figure 3.22).

Affected individuals underwent *GJB2* screening which identified no known pathogenic variants within the gene. A combination of genome-wide SNP mapping and exome sequencing studies was then undertaken in these families. Interrogation of the WES excluded variants in all genes currently reported to cause hearing loss.

A nonsense variant in a gene encoding a microtubule-associated molecule located on chromosome 13q was identified as the only candidate cause of the condition. The variant cosegregates as appropriate for an autosomal recessive condition in all family members, and is not listed in online genome databases (gnomAD, ExAC, 1000 genomes). Ongoing genotyping studies in the Amish has identified only two carriers of the variant in 300 Amish control chromosomes.

Although not previously linked to hearing loss the gene has been linked to a recessive form of intellectual disability (ID) [198, 199]. Interestingly this variant was also, initially, identified as a potential cause of ID in the Amish, through a novel, proof-of-principle study described in Chapter 5 that set out to identify potentially deleterious coincidentally carried heterozygous variants, that had yet to be reported in the community.

Further studies of this gene lie beyond the scope of this thesis, and are currently ongoing within the WoH research group.

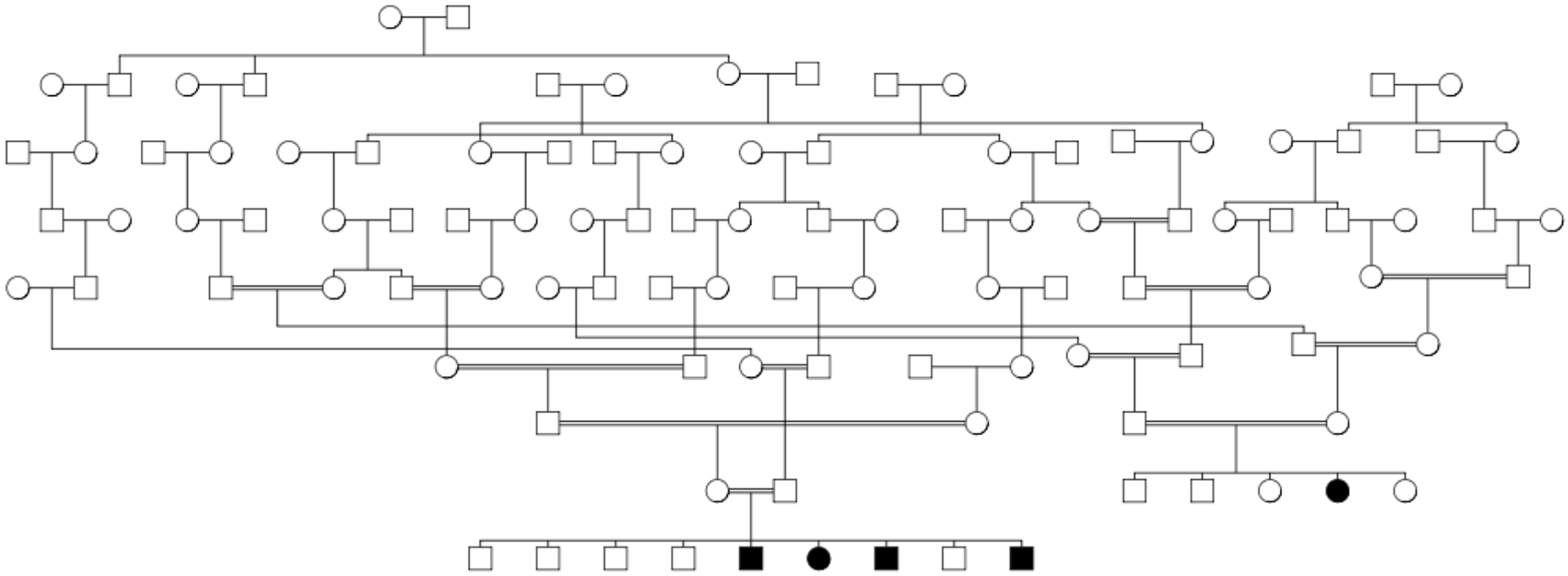


Figure 3.22: Simplified pedigree linking two Amish families comprising of five individuals affected with AR SNHL and neurodevelopmental delay.

3.6 Discussion

The work described in this chapter investigates the nature, aetiology and frequency of genetic causes of hearing loss in families from the Amish community.

The Amish hearing loss programme itself stems from a long running clinical-research study, the Windows of Hope programme which aims to learn more about the molecular basis of inherited disease, in this instance hearing loss, within the Amish community and provide important information to aid much needed diagnoses for affected individuals and their families. This is achieved through the translation of research findings to directly benefit patients in the form of improved diagnostic information, supporting genetic counselling and aiding the implementation, or development of novel, targeted therapies for use within this community and the general population.

As a result of the work outlined here, and through previous studies carried out by this group [177, 178, 184], 11 of the 19 families recruited to the Amish Hearing Loss Programme, with no previous genetic diagnosis, have now received confirmed genetic diagnoses (Figure 3.23). Additionally, the genetic cause of the observed hearing loss in two further families is under investigation and likely to be confirmed in the near future (section 3.5.5). Investigations are ongoing within the six families unfortunately still awaiting diagnoses.

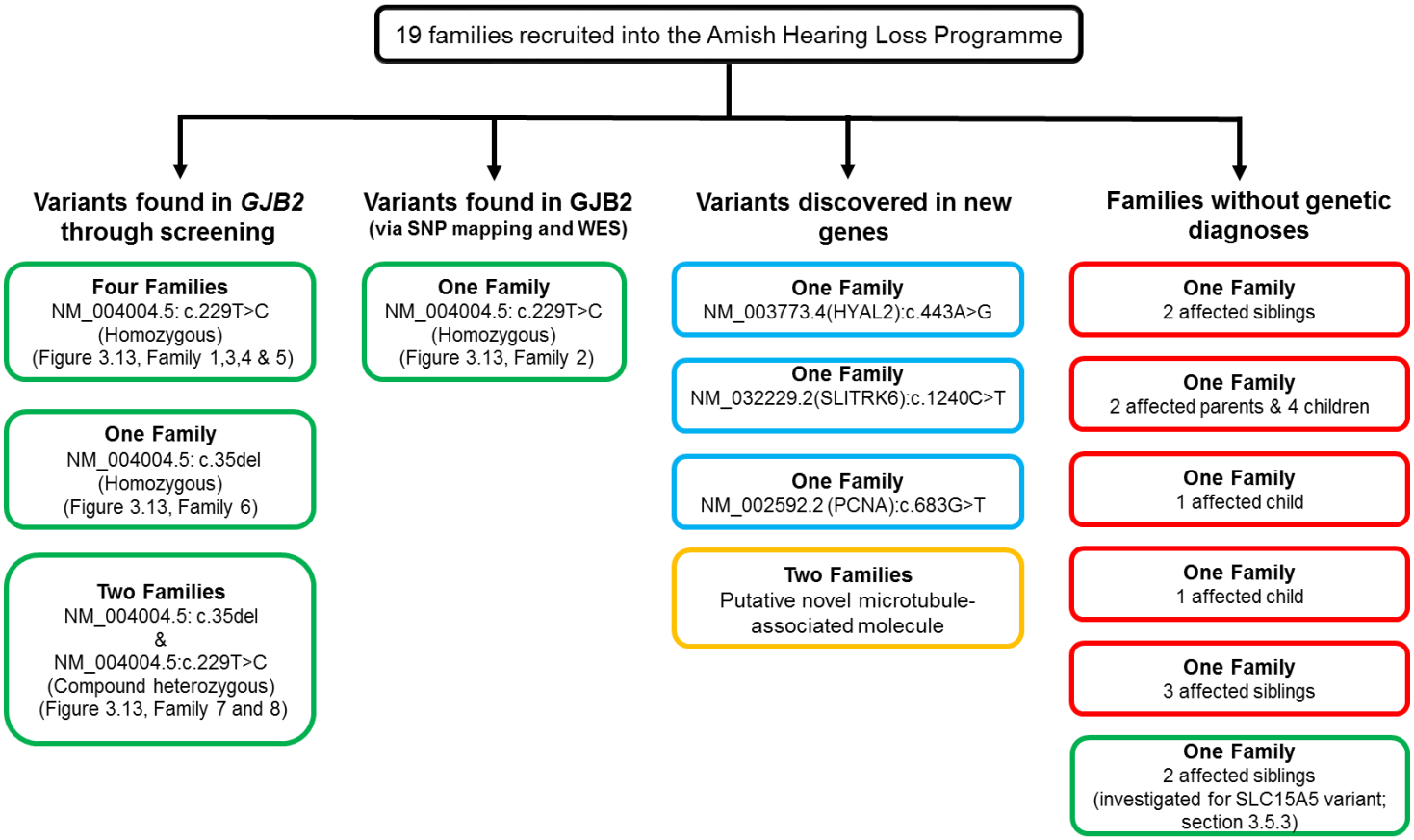


*Figure 3.23: Summary of families recruited into the Amish Hearing Loss Programme. Variants responsible for the hearing loss have been detailed where possible. **Green** boxes denote families where the investigations undertaken form part of this thesis. The **yellow** box highlights a family where the variant believed to be responsible was identified as part of this thesis (Chapter 5) despite the family studies being carried out by another group member (not included in this thesis). **Blue** boxes denote families where novel variants where discovered by other group members (not included in this thesis). **Red** boxes represent families that have not yet received a diagnosis with investigations ongoing.*

Certain characteristics of the Amish population, including origins from relatively few founder individuals, endogamy, large family size, and detailed genealogical records, enabling the construction of impressive family pedigrees empower genetic studies. However, these characteristics may also result in certain disorders, particularly those inherited in an autosomal recessive pattern, occurring more frequently in the community. These factors, in addition to the reducing cost and increasing availability of NGS technologies and high density whole genome SNP mapping (and associated linkage analysis) greatly facilitate the discovery of genes responsible for inherited disease, which might otherwise have been impossible in studies of other populations due to the genetic and environmental complexities of a condition. All of the conditions known in the Amish and described in the current study are recessively inherited with the exception of *COL1A2* and the mitochondrial disorders. As a result six new genes associated with HL in the Amish population (*PCNA, HYAL2, SLITRK6, HARS, LONP1, ST3GAL5*) have been discovered, several of which were identified by our group including a very recently discovered microtubule-associated molecule (section 3.5.5).

3.6.1 Identification of two distinct genetic disorders within the same Amish family

Neurodevelopmental disorders (NDD) are a group of disorders involving the abnormal development of the central nervous system. Affecting 1-3% of the population of children under 5 years of age it is one of the most common conditions presenting in paediatric clinics [200]. Due to its extreme heterogeneity and overlapping clinical outcomes of many distinct genetic causes of these conditions, reaching a specific diagnosis may be difficult. Although the diagnosis

of NDD has improved in recent years as a result of advances in genetic technologies [201] the cause of the condition remains undetermined in approximately half of affected individuals [202]. To ensure the best overall outcomes for affected children and their families it is important clinicians establish a diagnosis so they can implement the most appropriate therapeutic management strategy [200].

The Windows of Hope team were asked to assist in reaching a diagnosis for a family with six individuals displaying NDD, four of whom also had SNHL. Due to the broad range of genetic abnormalities known to result in a highly diverse spectrum of inherited neurodevelopmental disabilities, which often give rise to other disabilities, including visual and hearing impairment, it was not clear if the same, or two distinct genetic causes, were responsible for the difficulties observed in these children. Genome-wide SNP microarray identified a 0.56Mb microdeletion located on chromosome 16p11.2, previously shown to result in a clinically variable neurological phenotype [203], with dideoxy sequencing confirming the presence of a p.W77R GJB2 variant, a well-documented cause of AR-SNHL, common to all affected individuals. This confirmed the presence of two distinct genetic disorders in the same family.

Copy number variations (CNVs) within 16p11.2 are one of the most common structural chromosome disorders [188] with a prevalence of ~3/10,000 in the general population [204]. Recurrent ~600kb deletions or duplications are the most common genetic aetiologies of NDD and autism spectrum disorders (ASD), with a prevalence of ~1% in ASD patients [188]. The phenotype is characterised by a spectrum of neurodevelopment impairments including; developmental delay, language impairments, mild to moderate intellectual disability, schizophrenia,

altered body mass index, epilepsy and ASD [203, 205]. Whilst hearing impairment, both SNHL and conductive, has previously been reported in up to 11% of individuals with this microdeletion [204], a more recent study characterising the range and frequency of neurological variation within the phenotype did not associate hearing loss with this CNV. This study concluded that the a 16p11.2 deletion is characterised by;

- Highly prevalent (>75%) speech articulation abnormalities
- Hypotonia (low muscle tone) with hyporeflexia (below normal or absent reflexes.
- Poor agility
- Sacral dimples
- Seizures/epilepsy
- Large head size
- Chiari I/cerebellar tonsillar ectopia [188]

The affected individuals in this family were not did not display all of these phenotypes though they did display hyperactivity and social communication issues. The notable absence of any form of hearing impairment as a cardinal feature of this syndrome, indicated that it was unlikely to account for the hearing loss seen in the family which was subsequently shown to be due to *GJB2* mutation.

Due to the unique genomic architecture of the Amish population, as a result of a genetic bottleneck leading to the enrichment of certain disease-associated alleles, the co-occurrence of distinct two genetic disorders in the same family is not uncommon and is likely to occur more frequently than in the general population. The Windows of Hope team have assisted a number of families in

which three distinct inherited conditions are present amongst different family members, with some individuals being affected by all three disorders (Crosby/Baple personal communication). Due to this, and the outcome of this study all families recruited into the Amish Hearing Loss Programme are now screened for variants in *GJB2,* irrespective of any additional clinical phenotypes reported in affected individuals. Figure 3.24 summarises the screening strategy now implemented for new patients recruited to the programme.

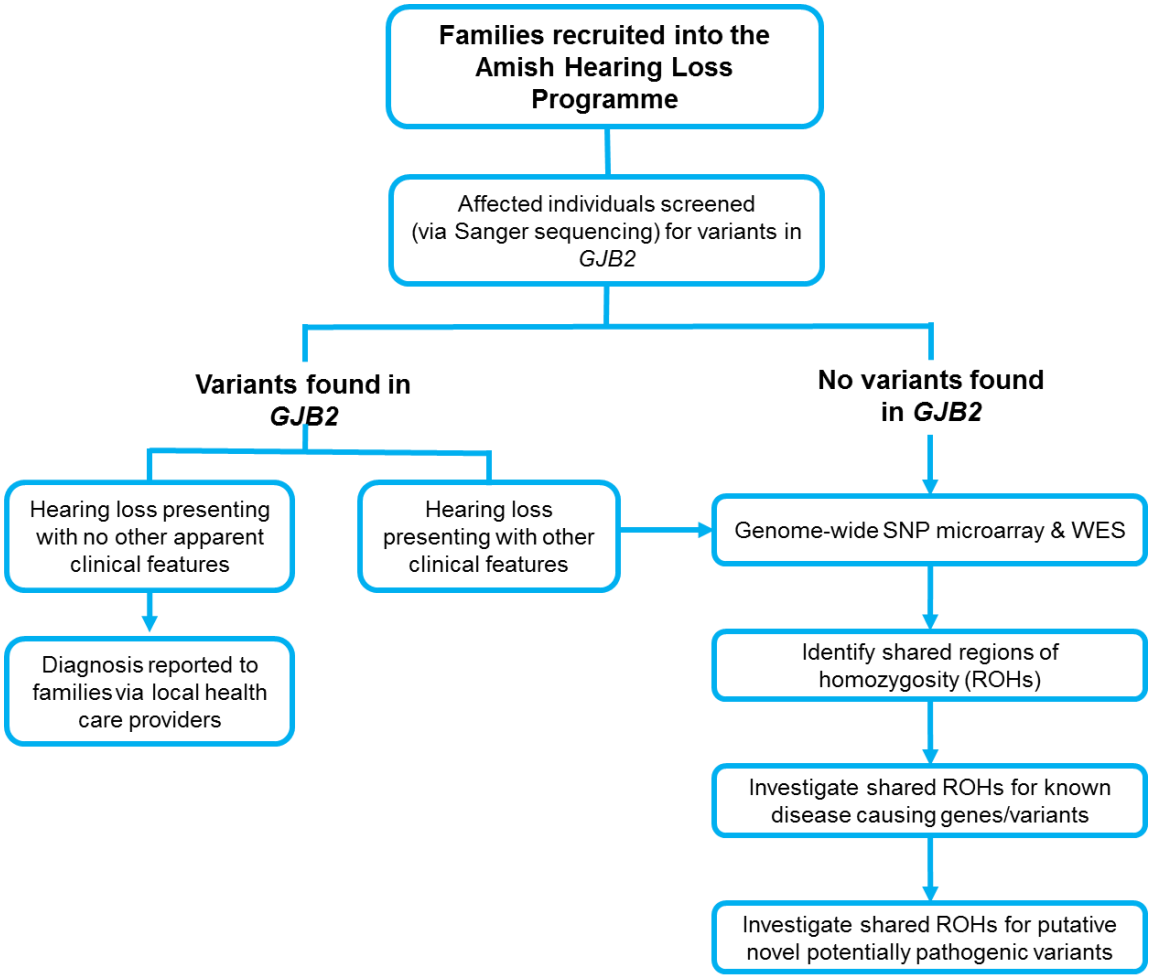


Figure 3.24: Updated screening strategy implemented for individuals recruited to the Amish Hearing Loss Programme.

3.6.2 GJB2 variants in the Amish occur on a distinct SNP genomic haplotype

In addition to these genes, other genetic causes of HL are also present in the Amish community. This work described in this thesis documents that *GJB2* variants are a common cause of NSHL, and within the HL cohort 8 of 19 families suffer hearing loss due to two previously reported *GJB2* variants (c.35delG and p.W77R). The c.35delG variant has also been reported in Hutterite [206] and Mennonite populations [207], which are distinct Anabaptist groups originating from Europe at the time of the radical reformation of the 16th century.

Mutations in *GJB2* are recognised as the most common cause of NSHL worldwide, and account for ~50% of NSHL cases in European populations [130]. More than 100 mutations located within *GJB2* have been identified [158], with the c.35delG mutation being the most common pathogenic variant in most Caucasian populations accounting for up to 70% of *GJB2* mutations [163]. Notably several studies have confirmed through haplotype analysis that the c.35delG variant represents a common ancient founder that arose in European and Middle Eastern populations, rather than a mutational hotspot [168, 169, 208]. The p.W77R *GJB2* variant was originally identified in a family of Israeli-Arab origin [189] and has subsequently been described in other populations in Europe, Australia, South America and the Middle East [209, 210]. The genetic data here confirms that both *GJB2* variants in the Amish occur on a distinct SNP genomic haplotype, indicating that each occurred, or more likely was introduced, via a single ancestor. Both gene mutations are also likely to represent the same ancestral mutations present in European and Middle Eastern populations [169, 208, 210].

3.6.3 Exclusion of the SLC15A5 variant as a cause of NSHL

Genetic studies in families with individuals with forms of hearing loss excluded from known causes of disease identified *SLC15A5* as a potential candidate gene for NSHL. Genome-wide SNP studies in two affected siblings identified only four notable (>1Mb) genomic regions of homozygosity common to both siblings. In parallel with this whole exome sequencing in a parental carrier identified only a single candidate variant, a nonsense variant in the *SLC15A5* gene, located in the largest homozygous region. The family identified originated from Ohio, the community from which most families in this study were recruited, in which the *SLC15A5* variant was detected at modest allele frequency. A notably higher allele frequency for the *SLC15A5* variant was identified in the Wisconsin community. Although these studies relate to a single family, our exome and whole genome SNP mapping studies excluded other known genetic causes of HL, and identified the *SLC15A5* as the sole candidate gene.

While it currently has no known function and is the first SLC15A family member to be linked to hearing loss another family member, SLC15A2, has been shown to be expressed in the otic vesicle of zebrafish. The otic vesicle is an embryonic structure that goes on to form the auditory and vestibular organ of the fish which is the homolog to the inner ear of mammals [191].

Work to characterise atypical solute carriers showed that SLC15A5 is only found in mammals which is suggestive of a function related to the specific features of vertebrates [194]. This study also revealed that SLC15A5 showed negligible expression in the pituitary, liver, spleen and thymus (Figure 3.25) and was one of only two genes to not be differentially expressed in the CNS. Additionally, a study carried out by Scheffer et al. in 2015 investigating gene expression in inner ear

mouse hair cells during development did report expression of SLC15A5 in utricular hair cells [211]. Furthermore, studies by our colleagues in Tübingen also detected SLC15A5 gene expression in mouse cochlea (section 3.5.3) supporting a possible role for SLC15A5 gene mutation as a cause of SNHL.

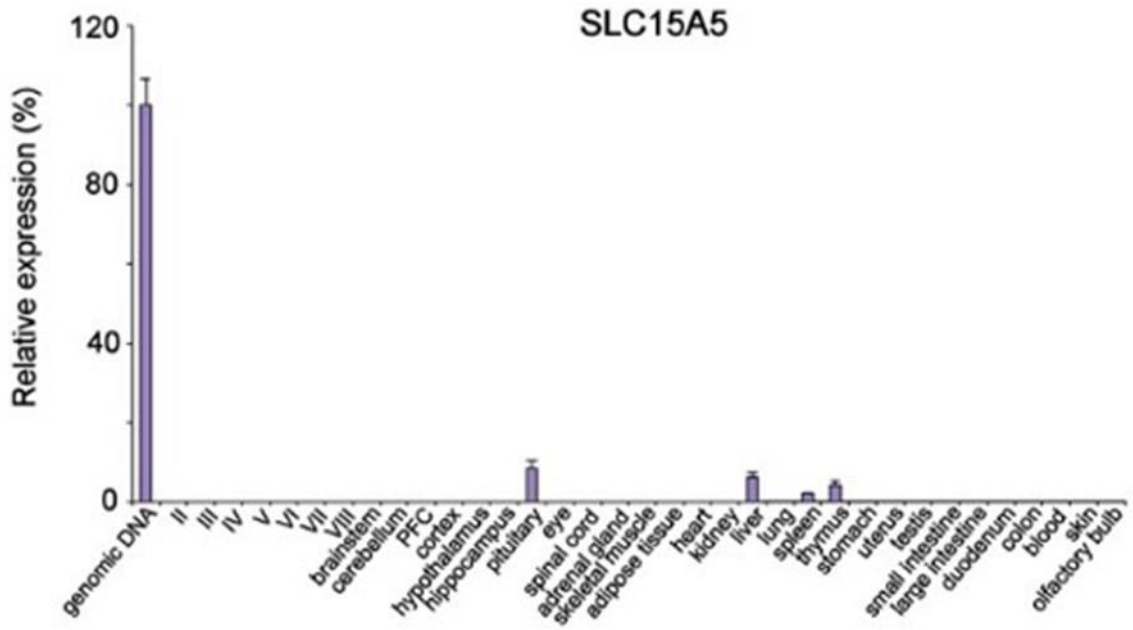


Figure 3.25: Results from quantitative real-time PCR data for SLC15A5. The data has been normalized against the detected expression levels for that particular gene in 25 ng of mouse genomic DNA (Taken from [194]).

Due to very little being known about SLC15A5 preliminary functional studies were undertaken to discover more about the expression and function of the molecule and its potential role in hearing loss. As expected for the nature of the variant *in silico* prediction software analysis predicted the candidate variant (NM_001170798.1:c.865G>T; p.Glu289Ter) to be pathogenic, due to the introduction of a premature termination codon (PTC), likely leading to protein truncation. While transcriptional outcomes were not assessed such aberrant mRNA transcripts may typically be subjected to the sophisticated quality control

monitoring mechanism, nonsense-mediated mRNA decay (NMD), that prevents the production of truncated polypeptides that could be toxic to normal cellular functions [212] resulting in the absence of any functioning protein product.

Initial immunocytochemistry experiments appeared to show a vesicular pattern inside the cell cytoplasm. This was not surprising due to the localisation of the SLC15A family member PEPT1 (SLC15A1) which was previously reported to localise primarily to the outer cell membrane and in adjacent vesicles [213]. Due to the lack of SLC15A5 in the cell membrane co-expression experiments were undertaken, alongside Rab7, a marker of recycling endosomes. The extensive, but not complete, overlap of SLC15A5 with Rab7 suggests it may be associated with lysosomes and/or endosomal trafficking and not the outer cell membrane. However, further investigations would need to be undertaken in order to reproduce and confirm these findings.

The lack of expression detected during immunohistochemical analysis of paraffin sections of inner ear mouse tissue at embryonic day 16.5 (E16.5) could have arisen due to a variety of reasons. Firstly the protein may not be expressed as this specific embryonic stage. Preliminary ISH experiments carried out by Dr John Chilton indicated that low levels of the protein were expressed in chick embryos at stage 20, corresponding approximately to E9.5 in a mouse, seven days prior to the sections used in these experiments. It would have been preferable to conduct these experiments at an early time point as it is possible, due to what is now know about *GJB2* [161], that a mechanism involved in the development hearing loss is likely to be disrupted during early development. Equally, the lack of staining could be a result of the SLC15A5 antibody not working on paraffin sections.

The inability to detect SLC15A5 protein via Western blotting could be due to a lack of SLC15A5 protein being expressed in the transfected HEK293 cells, which may have arisen due to the low transfection efficiency observed, or issues with antibody binding. However, this, coupled with a lack of detection in immunohistochemical experiments, indicates there may be an issue with the SLC15A5 antibody binding to the target sequence, epitope, on the protein. It was suggested that the target sequence used by the original (Origene) SLC15A5 antibody was quite long and, despite having good solubility in water, may condense into a folded, or aggregate structure, leading to variable success in different applications. To overcome this, a custom antibody targeting two regions of the protein could be used in future studies.

Another possible cause of the issues experienced with antibody binding may reflect a lack of conservation between human and murine protein sequences. Although it would be straightforward to identify regions that are similar between mouse and human, there would be a bias towards the more conserved membrane embedded portion of the protein, which means its use would be limited to experiments where the membranes are isolated/denatured. This may not help in determining the subcellular localisation of the protein, particularly given the initial results indicating the protein is not present on the outer cell membrane (Francesco Rao, Dundee Cell Products/Dr John Chilton, personal communication).

Whilst these studies were ongoing, exclusion of the *SLC15A5* variant as a cause of NSHL was achieved through a new collaborative arm providing additional genotyping data in other Amish communities as part of a high-volume sequencing program. This identified two individuals to be homozygous for the *SLC15A5*

variant, who apparently displayed no features of SNHL. Thus, while the specific cause of SNHL in this family remains unknown, this finding likely excludes the SLC15A nonsense variant as the likely cause.

This study provides has important lessons to consider when investigating potential, novel candidate pathogenic variants in endogamous communities. This study provides important lessons to consider when investigating potential, novel candidate pathogenic variants in endogamous communities, in which it is important to remain cautious when defining candidate variants, including those likely to be deleterious (nonsense), identified via small family studies. It is imperative that allele frequency datasets, both from within and outside the community, are thoroughly and regularly investigated, as conducted in this study. This permits the identification of additional variants in other families to aid confirmation of a gene as causative.

As demonstrated with this family, and other families within the Amish community [177], the interpretation and clinical significance of novel or rare variants may be challenging. This may be aided through the curation and dissemination of knowledge regarding rare gene variants.

Initiatives such as the Anabaptist specific variation database and MatchMaker Exchange (MME), launched in 2013, has significantly expedited the matching of unrelated cases with variants in the same gene and overlapping phenotypes. Whilst the genomic data collected by the Windows of Hope programme over the last 19 years is an invaluable source of information to the Amish community itself, its importance extends far beyond this into the general population. This data can be utilised by the “matchmaking community” in both a clinical and research setting to facilitate human disease gene identification, to aid genomic variant

classification and provide a valuable source of phenotypic and prognostic data on otherwise rare inherited disorders.

However, this study also demonstrates the utility of a community-specific dataset which shares, otherwise unpublished information on nonsense and frameshift variants that may have become enriched in the community but, like the *SLC15A5* variant, are known to be non-pathogenic. These concepts are explored and highlighted further in chapter five of this thesis.

3.6.4 Allele frequencies of SHL in the Amish population

Several syndromic presentations of SNHL have also been identified in the Amish population, including the previously well characterised conditions Usher syndrome [214] infantile Refsum disease [215] Jervell-Lange-Nielson (JLN) syndrome[181], *COL1A2*-related osteogenesis imperfecta [186], metabolic and mitochondrial disorders. JLN syndrome was described in the Amish population by [181] who reported a family with two affected siblings both with SNHL and long QT syndrome (LQTS), with the parents displaying borderline LQTS and normal hearing. DNA sequencing identified a homozygous 2bp deletion in *KCNQ1* predicted to result in a frameshift and premature termination as the likely cause in the affected siblings (NM_000218.2:c.451_452del; p.Leu151fs). While at low frequency, our studies confirmed the presence of this gene variant in both the Holmes and Geagua county Ohio Amish communities. Given the serious implications of sudden cardiac death associated with this disorder, it is crucially important to consider this genetic diagnosis in individuals with congenital SNHL in the Amish to ensure appropriate counselling, follow-up and management.

Several genetic causes of congenital SHL have also been identified in the Amish community in association with specific syndromic conditions. These include *PCNA, SLITRK6, YARS, GM3* synthase deficiency and CODAS (Cerebral, Ocular, Dental, Auricular and Skeletal anomalies) syndrome (see supplementary paper), each of which should be considered depending on the accompanying clinical features. Mitochondrial disorders associated with SNHL have also been described within the Amish population [216] and this group of disorders along with infantile Refsum disease [215], should be considered particularly in the context of SNHL associated with neurological and ophthalmological features. In Mennonite and Hutterite groups other conditions associated with SHL have been described, including mutation of *PCDH15* and *MYO7A* with Usher syndrome [214, 217], *ALMS1* in Alstrom syndrome [218, 219] and *EDNRB* in Waardenburg syndrome [220]. It is likely that other well characterised causes of both SHL and NSHL have been identified within the Amish population by local clinicians and researchers which remain unpublished.

3.6.5 Future work and considerations

As the data reviewing the causes and frequencies of hearing loss in the Amish community was being compiled for publication, ongoing studies identified a candidate new cause of SNHL and NDD in a large extended Amish family (section 3.5.5). This study investigated the molecular cause of hearing loss in two interlinking families in which known genetic causes of NSHL had been excluded, comprising five affected individuals, due to a nonsense mutation in a microtubule-associated molecule. The genetic power of this study was notably greater than the *SLC15A5* family study, due to the greater size of this family. Additionally, the

same p.Arg1197Ter nonsense variant was identified as a candidate cause of hearing loss and NDD by a collaborator of the Exeter team investigating families from the Middle East. Together, these studies provide strong evidence that the microtubule-associated molecule is responsible for this syndromic form of hearing loss. While beyond the remit of this thesis, future studies should be tailored to corroborate these findings and further explore the cellular and developmental role of this molecule about which little remains currently understood.

Such studies would entail undertaking molecular studies to more precisely define binding partners and subcellular localisation of the gene product to define its molecular role in hearing loss. In addition to this immunohistochemical and RT-PCR studies in mouse brain and inner ear at different developmental time points would be carried out to describe the pattern of gene expression in different tissues known to be linked to the development of hearing loss. Finally, mouse knockout studies could be conducted to investigate the phenotypical outcome in comparison with humans.

There are numerous significant benefits of elucidating the genetic basis of a condition such as HL in any community setting. These include providing a specific diagnosis for affected individuals and their families, identifying relevant therapies and treatments (for example consideration for CI), screening for the presence of other features that may develop in the condition, and to inform parents for reproductive counselling. As with other inherited conditions originally identified in the Amish, the identification and clinical and molecular definition of HL disorders in the community has been of notable benefit globally, enabling informed disease diagnosis, counselling and management for patients and families worldwide subsequently diagnosed with these conditions. With the exception of mutation of

PCNA, all conditions defined in the Amish have ultimately been described in other populations highlighting the global relevance and importance of studies of inherited conditions in such genetic isolates. Further studies to elucidate the molecular mechanisms of HL-related disorders is crucial, and will ultimately facilitate the development of new treatments to alleviate the burden associated with these disorders.

CHAPTER 4

DEFINING THE PHENOTYPE AND PATHOMOLECULAR BASIS OF A NOVEL FORM OF NEURODEVELOPMENTAL DISORDER ASSOCIATED WITH THE MISSENSE MUTATION OF SMAD NUCLEAR INTERACTING PROTEIN 1 (SNIP1)

4 Defining the phenotype and pathomolecular basis of a novel form of neurodevelopmental disorder associated with the missense mutation of Smad nuclear interacting protein 1 (*SNIP1)*

4.1 Introduction

As previously described, neurodevelopmental disorders (NDD) are a group of disorders involving the abnormal development of the central nervous system which often display a spectrum clinical features. Global developmental delay (GDD) is commonly observed in NDD and is defined as the failure of an individual to achieve developmental milestones within the expected age range. Whilst GDD can occur in isolation it is also often observed alongside other neurological (and potentially non-neurological) features including epilepsy and behavioural problems (autism spectrum disorder and attention-deficit hyperactivity disorder) [221].

A collaboration between five Amish-based clinics (The Windows of Hope, the Community Health Clinic, New Leaf Center, LaFarge Medical Clinic and The Clinic for Special Children) identified 35 individuals (20 males; 15 females) which, whilst displaying a broad spectrum of neurological phenotypes, also showed a striking number of overlapping clinical features. A review of the phenotypes displayed by these individuals identified global developmental delay, hypotonia, abnormal skull shape and the presence of seizures as the most consistent features. To investigate if a common genetic cause was responsible for the

difficulties experienced by these individuals a combination of autozygosity mapping and exome sequencing was undertaken on a small subset of families. This revealed a p.Glu366Gly missense mutation of *SNIP1* (Smad nuclear interacting protein 1), a widely expressed transcriptional suppressor of the TGF-β signal-transduction pathway, common to all affected individuals. This variant had previously been reported as a puatative novel candidate in three individuals, from a different Amish community, displaying phenotypes overlapping with individuals from our cohort including epilepsy and skull dysplasia [179]. This finding further corroborates the likely pathogenicity of this variant and its role in the observed disorder.

Due to its role as a transcription regulator, gene transcript studies where undertaken to elicit the effect of this variant on gene expression. These studies showed altered gene expression profiles for a number of molecules with well characterised roles in neurodevelopment, providing potential explanations for the source of the clinical phenotypes observed. This study consolidates phenotypic, genetic and gene expression data in support of the mutation of *SNIP1* as the cause of a novel autosomal recessive neurodevelopmental disorder in addition to providing insight into the molecular basis of the disorder and the role of SNIP1 as a transcription factor.

4.1.1 Cell signalling

Cell signalling, or signal transduction, is the fundamental process of cells communicating and responding to external cues in their environment [222]. Cells communicate via membrane-associated proteins or through the secretion and detection of molecules, such as hormones, cytokines, chemokines and growth

factors that influence cell behaviour and responses including proliferation, differentiation and metabolism [223].

The transforming growth factor–β (TGF-β) family is a well-studied example of a secreted growth factors that govern many developmental processes. Aberrant signalling within this pathway is well characterised, and shown to be associated with a number of human disorders including cancer, cardiovascular and musculoskeletal disease [223, 224].

4.1.2 TGF-β signalling pathway

In mammals, the TGF-β superfamily of growth factors are encoded by 33 genes that are widely expressed in a variety of tissues [223, 225]. They are essential throughout the life of an organism orchestrating critical processes ranging from gastrulation and the onset of body axis asymmetry in early embryonic development, to adult tissue homeostasis [226].

The TGF-β family contains a large group of cytokines , ~60 members [227], which can broadly be divided into four subgroups; TGF-βs that comprises of the three mammalian TGF-β isoforms, the Müllerian inhibitory factors (MIF), the activins/inhibins, and the bone morphogenetic proteins (BMP) and growth and differentiation factors (GDFs) which contains all 20 BMPs and the majority of GDFs, [226, 228] (Figure 4.1).

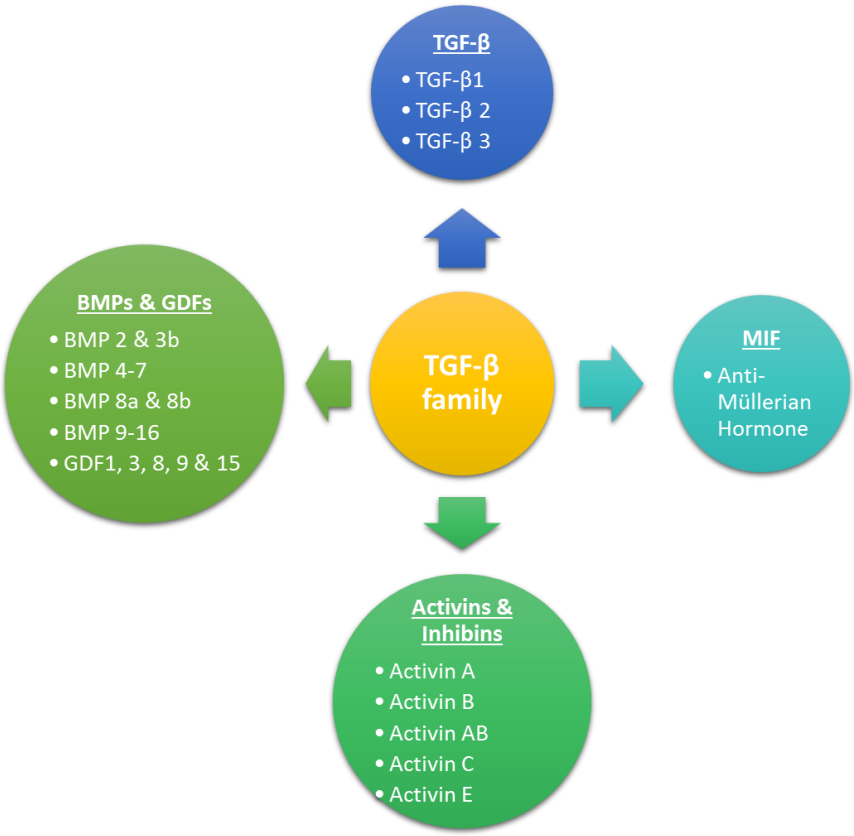


Figure 4.1: Members of the TGF-β family. Major families of the TGF-β superfamily include TGF-β, BMP, MIF and activin–inhibin. Image taken and modified from Drabsch & Dijke, 2012 [229].

Control of over the broad range of functions associated with members of the TGF-β family is achieved through signalling specificity and differential ligand affinities to different cell surface TGF-β receptors (TGFβR) as well as the co-regulation of the various pathways by a number of transcriptional co-activators and co-repressors [224]. TGF-β signal transduction includes Smad transcription factor mediated pathways (section 4.1.3: Smad proteins) which induce a transcriptional response, and non-Smad mediated pathways which evoke transcriptional responses as well as other direct cellular responses, not involving the regulation of transcription [224].

There are three main types of TGFβR present in all cell types, TGFβRI, TGFβRII and TGFβR-III. TGFβRI and TGFβRII mediate Smad dependent signal

transduction and contain a serine/threonine kinase domain, a transmembrane domain and a ligand-binding domain which associate in a homo- or heteromeric complexes and act as tetramers. [228, 230]. Binding of a homodimer TGF-β ligand to TGFβRII initiates the formation of a stable hetero-tetrameric complex with TGFβRI, where TGFβRI receptors are activated through phosphorylation by TGFβRII. Once activated TGFβRI receptors phosphorylate cytosolic Smad2 and Smad3 (receptor-regulated Smads, R-Smad) proteins at C-terminal serines. Activated R-Smads then form a trimer with Smad4, a common-partner Smad (co-Smad) enabling the complex to be imported into the nucleus. Inside the nucleus the R-Smad-co-Smad complex associates with sequence specific transcription factors at regulatory sequences in target genes regulating gene expression [231] (Figure 4.2B).

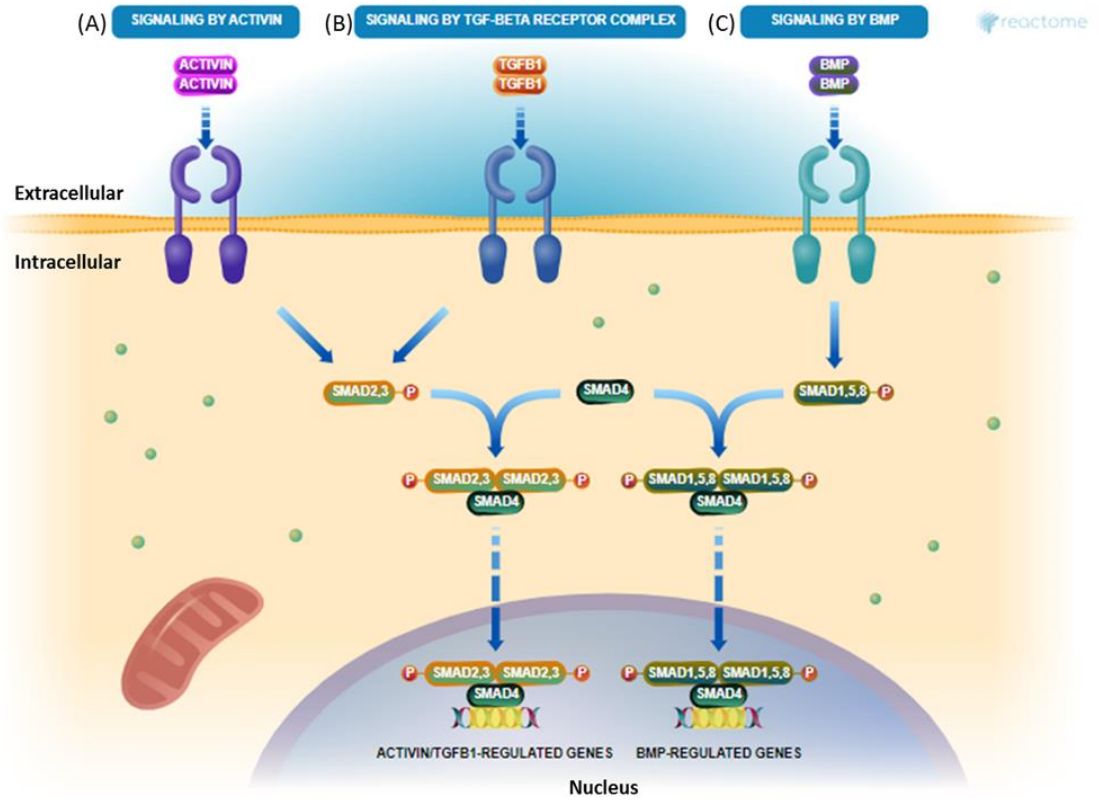


Figure 4.2: TGF-β signalling pathway. Summarising the different signal transduction pathways initiated by the different TGF-β ligands. A) Signalling by activin ligands. B) Signalling by TGF-β ligands and C) signalling by BMP ligands. Image taken and modified from (https://reactome.org/PathwayBrowser/#/R-HSA-9006936&DTAB=DT).

4.1.3 Smad proteins

Smad proteins are a small family of structurally similar intracellular proteins that act as transcription factors in the transforming growth factor-beta (TGF-β) pathway (**Error! Reference source not found.**B) [232]. The name originates from the contraction of the names of homologous genes, Sma and Mad, originally discovered in *Caenorhabditis elegans (C.elegans)* and *Drosophila melanogaster (Drosophila)* respectively.

The Smads are a well characterised signalling pathway initiated by activated TGF-β receptors. There are eight Smad family members in mammals [233] which can be grouped into three classes; receptor-regulated Smads (RSmads), common-partner Smads (Co-Smads), and inhibitory Smads (I-Smads). All Smad proteins are 500 amino acids in length and contain two globular domains; Mad-homology 1 (MH1) and Mad-homology 2 (MH2), joined by an unstructured linker (Figure 4.3). The MH1 domain, involved in nuclear import and cytoplasmic anchoring, DNA binding, and regulation of transcription, is highly conserved in all R-Smads and Smad4 but is not present in I-Smads. Due to the addition of 30 amino acids in the MH1 domain, encoded by an extra exon, Smad2 is unable to directly bind to DNA. The linker regions, whilst all containing important phosphorylation and recognition sites, are diverse in the different Smad proteins. For example the linker region in Smad3 contains a trans-activation (TA) domain which permits the binding of transcription co-regulators with the Smad4 linker region containing a nucleus export signal (NES). The MH2 domain, responsible for the regulation of Smad oligomerization, cytoplasmic anchoring and transcription of target genes, is highly conserved and considered to be one of the most versatile protein-interacting domains in signal transduction [232-234].

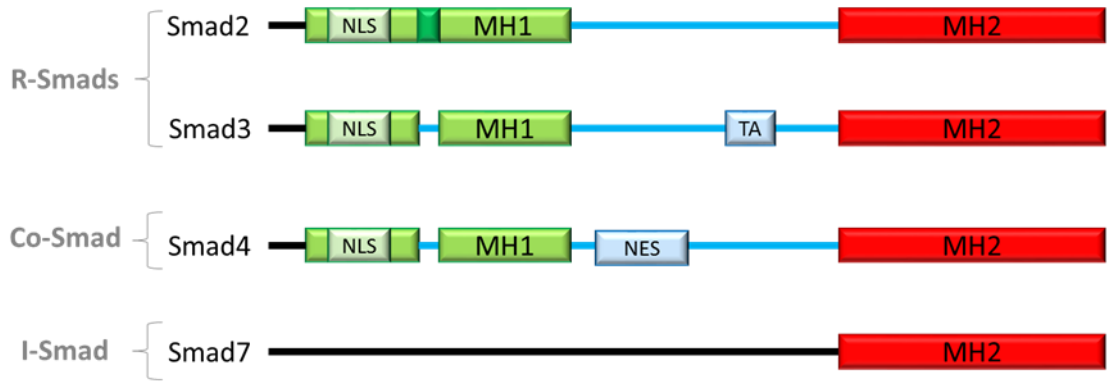


Figure 4.3: Domain structure of Smads. MH1 domain of Smad2 contains an additional 30 amino acids (dark green box). Smad3 contains a trans-activation (TA) in its linker region. Smad2, 3 and Smad4 contains a Nucleus Localization Signal (NLS) in their MH1 domain. Smad 7 lacks MH1 domain. Figure taken and modified from Samanta and Datta, 2012 [234].

As describe above, R-Smads are activated through phosphorylation by a TGFβRI receptor. Once activated, and in a complex with the co-Smad, Smad4, the complex is translocated into the nucleus where Smad4 associates with interacting transcription factors and p300/CBP enabling it to bind to the Smad binding element (SBE) in a target gene promoter driving transcription of these genes [235] (Figure 4.4).

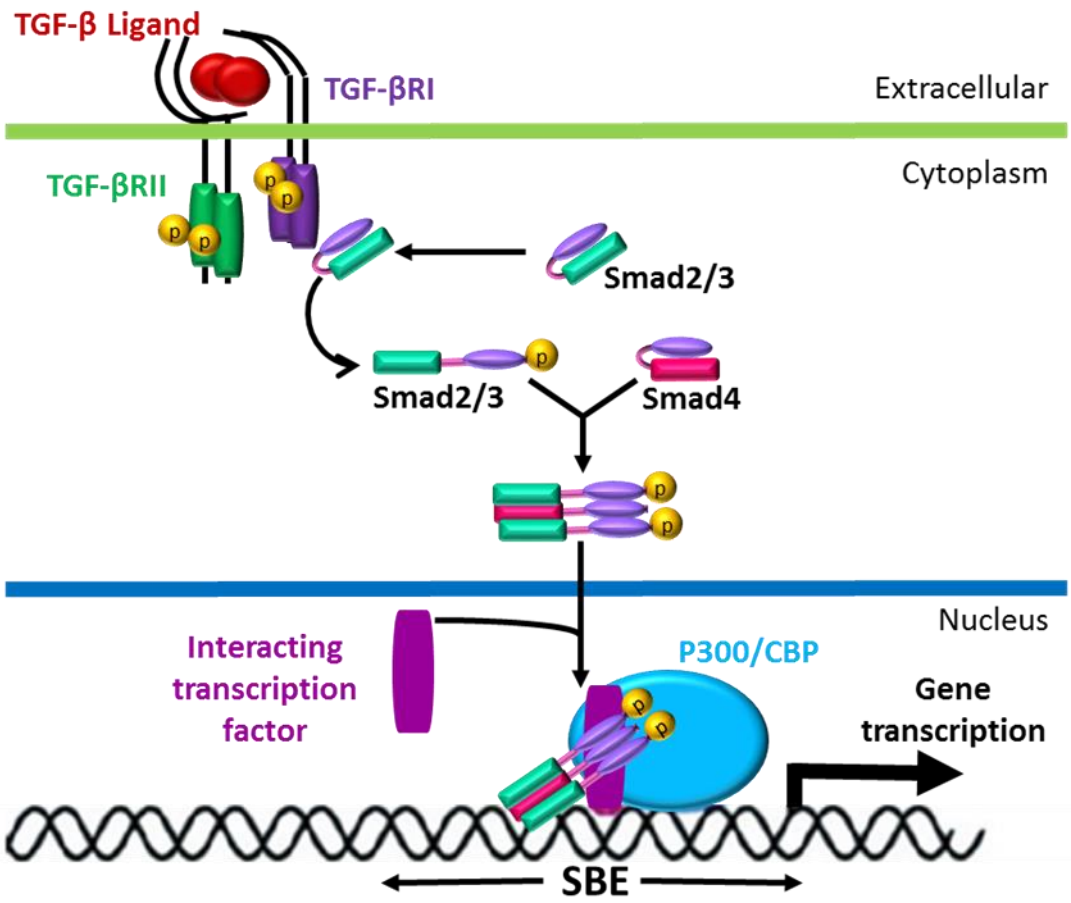


Figure 4.4: TGF-β/Smad signalling pathway from activation upon TGF-β ligand binding to translocation to the nucleus and effect on gene expression. Image taken and modified from Jiang et al. 2015 [235].

Defects in TGF-β proteins, or disruption of the Smad signalling pathway by other molecular means, has been linked to a number of human diseases including cancer [235], chondrodysplasias [236], chronic kidney disease (CKD) [237] and pulmonary hypertension [228, 236].

4.1.4 SNIP1

The *SNIP1* (Smad nuclear interacting protein 1) gene is located at 1p34.3 and encodes a widely expressed nuclear protein, SNIP1, which acts as a transcriptional suppressor of the TGF-β signal-transduction pathway. This evolutionary conserved protein, consisting of 396 amino acids, contains a two-part nuclear localisation signal (NLS) and a forkhead-associated (FHA) domain [238, 239] (Figure 4.5).

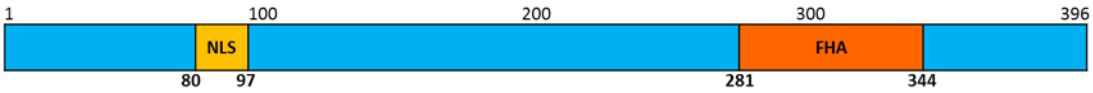


Figure 4.5: Schematic diagram of SNIP protein depicting the location of the two functional domains

The C-terminal FHA domain has been shown to bind to the N-terminus of c-Myc, a key regulator of cell proliferation and transformation, enhancing its transcriptional activity by stabilizing it against proteosomal degradation and by bridging the c-Myc/p300 complex [240]. Functional studies have demonstrated the N-terminal NLS inhibits Smad-dependent transcription by binding to p300/CBP co-activators [238, 241]. This interaction prevents Smad4 in the R-Smad-Co-Smad complex binding to the co-activators rendering it unable to bind to the Smad binding element (SBE) in target gene promoters [235], thus supressing transcription (Figure 4.6).

Orthologues of SNIP1 have been identified in organisms ranging from Homo sapiens to Caenorhabditis elegans (*C. elegans*), displaying varying degrees of conservation with 86% homology observed in murine SNIP1 compared to only

47% homology with the *C. elegans* homologue (C32E8.5) [242]. The most highly conserved domain within the SNIP1 protein is the FHA domain which is consistent with a role in growth regulation. Knock-down experiments in *C. elegans* showed that loss of SNIP1 resulted in embryonic lethality with growth defects and sterility observed in knockdown experiments involving adult worms [242].

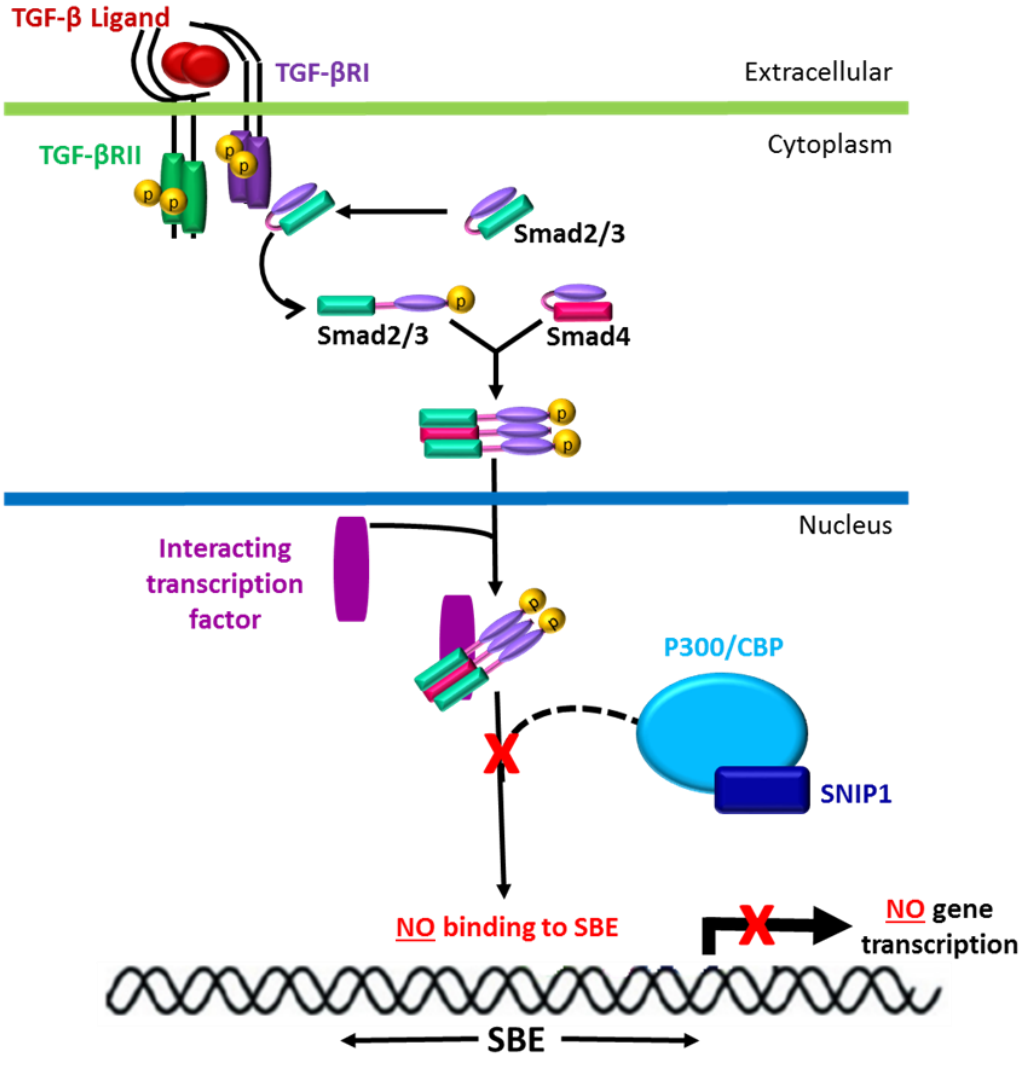


Figure 4.6: TGF-β/Smad signalling pathway from activation upon TGF-β ligand binding to translocation to the nucleus and repression of gene expression on SNIP1 binding to the p300/CBP transcription co-activator. Image taken and modified from Jiang et al. 2015 [235].

4.2 RESULTS

4.2.1 Identification of a pathogenic variant in SNIP1

Assuming that a founder mutation was responsible a genome-wide SNP microarray, (Illumina Human CytoSNP-12v2.1 330k array) performed using DNA from six affected individuals (Figure 4.7), was undertaken in conjunction with WES, performed using the DNA from one affected individual.

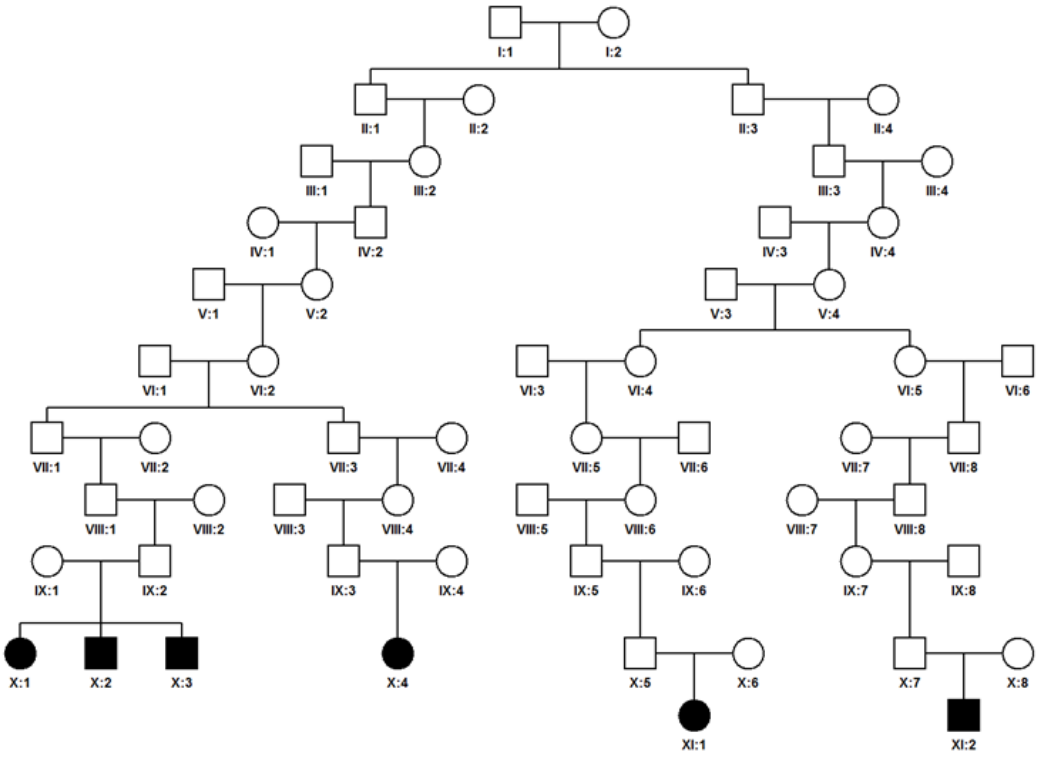


Figure 4.7: Simplified pedigree of the four Amish families initially investigated showing the six affected individuals initially genotyped using a genome-wide SNP microarray.

Inspection of resultant genotypes identified a single notable region of homozygosity of 1.65Mb on chromosome 1p34.3 shared by the affected individuals, delimited by SNP markers rs6667450 and rs10889902 (NC_000001.11: g.36, 492,230-38,143,653; (Figure 4.8) containing 17 genes and

likely to corresponding to the disease locus. No other notable regions (of >0.5Mb) of autozygosity were observed.

In parallel with the genome-wide SNP mapping, DNA from a single affected individual (Figure 4.7) underwent WES, to identify variants within this region as well as other candidate variants located genome-wide. After filtering the identified variants for call quality, potential pathogenicity, population frequency, and prioritisation for localisation within the candidate interval, a single potentially pathogenic sequencing variant (NM_024700.3: c.1097A>G; p.Glu366Gly) in exon four of *SNIP1* remained as the only candidate variant. Dideoxy sequence analysis of all family members confirmed that, as expected, all six affected individuals were homozygous for this variant, while the parents and unaffected siblings were all heterozygous. The variant, predicted to be disease causing in MutationTaster2 [94] and probably damaging in PolyPhen-2 [93], results in the substitution of a highly conserved (Figure 4.8) glutamic acid residue at position 366 for glycine (p.Glu366Gly).

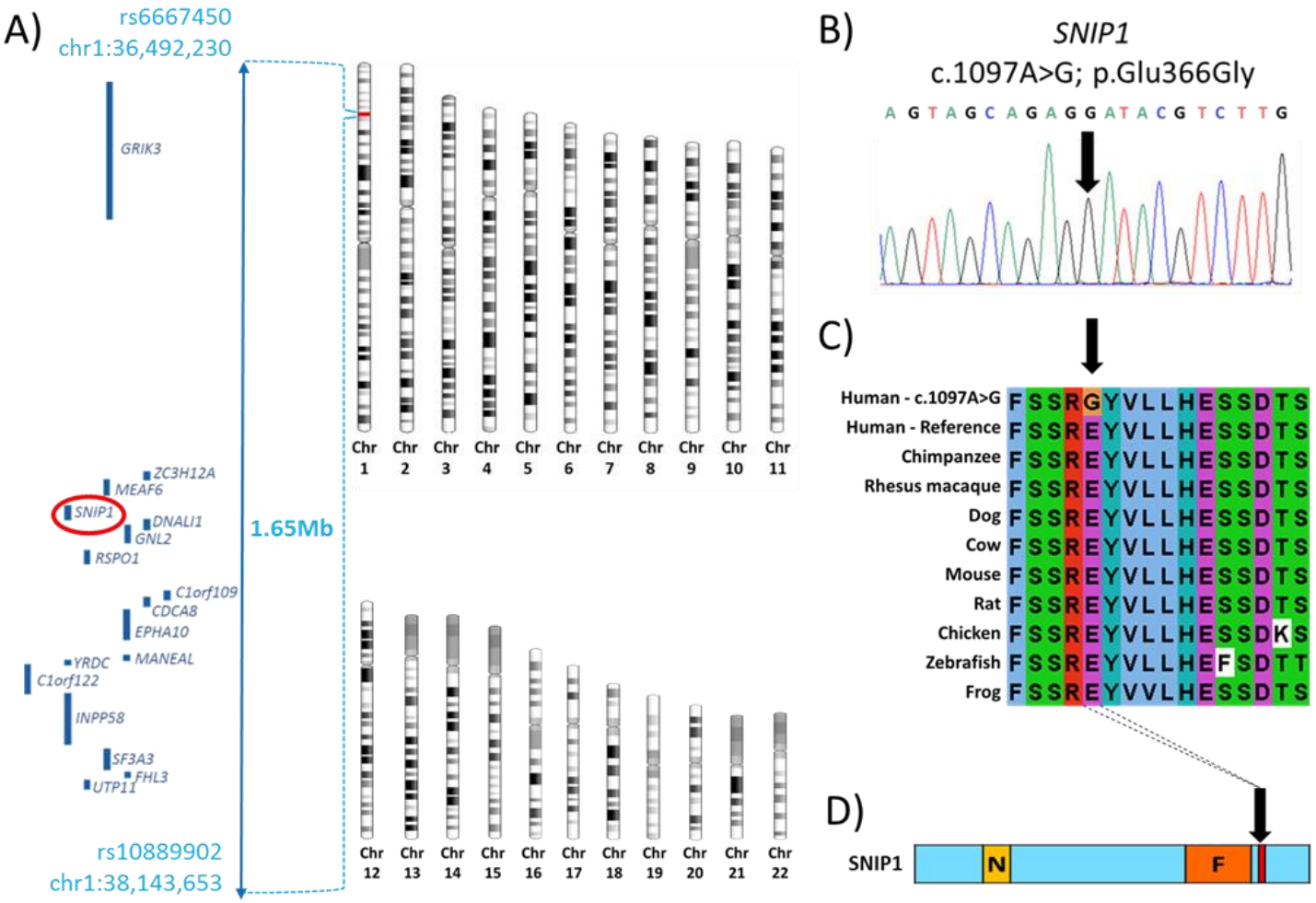


Figure 4.8: Pictorial representation of the (A) identified 1.65Mb disease locus on chromosome 1 containing 17 genes, with SNIP1 indicated (red circle). (B) sequence chromatogram corresponding to the SNIP1 c.1097A<G variant (C) multi-species amino acid alignment around the Glu366 region, showing stringent conservation of this region (D) schematic showing domain architecture of SNIP1 polypeptide sequence with regard to Glu366 , located alongside the forkhead-associated domain (F).

4.2.2 Defining the clinical phenotype of a neurological disorder displaying severe psychomotor delay with seizures, epilepsy and dysmorphic features

A collaboration between Windows of Hope, the Community Health Clinic, New Leaf Center, LaFarge Medical Clinic, and Clinic for Special Children identified 35 individuals originally diagnosed with a broad range of genetic disorders including glycogen storage disorders, sialidosis, Crouzon syndrome, mitochondrial disorder/myopathy, and GM3 synthase deficiency.

The spectrum of clinical features of this condition includes severe psychomotor delay with seizures, (decreased muscle tone), absent reflexes, and strabismus (improper alignment of the eyes) with horizontal nystagmus (involuntary eye movement), as well as neonatal hypotonia with poor feeding and characteristic dysmorphic features (Figure 4.9).

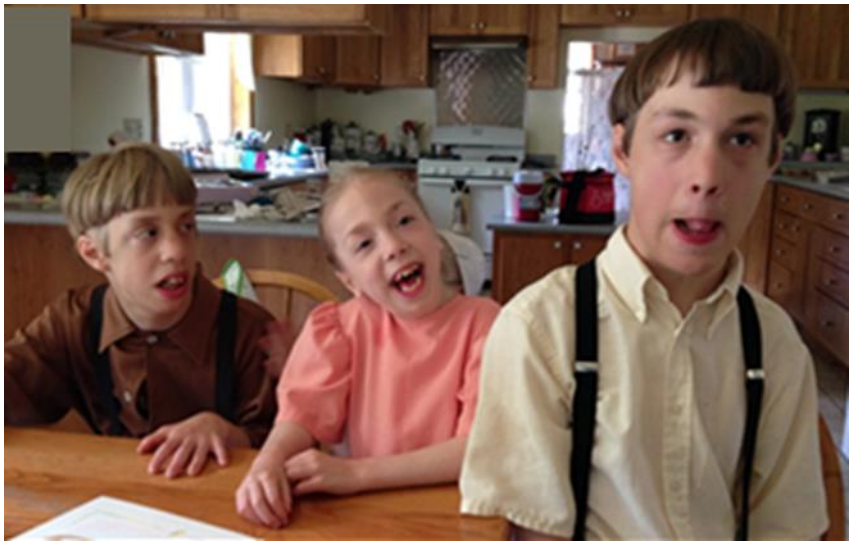


Figure 4.9: Three siblings displaying characteristic craniofacial features. Photograph provided by Dr Zineb Ammous from the Community Health Clinic, Topeka, US. Written consent granting permission for publishing was obtained locally.

Dr Zineb Ammous, a Clinical Geneticist from the Community Health Clinic, Topeka, US, conducted a survey of all 35 individuals and through historical and physical examination was able to precisely define the clinical phenotype (Table 4.1).

Table 4.1: The clinical presentation in the 33 affected individuals; ADHD, attention deficit hyperactivity disorder

	Percent Affected (n=35)
Central Nervous System	
Global developmental delays/ Non-verbal	100%
Hypotonia, hyporeflexia	100%
Seizures and/or epilepsy	100%
Abnormal brain MRI (ventriculomegaly, hypomyelination)	50%
Behavior problems (irritability, autistic spectrum, ADHD)	75%
Friendly affectionate personality	45%
Cardiopulmonary	
Upper respiratory disease (laryngomalacia, apnea, stridor)	75%
Congenital heart defects (e.g. VSD, ASD, CoA, Ao valve)	60%
Cardiomyopathy	12%
Vision and Hearing	
Horizontal nystagmus and/or Strabismus	45%
Failed newborn hearing, conductive loss	21%
Gastroenterology and Nutrition	
Feeding difficulties	100%
Small for gestation/Failure to thrive	54%
Aspiration with and without Gastrostomy tube	46%
Endocrinology	
Hypothyroidism	25%
Hypoglycemia	21%
Morphological Features	
Abnormal skull shape ("lumpy", craniosynostosis)	100%
Wide mouth, cupid bow upper lip	100%
High arched palate	100%
Micrognathia (Pierre Robin Sequence in some)	30%
Short hands, tapered fingers	54%
Spine abnormality (scoliosis, sacral dimple, tethered cord)	21%
Hernia (umbilical, inguinal)	21%
Congenital talipes equinovarus	13%

ADHD, attention deficit hyperactivity disorder. VSD, ventricular septal defect. ASD, atrial septal defect. CoA, coarctation of the aorta. Ao valve

All affected individuals investigated displayed an abnormal MRI showing an abnormal skull shape (skull dysplasia) and hypomyelination (Figure 4.10).

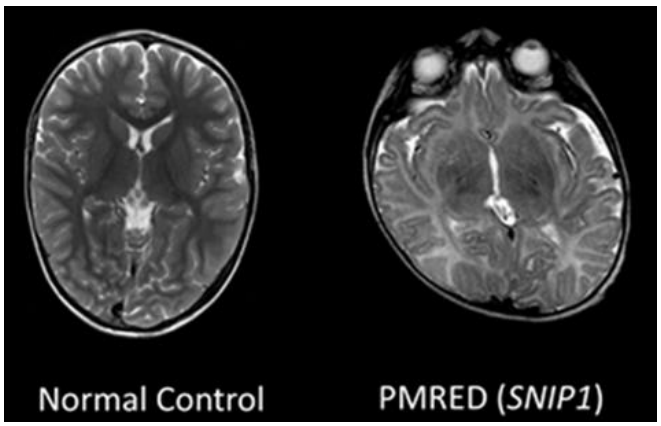


Figure 4.10: Axial brain MRI of a normal control (left) compared to an affected patient (right) showing skull dysplasia and hypomyelination.

Approximately 70% of patients suffered from respiratory complications such as laryngomalacia, stridor, and/or apnea which, in some cases, required a tracheostomy (Figure 4.11A). The condition also includes a number of additional features including talipes equinovarus (Figure 4.11B) and craniosynostosis (Figure 4.11C&D).

Figure 4.11: A) Infant patient with tracheostomy tube fitted as a result of respiratory complications. Atypical features talipes equinovarus (B) and craniosynostosis (C, D). Photograph provided by Dr Zineb Ammous from the Community Health Clinic, Topeka, US. Written consent granting permission for publishing was obtained locally.

4.2.3 Effect of SNIP1 mutation on gene expression and cellular pathways

The SNIP1 protein is an evolutionary conserved inhibitor of the TGF-β signal transduction pathway [238, 242]. However, it is not yet fully understood which genes it regulates specifically, nor which regulatory mechanisms it may be involved. To discern the impact of the SNIP1 p.Glu366Gly sequence variant on gene expression, and through which pathways it operates, whole transcriptome sequencing of RNA extracted from whole blood was undertaken on samples from six affected and six unaffected, age and sex matched control individuals.

Whole transcriptome sequencing enabled the characterisation of all RNA transcripts present within the samples by identifying differentially expressed genes and produced data on genes where expression was either increased (upregulated) or decreased (downregulated) in affected individuals (disease group) compared to the control group. This RNA-seq data underwent gene set enrichment and pathway analysis using the integrative web-based software application Enrichr [243] which collates a summary of the most biologically relevant enriched genes within a dataset [244]. This is done by comparing the frequency of individual annotations in the each gene list, from 102 available gene set libraries, with a reference list of genes. As this study utilised genome-wide, whole transcriptome sequencing the reference list contained all genes in the genome. For this analysis the “Reactome 2016” gene set library, the largest, most comprehensive and best characterised gene set list, was used to functionally annotate the differentially expressed genes within >2250 pathways. Table 4.2 and Table 4.3 summarise the top ten pathways, and associated genes, that were up- or down- regulated in the disease group compared to the control.

Table 4.2: Top ten pathways and associated genes identified by Enrichr as being upregulated in the disease group compared to controls

Enrichr code	Enrichr Pathway	Genes		P-value	Z-score	Combined score
R-HSA-181429	Serotonin neurotransmitter release cycle	*SYT1*	*SYN3*	0.002096	-2.01	12.41
R-HAS-212676	Dopamine neurotransmitter release cycle	*SYT1*	*SYN3*	0.003424	-2.03	11.56
R-HAS-266738	Developmental biology	*DUSP2* *CDH4* *EPAS1*	*CHL1* *LAMC1* *ROBO1*	0.009831	-2.37	10.94
R-HAS-500931	Cell-cell communication	*CDH4* *PARD3*	*PTK2*	0.01364	-1.99	8.55
R-HAS-373752	Netrin-1 signalling	*PTK2*	*ROBO1*	0.01112	-1.88	8.46
R-HAS-983695	Antigen activates B Cell Receptor (BCR) leading to generation of second messengers	*BLK*	*BLNK*	0.01379	-1.87	8
R-HAS-112310	Neurotransmitter release cycle	*SYT1*	*SYN3*	0.01611	-1.92	7.94
R-HSA-421270	Cell-cell junction organisation	*CDH4*	*PARD3*	0.02257	-1.94	7.35
R-HAS-000178	ECM proteoglycans	*PTPRS*	*LAMC1*	0.01858	-1.81	7.21
R-HAS-422475	Axon guidance	*DUSP2* *CHL1* *LAMC1*	*PTK2* *ROBO1*	0.04623	-2.17	6.68

Table 4.3: Top ten pathways and associated genes identified by Enrichr as being downregulated in the disease group compared to controls

Enrichr code	Enrichr Pathway	Genes		P-value	Z-score	Combined score
R-HSA-2672351	Stimuli-sensing channels	*TRPC3* *TRPC1* *ANO9*	*SGK1* *RPS27A*	0.0002381	-1.92	16
R-HSA-418890	Role of second messengers in netrin-1 signalling	*TRPC3*	*TRPC1*	0.001334	-2.16	14.29
R-HAS-69298	Association of licensing factors with the pre-replicative complex	*E2F3*	*RPS27A*	0.003057	-1.91	11.03
R-HAS-2173791	TGF-beta signalling EMT (epithelial to mesenchymal transition)	*F11R*	*RPS27A*	0.003482	-1.89	10.71
R-HAS-983712	Ion channel transport	*TRPC3* *TRPC1* *ANO9*	*SGK1* *RPS27A*	0.005615	-2	10.42
R-HAS-3295583	TRP channels	*TRPC3*	*TRPC1*	0.08426	-1.95	9.5
R-HAS-2559585	Oncogene induced senescence	*SGK1*	*RPS27A*	0.01358	-1.96	8.42
R-HAS-6804757	Regulation of TP53 degradation	*SGK1*	*RPS27A*	0.01612	-2.09	8.66
R-HAS-6806003	Regulation of TP53 expression and degradation	*E2F5*	*RPS27A*	0.01701	-2.11	8.58
R-HAS-382551	Transmembrane transport of small molecules	*RNASEL* *TRPC3* *TRPC1* *ABCC6*	*CYBRD1* *ANO9* *RPS27A* *SGK1*	0.01767	-2.066	8.31

The pathways described in these tables were selected by obtaining the highest combined score ranking (>5) of all the pathways analysed by this gene set library. The combined score multiples the log of the p-value, a statistical method using the Fisher's exact test to assess the probability of any gene belonging to any set (Fisher exact test based ranking), by the z-score, which computes the deviation from an expected rank (rank based ranking). This combined score ranking allows for the slight bias of the Fisher exact test which was reported to affect the ranking of terms based exclusively on the length of the gene sets in each of the gene-set libraries [243].

A number of the pathways identified through the Enrichr analysis are evidently, highly relevant to the observed phenotype. The two most impacted biological pathways, involving serotonin and dopamine release, showed the highest level of enrichment (highest combined score) within the upregulated gene list. Both serotonin and dopamine are monoaminergic neurotransmitters that have been associated in the regulation of seizures [245, 246], a cardinal characteristic feature of the condition observed in our cohort (**Error! Reference source not found.**). This makes two genes, *SYT1* (synaptotagmin 1) and *SYN3* (synapsin III), identified in these pathways interesting candidates.

CHL1 (cell adhesion molecule L1-lLike) and *ROBO1* (roundabout guidance receptor 1) also stand out as potential candidates as a result of their involvement in the developmental biology pathway. This pathway, which obtained the third highest combined score, is of interest due to the broad spectrum of phenotypic features of the disorder under investigation.

4.3 Discussion

The genetic and gene expression data described in this study strongly implicates mutation of *SNIP1* as the cause of a novel autosomal recessive neurodevelopmental disorder and offers insight into the potential underlying molecular mechanisms responsible for the observed phenotypes. The extensive review of available phenotypic data provided a unique opportunity to more precisely define the clinical spectrum and cardinal features of this novel disorder which was only possible due to the high frequency of the condition within the Amish community. These findings will ultimately aid diagnosis and treatment, and reduce the number of misdiagnoses of affected individuals within the Amish community.

4.3.1 SNIP1 variant (p.Glu366Gly) responsible for novel autosomal recessive neurodevelopmental disorder

This study was only possible due to the extremely high prevalence of this rare condition within the Amish, particularly the communities of Indiana where an allele frequency of 0.0870 has recently been estimated (Chapter 5). The large number of affected individuals with our cohort permitted the collection of robust genetic evidence supporting the pathogenicity of this *SNIP1* variant (NM_024700.3: c.1097A>G; p.Glu366Gly).

Although samples from only six affected individuals were available for genome-wide SNP mapping the data identified a single, common, very modestly sized genomic region (1.65Mb) on chromosome 1. Assuming a founder mutation was responsible for the condition this region was believed to be the likely disease locus. Exome sequencing was undertaken on one of the genome-wide SNP mapped individuals. As expected, the cross-referencing of this data with the SNP mapping

data identified a single candidate variant in *SNIP1* within the shared region of homozygosity. No other candidates, predicted to be deleterious by *in silico* prediction software tools, were identified within the region of interest or genome-wide. A total of 17 genes were found to be located within the homozygous critical region (Figure 4.8). Whilst none have previously been reported to cause seizures/epilepsy, hypotonia/hyporeflexia or dysmorphic facial features, two genes, *GRIK3* and *MANEAL*, within the region have been linked to a neurological phenotype.

A 2.6-Mb microdeletion in 1p34.3, encompassing *GRIK3*, has been associated with severe developmental delay, presenting with mild retrognathia and down-slanting palpebral fissures in a single female proband [247]. GRIK family members are described as playing an important role in synaptic potentiation, a crucial process for learning and memory. Therefore the reported haploinsufficiency of GRIK3, as result of the microdeletion, is likely to be the cause of the developmental delay in this individual [247]. Although *GRIK3* is present in the 1.65-Mb region of homozygosity shared by all the affected individuals in our cohort (Figure 4.8) there is only a small overlap between these two regions. This, in addition to the absence of variants within *GRIK3* in our exome sequencing data, excluded this gene as a possible candidate.

The second gene, *MANEAL*, was previously suggested as a novel cause of a complex infantile-onset neurodegenerative disorder observed in a single male proband, born to consanguineous parents from Saudi Arabia. A loss-of-function variant in *MANEAL* was found to co-occur with a homozygous splice defect in *OSTM1*, considered likely to be responsible for infantile malignant osteopetrosis.

The neurological features of this phenotype are characterised by developmental delay, optic nerve atrophy, dyskinetic movement disorder and neurodegeneration determined through the appearance of brain iron accumulation (NBIA)-like pattern on a brain MRI. However, this small family study was the first, and currently only, association of this gene with a human disease. As such further investigation and evidence will be required to confirm the clinical significance of this variant. Examination of our exome sequencing data and the absence of variants within *MANEAL*, excluded it as a possible candidate.

The finding that mutation of *SNIP1* is causative of the neurodevelopment disorder described in our patients is in agreement with a previous study defining five candidate genetic causes of NDD in the Amish, including *SNIP1*. This study identified three affected individuals from two sibships with a condition referred to as 'PMRED' (psychomotor retardation, epilepsy, and craniofacial dysmorphism). Exome sequencing identified the *SNIP1* c.1097A.G (p.Glu366Gly) alteration as a candidate genetic cause [179]. This, together with the data presented as part of this PhD thesis, corroborates the c.1097A.G *SNIP1* variant as the cause of this condition.

4.3.2 Defining the clinical phenotype of a neurological disorder displaying severe psychomotor delay with seizures, epilepsy and dysmorphic features

The work undertaken by Dr Zineb Ammous and our collaborating clinical teams has been instrumental in precisely defining the clinical phenotype of this condition, and enabling the cardinal features to be determined.

The authors of the PMRED study described the characteristic phenotype of this condition as neonatal hypotonia with poor feeding [179]. Both of which are universal features observed within our cohort. Puffenburger et al., also reported characteristic dysmorphic features including; a bulbous nose, wide mouth and tongue, broad jaw with protuberant angles, short hands, short tapered fingers, and broad thumbs [179] with cranial MRIs showing irregular skull surface, white matter abnormalities and ventriculomegaly. All of these features were observed in at least 50% of patients within our cohort (Table 4.1).

Neurological and neurodevelopmental difficulties were major features of the disorder described by Puffenburger. This included the presence of psychomotor delay, epilepsy and absent reflexes. Affected individuals did not learn to walk, develop speech or engage socially and developed seizures by 6 months of age, with either focal or generalised seizures of varying types that may be intractable. Electroencephalograms (EEGs) showed multifocal spike-wave discharges from central, occipital and temporal regions [179]. Whilst most individuals in our study achieved independent ambulation it was significantly delayed, compared to unaffected individuals, and all individuals are non-verbal however, several can communicate through signs, gesticulation or sounds. Hypo- or areflexia were also observed to be universal features with our cohort.

The three individuals described by Puffenburger et al. all displayed ophthalmic features described as strabismus with horizontal nystagmus [179]. Although not a universal feature in our cohort 45% of individuals we assessed displayed some degree of strabismus or nystagmus with myopia also being observed.

In addition to the features described above we identified a number of additional features, not identified in the Puffenburger study. Congenital cardiac defects are a common feature observed in 60% of individuals within our cohort. These defects include hypoplastic left heart syndrome (HLHS), with two individuals dying in the neonatal period, aortic stenosis and bicuspid aortic valve (BAV) of varying severity, coarctation (narrowing) of the aorta seen in one individual, aortic root dilatation, atrial septal defect (ASD), ventricular septal defect (VSD), patent ductus arteriosus (PDA), pulmonary artery stenosis and mitral valve regurgitation. In addition to these features 11% (4/35) of individuals presented with cardiomyopathy.

Several respiratory difficulties were also observed in the 35 patients assessed as part of this study. Around 75% of individuals displayed upper airway respiratory difficulties in the neonatal period, including laryngomalacia, pharyngomalacia and subglottic stenosis of variable severity causing a weak cry, stridor and apnoea where also observed. Several patients required tracheostomy procedures and supplementary home oxygen. Another reported respiratory feature was asthma.

Finally, approximately 25% of the individuals assessed as part of this study presented with endocrine features, that were not previously reported in the three PMRED patients including hypoglycaemia, hypothyroidism and, in one individual, dyshormonogenesis.

4.3.3 Gene expression and cellular pathway data analysis

The high frequency of this condition within the Amish provided the opportunity to investigate how mutation of *SNIP1* affects its role as a transcriptional regulator, and define the impact of gene mutation on these functions, by undertaking whole transcriptome sequencing. RNA samples were extracted from whole blood obtained from six affected individuals, which were age and gender matched to form

a control group of six samples in which the presence of the *SNIP1* gene alteration was excluded. While a relatively modest number of samples were available for whole transcriptome sequencing, we hypothesised that the unique genomic architecture and wider homogeneity of Amish individuals may likely be beneficial and increase the statistical power of this study, especially given the fact that all individuals are homozygous for the same founder gene mutation.

As with any neurological or neurodevelopmental disorder it was not possible to undertake gene expression profiling on the primary tissue of interest, the brain, instead whole blood, an accepted proxy, was used. Since the first blood-based transcriptomic study of the neurological disorder Huntington's disease in 2005 [248] many gene expression studies using whole blood have been undertaken to elicit the pathophysiological mechanisms of neurological conditions supporting whole blood as a useful proxy measure for gene expression in the central nervous system [249].

Further evidence in support of using whole blood to undertake gene expression studies to investigate the effect of the *SNIP1* mutation is the expression profile of SNIP1 reported by the Genotype-Tissue Expression (GTEx) project. The GTEx project was supported by the Common Fund of the Office of the Director of the National Institutes of Health (and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS) which set out to create a resource enabling the study of genetic variation and the regulation of gene expression in multiple reference human tissues [250]. Figure 4.12 shows the gene expression profile for SNIP1 in multiple human tissues.

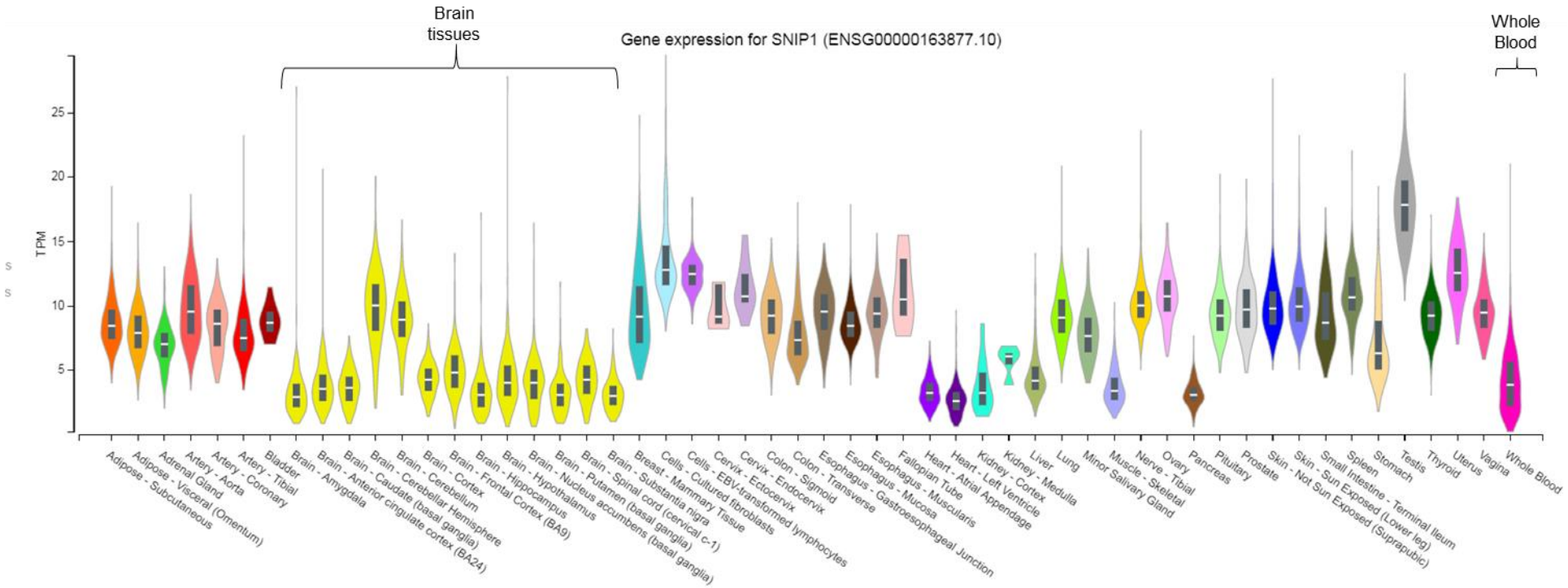


Figure 4.12: Gene expression for SNIP1 in multiple human tissues. The data used for this analysis obtained from: https://gtexportal.org/home/gene/SNIP1, the GTEx Portal on 21/10/2019 using GTEx Analysis Release V8 and dbGaP accession number phs000424.v8.p2 on 21/10/2019. Sorted according to tissue type with the outliner function switched off.

The data obtained from the GTEx portal indicates that whole blood provides a median TPM (transcript per million) for SNIP1 expression of 3.795 obtained from 755 samples. These values are within the broad range of median TPM units (2.929-9.948) obtained from the 13 different types of brain tissue analysed from a minimum of 139 samples as part of the GTEx project. It has been suggested that genes expressing more than two transcripts per million transcripts (TPM>2) are highly likely to be actively transcribed genes [251]. A reported TPM of 3.795 for SNIP1 expression in whole blood indicates that *SNIP1* is being actively transcribed in blood. This data supports the approach taken in this study and validates the use of RNA extracted from whole blood to investigate the impact of *SNIP1* mutation on transcription and gene expression.

Despite the limitations of sample size and the use of a proxy tissue sample, this gene expression study has been valuable in providing further insight into the key molecular signalling pathways potentially influenced by *SNIP1* mutation. Importantly, the developmental and biological pathways which were found to be most notably impacted by gene mutation were of direct relevance to the cardinal clinical features of the condition. Of immediate interest were the pathways involved in serotonin and dopamine release, which where the most highly enriched (highest combined score) within the upregulated gene list, and have well reported links associating their dysregulation with the onset of seizures, an invariable feature of *SNIP1*-related syndrome [245, 246].

Epilepsy affects ~65 million individuals globally [252] and whilst some epilepsies, or seizures, have well defined aetiologies, the underlying cause remains undefined in many cases. Correctly diagnosing an epilepsy syndrome has

significant implications with regard to treatment options [253]. A lack of knowledge regarding the underlying molecular basis of these syndromes is a contributing factor as to why epilepsy has been reported to be the leading neurological cause of reduced quality-adjusted life years (QALY), a measure of the quantity and quality of life as a result of healthcare interventions (Johnson 2019). Determining the underlying molecular cause of the SNIP1-related syndrome here, may provide an opportunity for clinicians to find the most suitable treatment option in order to achieve the greatest improvement in the quality of life experienced by affected individuals within our cohort, and others elsewhere affected by this condition.

Serotonin and dopamine and neurotransmitter release pathways

Both serotonin and dopamine are monoaminergic neurotransmitters that have been associated with the regulation of seizures [245, 246], a cardinal characteristic feature of this condition (Table 4.1). Two genes in the serotonin and dopamine and neurotransmitter release pathways, *SYT1* (synaptotagmin 1) and *SYN3* (synapsin III), are members of pre-synaptic protein families (the synaptotagmins and synapsins) responsible for regulating synaptic vesicle traffic and neurotransmitter release [254] by playing important roles in the process of normal calcium ion (Ca^{2+}) regulated neurotransmitter release [255, 256] (Figure 4.13).

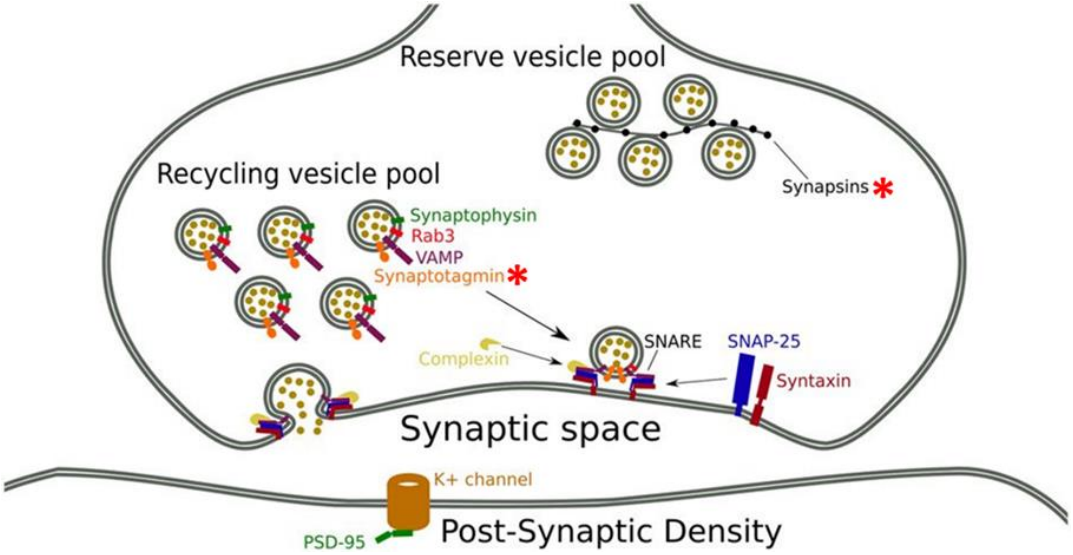


Figure 4.13: Showing the location of synaptic proteins in the synapses. Highlighted (*) are the two families of proteins identified due to genes of their family members being upregulated in our disease group (compared to controls) via enrichment analysis. Image taken (and modified) from Osimo et al. [257].

Normal brain function is critically reliant on the stringent regulation and timing of neurotransmitter release [256]. An action potential of +30mV within the presynaptic neurone will open voltage-gated Ca^{2+} channels, located in the presynaptic active zone (Figure 4.14A), initiating the release of a neurotransmitter filled synaptic vesicle. The influx of Ca^{2+} diffuses towards the synaptic vesicle and is detected by Ca^{2+} sensing proteins, from the synaptotagmin family (synaptotagmin 1 or synaptotagmin 2), located on the surface of the vesicle causing it to fuse with the presynaptic plasma membrane [258].

Synaptotagmin 1, encoded by the *SYT1* gene and responsible for the fast (millisecond) synchronous release of neurotransmitter [259], contains six different domains which are capable of acting mostly independently by sensing different molecules involved in different cell physiology pathways [260]. The C2A and C2B domains are the regions of the protein responsible for detecting Ca^{2+} ions; the C2A domain is capable of binding three Ca^{2+} ions with the C2B domain binding two Ca^{2+} ions (Figure 4.14B) [260]. The binding of Ca^{2+} to synaptotagmin 1

causes a conformation change within the protein that permits its interaction with the membrane fusion machinery [258]. Along with syntaxin binding protein 1 (MUNC18), Ca^{2+}-dependent activator proteins for secretion (CAPs) and the complexins (Figure 4.13), synaptotagmin 1 regulates the assembly of the SNARE (soluble N-ethylmaleimide-sensitive factor attachment) complex which is responsible for mediating exocytosis [261], the fusion of the synaptic vesicle with the presynaptic membrane and subsequent release of neurotransmitter into the synaptic cleft [262].

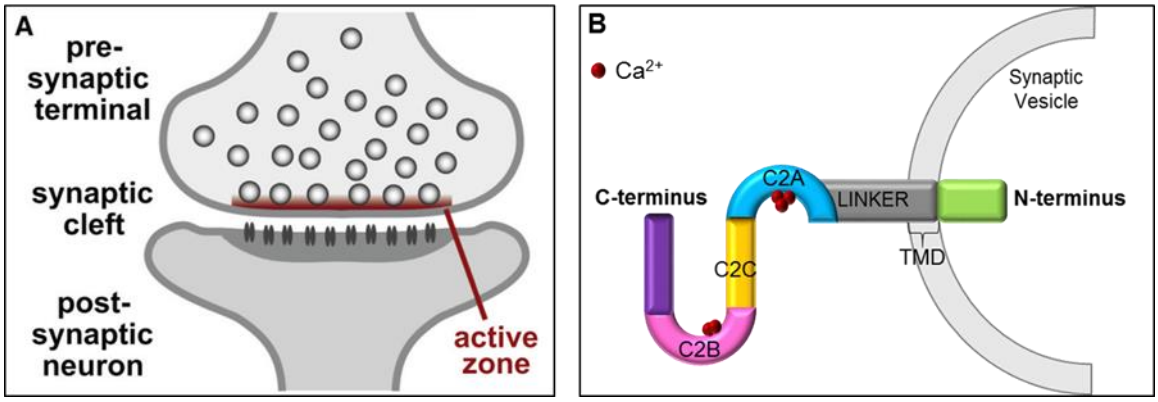


Figure 4.14: A) Schematic drawing of a synapse indicating the location of the active zone (image adapted from Südhof [263]). B) A schematic view of synaptotagmin1, with each functional domain is coloured differently, showing the location of Ca^{2+} binding to the C2A and C2B domains (image replicated and modified from Brachya et al. [260]).

In 2018, Baker et al. described 11 individuals with de novo heterozygous missense mutations in *SYT1* with neurodevelopmental delay (NDD) displaying a range phenotypic features (**Appendix H**), many overlapping those seen in *SNIP1*-related syndrome patients (Table 4.4).

Important differences between the presentations of these two disorders, is the absence of seizures and skull dysplasia in individuals affected by *SYT1*-

associated NDD. These cardinal features of *SNIP1*-related syndrome may relate to a more global developmental role of SNIP1.

Whilst other non-neurological features, such as congenital heart defects and respiratory issues are common to both *SYT1*-associated and *SNIP1*-related disorders endocrinological features, such as hypothyroidism and hypoglycaemia appear to be completely absent from the *SYT1*-associated disorder (*Table 4.4*).

Table 4.4: Comparison of the clinical presentation in the 33 affected individuals and SYT1- associated NDD.

Clinical presentation of the 35 affected individuals in Amish cohort	SYT1-associated NDD
Central Nervous System	
Global developmental delays/ Non-verbal	✓
Hypotonia, hyporeflexia	✓
Seizures and/or epilepsy	
Abnormal brain MRI (ventriculomegaly, hypomyelination)	
Behavior problems (irritability, autistic spectrum, ADHD)	
Friendly affectionate personality	
Cardiopulmonary	
Upper respiratory disease (laryngomalacia, apnea, stridor)	✓
Congenital heart defects (e.g. VSD, ASD, CoA, Ao valve)	✓
Cardiomyopathy	
Vision and Hearing	
Horizontal nystagmus and/or Strabismus	✓
Failed newborn hearing, conductive loss	
Gastroenterology and Nutrition	
Feeding difficulties	✓
Small for gestation/Failure to thrive	
Aspiration with and without Gastrostomy tube	
Endocrinology	
Hypothyroidism	
Hypoglycemia	
Morphological Features	
Abnormal skull shape ("lumpy", craniosynostosis)	
Wide mouth, cupid bow upper lip	
High arched palate	
Micrognathia (Pierre Robin Sequence in some)	
Short hands, tapered fingers	
Spine abnormality (scoliosis, sacral dimple, tethered cord)	✓
Hernia (umbilical, inguinal)	
Congenital talipes equinovarus	✓

Interestingly, SYT1 over expression, a possible outcome from the upregulation of SYT1 in our disease group, has been reported to repress tumor necrosis factor-alpha (TNF-α)-dependent nuclear factor-kappa beta (NF-κβ) transcriptional activation [264]. Notably, the NF-κβ signalling pathway is the same pathway the SNIP1 protein has been shown strongly to inhibit [238].

Synapsins are a family of evolutionary conserved, neuron-specific phosphoproteins [265]. Being present in all synapses in the brain, with the exception of ribbon synapses, they comprise ~1% of all proteins found in the brain making them one of most abundant families of synaptic proteins [266]. As a result of alternative splicing three synapsin genes, *SYN1*, *SYN2* and *SYN3*, encode a number of synapsin proteins generating distinct isoforms displaying isoform-specific distribution [255] [265]. Amino acid analysis of synapsin I and II, the first of the synapsins to be identified, showed that the N-terminus of each protein is highly preserved between the different isoforms whereas the C-terminus is more variable (Figure 4.15) reflecting their different functional properties and distribution [267, 268].

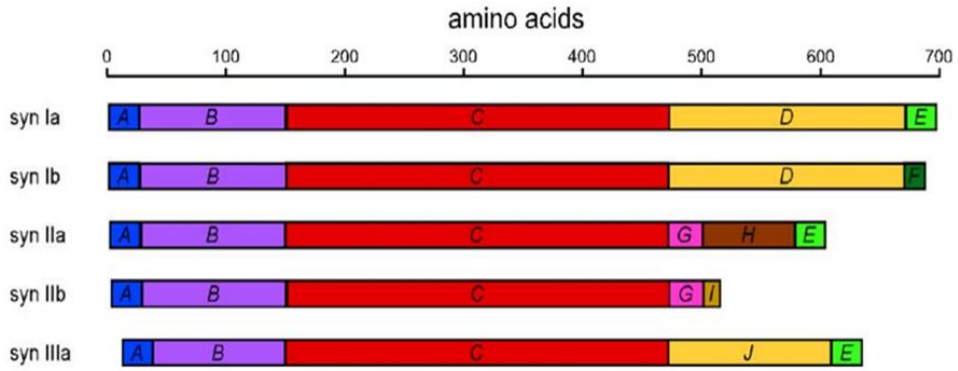


Figure 4.15: The synapsin family protein domains. Domain A, a short N-terminal region, is shared by all synapsin isoforms and contains a phosphorylation site that controls the reversible association with synaptic vesicles. Domain B is rich in small amino acids, varies between isoforms and is considered as a linker region connecting domain A to domain C. Domain C is a large region (~300 amino acids) believed to stabilise the interaction with the synaptic vesicle by penetrating its lipid bilayer. After domain C, the amino acid sequence diverges in the different synapsin gene products. However, all isoforms bear a proline-rich domain within the C-terminal region (within domains D, G, H or J). The amino acid scale is shown along the top. Image taken and modified from Cresca et al. [268].

The synapsin proteins have five major functions (Figure 4.16) with synapsins I and II playing similar roles in cellular processes. The later characterised synapsin III isoforms are predominantly expressed in early neuronal development but do maintain a role in synaptogenesis, neurogenesis and neuronal plasticity [266]. The synapsin proteins act over longer period of time, compared to the synaptotagmins, ranging from minutes to days when undertaking a role involved in in neuronal plasticity [260].

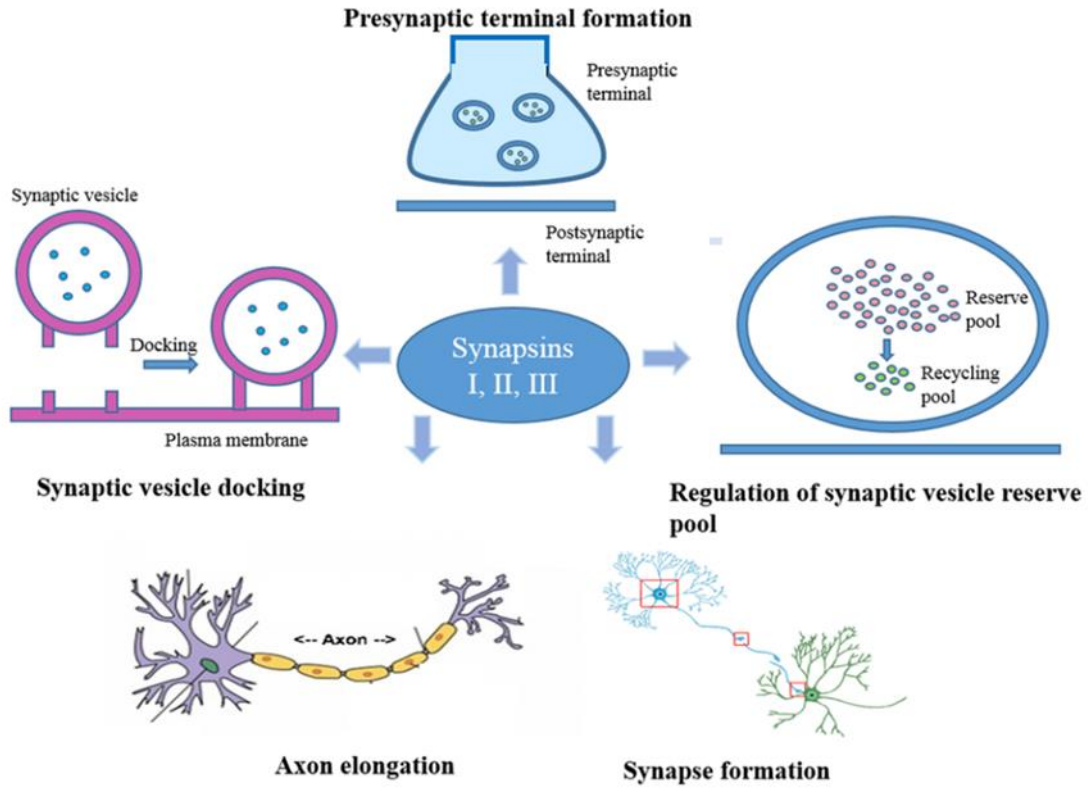


Figure 4.16: The major functions of synapsins taken from Mirza and Zahid, 2018 [266].

Mutation of synapsin genes have been associated with a number of neurological disorders including; Alzheimer's disease, schizophrenia, bipolar disorder, multiple sclerosis, Huntington's disease and of most relevance to this study, epilepsy [266]. Molecular analysis has implicated both synapsin I and II with the development of epilepsy. Nonsense and missense mutation of *SYN1* has been reported in individuals with autism spectrum disorders (ASDs) and epilepsy [269] and a polymorphism (rs37733634) in *SYN2* was shown to be significantly associated with idiopathic generalized epilepsy (IGE) [270].

Mutation of *SYN3* has not yet been reported to cause epilepsy. However, a study conducted in 2006 proposed synapsin III as a candidate for familial partial epilepsy with variable foci based on protein-protein interactions, described in the human protein reference database (HPRD), with synapsin I, a protein previously

linked with epilepsy [271]. Furthermore, in 2015 Zaltieri et al. presented evidence that synapsin III is involved in the regulation of dopamine neuron synaptic function [272]. The link between altered dopaminergic systems and epilepsy is well known [246] with a direct link between increased hippocampal extracellular concentrations and seizure activity previously demonstrated [273]. Taken together this information places SYN3 upregulation, as a result of mutation in SNIP1, as a strong possible cause of the observed epilepsy and/or NDD aspect of the phenotype in the affected individuals within our cohort.

Interrogation of the disorders associated with these genes identifies putative treatments and therapies that may be of benefit to individuals with SNIP1-associated syndrome [256, 271]. In 2018, Baker and Gordan et al. described the use of pramipexole, a non-ergot dopamine agonist (NEDA) widely used in the treatment of Parkinson's disease (PD), in a single patient open-label treatment experiment to treat SYT1-associated disorder. It was reportedly associated with; rapid and sustained reduction of the movement disorder, reduction in the frequency and severity of agitated and self-injurious behaviours, and increased responsiveness to social and environmental stimuli. In addition a reduction in EEG abnormalities was observed. Although this treatment has not yet been trialled on a second patient and further validation is required to assess the true effectiveness of this treatment, it is a promising finding that has the potential to reduce some of the clinical features of SNIP1-associated syndrome patients [256]. Additionally, the identification of serotonin and dopamine release pathways as the possible cause of epilepsy in our patients suggests the use of anti-epileptic drugs targeting these pathways may alleviate symptoms in SNIP1-associated syndrome individuals. For example phenytoin, carbamazepine, valproic acid,

lamotrigine and zonisamide have all been shown to cause an increase in extracellular serotonin levels inhibiting the onset of many types of seizures [274].

Developmental biology pathways

Due to the broad spectrum of phenotypic features another enriched pathway of interest is the developmental biology pathway which achieved the third highest combined score, for upregulated pathways. Interestingly, a few genes identified have been linked to disorders displaying overlapping characteristics of SNIP1-associated syndrome. The gene with the most notable overlap of phenotypic features is CHL1 (Cell Adhesion Molecule L1 Like). In 2012, Chen et al. described an infant with partial monosomy 3p (3p26.2 --> pter) and partial trisomy 5q (5q34 --> qter) presenting with psychomotor retardation, developmental delay, clinodactyly of the thumb, coarctation of the aorta, patent ductus arteriosus, peripheral pulmonary stenosis, atrial septal defect, microcephaly, brachycephaly, a small oval face, almond-shaped eyes, a down-turned mouth, a widened nasal bridge, hypertelorism, epicanthic folds, long philtrum and low-set large ears [275]. Important differences between the presentation of this individual and SNIP1-associated syndrome is the absence of craniosynostosis, which to date has been observed in all our patients. Additionally, a number of CNVs impacting the CHL1 gene have been linked to learning and language difficulties, another common feature of SNIP1-associated disease [276-278]. The prominent overlap of phenotypes may be suggestive of a link between a mutation in SNIP1 affecting the CHL1 protein and its associated functions.

The *CHL1* (close homolog of L1) gene, located at 3p26.3, encodes a protein belonging to the L1 family of neural cell adhesion molecules (NCAM), forming

part of the larger immunoglobulin superfamily (IgSF) [276, 277]. Members of the L1 family, often referred to as immunoglobulin cell adhesion molecules (IgCAMs), are essential for normal brain development. IgCAMs facilitate a range of developmental processes including cell proliferation and migration, neuritogenesis, axonal fasciculation, synaptogenesis, and stabilization of synapses. Disruption of IgCAM genes and subsequent disturbance of these processes has been linked to a number of neuropsychiatric disorders such as schizophrenia and mood disorders [279].

Members of the L1 family may either be a transmembrane glycoprotein, such as CHL1, or linked to the cell surface by a glycosylphosphatidyl inositol (GPI) anchor. They are characterised by the presence of six Ig-like domains located at the N-terminal and at least four fibronectin type III (FN3) repeats (Figure 4.17) [280].

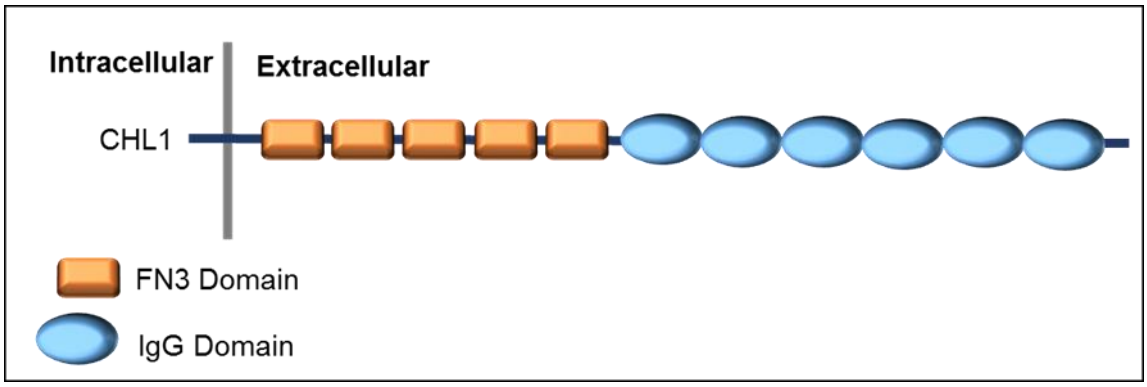


Figure 4.17: The schematic structure of CHL1 demonstrating the characteristic structure of a transmembrane immunoglobulin cell adhesion molecules (IgCAMs) containing six Ig-like domains and five FN3 repeats. Image modified from Irintchev and Schachner, 2012 [279].

CHL1, is expressed in most neurones within the central nervous system (CNS) and display higher levels of expression in embryonic brains compared to adult brains [279]. CHL1 regulates neurite overgrowth, axonal guidance, migration,

differentiation of neurones and in the mature brain accumulates at the axonal membrane to regulate synapse function [276]. Murine models of CHL1 function have shown *Chl1* deficiency causes aberrant neurotransmission, motor co-ordination and behaviour providing further evidence of its importance in normal brain functioning [276].

A further gene of interest enriched in the developmental biology pathway is *ROBO1* (roundabout guidance receptor 1) a transmembrane receptor involved in signal transduction, located at 3p12. Like CHL1, ROBO1 is a member of the NCAM family however, its role in axon guidance defines a separate, novel IgSF subfamily [281].

SLIT-Roundabout (SLIT/ROBO) signalling is now known to be involved a number of processes including, kidney induction and heart tube formation [282]. There are four members of the ROBO family (ROBO1, ROBO2, ROBO3 and ROBO4) three of which are expressed in human brain cells each containing five Ig domains, three FN3 domains and four conserved cytoplasmic domains. ROBO2 and ROBO3 are only seen in the nervous system but not in the vascular system where ROBO1 is known to be expressed in both systems making it an interest candidate for both the neurological and cardiac phenotypes observed in SNIP1-related syndrome [282, 283] (Figure 4.18).

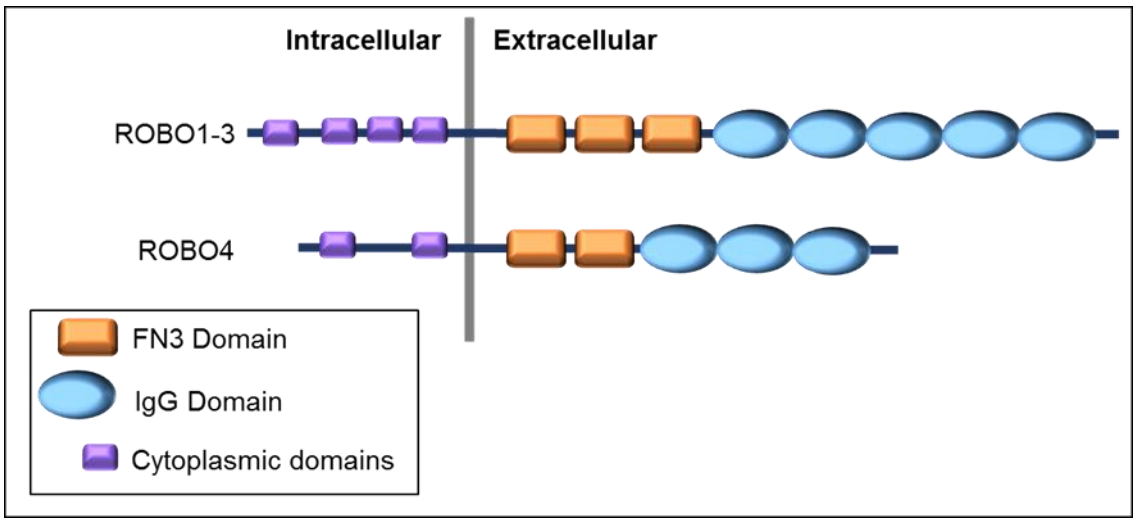


Figure 4.18: The schematic structure of ROBO family members. ROBO1-3 showing the typical structure wiht five Ig domains, three fibronectin type III (FN3) repeats and four conserved cytoplasmic domains and ROBO4 comprising of only three Ig domains, two FN3 domains and two cytoplasmic domains. Image modified from Ypsilanti et al., 2010 [283] and Yadva and Narayan, 2014 [282].

Disruption of *ROBO1* expression has been linked to learning difficulties particularly dyslexia susceptibility [284]. Although SNIP1-associated syndrome affected individuals experience learning difficulties the clinical feature of most relevance is that the disruption of ROBO1 has been reported to give rise to a relevant cardiac phenotype [285].

Loss of function (LOF) *ROBO1* variants were identified in three unrelated probands displaying different cardiac phenotypes including; ventricular septal defect (VSD), tetralogy of Fallot (a combination of four heart defects; VSD, pulmonary valve stenosis, a misplaced aorta and right ventricular hypertrophy) and congenital heart disease (CHD). Additionally, two mouse models with *Robo1* variants have been found to exhibit an atypical slit-Robo signalling pathway leading to the development of septation and outflow tract defects and, interestingly, craniofacial abnormalities [285], a cardinal feature observed in our cohort.

Although requiring further validation, the identification of molecules known to be involved with brain development (CHL1 and ROBO1) and neurotransmitter release (SYT1 and SYN3), previously linked to neurodevelopment and neurological condition and displaying considerable overlap with SNIP1-associated syndrome phenotypes provides insight into the function of SNIP1 as a transcriptional regulator.

4.3.4 Future work

The genetic and transcriptomic studies defined here have laid important foundations for a greater understanding of the pathomolecular basis of a novel form of neurodevelopmental delay. Currently the p.Glu366Gly amino acid alteration identified in the Amish is the only known cause of this condition. As no other families are reported elsewhere with this or other *SNIP1* alterations, it is important to work with genome sequencing teams worldwide to identify additional families with this and other candidate *SNIP1* mutations to learn more about genotype-phenotype outcomes associated with this condition.

An additional important next step is the validation and further investigation of the genes and developmental pathways identified as altered through whole transcriptome sequencing. This could be achieved through the use of a Taqman low density array (TLDA) which will perform real-time PCR detection and relative quantitation of expression for the genes of interest (targets). The TLDA platform is a closed system utilising a validated singleplex PCR methodology. The TLDA card, capable of running 24 duplicates, is designed with targets pre-allocated to reaction wells enabling fast and simultaneous detection of multiple gene targets [286]. Patient and control samples for these studies have been obtained, and this work is now in progress.

In parallel with this functional studies, using (for example) murine models and transgenic cell lines, could be conducted to assess the expression of *SNIP1* in the brain and to more precisely describe the binding partners and subcellular localisation of the SNIP1 protein, and define its molecular role including potential downstream impacts. Due to its likely involvement in developmental pathways, studies in embryonic mice, <E8.0, may determine its potential role in formation of the nervous plate and heart tube, given the presentation of related neurological and cardiac phenotypes. Additionally, as craniosynostosis is a cardinal clinical sign of this condition these studies may identify molecules of relevance to skull development and suture formation/closure, and provide important insights into these poorly understood biological processes.

There are numerous potential benefits to this work involving the broadening of the clinical spectrum of this disorder and elucidating the functional outcomes of mutations in *SNIP1*. The *SNIP1* mutation has now been incorporated into regional molecular diagnostic sequencing panels utilised by the Amish communities. This will enable early diagnosis, aiding early intervention and patient management as well as preventing the future misdiagnosis of this condition. Together this will ensure affected individuals and their families are provided with adequate genetic counselling and access to potential therapies and treatments that may help reduce the observed phenotypes of *SNIP1*-related syndrome and improve the quality of life for those affected by the disorder. The enrichment of *SNIP1*-related syndrome within the Amish communities of Indiana has provided a unique opportunity to not only confirm the genetic cause but to investigate the functional consequences of the identified mutation which can be used to inform possible future treatment options.

CHAPTER 5

INTERROGATION OF AMISH AGGREGATED EXOME DATA TO IDENTIFY POTENTIALLY DELETERIOUS COINCIDENTAL HETEROZYGOUS SEQUENCE VARIANTS AND DETERMINE THE ALLELE FREQUENCIES OF PATHOGENIC VARIANTS SEEN WITHIN THE VARIOUS AMISH COMMUNITIES

5 Interrogation of Amish aggregated exome data to identify potentially deleterious coincidental heterozygous sequence variants and determine the allele frequencies of pathogenic variants seen within the various Amish communities

5.1 Aims

Over recent years the WoH Project has accumulated extensive single nucleotide polymorphisms (SNP) and exome datasets from patients and individuals from the Amish community. The aim of this project was to begin to explore these datasets more widely, to describe and characterise the architecture of the Anabaptist genome. The overarching objectives of the study were;

- To interrogate exome datasets to identify potentially deleterious autosomal recessive variants in genes already known to cause disease, coincidentally carried by individuals enrolled on the study, to learn more about the spectrum of inherited diseases present in the community.
- To more precisely determine the allele frequencies (AF) of the pathogenic variants identified across the various Amish communities.
- To share information gained from the analysis of the Anabaptist genome within the healthcare system locally and with the scientific and clinical communities more widely.

5.2 Introduction

Sequence variations (variants) occur throughout the genome. Some arising at positions that alter the DNA sequence in a way that changes the amino acid sequence of the encoded protein (non-synonymous) where other variants will have no effect on a resultant protein (synonymous). Non-synonymous variants effect the function of protein and therefore the phenotypic characteristic of an individual, often effecting susceptibility towards a genetic disorder [287].

There are several types of non-synonymous variants including; nonsense, frameshift and missense. Nonsense variants are single nucleotide changes that result in the introduction of a premature stop codon (pre termination codon, PTC). Frameshift variants are caused by the insertion or deletion (indels) of a number of nucleotides within a DNA sequence altering the reading frame (triplet codon pattern). Missense variants are point mutations that do not introduce a PTC but still alter the amino acid sequence of resultant protein and have the potential to affect its functionality. Nonsense and frameshift variants are considered the most likely to be disease causing (pathogenic) due to their ability to dramatically alter the function of a protein. In some cases, due to a cellular quality control mechanism known as nonsense-mediated decay (NMD), an aberrant protein is abolished completely through degradation of its mRNA transcript [288].

5.2.1 Variant annotation

Online genomic databases are one way the scientific and clinical communities have gone about collating information on variants identified in many laboratories across the globe to facilitate data sharing and enable standardisation of variant classification. However, determining the clinical significance of (particularly novel) sequence variants which is vitally important for rare disease diagnostics, is still extremely challenging [289].

In 2015, the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) published standards and guidelines as an educational resource to aid clinical laboratory geneticists in the interpretation of sequence variants [289]. Although compliance to these standards is voluntary its use has been widely endorsed globally proving instrumental in the promotion of consistency and agreement of sequence variant classification across the clinical genetics community [290]. The ACMG/AMP guidelines incorporate several types of evidence that are used to classify variants. Figure 5.1 summaries how this evidence is used during the process of variant call file (VCF) annotation.

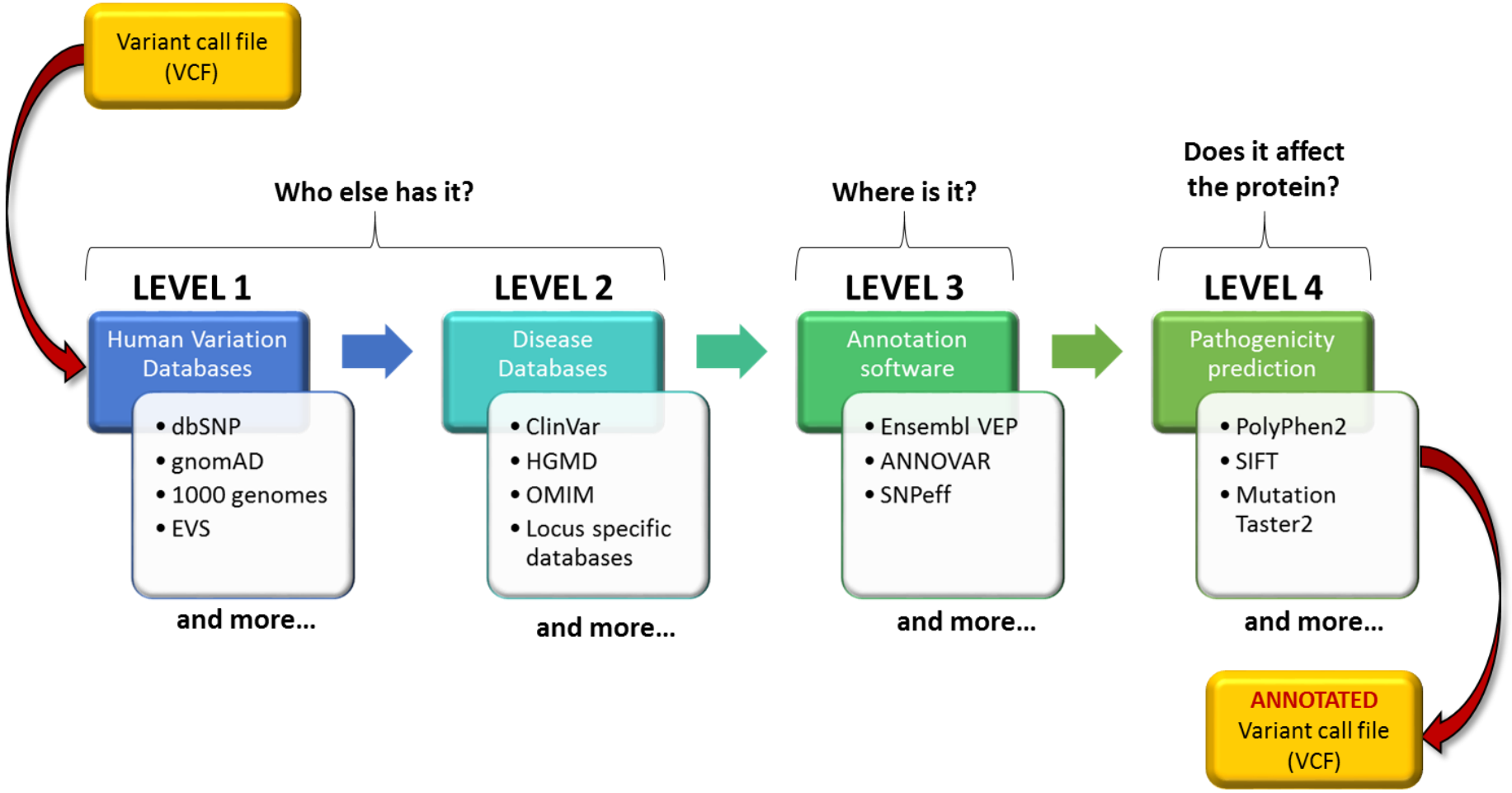


Figure 5.1: Schematic showing an overview of the different sources of information used during the process of variant call file annotation.

The first step in annotating a VCF file involves determining if a variant has been previously reported, at what frequency and in which populations (Figure 5.1). There are a number of publically available databases containing this information which can be used depending on which is best suited to the research or clinical needs of a particular study [291]. The two commonly used databases are the Single Nucleotide Polymorphism Database (dbSNP) [292] and the Genome Aggregation Database (gnomAD) [89].

The Single Nucleotide Polymorphism Database (dbSNP)

The Single Nucleotide Polymorphism database (dbSNP) was established in 1999 by the National Centre for Biotechnology Information (NCBI) as a freely available catalogue of simple genetic polymorphisms [293]. Collated as a tool to aid the understanding of human variation and molecular genetics for use in a broad range biological applications including large scale association studies, gene mapping, functional analysis, pharmacogenomics and evolutionary biology [287].

The name of this database, dbSNP, is slightly contradictory given the variety of variants included in the database including; single-base nucleotide substitutions (SNPs), small-scale multi-base deletions or insertions (indels), retroposable element insertions and microsatellite repeat variations (short tandem repeats or STRs). The name probably reflects the fact that single nucleotide variants are the largest class of variants with in dbSNP, comprising ~99.7%, of the database [287, 292, 293]. However, this too is confusing as single nucleotide variants are not always polymorphisms. A polymorphism, by definition, is a variant that occurs in >1% of the population [294]. The spectrum of variants included in dbSNP

ranges from neutral polymorphisms to disease-causing clinical mutations and provides information on clinical actionability.

In terms of variant annotation the dbSNP database is a useful source of information reporting; the context of a variant (showing the surrounding sequence), the frequency of the variant in different populations and the experimental method used to assess the possible functional implications the variant [287].

The Genome Aggregation Database (gnomAD)

The Genome Aggregation Database (gnomAD) is a large-scale reference data set cataloguing genetic diversity observed across 125,748 human exomes and 15,708 human genomes [89]. This data set builds on the work of its predecessor the Exome Aggregation Consortium (ExAC) which collated variant data from 60,706 exomes [295].

GnomAD has now become a vital tool for both the clinical interpretation of variants. As the largest database of high-quality variant calls it permits the filtering of candidate variants by frequency at the highest resolution, which is particularly important for analysis low-frequency variants [295].

An important point to consider when using frequency data from gnomAD is the origin of the exome and genome data included in the database. Although individuals, and first-degree relatives, known to be affected by severe paediatric disease have been removed, included individuals were sequenced as part of a number of different disease-specific and population studies. This means the database may contain individuals with severe disease which should be taken into

consideration when investigating the number of reported homozygotes for a given variant to assess its likely pathogenicity.

Two other databases; ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/) [296] at the National Centre for Biotechnology (NCBI), an open access database, and Human Gene Mutation Database (HGMDPro), a commercially available database (http://www.hgmd.cf.ac.uk/ac/index.php), [297] are also now routinely utilised by researchers, clinical laboratory staff and clinicians to ascertain and aid interpretation of genetic variation [298].

ClinVar

Initially released in 2013, populated predominantly by variants described in OMIM® and GeneReviews™, this freely accessible archive of germline and somatic variants now holds over 600,000 submitted records from ~1000 submitters, from 60 countries on five continents [299], providing information on >430,000 unique variants [300]. Although not the first centralised database of human genetic variation, ClinVar was the first to aggregate data from many different sources, including clinical testing laboratories [300], and to include both germline and somatic variants of any size, type or genomic location [301] unlike other databases such as HGMD or the Catalogue of Somatic Mutations in Cancer (COSMIC) [302].

Another benefit of ClinVar, particularly for those in a clinical setting, is its use of the specific standardised terminology recommended by the American College of Medical Genetics and Genomics (ACMG) to classify the clinical significance of variants causing Mendelian disorders as; “pathogenic,” “likely pathogenic”, “uncertain significance”, “likely benign”, or “benign” [95]. Although the accuracy

of individually submitted variant interpretations is not evaluated, ClinVar calculates an aggregated interpretation based on those provided by each submitter and reviews the level of supporting evidence used to ascertain the clinical significance by allocating a review status to each variant, taking into account the type of submitter [301]. The review status is summarised on each variation report in the form of gold stars. Table 5.1 provides definitions of each review status and the corresponding number of gold stars.

Whilst this process ensures the quality and reliability of submitted variants included within the database the requirement of expert panels and steering committees to assess evidence and award any status higher than two stars may hinder variants obtaining relevant review statuses within appropriate timeframes. An example of this would be the *GJB2* gene variant, NM_004004.5 (*GJB2*):c.35delG; p.Gly12Valfs, which although first identified as a genetic cause of hearing loss in 1997 [155] and subsequently being found to be responsible for ~50% of all hearing loss in European populations (Smith, Bale et al. 2005), was only recently (September 20 2018) granted a three star review status by an expert panel.

Table 5.1: ClinVar review status, assignment of stars and description of each when each status is awarded.

Gold stars	Review status	Description
Four	Practice guideline	Evidence for a variant has been included in practice guidelines for a particular genetic disorder. Evidence is reviewed by the ClinGen Steering Committee who award practice guideline status for clinical assertions.
Three	Reviewed by expert panel	The evidence for a particular variant has been reviewed by expert panel and a consensus of clinical significance has been agreed.
Two	Criteria provided Multiple submitters No conflicts	Obtained when two (or more) submitters with assertion criteria* and evidence independently provide the same interpretation.
One	Criteria provided Conflicting interpretations	Assigned to a variant when multiple submitters provide assertion criteria* and evidence but there are conflicting interpretations. The independent values are enumerated for clinical significance.
One	Criteria provided Single submitter	Allocated to variants where one submitter has provided an interpretation with assertion criteria* and evidence
None	No assertion for the individual variant	The allele has not interpreted directly in any submission for example, it has been submitted to ClinVar only as a component of a haplotype or a genotype.
None	No assertion criteria provided	The variant submission included an interpretation but no assertion criteria* or evidence was provided.
None	No assertion provided	The allele was included in a submission that did not provide an interpretation.

*Assertion criteria refers to a publication that describes the criteria used by the submitter to assign the variant to an ACMG category

ClinVar has been instrumental in facilitating the sharing of variant interpretations, particularly within the clinical genetics community. Its use is encouraging communication and debate regarding differences in variant interpretations in

different laboratories as well as helping promote the standardisation of reporting the clinical significance of variants [300].

HGMD

First made publicly available in April 1996 [303], HGMD was the first comprehensive and publicly available database of germline mutations underlying human inherited disease [304]. It was recently reported to contain >203,000 different genetic lesions, in over 8000 genes that have all been manually curated from more than 2600 journals. [305].

HGMD is available in two versions. The public version of HGMD is freely available to registered users, but is updated with newly reported mutations twice a year. The professional version is only available via paid subscription (via QIAGEN) but is updated more regularly being curated and populated with newly reported mutations every quarter. It also includes a number of additional functions making it possible to interrogate particular variants, providing links relevant publications, and is available as a flat file permitting its use in exome and genome VCF annotation [305].

Unlike ClinVar, HGMD does not use standardised ACMG terminology to classify variants. Instead variants in HGMD are assigned to one of six classes (Table 5.2) depending on the strength of the clinical and functional evidence provided in the peer-reviewed scientific literature in which the variant was described. A further difference between HGMD and ClinVar is that HGMD do not regularly review or remove variants, or genes, no longer considered to be disease-causing meaning it contains a lot of historical data providing out of date interpretations. One example is a missense variant of *CASR*, linked to familial hypoparathyroidism

and hypocalciuric hypercalcemia, which is still reported as DM in HGMD but was awarded a two star reviewed clinical significance of benign/likely benign in ClinVar in February 2018. In addition a number of genes, *CDKN1A*, *CDKN2B* and *CDKN2C*, are also still regarded as containing disease-causing mutations, even though the genes themselves are no longer considered disease causing.

Table 5.2: Summary of the six variant classes used by HGMD [305].

Variant Class	Symbol	Class description
Disease-causing mutations	DM	Variants that are very likely to be causing the observed clinical phenotype
Probable/possible pathological mutations	DM?	Variants where there is some degree of uncertainty regarding the interpretation of clinical significance in the scientific report
Disease-associated polymorphisms	DP	Variants with evidence to support a significant association with a disease/clinical phenotype in addition to evidence that the variant is likely to be of some functional relevance although there may not currently be any direct evidence of a functional effect
Functional polymorphisms	FP	A direct functional effect has been demonstrated but there is currently no reported associated disease
Disease-associated polymorphisms with supporting functional evidence	DFP	Variants should not only have been reported to be significantly associated with disease, but should also display direct evidence of being of functional relevance
Retired records	R	Variants that have been removed from HGMD. • Found to have been erroneously included • If the variant has been subject to retraction/correction in the literature resulting in the record becoming obsolete, merged or otherwise invalid

Despite this HGMD data is considered an extremely useful resource and is used extensively to aid classify variants according to their pathogenicity and to perform meta-analyses on different types of gene mutation causing human inherited disease improving our understanding of the mutation spectrum and molecular mechanisms underlying hereditary disease. With more than 100,000 registered users of its public version, and in excess of 7.8 million queries successfully served since 2007, it is an essential tool for many researchers and diagnostic laboratories when annotating sequencing data so much so that NHS England currently holds a licence accessible to all NHS clinical scientists and clinicians [305].

5.2.2 International effort to share genomic variant data

The development and maintenance of online genomic databases is not only important for obtaining evidence to determine the clinical significance of genetic variants or reviewing the mutational spectrum of specific genes, it is a vital repository of information for use in bioinformatic projects that will underpin personalised genomic medicine. Despite the considerable progress in disease gene identification since the advent of NGS technologies, the majority of annotated genes have yet to be assigned a function, in the context of human disease traits [306]. In recent years there has been an enormous international effort within the genomics community to identify the best ways of sharing genomic data, including benign variants, to maximise the utility of the available data and increase our knowledge of annotated genes by more precisely defining their role in disease. Several programmes, at both a national and international level, have been established in an effort to achieve a comprehensive understanding of the molecular basis of disease biology and disease gene function [306].

In the US, two National Institutes of Health (NIH) projects, ClinVar and ClinGen, have formed a partnership to improve knowledge of clinically relevant genomic variation through the sharing, archiving, curation and dissemination of genomic data. ClinVar, as detailed in section 2175.2.1 is an online database that collates information on genetic variants and its relationship to human health. Whereas ClinGen is a central clinical genome resource that works to define the clinical relevance of genes and variants for use in precision medicine and research. The overall aim of this partnership is to “improve patient are through genomic medicine” (Figure 5.2).

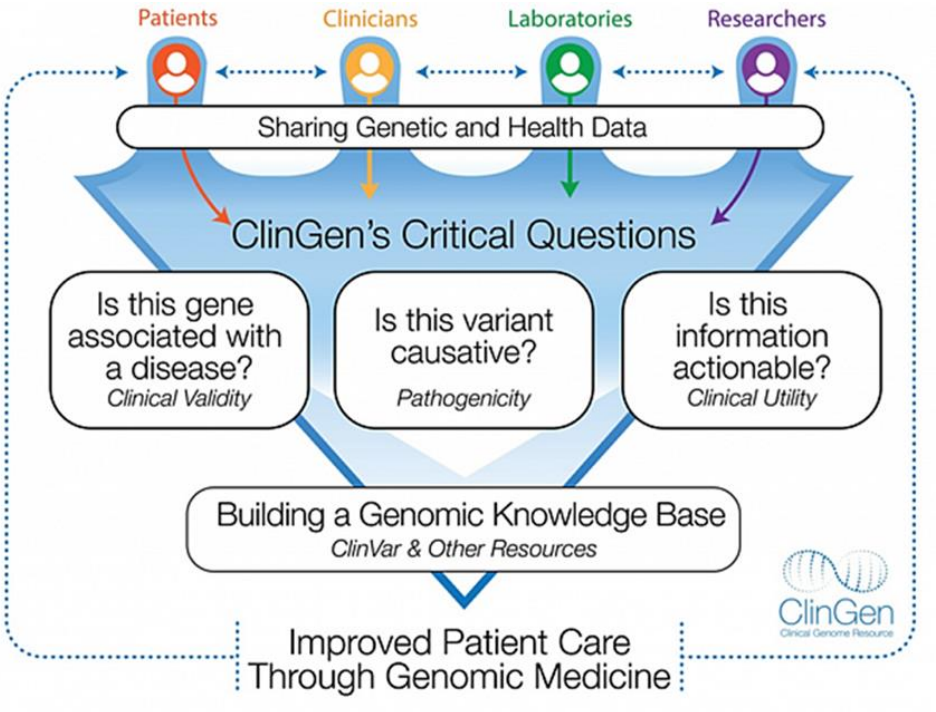


Figure 5.2: Schematic illustrating the overall aims of the NIH ClinGen project describing it hopes to improve patient care through genomic medicine reproduced from https://www.clinicalgenome.org/about.

Other NIH-supported programmes include the establishment of Centres for Mendelian Genomics (CMGs) which have been responsible for the development

of a number of gene-matching tools including; GeneMatcher, MyGene2 and Matchbox. Gene-matcher tools were designed to connect clinicians with researchers in human and model organism genetics [306]. These gene-matcher tools are now able to communicate through the use of a common application programming interface (API) [307] hosted by the Matchmaker Exchange (MME). This initiative, launched in October 2013, aimed to internationally unify efforts in gene-phenotype matching by facilitating the interaction between multiple disconnected projects (Figure 5.3) to by providing a robust and systematic approach to rare disease gene discovery.

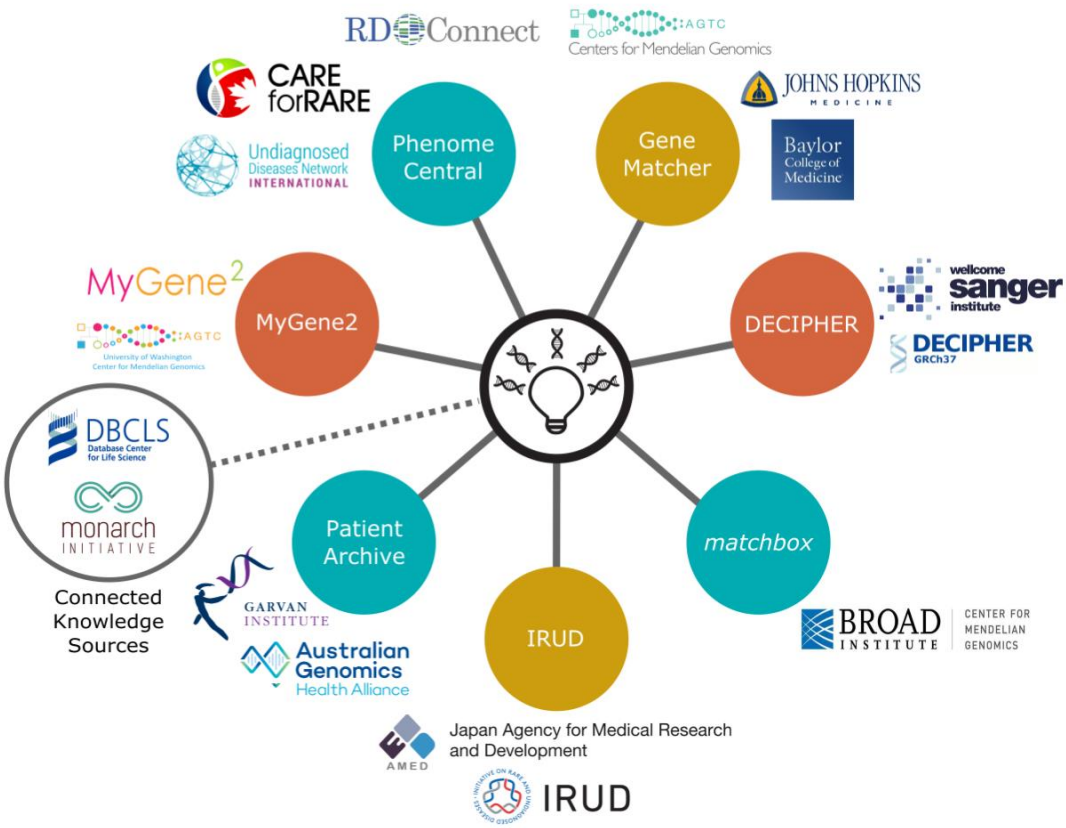


Figure 5.3: Member Organizations of the MME project reproduced from https://www.matchmakerexchange.org/

One example of the power of the MME initiative is demonstrated through the work of Bruel et al. who over a period of 2.5 years used the application to successfully matched 84% of the genes they submitted (60/71) and enabling confirmation of the pathogenicity of 39% of these matched genes (23/60) [308].

5.2.3 Contribution of Amish knowledge to the international genetic community

As well as greatly expediting disease diagnosis, the increased availability and widespread use of NGS technology for clinical genetic testing generates large amounts of genomic data on the individuals investigated. When undertaken in a community setting, this provides an opportunity to greatly advance knowledge of the architecture of genetic disease of direct relevance to that population. Research undertaken by the WoH and other research groups, including research undertaken by the Clinic for Special Children (CSC), has led to the accumulation of a large body of genomic information regarding known and candidate new pathogenic variants present within the Anabaptist communities. Together with other groups working with Anabaptist communities, WoH is leading the creation of an “Anabaptist specific mutation database” to support clinical services managing patients. To increase the clinical utility of Anabaptist variant data it is being shared with existing public variant databases, including ClinVar, to aid variant interpretation globally.

Genomic data from the Anabaptist community, which display a higher incidence of some rare genetic variants due to the presence of founder mutations, is beneficial in a number of ways. For example, it may enable the matching of cases with families with similar phenotypic and genotypic profiles located elsewhere. An

early example is that of Ellis-van Creveld (EVC) syndrome, which was first described in 1940 by Richard Ellis and Simon Van Creveld [2]. Ellis-van Creveld syndrome is a rare form of dwarfism involving skeletal and chondroectodermal dysplasia with an incidence in non-Amish populations of 7/1000000. However, the condition is present at greatly increased frequency in some Amish Demes reaching ~1/5000 [309] where it has been possible to trace the lineage of the variant back to a single founder couple, Samuel King and his wife, who immigrated to Eastern Pennsylvania in 1744 [310]. After the initial discovery and description of this condition in the Amish, other cases of EVC syndrome occurring globally were more easily recognised; a number of other more recent examples of this stem from the work of the WoH Project including GM2 and GM3 synthase deficiencies, *HERC2* (Blue eye delay) syndrome, *KPTN* (MASD) syndrome, Troyer (*SPG20*) syndrome and Mast (*SPG21*) syndrome. Thus, the willingness of the Amish communities to partake in genetic studies, such as those conducted by WoH, has been paramount to the increased understanding within the medical genetic field of rare genetic disease.

5.3 Results

5.3.1 Identifying potentially deleterious coincidental heterozygous sequence variants

Exome data from 26 individuals with moderate-severe intellectual disability with/without additional syndromic features was de-identified, aggregated and analysed to identify potentially deleterious heterozygous variants, not thought to be responsible for the difficulties experienced by the individuals but being coincidentally carried. This work was undertaken as a proof-of-principle study to ascertain the utility of such an approach for facilitating the identification of candidate disease associated variants within a community setting (summarised in Figure 5.4). Variants were initially filtered to prioritise only heterozygous variants predicted to be of high functional impact (nonsense and frameshift) that passed stringent quality control constraints (PASS, Filter [VCF]). Only variants with high Phread scores (Q scores) were included in the analysis in order to reduce the likelihood of false positive variant calls.

As an additional quality metric only variants with good quality calls were included in our analysis. This involved reviewing the allele depths for each heterozygous call in each individual to ensure the proportion of the alternative allele was ~50%, compared to the reference allele. Concerns over the quality of a heterozygous call, due to uneven allele ratios, were investigated using the Integrative Genomics Viewer (IGV) software. IGV, a desktop visualisation tool for large integrated genomic datasets [311], permits the interrogation of variant calls by displaying the coverage and quality of read alignments.

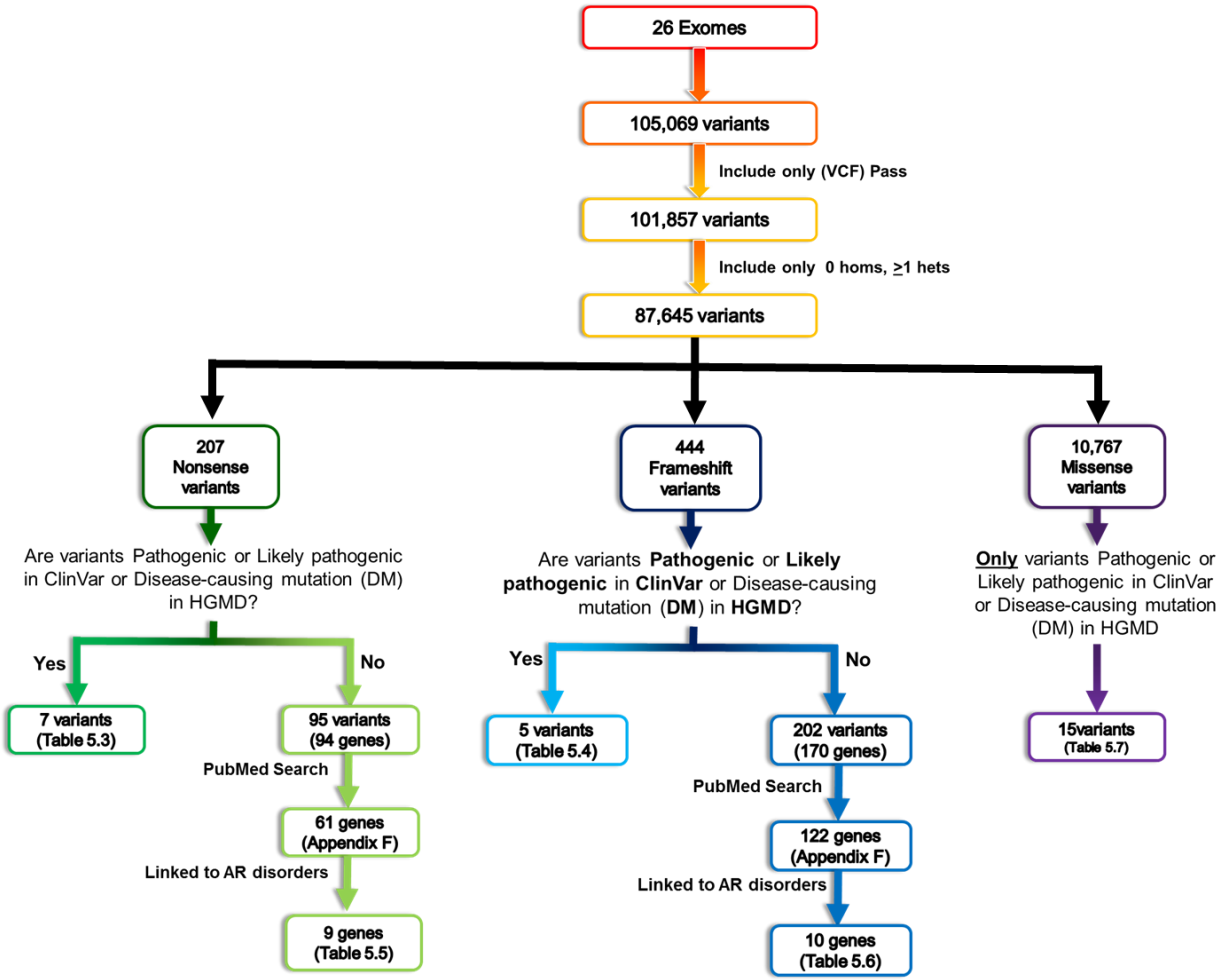


Figure 5.4: Summary of the Amish exome data analysis including filtering criteria, number of variants and location of results.

Variants previously described as associated with human genetic disorders

Initial findings identified seven nonsense variants, responsible for autosomal recessive disease, reported as pathogenic, or likely pathogenic, in ClinVar or described as a disease-causing mutation (DM) in HGMDPro (Table 5.3). A further five frameshift variants reported as pathogenic, or likely pathogenic, in ClinVar or described as a disease-causing mutation (DM) in HGMD were also identified (Table 5.4). Eight, of a total of 12 (nonsense and frameshift) variants identified in our cohort had previously been seen in the Amish community (shown in blue in Table 5.3 and Table 5.4), six of which (*KPTN*, *WDR73*, *SPG21*, *TMCO1*, *SLITRK6* and *ST3GAL5*) where identified through the work of the WoH.

Phenotype/ Disorder	Gene	MOI	HGVS Nomenclature	dbSNP	gnomAD	
					Freq. (March 2019)	No. of homs
Amish infantile epilepsy syndrome	*ST3GAL5*	AR	NM_003896.3: c.862C>T; p.(Arg288Ter)	rs104893668	0.00005425	0
Craniofacial dysmorphism, skeletal anomalies and mental retardation	*TMCO1*	AR	NM_019026.4: c.292_293del; p.(Ser98Ter)	rs752176040	0.0002768	0
Deafness and myopia	*SLITRK6*	AR	NM_032229.2: c.1240C>T; p.(Gln414Ter)	rs587777069	0.000008884	0
Glycogen storage disease type 1A	*G6PC*	AR	NM_000151.3: c.1039C>T; p.(Gln347Ter)	rs80356487	0.0003813	0
Mental retardation, autosomal recessive 41	*KPTN*	AR	NM_007059.3: c.776C>A; p.(Ser259Ter)	rs374298314	0.00007807	0
Coenzyme Q10 deficiency, primary, 8	*COQ7*	AR	NM_016138.4 c.422T>A p.(Val141Glu)	rs864321686	.	.
Intellectual disability	*FRY*		NM_023037.2: c.3589C>T; p.(Arg1197Ter)	.	.	.

Of the four variants not yet reported in the Amish, three have strong evidence in support of their pathogenicity due to the fact they are reported in ClinVar as pathogenic; with two variants, located in *CYP1B1* and *RAD50,* receiving two star

review statuses indicating multiple submitters have independently provided evidence in supporting the same interpretation. However, the pathogenicity of the

Phenotype/ Disorder	Gene	MOI	HGVS Nomenclature	dbSNP	gnomAD		
					Freq. (March 2019)	No. of homs	Clinic sig.
Mast syndrome	*SPG21*	AR	NM_016630.6: c.601dupA; p.(Thr201Asnfs)	rs387906275	0.0001147	0	Pathog
Familial partial lipodystrophy	*LIPE*	AR	NM_005357.2: c.3203_3221del; p.(Val1068Glyfs*102)	rs587777699	.	.	Pathog Like pathog
Epidermolysis bullosa, Herlitz	*LAMC2*	AR	NM_005562.2: c.2006_2012del; p.(Ile669Lysfs)	rs778012079	.	.	Like pathog
Buphthalmos	*CYP1B1*	AR	NM_000104.3: c.1064_1076del; p.(Arg355Hisfs*69)	rs72549380	0.0004189	0	Pathog
Nijmegen breakage syndrome-like disorder	*RAD50*	.	NM_005732.3: c.2156dupT; p.(Glu723Glyfs*5)	rs397507178	0.0004818	0	Pathog

fourth variant in *FRY* is uncertain as there is insufficient evidence to support it being a disease gene as it is not reported in ClinVar and has only been observed in one family, in the literature, where it was described in one large family with poor phenotype data.

*Table 5.3: Nonsense variants classified as pathogenic or likely pathogenic in ClinVar or with a disease-causing mutation (DM) in HGMD® reported to cause an AR disorder/phenotype. Variants previously reported in the Amish are highlighted in **blue**. The variant in FRY is shown in grey there is less robust evidence in support of its pathogenicity.*

Phenotype/ Disorder	Gene	MOI	HGVS Nomenclature	dbSNP	gnomAD		ClinVar			HGMD	
					Freq. (March 2019)	No. of homs	Clinical sig.	Review Status (March 2019)	Last Reviewed	Accession No.	Variant Class
Amish infantile epilepsy syndrome	*ST3GAL5*	AR	NM_003896.3: c.862C>T; p.(Arg288Ter)	rs104893668	0.00005425	0	Pathogenic	1	Apr 24, 2018	CM043092	DM
Craniofacial dysmorphism, skeletal anomalies and mental retardation	*TMCO1*	AR	NM_019026.4: c.292_293del; p.(Ser98Ter)	rs752176040	0.0002768	0	Pathogenic	2	Jul 5, 2017	CD100153	DM
Deafness and myopia	*SLITRK6*	AR	NM_032229.2: c.1240C>T; p.(Gln414Ter)	rs587777069	0.000008884	0	Pathogenic	0	Nov 24, 2014	CM133801	DM
Glycogen storage disease type 1A	*G6PC*	AR	NM_000151.3: c.1039C>T; p.(Gln347Ter)	rs80356487	0.0003813	0	Pathogenic	2	May 3, 2018	CM940797	DM
Mental retardation, autosomal recessive 41	*KPTN*	AR	NM_007059.3: c.776C>A; p.(Ser259Ter)	rs374298314	0.00007807	0	Pathogenic	1	Sep 14, 2017	CM140085	DM
Coenzyme Q10 deficiency, primary, 8	*COQ7*	AR	NM_016138.4 c.422T>A p.(Val141Glu)	rs864321686	.	.	Pathogenic	0	Nov 1, 2015		
Intellectual disability	*FRY*		NM_023037.2: c.3589C>T; p.(Arg1197Ter)	.	.	.	.	.	.	CM118305	DM

Table 5.4: Frameshift variants classified as pathogenic or likely pathogenic in ClinVar or with a disease-causing mutation (DM) in HGMD® reported to cause an AR disorder/phenotype. Variants previously reported in the Amish are highlighted in blue.

Phenotype/ Disorder	Gene	MOI	HGVS Nomenclature	dbSNP	gnomAD		ClinVar			HGMD	
					Freq. (March 2019)	No. of homs	Clinical sig.	Review Status (March 2018)	Last Reviewed	Accession No.	Variant Class
Mast syndrome	*SPG21*	AR	NM_016630.6: c.601dupA; p.(Thr201Asnfs)	rs387906275	0.0001147	0	Pathogenic	1	Mar 5, 2018	CI033303	DM
Familial partial lipodystrophy	*LIPE*	AR	NM_005357.2: c.3203_3221del; p.(Val1068Glyfs*102)	rs587777699	.	.	Pathogenic/ Likely pathogenic	1	Jun 2, 2016	CD146280	DM
Epidermolysis bullosa, Herlitz	*LAMC2*	AR	NM_005562.2: c.2006_2012del; p.(Ile669Lysfs)	rs778012079	.	.	Likely pathogenic	1	Aug 26, 2016	CD068382	DM
Buphthalmos	*CYP1B1*	AR	NM_000104.3: c.1064_1076del; p.(Arg355Hisfs*69)	rs72549380	0.0004189	0	Pathogenic	2	Apr 30, 2018	CM014174	DM
Nijmegen breakage syndrome-like disorder	*RAD50*	.	NM_005732.3: c.2156dupT; p.(Glu723Glyfs*5)	rs397507178	0.0004818	0	Pathogenic	2	Aug 1, 2018	.	.

Variants not previously described as associated with human genetic disorders

There were 94 genes containing 95 nonsense variants and 170 genes containing 202 frameshift variants within the aggregated exome data that had not previously been reported in either ClinVar or HGMD. To assess the potential pathogenicity of these variants, the genes in which these variants were located were investigated via a PubMed literature review. This literature review involved looking for any association with a human disease including model organisms or cell line studies which proposed an associated link with a human disease phenotype.

This analysis indicated 61 genes, housing nonsense variants, and 122 genes, housing frameshift variants, within our cohort, where a putative link to a human disease phenotype had been proposed in the literature (**Appendix I**). Out of these 61 genes only nine genes, containing a nonsense variant in our cohort, were putatively linked to an autosomal recessive disorder reported in humans. However, some of these genes have been linked to more than one phenotype/disorder in the literature, such as *TIMP4*, which has been associated with both focal epilepsy and Kawasaki disease (Table 5.5). Similarly, ten genes out of the 122 identified in our cohort as containing a frameshift variant were linked to 11 autosomal recessive disorders (Table 5.6).

Table 5.5: Candidate novel heterozygous nonsense variants identified in genes previously associated with an autosomal recessive disorder in humans identified in our Amish aggregated exome dataset.

Phenotype	MOI	*Gene*	Identified in Amish Exome Cohort		Variants in Literature		First Author, Year
			Chromosomal variant (hg38)	cNomen	Chromosomal variant (hg38)	cNomen	
Carboxylesterase 1 deficiency	AR	*CES1*	NC_000016.10: g.55819579G>A	NM_001025194.1: c.859C>T; p.Arg288Ter	NC_000016.10: g.55823658C>T	NM_001025194.1: c.428G>A;	Zhu,2008
Coenzyme Q10 deficiency, primary, 8	AR	*COQ7*	NC_000016.10: g.19072005C>T	NM_016138.4: c.151C>T; p.Arg51Ter	NC_000016.10: g.19075775T>A	NM_016138.4:c.422T>A; p.Val141Glu	Freyer, 2015
Glutamate pyruvate transaminase polymorphism	AR	*GPT*	NC_000008.11: g.144506253T>G	NM_005309.2: c.978T>G; p.Tyr326Ter	NC_000008.10: g.145729727C>G	NM_005309.2:c.40C>G; p.His14Asp	Sohocki, 1997
Thyroid hormone metabolism, abnormal	AR	*SECISBP2*	NC_000009.12: g.89328674C>T	NM_024077.4: c.589C>T; p.Arg197Ter	NC_000009.12: g.89348095G>A	NM_024077.4:c.1619G>A; p.Arg540Gln	Dumitrescu, 2011 Dumitrescu, 2005
					NC_000009.12: g.89341356A>T	NM_024077.4:c.1312A>T; p.Lys438Ter	
					NC_000009.12: g.89338609G>A	NM_024077.4: c.1212+29G>A	
Focal epilepsy	AR	*TIMP4*	NC_000003.12: g.12153599_12153602del	NM_003256.3: c.588_591del; p.Cys197Ter	NC_000003.12: g.12159406C>T	NM_003256.3: c.-566G>A	Haerian, 2015

Kawasaki disease	AR	*TIMP4*	NC_000003.12: g.12153599_12153602del	NM_003256.3: c.588_591del; p.Cys197Ter	NC_000003.12: g.12159406C>T	NM_003256.3: c.-566G>A	Ban, 2009
Autosomal recessive hearing loss	AR	*TMPRSS5*	NC_000011.10: g.113700030C>T	NM_001288751.1: c.11G>A; p.Trp4Ter	NC_000011.10: g.113690928C>A NC_000011.10: g.113689810G>T NC_000011.10: g.113690928C>T	NM_030770.3:c.976G>T; p.Ala326Ser NM_030770.3:c.1314C>A; p.Tyr438Ter NM_001288751.1:c.949G>T; p.Ala317Ser	Guipponi, 2008
CAKUT) and VACTERL	AR	*TRAP1*	NC_000016.10: g.3658137G>A	NM_016292.2: c.2107C>T; p.Arg703Ter	NC_000016.10: g.3664437C>T	NM_016292.2: c.1406G>A; p.Arg469His	Skinner, 2014 Westland, 2014
Familial Focal Segmental Glomerulosclerosis (FSGS)	AR	*TTC21B*	NC_000002.12: g.165907746G>A	NM_024753.4: c.2500C>T; pGln834Ter	NC_000002.12: g.165941111G>A	NM_024753.4:c.626C>T; p.Pro209Leu	Huynh Cong, 2014 Bullich,2017
Short-rib thoracic dysplasia 4 with or without polydactyly	AR	*TTC21B*	NC_000002.12: g.165907746G>A	NM_024753.4: c.2500C>T; pGln834Ter	NC_000002.12: g.165929290G>A NC_000002.12:g.165911404A>G	NM_024753.4:c.1231C>T; p.Arg411Ter NM_024753.4:c.2384T>C; p.Leu795Pro	Davis, 2011

Parkinson disease 23, autosomal recessive, early onset	AR	*VPS13C*	NC_000015.10: g.61929675C>A	NM_020821.2: c.6112G>T; p.Gly2038Ter	NC_000015.10: g.61915631A>C	NM_020821.2: c.8445+2T>G	Lesage, 2016
					NC_000015.10 :g.61882652C>A	NM_020821.2:c.9568G>T; p.Glu3190Ter	
					NC_000015.10: g.61958608C>G	NM_020821.2:c.4165G>C; p.Gly1389Arg	

For variants reported in more than one transcript in our data, only the canonical (Refseq) transcript has been used. For transcripts that did not have a Refseq transcript available the NCBI database was used to find the longest isoform was used as the predicted effect of the variant was the same in all transcripts. Variants found in the literature that are reported in ClinVar as pathogenic or likely pathogenic are highlighted in red.

Table 5.6: Candidate novel heterozygous frameshift variants identified in genes previously associated with an autosomal recessive disorder in humans identified in our Amish aggregated exome dataset.

Phenotype	MOI	Gene	Variant identified		Variants in Literature		First Author, Year
			Chromosomal variant (hg38)	cNomen	Chromosomal variant (hg38)	cNomen	
Reticular dysgenesis	AR	*AK2*	NC_000001.11: g.33013400_33013401insATGTC	NM_001625.3: c.500_501insGACAT; p.Ile167Metfs*8	NC_000001.11: g.33036828T>C NC_000001.11: g.33014526T>C NC_000001.11: g.33013353A>T NC_000001.11:g .33013345G>A NC_000001.11: g.33021616G>A NC_000001.11: g.33013204T>A NC_000001.11: g.33036804C>A	NM_001625.3:c.1A>G; p.Met1Val NM_001625.3:c.494A>G; p.Asp165Gly NM_001625.3:c.548T>A; p.Leu183Ter NM_013411.4:c.556C>T; NM_013411.4:c.307C>T; p.Arg103Trp NM_001625.3:c.697A>T; p.Lys233Ter NM_001625.3:c.25G>T; p.Glu9Ter	Pannicke, 2009
Nephronophthisis 15	AR	*CEP164*	NC_000011.10: g.117351942del	NM_014956.4: c.347del; p.Lys116Argfs*22	NC_000011.10: g.117338618A>C NC_000011.10: g.117351872C>T NC_000011.10: g.117381864C>T NC_000011.10: g.117387204C>T	NM_014956.4:c.32A>C; p.Gln11Pro NM_014956.4:c.277C>T; p.Arg93Trp NM_014956.4:c.1573C>T; p.Gln525Ter NM_014956.4:c.1726C>T; p.Arg576Ter	Chaki, 2012

Adenocarcinoma of lung, response to tyrosine kinase inhibitor in	AR	*EGFR*	NC_000007.14: g.55191810_55191811del	NM_005228.3: c.2561_2562del; p.Thr854Argfs*42	NC_000007.14: g.55191822T>G NC_000007.14: g.55174014G>T NC_000007.14: g.55174014G>A NC_000007.14: g.55181378C>T	NM_005228.4:c.2573T>G; p.Leu858Arg NM_005228.4:c.2155G>T; p.Gly719Cys NM_005228.4:c.2155G>A; p.Gly719Ser NM_005228.4:c.2369C>T; p.Thr790Met	Lynch, 2004 Kobayashi, 2005
Inflammatory skin and bowel disease, neonatal, 2	AR	*EGFR*	NC_000007.14: g.55191810_55191811del	NM_005228.3: c.2561_2562del; p.Thr854Argfs*43	NC_000007.14: g.55157738G>A	NM_005228.4:c.1283G>A; p.Gly428Asp	Campbell, 2014
Retinitis Pigmentosa 25	AR	*FAM46A*	NC_000006.12: g.81752072_81752073insGCCG	NM_017633.2: c.69_70insCGGC; p.Gly24Argfs*171	NC_000006.12: g.81752068C>T	NM_017633.2:c.74G>A; p.Gly25Asp	Barragan, 2007
Encephalopathy due to defective mitochondrial and peroxisomal fission 2	AR	*MFF*	NC_000002.11: g.228205096ins28	NM_001277061.1: c.518_518+1ins28; p.Trp174Profs*20	NC_000002.12: g.227330777C>T NC_000002.12: g.227355756C>T	NM_020194.5:c.190C>T; p.Gln64Ter NM_001277061.1:c.892C>T; p.Arg298Ter	Shamseldin, 2012 Koch, 2016
Xanthinuria, type II	AR	*MOCOS*	NC_000018.10: g.36205146_36205147del	NM_017947.2: c.1088_1089del; p.Leu363Profs*16	NC_000018.10: g.36213402C>T NC_000018.10: g.36195283G>C NC_000018.10: g.36260092C>T	NM_017947.3:c.1255C>T; p.Arg419Ter NM_017947.3:c.169G>C; p.Ala57Pro NM_017947.3:c.2326C>T; p.Arg776Cys	Ichida, 2001 Zhou, 2015 Yamamoto, 2003 Peretz, 2007
Iminoglycinuria, Digenic	AR	*SLC6A18*	NC_000005.10: g.1243571dup	NM_182632.2: c.1148dup; p.Leu384Profs*54	n/a	IVS1, G-A, +1	Bröer, 2008
Osteogenesis Imperfecta, Type XII	AR	*SP7*	NC_000012.12: g.53329306_53329307del	NM_001173467.2: c.135_136del; p.Lys46Alafs*7	n/a	1-BP DEL, 1052A	Lapunzina, 2010

Uncombable hair syndrome 3	AR	*TCHH*	NC_000001.11: g.152111961_152111962del	NM_007113.3: c.1255_1256del; p.Leu419Glufs*258	NC_000001.11: g.152112226G>A	NM_007113.3:c.991C>T; p.Gln331Ter	Basmanav, 2016 Wu, 2016
					NC_000001.11: g.152110849A>C	NM_007113.3:c.2368T>G; p.Leu790Val	
					NC_000001.11: g.152110849A>C	NM_007113.3:c.2368T>G; p.Leu790Val	
Night blindness, congenital stationary (complete), 1C, autosomal recessive	AR	*TRPM1*	NC_000015.10: g.31002174del	NM_001252020.1: c.4577del; p.Tyr1526Serfs*37	n/a	IVS16DS, T-C, +2	Li, 2009 Audo, 2009 van Genderen, 2009 Nakamura, 2010
					n/a	1-BP DEL, 412G	
					n/a	36.4-KB DEL, EX2-7	
					NC_000015.10: g.31028454A>T	NM_002420.5:c.3105T>A; p.Tyr1035Ter	
					NC_000015.10: g.31070213G>A	NM_002420.5:c.31C>T; p.Gln11Ter	
					NC_000015.10: g.31068010A>G	NM_002420.5:c.296T>C; p.Leu99Pro	
					NC_000015.10: g.31042140G>T	NM_002420.5:c.1832C>A; Pro611His	
					NC_000015.10: g.31042102G>A	NM_002420.5:c.1870C>T; p.Arg624Cys	
					NC_000015.10: g.31032930G>T	NM_002420.5:c.2645C>A; p.Ser882Ter	

*For variants reported in more than one transcript in our data, only the canonical (Refseq) transcript has been used. For transcripts that did not have a Refseq transcript available the NCBI database was used to find the longest isoform was used as the predicted effect of the variant was the same in all transcripts. Variants found in the literature that are reported in ClinVar as pathogenic or likely pathogenic are highlighted in **red**.*

Next, heterozygous missense variants identified within the aggregated Amish exome dataset were analysed. Given the large number of variants (10,767) only those previously reported as pathogenic or likely pathogenic in ClinVar were initially prioritised for further evaluation. A total of 32 missense variants were identified as pathogenic. However, upon further investigation 16 had been reclassified in ClinVar (to “conflicting interpretations of pathogenicity”, “benign”, “uncertain significance” or “other”) with one variant in a gene now reported to show an X-linked mode of inheritance, so these were subsequently removed from this dataset. Table 5.7 summarises the remaining 15 missense variants, found in 14 different human disease genes, grouped by primary system affected by the disorder. Additional information in the summary table includes; associated disorders, ClinVar significance, and the frequency (in European, non-Finnish; which was used due to the Finnish population being a population isolate with a different enrichment of variants compared to the rest of Europe) and number of homozygotes (Total) reported in gnomAD.

Eight of the variants identified have previously been reported in the Amish community, indicating that there are seven additional rare (AF ~1% or less) heterozygous missense variants previously associated with an autosomal recessive disease in humans present within the Amish community that have yet to be reported to cause disease within the community. Further work is required to establish the pathogenicity of these variants.

*Table 5.7: Rare (AF ~1% or less) heterozygous missense variants identified in the aggregated Amish exome data that have previously been associated with autosomal recessive disease in humans and are reported as pathogenic or likely pathogenic in ClinVar. Grouped by primary system affected. Variants previously reported in the Amish in association with the disease are highlighted in **blue**.*

Phenotype	MOI	Gene	HGVS Nomenclature	ClinVar			GnomAD	
				Clinical significance:	Review Status (March 2019)	Last Reviewed	Feq.	No. of Homs
Audiology								
Deafness, autosomal recessive 1A	AR	*GJB2*	NM_004004.5:c.229T>C;p.(Trp77Arg)	Pathogenic	2	Oct 4, 2017	0.00002638	0
Gastroenterology								
Congenital glucose-galactose malabsorption	AR	*SLC5A1*	NM_000343.3:c.1673G>A;p.(Arg558His)	Pathogenic	1	Dec 9, 2016	0.00003099	0
Haematology/Immunology								
Myeloperoxidase deficiency	AR	*MPO*	NM_000250.1:c.995C>T;p.(Ala332Val)	Pathogenic	0	May 1, 2004	0.01791	45
Myeloperoxidase deficiency	AR	*MPO*	NM_000250.1:c.752T>C;p.(Met251Thr)	Pathogenic	0	Nov 15, 1997	0.01365	26
von Willebrand disease, recessive form	AR	*VWF*	NM_000552.3:c.2561G>A;p.(Arg854Gln)	Pathogenic	2	Nov 21, 2018	0.005343	5
Pyruvate kinase deficiency, amish type	AR	*PKLR*	NM_000298.5:c.1436G>A;p.(Arg479His)	Pathogenic	1	Apr 14, 2017	0.00009292	0
Hepatology								
Progressive intrahepatic cholestasis	AR	*ATP8B1*	NM_005603.4:c.923G>T;p.(Gly308Val)	Pathogenic	1	Dec 19, 2017	0.000008795	0
Metabolic								
Iminoglycinuria, digenic, hyperglycinuria	AR	*SLC36A2*	NM_181776.2:c.260G>T;p.(Gly87Val)	Pathogenic	0	Dec 1, 2008	0.01259	24
Glutaryl-CoA oxidase deficiency	AR	*SUGCT*	NM_001193311.1:c.1006C>T;p.(Arg336Trp)	Pathogenic	2	Oct 31, 2018	0.008083	9
Neurology								
Ataxia, spastic, 4	AR	*MTPAP*	NM_018109.3:c.1432A>G;p.(Asn478Asp)	Pathogenic	0	Dec 1, 2014	-	-
Psychomotor retardation, epilepsy, and craniofacial dysmorphism	AR	*SNIP1*	NM_024700.3:c.1097A>G;p.(Glu366Gly)	Pathogenic	0	Jan 1, 2012	-	-
Cohen syndrome	AR	*VPS13B*	NM_017890.4:c.8459T>C;p.(Ile2820Thr)	Pathogenic	0	Jul 21, 2016	-	-
Trichothiodystrophy, nonphotosensitive 1	AR	*MPLKIP*	NM_138701.3:c.430A>G;p.(Met144Val)	Pathogenic	0	Mar 1, 2005	0.00005277	0
Renal								
Familial renal glucosuria	AR	*SLC5A2*	NM_003041.3:c.1961A>G;p.(Asn654Ser)	Pathogenic	0	Feb 1, 2004	0.007871	10
Respiratory								
BCHE, dibucaine-resistant I, postanesthetic apnea	n/a	*BCHE*	NM_000055.2:c.293A>G;p.(Asp98Gly)	Pathogenic/ Likely pathogenic	2	Oct 31, 2018	0.01766	36

5.3.2 Determining the AF of pathogenic variants seen within the various Amish communities

In order to learn more about the prevalence of 165 pathogenic variants known to be associated with the disease and present in the Amish communities (**Appendix A**), and as part of wider strategy to develop new more streamlined genotyping approaches for disease diagnosis in the community, 171 distantly related anonymised Amish individuals from Ohio (Holmes County), Ohio (Geauga County), Indiana, and Wisconsin communities were genotyped using a multiplexed amplicon, PLEXseq [102] sequencing approach. As expected, initial findings showed remarkably divergent allele frequencies for these variants reflecting the distinct ancestral histories of each Amish community.

Each of the 171 samples were run in triplicate on the PLEXseq panel to enable validation. For a sample to be included in the allele frequency calculation for a given variant, at least two out of the three repeats needed to be concordant. Samples that did not meet this criteria, due to failed or conflicting genotypes, were not counted.

The two most commonly occurring variants in our dataset are responsible for hereditary hemochromatosis NM_000410.3 (HFE):c.187C>G and NM_000410.3 (HFE):c.845G>A (Table 5.8), and were observed at an AF of 0.1579 and 0.1140 respectively. This is corroborated by the high AF observed within our cohort and the AF of 0.1443 for European (non-Finnish) reported in gnomAD.

Table 5.8: Allele frequency analysis of the most commonly observed variants within different Amish communities.

Disorder	Gene	MOI	Variant	Region	AF	gnomAD Freq. (March 2019)
Hereditary hemochromatosis	*HFE*	AR	c.187C>G; p.His63Asp (NM_000410.3) chr6:g.26090951C>G	Indiana Ohio Holmes Ohio Geauga Wisconsin **TOTAL**	0.1739 0.1176 0.1818 0.2000 **0.1579**	0.1443
Hereditary hemochromatosis	*HFE*	AR	c.845G>A; p.Cys282Tyr (NM_000410.3) chr6:g.26092913G>A	Indiana Ohio Holmes Ohio Geauga Wisconsin **TOTAL**	0.0652 0.1397 0.0909 0.1400 **0.1140**	0.0576

gnomAD frequencies refers to the European (non-Finnish) frequency. MOI; Mode of inheritance.

Unlike the *HFE* variants that are observed in all regions investigated, a number of variants were only observed in one region within our cohort. Variants seen in only one Amish community, but reported in more than one person are summarised in Table 5.9.

Table 5.9: Allele frequency analysis of heterozygous variants only observed within one Amish community within our cohort.

Disorder	Gene	MOI	Variant	Region	AF	gnomAD Freq. (March 2019)
Bardet-Biedl syndrome 1	*BBS1*	AR	c.1169T>G; p.Met390Arg (NM_024649.4) chr11:g.66526181T>G	Indiana Ohio Holmes Ohio Geauga Wisconsin **TOTAL**	- - - 0.0600 **0.0088**	0.002773
Glycogen storage disease 1a	*G6PC*	AR	c.1039C>T; p.Gln347Ter (NM_000151.3) chr17:g.42911391C>T	Indiana Ohio Holmes Ohio Geauga Wisconsin **TOTAL**	- 0.0221 - - **0.0088**	0.000381
Galactosemia	*GALT*	AR	c.563A>G; p.Gln188Arg (NM_000155.3) chr9:g.34648170A>G	Indiana Ohio Holmes Ohio Geauga Wisconsin **TOTAL**	- - 0.0185 - **0.0059**	0.002663
Glutaric aciduria, type 1	*GCDH*	AR	c.1262C>T;p.Ala421Val (NM_000159.30 chr19:g.12899486C>T	Indiana Ohio Holmes Ohio Geauga Wisconsin **TOTAL**	**-** 0.0294 **-** **-** **0.0113**	0.0002867
Non-syndromic intellectual disability, autism, and gait disturbance	*HERC2*	AR	c.1781C>T; p.Pro594Leu (NM_004667.5) chr15:g.28265707G>A	Indiana Ohio Holmes Ohio Geauga Wisconsin **TOTAL**	- 0.0368 - - **0.0146**	0.00001548
McKusick Kaufman syndrome	*MKKS*	AR	c.724G>T;p.Ala242Ser (NM_018848.3) chr20:g.10412791C>A	Indiana Ohio Holmes Ohio Geauga Wisconsin **TOTAL**	- - 0.0182 - **0.0058**	0.009546
McKusick Kaufman syndrome	*MKKS*	AR	c.250C>T; p.His84Tyr (NM_018848.3) chr20:g.10413265G>A	Indiana Ohio Holmes Ohio Geauga Wisconsin **TOTAL**	- - 0.0182 - **0.0058**	0.000008801

Ataxia-telangiectasia-like disorder 2	*PCNA*	AR	c.683G>T; p.Ser228Ile (NM_002592.2) chr20:g.5115472C>A	Indiana Ohio Holmes Ohio Geauga Wisconsin **TOTAL**	- 0.0147 - - **0.0058**	0.00001759
Phenylketonuria	*PAH*	AR	c.284_286del; p.Ile95del (NM_000277.1) chr12:g.102894801_102894803del	Indiana Ohio Holmes Ohio Geauga Wisconsin **TOTAL**	**-** 0.0179 **-** **-** **0.0056**	0.00005437
Pyruvate kinase deficiency	*PKLR*		c.1436G>A; p.Arg479His (NM_181871.3) chr1:g.155293177C>T	Indiana Ohio Holmes Ohio Geauga Wisconsin **TOTAL**	- - 0.0181 - **0.0058**	0.0001129
Limb-girdle muscular dystrophy	*SGCB*	AR	c.452C>G;p.Thr151Arg (NM_000232.4) chr4:g.52028899G>C	Indiana Ohio Holmes Ohio Geauga Wisconsin **TOTAL**	0.0435 0 0 0 **0.0058**	0.00006498
Limb-girdle muscular dystrophy	*SGCB*	AR	c.271C>T; p.Arg91Cys (NM_000232.4) chr4:g.52029836G>A	Indiana Ohio Holmes Ohio Geauga Wisconsin **TOTAL**	- 0.0147 - - **0.0058**	0.00009799
Crigler-Najjar syndrome	*UGT1A1*	AR	c.222C>A;p.Tyr74Ter (NM_000463.2) chr2:g.233760509C>A	Indiana Ohio Holmes Ohio Geauga Wisconsin **TOTAL**	- - 0.0182 - **0.0058**	0.000008793

gnomAD frequencies refers to the European (non-Finnish) frequency. MOI; Mode of inheritance. AR; Autosomal recessive.

5.3.3 Community allele frequency data confirming rare disease

Interrogation of our aggregated Amish exome dataset, which has now been expanded from an initial 26 exomes to the current 117, has allowed the AFs of variants within the community to be determined. Knowledge of the AFs within the community of variants in genes not yet associated with human disease or where the evidence for disease association is not yet conclusive, is hugely advantageous as it has the capacity to confirm, or refute the disease gene association. This study has been able to conclusively confirm the pathogenicity of three candidate pathogenic variants within the Amish community and consolidate three genes as a cause of human disease (Table 5.10).

Table 5.10: Allele frequency data for variants in putative disease genes in different Amish settlements

Disorder/ Phenotype	Gene	Variant [GRCh38)	PLEXseq Data				Aggregated Amish exome data			gnomAD Freq. (March 2019)
			Region	AF	No.of Hets	No. of Homs	AF (No. of exomes)	No.of Hets	No. of Homs	
Hydranencephaly with renal aplasia-dysplasia	CEP55	c.514dup; p.Ile172Asnfs (NM_001127182) chr 10:g.93507042dup	Indiana	-	0	0	0.0087 (115)	1	0	-
			Ohio Holmes	0.0221	3	0				
			Ohio Geauga	0.0273	3	0				
			Wisconsin	0.0200	1	0				
			Total	0.0205	7	0				
Situs inversus (SI) and male infertility	MNS1	c.407_410del;p.Glu136Glyfs*16 (NM_018365.2) chr15:g.56446887_56446890Del	Indiana	n/a	n/a	n/a	0.0517 (114)	7	0	-
			Ohio Holmes	n/a	n/a	n/a				
			Ohio Geauga	n/a	n/a	n/a				
			Wisconsin	n/a	n/a	n/a				
			Total	n/a	n/a	n/a				
Psychomotor retardation, epilepsy, and craniofacial dysmorphism	SNIP1	c.1097A>G; p.Glu366Gly (NM_024700.3) chr1:g.37537842T>C	Indiana	0.0870	4	0	0.0571 (116)	6	0	-
			Ohio Holmes	0.0441	6	0				
			Ohio Geauga	0.0182	2	0				
			Wisconsin	-	0	0				
			Total	**0.0351**	**12**	**0**				

5.4 Discussion

The genetic data presented here forms part of a proof-of-principle study which aims to characterise the spectrum and frequencies of inherited diseases in the Amish community to aid the identification of novel disease genes, disease diagnosis, clinical management and genetic counselling. This involved examining aggregated exome sequencing data from 26 Amish individuals to identify potentially deleterious heterozygous variants coincidentally carried by individuals from the Amish community and utilising a multiplexed amplicon sequencing approach to study the prevalence of 165 pathogenic variants by calculating the allele frequencies in a cohort of 171 Amish individuals.

5.4.1 Analysis of Amish aggregated exome data

Analysis of exome sequencing data revealed the presence of 12 variants, reported to be pathogenic in ClinVar or disease-causing in HGMD, coincidentally carried and not thought to be responsible for the difficulties experienced by the individuals within the cohort. Seven of these variants are already well established as pathogenic within the Amish community confirming that this method is capable of accurately identifying causative variants. Interestingly this analysis also identified seven variants reported to be pathogenic in ClinVar/HGMDPro that have not yet been observed within the community, but are clearly present. This provides an opportunity for health care providers in the community to more readily link observed phenotypes with genetic disorders and the responsible genotypes. For example, although not yet described as pathogenic in ClinVar or HGMD a frameshift variant in *GJC3* (NM_181538.2:c.329dup; p.Glu111Glyfs*67) was identified in our exome sequencing data (**Appendix I**). A literature review of this gene discovered heterozygous missense mutation (c.807A>T; p.Glu269Asp) of

GJC3, which encodes the gap junction protein connexin 29, is reported to cause NSHL [158, 312, 313] . It would therefore be pertinent for current and new patients recruited to the Amish Hearing Loss Programme to be screened for the frameshift *GJC3* variant identified in our cohort.

One of the most promising applications of this type of study is its potential to identify variants that will enable early intervention and deliver improved outcomes for affected individuals. For example, individuals homozygous for the identified *CYP1B1* variant (NM_000104.3:c.1064_1076del; p.Arg355Hisfs*69) are likely to experience autosomal recessive buphthalmos as a result of congenital (infantile) glaucoma. This rare variant has an AF of 0.0003648 in gnomAD (European, non-Finnish), with no reported homozygotes, supporting pathogenicity. Buphthalmos is congenital enlargement of the eye, which requires early surgical treatment to ensure the preservation of existing vision [314].

Another variant of interest is a nonsense variant (NM_016138.4:c.422T>A; p.Val141Glu) identified in *COQ7* which is reported in ClinVar as pathogenic and responsible for a primary coenzyme Q10 (CoQ10) deficiency disorder. This autosomal recessive multisystem disorder presents with neurologic manifestations, including fatal neonatal encephalopathy with hypotonia, a late-onset slowly progressive multiple-system atrophy-like phenotype and may include dystonia, spasticity, seizures, and intellectual disability. A diagnosis of this deficiency disorder can be established through a genetic test confirming the presence of biallelic pathogenic variants. Early diagnosis provides the opportunity to implement early treatment, in the form of high, dose oral CoQ10. Dietary supplementation is reported to limit disease progression including the progression of renal disease and onset of a neurological phenotype [315]. This

disorder has not yet been reported within the Amish community, but the presence of heterozygotes for this variant within our exome dataset indicates that it is present and should be considered as a potential diagnosis in individuals presenting with a potentially consistent phenotype.

This approach for identifying coincidentally carried variants has the potential to identify other variants underlying treatable disorders that which could then be included in an Amish targeted genetic newborn screening (NBS) programme.

The importance and successful application of NBS is best demonstrated through its use to screen for phenylketonuria, an error of amino acid metabolism, characterised by mutations of the phenylalanine hydroxylase (*PAH*) gene, which if not treated it can result in in profound and irreversible mental disability. However, early detection, through NBS, and implementation of a phenylalanine-restricted diet soon after birth can stop levels of phenylalanine becoming raised in the blood preventing the neuropsychological deficits [5].

Since its introduction in the US in 1963 [316], to screen for phenylketonuria, NBS has helped diagnose millions (1 in every 320 new-borns) of potentially severe or lethal conditions before clinical symptoms are observed. Each year in the US, 99.9% of the ~4 million infants born are screened [317]. However, within Amish, and Mennonite, families the proportional of all children in a family receiving NBS has been reported to be as low as 40% despite the majority of families recognising its importance [318]. The successful generation and implementation of an Amish (or Anabaptist) specific NBS would require support from the Amish communities and collaboration between the local health care providers, to ensure its accessibility, and clinical research partnerships, such as WoH and CSC, to ensure the inclusion of variants with the greatest clinical utility.

This pilot, proof-of-principle study demonstrates the clinical value of examining exome sequencing data from a community such as the Amish to identify coincidentally carried variants and its capacity to detect known and novel pathogenic variants that have yet to be reported but are clearly present within the community. Information gained from this study can be used to inform clinicians about the nature and spectrum of disorders present within the different Anabaptist communities, accelerating genetic diagnosis and permitting the development of health policies and the implementation of early targeted treatments for affected individuals within these communities.

5.4.2 Determining allele frequencies

Allele frequencies of 165 variants were investigated in 171 unaffected individuals from different Amish demes. The findings of this study revealed different carrier frequencies for most variants in the different communities. This is likely to reflect the specific migration events that occurred within the population since the genetic bottleneck event in the 18th and 19th century, when individuals fled Europe heading for the US, to the subsequent relocations of affiliations as the Amish population expanded. Knowledge of carrier frequencies within different regions can be used to estimate disease prevalence and improve the clinical management of disorders by increasing awareness of the disorder amongst clinicians and the local Amish community, tailoring diagnostic services, facilitating the planning and dissemination of healthcare resources and implementing effective treatment strategies.

This study as has also been beneficial in highlighting the co-occurrence of more than one homozygous pathogenic variant within the same individual/family and how it can impede a clinical diagnosis. A retrospective analysis of 2076 un-related patients with a molecular diagnosis conducted in 2017 found that 4.9% of individuals had received a dual diagnosis, where two or more disease loci are responsible for their clinical characteristics [319].

A notable example of this issue, from the literature, is Fitzsimmons syndrome which was described in 1987 by Fitzsimmons and Guilbert on diagnosing twins presenting with progressive spastic paraplegia, brachydactyly with cone shaped epiphyses, short stature, dysarthria, and “low-normal” intelligence. Exome sequencing conducted in 2009, comparing one of the twins with the only other reported case of this syndrome in the literature found no single genetic cause shared by the affected individuals. Instead multiple genetic causes were identified. The twins were found to have heterozygous mutation of the *SACS* gene, a known cause of AR spastic ataxia of Charlevoix Saguenay in addition to heterozygous mutation in *TRPS1*, a gene known to cause Trichorhinophalangeal syndrome type 1 (TRPS1 type 1) which is responsible for the brachydactyly feature. The singleton was found to have a mutation in the *TBL1XR1* gene, believed to be the cause of their cognitive impairment and autistic features but no underlying genetic cause was found to be the cause of their spasticity or brachydactyly [320].

Due to the unique genetic architecture of the Amish community, resulting from the limited number of founder individuals, it can be assumed that the proportion of individuals possessing multiple disease loci within this endogamous population

will be higher than reported in unrelated patients due to the enrichment of certain variants within the population.

This phenomena was detected in a family with two siblings each with features suggestive of a syndromic disorder, initially believed to be unrelated. The male sibling, presented with intellectual disability and dysmorphic facial features and was found to have a pathogenic variant in *MTPAP* (NM_018109.3:c.1432A>G; p.Asn478Asp) known to cause spastic ataxia [321]. He has subsequently developed spasticity and ataxia consistent with this disorder. The female sibling who had entirely normal development, presented with pectus excavatum, single palmar creases and aortic stenosis, was found to be homozygous for a variant in *HYAL2* (NM_003773.4:c.443A>G) previously linked to cor triatriatum sinister (and orofacial clefting) [184] confirming the presence of two distinct genetic disorders within the family. After discovering this second pathogenic variant within the family the male sibling, was tested for the *HYAL2* variant and also found to be homozygous, explaining his dysmorphic facial features. The father of these children complained of palpitations and dizzy spells and was subsequently diagnosed with hypertrophic cardiomyopathy (HCM), and was found to be heterozygous for the Amish founder variant in *MYBPC3* (NM_000256.3:c.3330+2T>G) associated with HCM. This finding confirmed the presence of three distinct genetic disease loci/disorders present within this single family.

5.4.3 Community allele frequency data confirming rare disease genes

The work outlined in this chapter, and subsequent expansion of this dataset, has permitted the rapid interrogation of aggregated Amish exomes allowing the AFs of variants within the community to be determined. The impact of these findings

is increased by aligning the exome data with metadata, providing information on which community and church group an individual originates, enabling the AFs to be evaluated in the context of the different regions.

From a clinical perspective, this information is extremely beneficial as it is assisting local clinicians in confirming disorders which are already recognised in the different communities, but even more importantly, identifying disorders which have not previously been reported within the Amish community and may be being missed by the clinics. From an academic perspective, it has the proven capacity to confirm, or refute, the pathogenicity of novel, putative pathogenic variants within a community.

This study has conclusively confirmed the pathogenicity of three novel pathogenic variants within the Amish community and consolidated three putative human disease genes, for which the evidence of association with the reported disease phenotype was previously limited.

Firstly, the pathogenicity of a homozygous founder frameshift variant in *CEP55* (NM_018131.4: c.514dup; p.Ile172Asnfs*17) as the cause of hydranencephaly and renal dysplasia, present in two siblings with a lethal foetal disorder [322] was confirmed (**Appendix J**).

Two recent studies of single families, one [323] the other of Canadian Mennonite ancestry [324] reported loss of function mutations in *CEP55* and corresponding phenotypes, with foetuses presenting with Meckel-like syndrome and MARCH (multinucleated neurones, anhydramnios, renal dysplasia, cerebellar hypoplasia and hydranencephaly) syndrome. This, alongside the increased frequency of this variant in the Amish community, has enabled us to learn more about the clinical features of this disorder, and has corroborated mutation of *CEP55* as a cause of hydanencephaly and renal dysplasia.

A further example is the confirmation of the missense *SNIP1* variant (NM_024700.3:c.1097A>G), discussed in chapter 4, as the underlying molecular cause of psychomotor retardation, epilepsy, and craniofacial dysmorphism (PMRED). This variant has a high allele frequency, ~6%, in our aggregated exome data, being particularly enriched in the communities of Indiana, but is yet to be reported in gnomAD. The high prevalence of this variant has been instrumental in being able to precisely define the clinical phenotype and confirm pathogenicity of the variant and disease gene association.

Finally, the pathogenicity of the *MNS1* variant, NM_018365.2: c.407_410del;p.Glu136Glyfs*16, in the Amish, and mutation of *MNS1* as a cause of laterality defects (situs inversus, SI) and male infertility in humans has been confirmed by our studies (**Appendix K**). Due to the nature of this condition the presence or absence of SI in homozygous individuals is a randomised event, attributed to the inability of dysfunctional embryonic nodal cilia to perform normal rotation [325]. This means that a large number of individuals need to be genotyped and phenotyped to clearly demonstrate a causative link between homozygosity for a particular variant and SI. This is made possible within the Amish due to the large family sizes that are typically seen within this founder community. The AF of the *MNS1* founder variant within the Amish, determined through interrogation of aggregated Amish exomes, was also instrumental in collating sufficient genetic evidence to support the pathogenicity of the *MNS1* variant. This data, taken together with the previous report of an SI and male infertility phenotype observed in *Msn1* knock-out mice [326] confirmed the association of *Msn1* mutation with SI and male infertility in humans.

One example of the utility of this type of aggregated exome data for refuting the potential pathogenicity of a variant and enabling the exclusion of a potential candidate genetic cause of disease is illustrated by the nonsense *SLC15A5* variant (NM_001170798.1:c.865G>T; p.Glu289Ter) investigated as a potential novel cause of NS-SNHL in chapter 3. *SLC15A5* was excluded as a candidate cause of hearing loss due to additional genotyping data, generated via a similar high-volume sequencing approach by our collaborators working within the Pennsylvanian Amish communities. The high frequency of the *SLC15A5* variant within the Pennsylvanian communities, and subsequent detection of homozygotes, lead to its rejection as a candidate cause of hearing loss.

This data also demonstrated the importance of considering the Amish as distinct communities and church groups with differing genetic backgrounds, as opposed to one large population. Whilst variants common among all Amish communities exist, likely originating from a shared European ancestor, the distinct migration patterns and geographical isolation of different communities since the original migration to the US, has given rise to large differences between AFs within each community. Another benefit of the Pennsylvanian dataset is the large number of individuals sequenced within a single community. The AF of the *SLC15A5* variant may not be enriched within in the Pennsylvanian community, when compared to other Amish communities, but the increased number of individuals for whom aggregated data is available increases the likelihood of a homozygotes being detected is increased.

In addition to the aggregate exome data, the PLEXseq approach allows us to rapidly genotype variants of interest within the Amish and obtain community specific AFs. The flexible nature of this platform allows for the rapid inclusion of

new candidate variants, seen within the community, which can then be scrutinised. Moving forward this approach will be used as a platform to provide newborn screening and diagnostic testing for affected individuals with the data generated also being anonymously aggregated in order for the AFs of known and putative disease associated variants to be determined. This data can then be used to determine the spectrum of disease across the different communities which can be used to help inform local health policies.

One example of where this data has already proved to be valuable for this purpose is in determining the likely presence of individuals with undiagnosed cartilage hair hypoplasia in the Wisconsin communities, due to the extremely high allele frequency of the *RMRP* founder mutation (NR_003051.3:g.70A>G) within this community [327].

Cartilage hair hypoplasia (CHH) is a rare, autosomal recessive, metaphyseal chondrodysplasia characterized by sparse hair, short stature and short limbs in combination with mild to moderately severe cellular immunodeficiency and erythropoiesis [328-330]. Individuals present with a highly variable phenotype and can been misdiagnosed clinically with achondroplasia, due to a similar physical appearance of short stature and short limbs. However, CHH affected individuals are at risk of severe immunodeficiency, which depending on its severity needs to be carefully monitored and treated. Individuals displaying severe immunodeficiency disorder (SCID) will typically require a bone marrow transplant. Whereas some individuals, displaying milder immunodeficiency, may be more susceptible to contracting infections. For these individuals contracting varicella (chickenpox) comes with a very high risk mortality, especially in undiagnosed individuals that are unlikely to receive the immediate, high-dose acyclovir treatment [331]. CHH can be detected through a new born T-cell

Receptor Excision Circles (TREC) test which measures the number of circular DNA molecules (TRECs) formed within developing T-cells. A healthy infant blood sample will have one TREC per 10 T-cells, reflecting a high rate of T-cell generation whereas a sample from an infant presenting with SCID will lack TREC completely [332].

Knowledge about the high allele frequency of the Amish *RMRP* founder mutation within the Wisconsin Amish has been helpful for healthcare providers serving this community who have worked with the community to provide community sensitive education and information about the condition and promote uptake of newborn screening.

The work outlined in this chapter highlights the value of knowledge about known and newly defined pathogenic variants within a community. This information can now be used to develop a comprehensive genetic testing platform tailored to the specific genome of the Amish community that has the potential to be expanded as new information about inherited disorders arises. This will provide an immensely powerful tool for clinicians and local health care providers to utilise to support the delivery of improved healthcare outcomes for members of the Amish community.

5.4.4 Future work and considerations

Due to the success of the proof-of-principle study identifying potentially deleterious coincidental heterozygous sequence variants the immediate next step is to broaden the parameters of the study by expanding the dataset to include exome sequencing data from more individuals. Currently at least 150 exomes will be included in the next round of analysis which will help us learn even

more about known pathogenic variants present in the community which are yet to be reported.

The information from this study will also inform variant selection for the next

Phenotype/ Disorder	Gene	MOI	HGVS Nomenclature	dbSNP	gnomAD	
					Freq. (March 2019)	No. of homs
Amish infantile epilepsy syndrome	*ST3GAL5*	AR	NM_003896.3: c.862C>T; p.(Arg288Ter)	rs104893668	0.00005425	0
Craniofacial dysmorphism, skeletal anomalies and mental retardation	*TMCO1*	AR	NM_019026.4: c.292_293del; p.(Ser98Ter)	rs752176040	0.0002768	0
Deafness and myopia	*SLITRK6*	AR	NM_032229.2: c.1240C>T; p.(Gln414Ter)	rs587777069	0.000008884	0
Glycogen storage disease type 1A	*G6PC*	AR	NM_000151.3: c.1039C>T; p.(Gln347Ter)	rs80356487	0.0003813	0
Mental retardation, autosomal recessive 41	*KPTN*	AR	NM_007059.3: c.776C>A; p.(Ser259Ter)	rs374298314	0.00007807	0
Coenzyme Q10 deficiency, primary, 8	*COQ7*	AR	NM_016138.4 c.422T>A p.(Val141Glu)	rs864321686	.	.
Intellectual disability	*FRY*		NM_023037.2: c.3589C>T; p.(Arg1197Ter)	.	.	.

stages of the PLEXseq genetic testing panel development. Allele frequencies of the known pathogenic variants not yet reported (

Table 5.3,

Table 5.4 and Table 5.7) will be determined within the different Amish demes. The information about possible underlying genetic causes of disease within the community can then be used by clinicians and local health care providers.

With regard to determining the allele frequencies of the most commonly occurring pathogenic variants seen within various Amish communities the next step for this study is to increase the number of control samples included PLEXseq genetic testing panel and, where possible, more evenly represent the various Amish

demes. In addition, the number of variants assessed can be increased with primers being modified for variants that failed to produce genotypes

Phenotype/ Disorder	Gene	MOI	HGVS Nomenclature	dbSNP	gnomAD		
					Freq. (March 2019)	No. of homs	Clinic sig.
Mast syndrome	*SPG21*	AR	NM_016630.6: c.601dupA; p.(Thr201Asnfs)	rs387906275	0.0001147	0	Pathog
Familial partial lipodystrophy	*LIPE*	AR	NM_005357.2: c.3203_3221del; p.(Val1068Glyfs*102)	rs587777699	.	.	Pathog Like pathog
Epidermolysis bullosa, Herlitz	*LAMC2*	AR	NM_005562.2: c.2006_2012del; p.(Ile669Lysfs)	rs778012079	.	.	Like pathog
Buphthalmos	*CYP1B1*	AR	NM_000104.3: c.1064_1076del; p.(Arg355Hisfs*69)	rs72549380	0.0004189	0	Pathog
Nijmegen breakage syndrome-like disorder	*RAD50*	.	NM_005732.3: c.2156dupT; p.(Glu723Glyfs*5)	rs397507178	0.0004818	0	Pathog

(NM_020435.3 (GJC2):c.203A>G and NM_152743.3 (*BRAT1*): c.638dup) in the pilot study. Increasing the cohort will enable a more accurate representation of the frequencies of variants present in the different communities.

Although beyond the scope of the current study, this approach of exome sequencing data analysis has the potential to validate the clinical significance of variants in online databases, such as ClinVar and HGMD, and variants reported in the literature as disease causing, for example by identifying healthy homozygotes for previously reported pathogenic variants in genes associated with fully penetrant congenital onset disorders.

Aggregation databases such as gnomAD are a fantastic resource for the interpretation of novel variants by providing extensive frequency data including the number of observed homozygotes. It could be beneficial for them to include

more data from population isolates If this data were included it will be important that the population of origin is clearly delineated (as with the Finnish and Ashkenazi Jewish) to avoiding skewing the frequency data of the more exogamic populations.

Analysis of genotyping data from the Amish community also affords the opportunity to gain insight into pathogenic polymorphisms by expanding current knowledge of globally common single nucleotide variants (SNVs) that whilst in isolation are benign, have the potential to cause recessive disease when occurring in association with a loss of function (LoF) variant in a complex compound heterozygous fashion (Figure 5.5).

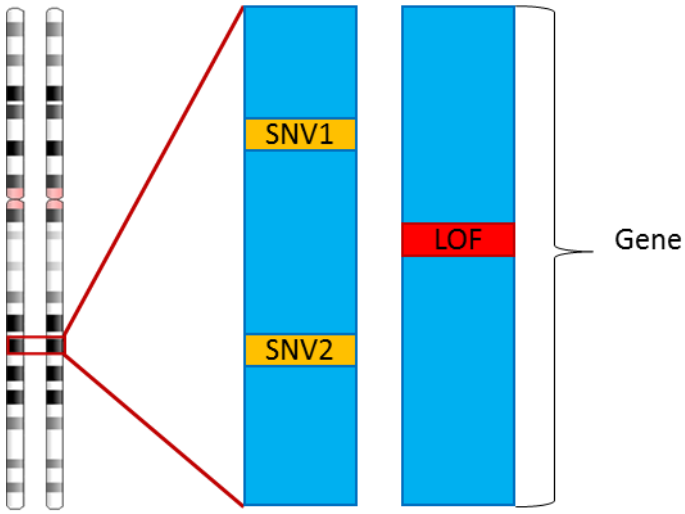


Figure 5.5: Schematic representation of a common SNVs causing recessive disease when occurring in combination with a LoF variant in a complex compound heterozygous fashion.

Finally, the work of the WoH project over the last 15years has generated a vast, and continually increasing, SNP mapping and exome sequencing dataset that could be utilised to further understand the link between autozygosity and disease

by looking at the regions in which founder mutations are located and determining a relationship to aid the identification of future novel variants.

These studies demonstrate the numerous and wide-ranging benefits of community based studies to the global scientific community and wider society. The analysis of exome sequencing data from the Amish population has potentially huge benefits in aiding variant interpretation which in turn will assist in the delivery of genetic diagnoses, genetic counselling and therapeutic treatments for individuals and families affected by rare genetic disease both within and outside the Amish community.

CAHPTER 6

Final discussion and future work

6 Final discussion and future work

The work of this thesis highlights the importance of genetic studies conducted within population isolates such as the Amish, and the huge benefits that this research offers to the patients and families involved and to clinical genetics services and scientific knowledge globally. These benefits are particularly evident when determining the molecular causes and defining the clinical phenotypes of otherwise rare genetic disorders that have, due to the present of founder mutations, become enriched within these communities.

The WoH project, to which the studies outlined here contribute, is a long-running, non-profit community genetic research program. The study aims to discover the spectrum, nature and molecular basis of inherited diseases in the Amish, and ensure information gained from these studies directly benefits its members by improving healthcare outcomes. The implementation of community-appropriate educational programmes, for both local healthcare providers and members of the community, is transforming molecular diagnostic and counselling services regionally by enabling substantial cost savings through supporting the provision of early and accurate diagnosis, and the introduction of targeted disease-specific clinical management strategies.

A longer term initiative, now being implemented in clinical-diagnostic labs locally who serve the community, is the introduction of a community specific genetic newborn (NBS) programme. The current USA NBS service entails a primarily metabolic screen for a variable number of disorders (State dependent) from a heel prick sample taken from a newborn collected on filter paper. Conversely, the targeted Amish NBS in development which can be performed utilising the same

filter paper heel prick sample permits the screening of hundreds of pathogenic founder mutations and offers a flexible cost-effective genetic testing platform which may be readily expanded to incorporate new disease-associated variants as they are discovered. This programme is made possible by the founder mutation basis of inherited diseases in the community, and knowledge of the frequency and nature of the specific mutations responsible.

One notable example of the benefits of a genetic-based approach to NBS is in the diagnosis of propionic acidemia, an AR neonatal onset metabolic disorder caused by mutation of the *PCCB* (propionyl-CoA carboxylase subunit beta) gene [333]. The resulting dysfunction of the mitochondrial enzyme propionyl CoA carboxylase leads to the accumulation of propionyl-CoA which inhibits mitochondrial metabolism, a vital process in the critic acid cycle (an essential pathway used in aerobic respiration) [334]. Affected individuals are usually asymptomatic at birth but present with poor feeding, vomiting and fatigue during the first few days of life. If left untreated, at its most severe, an otherwise healthy newborn can start to display lethargy and seizures followed by coma and eventually death [334, 335]. However, nutritional intervention in the form of a propiogenic amino acid restricted diet, which prevents propionyl-CoA accumulation and allows affected individuals to maintain as close to normal plasma concentrations of propionyl-CoA, can prevent the onset of the otherwise observed difficulties [335]. It is widely accepted that pre-symptomatic diagnosis of propionic acidemia and propiogenic amino acid restriction significantly reduces morbidity and mortality rates delivering more positive clinical and neurological outcomes [333]. Although newborn screening in the form of a metabolic assay is currently available, it has been reported that >60% of individuals are already

symptomatic at the time of diagnosis, or may not be detected by the metabolic test [333]. The introduction of an accurate and sensitive genetically-based test, stemming from knowledge of founder mutation underlying the condition, will provide a greatly facilitate timely and accurate diagnosis, enabling earlier treatment and surveillance aimed at reducing potential preventable morbidity and mortality. Additionally, the genetic-based NBS approach allows the inclusion of inherited disorders which do not necessarily involve metabolic outcomes (e.g. SNIP1-associated syndrome), so that patients and families may benefit from early intervention and treatment approaches as they are developed.

The growing knowledge of the full spectrum of inherited diseases present across all Amish communities will be instrumental in informing decisions regarding a variants' inclusion within genetic NBS. As this approach becomes established, it may be considered prudent to develop regional specific genetic NBS platforms to detect the inherited conditions particular to individual Amish communities, such as the inclusion of variants responsible for cartilage hair hypoplasia and Troyer syndrome (Ohio Amish).

One example of how regional-specific candidates can be identified is by establishing the allele frequencies (AF) of pathogenic variants known to be present in the Amish. In this study, the allele frequencies of *GJB2* variants, and other founder gene mutations linked to hearing loss were determined to learn more about their prevalence in different Amish communities. As expected, remarkably divergent AF for each gene were observed reflecting the distinct ancestral histories of each Amish community. Notably the *KCNQ1* founder mutation, responsible for Jervell and Lange-Nielsen syndrome (JLNS) an autosomal recessive syndromic form of congenital profound SN hearing loss

associated with a prolonged QT interval on ECG and potentially fatal tachyarrthymias if undiagnosed and untreated, is present in both the Holmes County and Geauga Ohio Amish communities making it a clear candidate for inclusion on a genetic NBS for these communities

A further example of a potential candidate for consideration of inclusion on a genetic NBS platform is the *SNIP1* variant, conclusively demonstrated through this study to be causative of a SNIP1-associated syndrome, given its high prevalence within the Ohio, Indiana and Wisconsin Amish communities. The high prevalence of this variant enabled the investigation of 33 individuals, presenting with this novel complex neurological disorder enabling the clinical phenotype, to be defined which has laid important foundations for a greater understanding of the underlying biological mechanisms responsible for the condition. Demonstrating the power of interrogating aggregated sequencing data to expand current knowledge of the spectrum of disease within the Amish community. The identification of the precise molecular cause of this disorder is fundamental to accurately diagnose this syndrome in future affected individuals, and in the development of prospective targeted treatments for the complications associated with this disorder such as the severe intractable seizures that afflict the majority of affected children.

Additional studies, conducted as part of the Amish hearing loss programme, aiming to investigate the nature, aetiology and frequency of genetic causes of hearing loss within the Amish community, has successfully provided genetic diagnoses to families, and allowed local healthcare providers to implement targeted diagnostic and clinical management strategies, including overseeing the monitoring and screening of additional features that may develop in syndromic

forms of the condition. The link between diagnostic delay in hearing loss, and substantially increased negative outcomes for affected individuals, has been widely reported [136-138] highlighting the importance of timely and accurate precise diagnoses particularly with regard to the monitoring of syndromic forms and implementation of protective management strategies. As knowledge of the spectrum of gene mutations within the Amish communities is expanded new variants can be readily included into the NBS platform facilitating early diagnosis and treatment. Additional benefits of such an approach include: providing the families of affected individuals with accurate information about the prognosis and progression of the condition and enabling the provision of appropriate reproductive counselling advising on recurrence rates in future offspring. In summary, the implementation of a genetic NBS within the Amish community has enormous potential to transform the diagnosis of inherited conditions and subsequent implementation of strategic care plans, including targeted treatments, for affected individuals. This will reduce the economic burden of diseases within communities by delivering substantial cost savings and improve healthcare outcomes.

A further possible utility of a genetic testing approach to screening within the Amish relates to carrier testing in adults, and the potential for detection of adult onset diseases such as a form of hypertrophic cardiomyopathy (HCM), due to an *MYBPC3* gene mutation responsible for which is at extremely high frequency across many Amish communities. HCM is a relatively common autosomal dominant disorder, displaying variable expressivity and incomplete penetrance [336]. Clinical presentations of HCM range from asymptomatic hypertrophic (enlarged), non-dilated left ventricle to progressive heart failure or sudden cardiac

death (SCD) [337, 338]. Although considered the most common cause of SCD in young people globally, the majority of affected individuals do not experience substantial symptoms and often remain undiagnosed [336, 339]. Unlike propionic acidemia, HCM is a common global disease and is reported to affect 1 in 500 people in more than 50 countries [338, 339]. In 2015, it was suggested that potentially pathogenic *MYPBC3* variants may be carried by more than 60 million people, in part due to the high occurrence of founder mutations present in different populations [336].

The expansion of a targeted population genetic testing approach to provide carrier testing for consenting adults would enable the identification of individuals at risk of developing HCM symptoms (or indeed diagnose already symptomatic individuals), so that the appropriate management and treatment options can be put in place. The WoH and the Amish communities are working alongside the Cleveland Clinic and other Cardiology departments serving the Amish to develop a clinical care programme for adults affected with HCM, due to the Amish *MYBPC3* mutation, at a significantly reduced cost. Additionally, due to the high frequency of HCM in the Amish arising from the same *MYBPC3* founder mutation, ongoing studies are expanding the knowledge of the molecular basis of the condition by defining genetic variants and modifiers which may be of global benefit to families with this condition.

Whilst, determining the carrier status of individuals can be clinically actionable, as described with *MYBPC3*, and should be shared immediately, the identification of carrier status for other conditions require no clinical intervention but will inform an individual's reproductive risk [340]. It is the recommendation of many international guidelines that carrier testing for this purpose should be postponed until an individual can be actively involved in the decision making process [340].

In these cases, although the genetic data identifying individuals as carriers would be available from a NBS there is no immediate need to share this information. The Amish community may wish to introduce a programme that permits the sharing of an individual's carrier status only when later requested, as with the Dor Yeshorim service used within the Jewish community.

Dor Yeshorim (also called Committee for Prevention of Genetic Diseases) was screening service to minimise, and eventually eradicate, fatal and debilitating recessive genetic disorders from their communities. The system involves individuals being tested, often through high school programmes [340], for a number of genetic disorders observed at significantly higher frequencies than other populations including; TaySachs disease, cystic fibrosis, Gaucher disease type I, Canavan disease, familial dysautonomia, Bloom syndrome, Fanconi anemia, glycogen storage disease type 1a, mucolipidosis type IV, and Niemann–Pick disease type A [341]. The resulting genotyping data is confidential held until requested. When two individuals wish to start a relationship they submit unique ID numbers to the service which informs the couple if any offspring are likely to have one the genetic conditions. The exact genotypes any one individual are never disclosed in an effort to remove the possibility of discrimination or stigmatisation. Although one option, the Amish communities do not necessarily need to implement a community-wide policy on the sharing of carrier status information gained from targeted community testing programmes. It may be decided that the sharing of this information is a decision for individual families or church groups to agree upon.

The value of studying genetic disorders, particularly those of a rare or recessive nature, within population isolates is now widely recognised. The work outlined

here demonstrates the capacity of community genetic studies, undertaken in the Amish communities, to elucidate the molecular cause of genetic disorders and identify novel disease genes. Findings from these studies are of major clinical relevance to both individuals within the community, families affected by rare genetic disorders worldwide and to the global clinical genetics community. The WoH is extremely privileged to work within the Amish communities with the most important outcomes of these studies remaining the translational benefits they provide to members within the community.

7 APPENDIX

7.1 Appendix A - Amish Genome Project – Plexseq variant list

Transcript Variant	Chromosomal Variant [hg38]
NM_005957.4(MTHFR):c.1129C>T	NC_000001.11:g.11794766G>A
NM_001166120.1(HSD3B2):c.35G>A	NC_000001.11:g.119415454G>A
NM_001111.4(ADAR):c.3019G>A	NC_000001.11:g.154588125C>T
NM_001005741.2(GBA):c.1226A>G	NC_000001.11:g.155235843T>C
NM_181871.3(PKLR):c.1436G>A	NC_000001.11:g.155293177C>T
NM_170707.3(LMNA):c.568C>T	NC_000001.11:g.156134457C>T
NM_001012331.1(NTRK1):c.1614+1G>A	NC_000001.11:g.156876211G>A
NM_003126.2(SPTA1):c.6154del	NC_000001.11:g.158620433del
NM_019026.4(TMCO1):c.292_293del	NC_000001.11:g.165768200_165768201del
NM_000130.4(F5):c.1601G>A	NC_000001.11:g.169549811C>T
NM_014625.2(NPHS2):c.413G>A	NC_000001.11:g.179561327C>T
NM_014053.3(FLVCR1):c.361A>G	NC_000001.11:g.212858813A>G
NM_020435.3(GJC2):c.203A>G	NC_000001.11:g.228157961A>G
NM_013411.4(AK2):c.622T>G	NC_000001.11:g.33013279A>C
NM_024700.3(SNIP1):c.1097A>G	NC_000001.11:g.37537842T>C
NM_005857.4(ZMPSTE24):c.54dup	NC_000001.11:g.40258325dup
NM_015506.2(MMACHC):c.271dup	NC_000001.11:g.45507545dup
NM_031475.2(ESPN):c.752G>A	NC_000001.11:g.6440702G>A
NM_031475.2(ESPN):c.1015C>T	NC_000001.11:g.6444505C>T
NM_000016.5(ACADM):c.199T>C	NC_000001.11:g.75732724T>C
NM_000016.5(ACADM):c.287-30A>G	NC_000001.11:g.75733498A>G
NM_000016.5(ACADM):c.985A>G	NC_000001.11:g.75761161A>G
NR_023343.1(RNU4ATAC):n.51G>A	NC_000002.12:g.121530930G>A
NM_005199.4(CHRNG):c.459dup	NC_000002.12:g.232541482dup
NM_000463.2(UGT1A1):c.222C>A	NC_000002.12:g.233760509C>A
NM_022437.2(ABCG8):c.1720G>A	NC_000002.12:g.43875377G>A
NM_000341.3(SLC3A1):c.1136+2T>C	NC_000002.12:g.44301129T>C
NM_000341.3(SLC3A1):c.1354C>T	NC_000002.12:g.44312607C>T
NM_003896.3(ST3GAL5):c.862C>T	NC_000002.12:g.85844542G>A
NM_020184.3(CNNM4):c.1813C>T	NC_000002.12:g.96799188C>T
NM_001079878.1(CNGA3):c.1126A>G	NC_000002.12:g.98396350A>G
NM_153240.4(NPHP3):c.2104C>T	NC_000003.12:g.132696798G>A
NM_000532.4(PCCB):c.1606A>G	NC_000003.12:g.136330012A>G
NM_001281724.2(BTD):c.1336G>C	NC_000003.12:g.15645186G>C
NM_001281724.2(BTD):c.1374A>C	NC_000003.12:g.15645224A>C
NM_001281724.2(BTD):c.1465T>C	NC_000003.12:g.15645315T>C
NM_000404.2(GLB1):c.902C>T	NC_000003.12:g.33051895G>A
NM_015175.2(NBEAL2):c.881C>G	NC_000003.12:g.46991644C>G
NM_002292.3(LAMB2):c.440A>G	NC_000003.12:g.49132135T>C
NM_003773.4(HYAL2):c.443A>G	NC_000003.12:g.50320047T>C
NM_001130713.2(LEF1):c.133G>C	NC_000004.12:g.108167635C>G
NM_000142.4(FGFR3):c.742C>T	NC_000004.12:g.1801837C>T
NM_001151.3(SLC25A4):c.523del	NC_000004.12:g.185145175del
NM_000128.3(F11):c.1327C>T	NC_000004.12:g.186285660C>T
NM_000232.4(SGCB):c.452C>G	NC_000004.12:g.52028899G>C
NM_000232.4(SGCB):c.271C>T	NC_000004.12:g.52029836G>A
NM_153717.2(EVC):c.1886+5G>T	NC_000004.12:g.5793722G>T
NM_001193376.1(TERT):c.1710G>C	NC_000005.10:g.1282488C>G
NM_001369.2(DNAH5):c.10815del	NC_000005.10:g.13753290del
NM_001369.2(DNAH5):c.4348C>T	NC_000005.10:g.13865675G>A
NM_002109.5(HARS):c.1361A>C	NC_000005.10:g.140674776T>G
NM_001044.4(SLC6A3):c.1409A>G	NC_000005.10:g.1409115T>C
NM_001044.4(SLC6A3):c.1408T>A	NC_000005.10:g.1409116A>T

NM_001044.4(SLC6A3):c.1269+1G>A	NC_000005.10:g.1411242C>T
NM_002185.3(IL7R):c.2T>G	NC_000005.10:g.35856979T>G
NM_022132.4(MCCC2):c.295G>C	NC_000005.10:g.71599672G>C
NM_022132.4(MCCC2):c.517dup	NC_000005.10:g.71604361dup
NM_022132.4(MCCC2):c.687A>C	NC_000005.10:g.71626702A>C
NM_003309.3(TSPYL1):c.457dup	NC_000006.12:g.116279374dup
NM_000410.3(HFE):c.187C>G	NC_000006.12:g.26090951C>G
NM_000410.3(HFE):c.845G>A	NC_000006.12:g.26092913G>A
NM_012434.4(SLC17A5):c.115C>T	NC_000006.12:g.73644583G>A
NM_001008844.1(DSP):c.699G>A	NC_000006.12:g.7562753G>A
NM_000492.3(CFTR):c.1521_1523del	NC_000007.14:g.117559592_117559594del
NM_000492.3(CFTR):c.3302T>A	NC_000007.14:g.117611743T>A
NM_000492.3(CFTR):c.3773dup	NC_000007.14:g.117642493dup
NM_000492.3(CFTR):c.1364C>A	NC_000007.14:g.117652877 C>G
NM_014141.5(CNTNAP2)c.3709del	NC_000007.14:g.148383882del
NM_152743.3(BRAT1):c.638dup	NC_000007.14:g.2543755dup
NM_138701.3(MPLKIP):c.430A>G	NC_000007.14:g.40133169T>C
NM_024728.2(SUGCT):c.895C>T	NC_000007.14:g.40459197C>T
NM_017802.3(DNAAF5):c.2384T>C	NC_000007.14:g.780097T>C
NM_004912.3(KRIT1):c.47G>C	NC_000007.14:g.92242089C>G
NM_000089.3(COL1A2):c.2098G>T	NC_000007.14:g.94420251G>T
NM_000498.3(CYP11B2):c.104_108del	NC_000008.11:g.142917733_142917737del
NM_018105.2(THAP1):c.135_139delinsGGGTTTA	NC_000008.11:g.42839314-42839318delins TAA ACCC
NM_018972.2(GDAP1):c.692C>T	NC_000008.11:g.74363051C>T
NM_017890.4(VPS13B):c.9260dup	NC_000008.11:g.99823833dup
NM_001127610.1(BAAT):c.226A>G	NC_000009.12:g.101371179T>C
NM_000113.2(TOR1A):c.907_909del	NC_000009.12:g.129814062_129814064del
NM_012144.3(DNAI1):c.48+2dup	NC_000009.12:g.34459055dup
NM_000155.3(GALT):c.563A>G	NC_000009.12:g.34648170A>G
NM_000155.3(GALT):c.940A>G	NC_000009.12:g.34649445A>G
NR_003051.3(RMRP):n.71A>G	NC_000009.12:g.35657948T>C
NM_000170.2(GLDC):c.2186del	NC_000009.12:g.6556169del
NM_004817.3(TJP2):c.143T>C	NC_000009.12:g.69216367T>C
NM_001144914.1(FGFR2):c.758C>G	NC_000010.11:g.121520160G>C
NM_018109.3(MTPAP):c.1432A>G	NC_000010.11:g.30313926T>C
NM_000124.3(ERCC6):c.2709+1G>T	NC_000010.11:g.49473476C>A
NM_000124.3(ERCC6):c.1293_1320del	NC_000010.11:g.49524110_49524137del
NM_004273.4(CHST3):c.1298C>T	NC_000010.11:g.72008329C>T
NM_001127182.1(CEP55):c.514dup	NC_000010.11:g.93507042dup
NM_000051.3(ATM):c.1564_1565del	NC_000011.10:g.108251029_108251030del
NM_000051.3(ATM):c.5932G>T	NC_000011.10:g.108312424G>T
NM_000051.3(ATM):c.6200C>A	NC_000011.10:g.108317374C>A
NM_000482.3(APOA4):c.552_749dup	NC_000011.10:g.116821309_116821506dup
NM_000360.3(TH):c.698A>G	NC_000011.10:g.2167030T>C
NM_000218.2(KCNQ1):c.451_452del	NC_000011.10:g.2527992_2527993del
NM_000448.2(RAG1):c.2974A>G	NC_000011.10:g.36576278A>G
NM_000256.3(MYBPC3):c.3330+2T>G	NC_000011.10:g.47333192A>C
NM_024649.4(BBS1):c.1169T>G	NC_000011.10:g.66526181T>G
NM_007103.3(NDUFV1):c.640G>A	NC_000011.10:g.67610510G>A
NM_006019.3(TCIRG1):c.1228G>A	NC_000011.10:g.68047495G>A
NM_002335.3(LRP5):c.1225A>G	NC_000011.10:g.68386525A>G
NM_002335.3(LRP5):c.1275G>A	NC_000011.10:g.68386575G>A
NM_000277.1(PAH):c.1315+1G>A	NC_000012.12:g.102840399C>T
NM_000277.1(PAH):c.1066-11G>A	NC_000012.12:g.102843790C>T
NM_000277.1(PAH):c.782G>A	NC_000012.12:g.102852875C>T
NM_000277.1(PAH):c.284_286del	NC_000012.12:g.102894801_102894803del
NM_000431.2(MVK):c.803T>C	NC_000012.12:g.109591275T>C
NM_000431.2(MVK):c.1174G>A	NC_000012.12:g.109596560G>A
NM_002150.2(HPD):c.1005C>G	NC_000012.12:g.121839998G>C

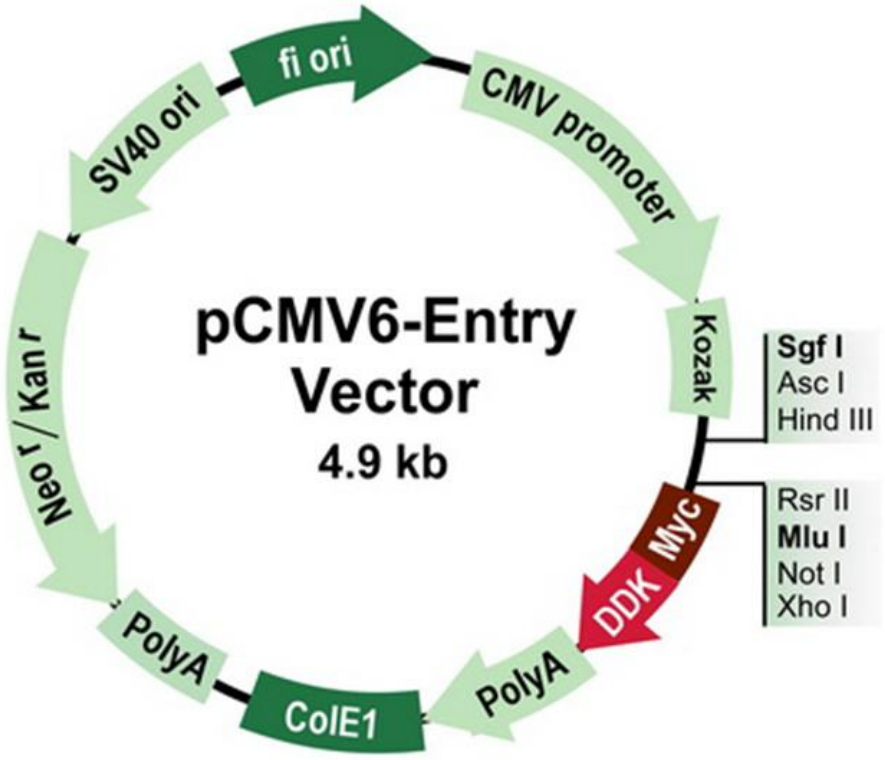
NM_002150.2(HPD):c.479A>G	NC_000012.12:g.121849726T>C
NM_002150.2(HPD):c.85G>A	NC_000012.12:g.121857765C>T
NM_001170798.1(SLC15A5):c.865G>T	NC_000012.12:g.16244690C>A
NM_000719.6(CACNA1C):c.1216G>A	NC_000012.12:g.2504944G>A
NM_001478.4(B4GALNT1):c.1514G>A	NC_000012.12:g.57626832C>T
NM_001065.3(TNFRSF1A):c.362G>A	NC_000012.12:g.6333477C>T
NM_003805.3(CRADD):c.382G>C	NC_000012.12:g.93850053G>C
NM_000282.3(PCCA):c.2017G>C	NC_000013.11:g.100515544G>C
NM_004004.5(GJB2):c.229T>C	NC_000013.11:g.20189353A>G
NM_004004.5(GJB2):c.35del	NC_000013.11:g.20189547del
NM_001142296.1(SPG20):c.1110del	NC_000013.11:g.36329416del
NM_003991.3(EDNRB):c.828G>T	NC_000013.11:g.77901181C>A
NM_032229.2(SLITRK6):c.1240C>T	NC_000013.11:g.85795269G>A
NM_001163940.1(PYGL):c.1518+1G>A	NC_000014.9:g.50913028C>T
NM_000295.4(SERPINA1):c.1096G>A	NC_000014.9:g.94378610C>T
NM_004667.5(HERC2):c.1781C>T	NC_000015.10:g.28265707G>A
NM_173087.1(CAPN3):c.2030G>A	NC_000015.10:g.42410926G>A
NM_016630.6(SPG21):c.601dup	NC_000015.10:g.64969323dup
NM_032856.3:(WDR73)c.888del	NC_000015.10:g.84643719del
NM_000339.2(SLC12A3):c.1924C>G	NC_000016.10:g.56885363C>G
NM_001270974.2(HYDIN):c.2047G>T	NC_000016.10:g.71067318C>A
NM_000151.3(G6PC):c.1039C>T	NC_000017.11:g.42911391C>T
NM_000342.3(SLC4A1):c.2422C>T	NC_000017.11:g.44251478G>A
NM_001126121.1(SLC25A19):c.530G>C	NC_000017.11:g.75278265C>G
NM_005603.4(ATP8B1):c.923G>T	NC_000018.10:g.57695188C>A
NM_000159.3(GCDH):c.1262C>T	NC_000019.10:g.12899486C>T
NM_001126335.1(SLC7A9):c.1166C>T	NC_000019.10:g.32842226G>A
NM_001126335.1(SLC7A9):c.201C>T	NC_000019.10:g.32864663G>A
NM_000285.3(PEPD):c.793C>T	NC_000019.10:g.33411697G>A
NM_004646.3(NPHS1):c.3250del	NC_000019.10:g.35831679del
NM_004646.3(NPHS1):c.1481del	NC_000019.10:g.35846154del
NM_000709.3(BCKDHA):c.1312T>A	NC_000019.10:g.41424582T>A
NM_007059.3(KPTN):c.776C>A	NC_000019.10:g.47479874G>T
NM_007059.3(KPTN):c.714_731dup	NC_000019.10:g.47479919_47479936dup
NM_001126132.2(TNNT1):c.505G>T	NC_000019.10:g.55137209C>A
NM_018848.3(MKKS):c.724G>T	NC_000020.11:g.10412791C>A
NM_018848.3(MKKS):c.250C>T	NC_000020.11:g.10413265G>A
NM_001257137.1(ITCH):c.394dup	NC_000020.11:g.34413798dup
NM_015474.3(SAMHD1):c.1411-2A>G	NC_000020.11:g.36904251T>C
NM_015474.3(SAMHD1):c.428G>A	NC_000020.11:g.36935110C>T
NM_153638.2(PANK2):c.930_936del	NC_000020.11:g.3908227_3908233del
XM_006723679.1(ADA):c.646G>A	NC_000020.11:g.44623039C>T
NM_002592.2 (PCNA):c.683G>T	NC_000020.11:g.5115472C>A
NM_130445.3(COL18A1):c.3514_3515del	NC_000021.9:g.45510082_45510083del
NM_000343.3(SLC5A1):c.1673G>A	NC_000022.11:g.32104793G>A
NM_020461.3(TUBGCP6):c.5458T>G	NC_000022.11:g.50217738A>C
NM_000133.3(F9):c.1025C>T	NC_000023.11:g.139561710C>T
NM_000397.3(CYBB):c.1222G>A	NC_000023.11:g.37805076G>A
NM_000397.3(CYBB):c.1335C>A	NC_000023.11:g.37806407C>A
NM_000531.5(OTC):c.422G>A	NC_000023.11:g.38401310G>A
NM_001145252.1(CFP):c.379T>G	NC_000023.11:g.47628126A>C
Mitochondrial	NC_012920.1:m.13513G>A
Mitochondrial	NC_012920.1:m.3243A>G

7.2 Appendix B – Primer sequences

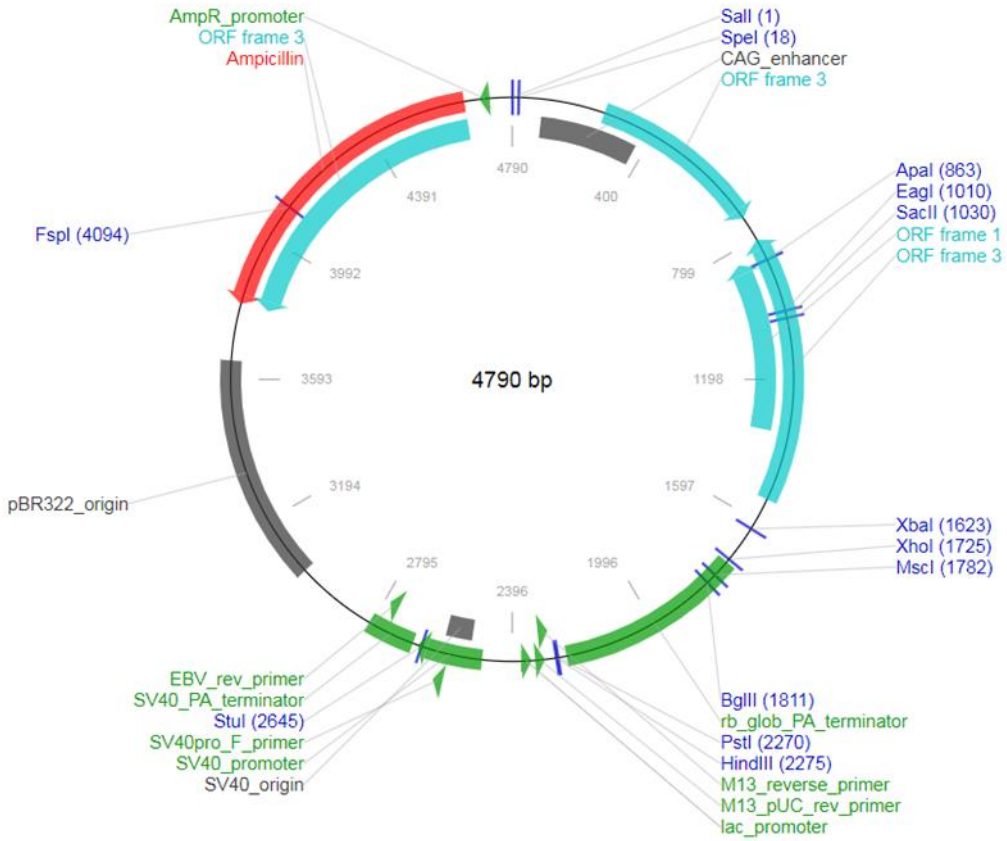
Gene (Variant)	Forward sequence	Reverse sequence
GJB2 (chr13:20189348-20189358 T/C)	GCATTCGTCTTTTCCAGAGC	CATGGAGAAGCCGTCGTACA
GJB2 (chr13:20,189,542-20,189,552 G/-)	GTGGCCTACCGGAGACATGA	CTCATCCCTCTCATGCTGTC
SNIP1 (chr1:37,537,837-37,537,847 A/G	ACAGTTGGCCGAAGAGTGAA	TCAATCAAAGACTTCCAAGAAGG

7.3 Appendix C – details of expression vector maps

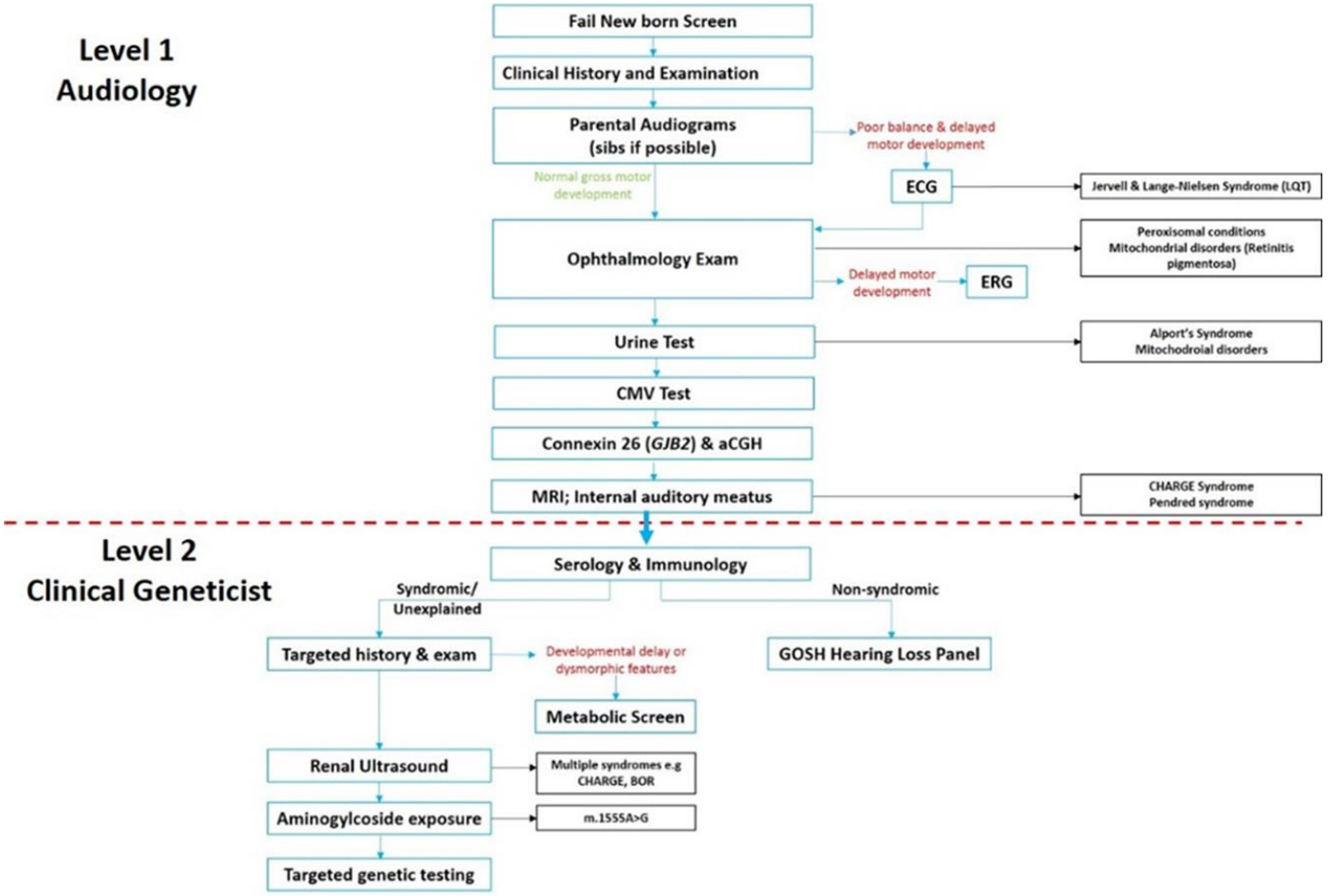
7.3.1 PCMV6-Entry, mammalian vector with C-terminal Myc- DDK Tag



7.3.2 pCAGGS vector map



7.4 Appendix D - Flow diagram summarising the BAAP guidelines for aetiological investigation into bilateral severe to profound permanent childhood hearing impairment



7.5 APPENDIX E - Most common syndromes with hearing loss as a cardinal feature (grouped by inheritance pattern)

Syndrome	Genes	MOI	Incidence	Hearing Impairment			Clinical Features
				Type	Onset	Severity	
Waardenburg Syndrome	*PAX3* *MITF* *EDNRB* *EDN3* *SOX10*	AD	1 in 40,000	WS1: Sensorineural	Congenital	Variable	Most common type of AD SHL Pigmentary abnormalities of skin, hair (white forelock 1), & eyes (heterochromia iridis) Subtype characteristics: WS1. Dystopia canthorum 2 WS2. Absence of dystopia canthorum; many other features shared w/WS1 WS3. Upper-limb abnormalities WS4. Hirschsprung disease
Branchio-Oto-Renal Syndrome	*EYA1* *EYA2* *EYA3* *SIX1* *SIX5*	AD	1 in 40,000	mixed (50%) conductive (30%) sensorineural (20%)	~25% progressive loss	35% Severe	Second most common type of AD SHL. BOR 3: Branchial cleft cysts or fistulae, malformations of external ear incl preauricular pits, & renal anomalies. BOS 3: Same features as BOR syndrome but w/out renal involvement. High penetrance; extremely variable expressivity
Neurofibromatosis Type 2	*NF2*	AD	1 in 33,000	Sensorineural	~3rd decade	Generally unilateral & gradual; can be bilateral & sudden	Hearing loss secondary to bilateral vestibular schwannomas; a rare, potentially treatable type of deafness Risk for a variety of other tumors incl meningiomas, astrocytomas, ependymomas, & meningioangiomatosis
Treacher Collins Syndrome	*TCOF1* *POLR1D* *POLR1C*	AD	1 in 50,000	Conductive			A condition that affects the development of bones and other tissues of the face including micrognathia. Other features may include cleft palate, eye abnormalities, and hearing loss.
Stickler Syndrome	*COL2A1* *COL11A1* *COL11A2*	AD	1 in 7,500 to 9,000 newborns	Conductive Sensorineural	Variable	Variable	Connective tissue disorder that can include myopia, cataract, & retinal detachment; Midfacial underdevelopment & cleft palate (either alone or as part of the Robin sequence) Mild spondyloepiphyseal dysplasia and/or precocious arthritis
Alport syndrome	COL4A3 COL4A4 COL4A5	AD	1 in 50,000 newborns	Sensorineural	Typically after age 10yrs	Varying severity Progressive	Renal, cochlear, ocular involvement In the absence of treatment, renal disease progresses from microscopic hematuria to proteinuria, progressive renal insufficiency, ESRD

Syndrome	Genes	MOI	Incidence	Hearing Impairment			Clinical Features
				Type	Onset	Severity	
Stickler Syndrome	*COL9A1* *COL9A2* *COL9A3*	AR	1 in 7,500 to 9,000 newborns	Conductive Sensorineural	Variable	Variable	Connective tissue disorder that can include myopia, cataract, & retinal detachment; Midfacial underdevelopment & cleft palate (either alone or as part of the Robin sequence) Mild spondyloepiphyseal dysplasia and/or precocious arthritis
Usher syndrome type I	MYO7A USH1C CDH23 PCDH15 USH1G CIB2	AR		Sensorineural	Congenital	Severe to profound	Abnormal vestibular function affected persons find traditional amplification ineffective & usually communicate manually because of the vestibular deficit, developmental motor milestones for sitting & walking always reached at later-than-normal ages
Usher syndrome type II	ADGRV1 WHRN USH2A	AR		Sensorineural	Congenital	Mild to severe	Normal vestibular function Hearing aids provide effective amplification for these persons; communication usually oral
Usher syndrome type III	CLRN1 HARS	AR		Sensorineural	Congenital	Progressive	Progressive deterioration of vestibular function
Pendred syndrome	SLC26A4	AR	No data*[1]	Sensorineural	Congenital	Usually (though not invariably) severe to profound	2nd most common type of AR SHL Hearing & euthyroid goiter Deafness associated w/an abnormality of the bony labyrinth (Mondini dysplasia or dilated [enlarged] vestibular aqueduct) Goiter not present at birth, develops in early puberty (40%) or adulthood (60%)
Jervell and Lange-Nielsen Syndrome	KCNQ1 KCNE1	AR	1.6 to 6 per 1 million people worldwide	Sensorineural	Congenital	Profound	3rd most common type of AR HL Deafness & prolongation of the QT interval as detected by ECG (abnormal QTc [c=corrected] >440 msec) Syncopal episodes; sudden death
Biotinidase deficiency	BTD	AR	1 in 60,000 newborns	Sensorineural	Variable	Variable	If not recognized & corrected by daily addition of biotin to the diet, affected persons develop neurologic features (e.g., seizures, hypertonia, developmental delay, ataxia) & visual problems Some degree of HL is present in ≥75% of children who become symptomatic Cutaneous features (e.g., skin rash, alopecia, conjunctivitis)
Refsum disease	PHYH PEX7	AR	No data*[2]	Sensorineural	Variable	Severe Progressive	Anosmia & early-onset retinitis pigmentosa – both universal findings w/variable combinations of neuropathy, deafness, ataxia, & ichthyosis
Alport syndrome	COL4A3 COL4A4 COL4A5	AR	1 in 50,000 newborns	Sensorineural	Typically after age 10yrs	Varying severity Progressive	Renal, cochlear, ocular involvement In the absence of treatment, renal disease progresses from microscopic hematuria to proteinuria, progressive renal insufficiency, ESRD

Syndrome	Genes	MOI	Incidence	Hearing Impairment			Clinical Features
				Type	Onset	Severity	
Alport syndrome	COL4A3 COL4A4 COL4A5	X-Linked	1 in 50,000 newborns	Sensorineural	Typically after age 10yrs	Varying severity Progressive	Renal, cochlear, ocular involvement In the absence of treatment, renal disease progresses from microscopic hematuria to proteinuria, progressive renal insufficiency, ESRD
Deafness-dystonia-optic neuronopathy syndrome (Mohr-Tranebjaerg syndrome)	TIMM8A	X-Linked	No data*[3]	Sensorineural	Early childhood	Progressive; pre- or postlingual	Visual disability, dystonia, fractures, intellectual disability

7.6 APPENDIX F - Hearing loss in the Amish

7.6.1 Causes of SNHL in the Amish Community

PCNA

The proliferating cell nuclear antigen (*PCNA*) gene is located on chromosome 20p12.3 and contains 6 exons [342]. PCNA is essential for DNA replication and repair [343], playing a central role at the DNA replication fork by recruiting necessary enzymes [344], and in accordance with such a fundamental cellular role is notably highly conserved through evolution. Our group identified a hypomorphic missense amino acid variant in the *PCNA* gene as a cause of sensorineural syndromic hearing loss (SHL) in the Ohio Amish population [177]. Hearing loss occurs as part of a more clinically complex DNA damage repair disorder with some overlap with xeroderma pigmentosum (XP) and Cockayne syndrome (CS). Other syndromic clinical features associated with the *PCNA* variant include ocular and cutaneous telangiectasia, premature aging, photophobia, photosensitivity with predisposition to sun-induced malignancy, short stature, learning difficulties and neurodegeneration with cerebellar atrophy [177]. This rare cause of SHL was originally identified in a single extended Amish family comprising four affected individuals (three siblings and a more distant cousin), and remains the only *PCNA* mutation described in a human inherited disorder. The causative mutation was identified using a combined strategy of homozygosity mapping and whole exome sequencing (WES), which identified a single candidate gene variant in *PCNA* (c.683G>T; p.Ser228Ile) in a particularly small (0.77Mb) autozygous region containing just six genes, which comprised the only autozygous region shared by all affected individuals. More recently a further affected Amish individual, from Wisconsin, has since been identified and

confirmed to be homozygous for the same p.Ser2281Ile founder mutation, validating this gene as causative of this condition [345]. To date this variant has only been identified, by our group, in the Holmes County communities of Ohio at an allele frequency (AF) of 0.015% which is significantly higher than the 0.000007955% reported in gnomAD [295] indicating this variant is enriched within this community.

PCNA functions as an essential sliding clamp protein during DNA replication and repair processes. A significant impairment of PCNA protein function is likely to be incompatible with life, and our functional studies determined that the causative variant likely affects only specific functional aspects of molecular function [177]. SNHL in this condition was universal, of prelingual onset and of moderate to profound severity and worse at higher frequencies. Early onset of SNHL is a distinguishing factor from other DNA repair disorders in which SNHL typically displays an older age of onset, and is not a universal feature. It should be noted however, that due to its complex tertiary structure, and the large number of molecular interacting partners involved in the diverse functional roles of PCNA, other sequence alterations identified in other functional domains of PCNA may well exhibit clinical outcomes distinct to those associated with the p.Ser228Ile variant present in the Amish.

SLITRK6

Gene SLIT and NTRK-like family member 6 (*SLITRK6*), located on chromosome 13q31.1, comprises a single coding exon [346]. The gene encodes an integral membrane protein with leucine rich repeat domains and is expressed in specific brain regions and the neural retina, in which it is thought to play a role in synaptogenesis [178]. Our group identified the genetic basis of an autosomal

recessive deafness and myopia syndrome characterised by severe congenital myopia and SNHL, associated with sequence variants in *SLITRK6* identified in three separate families (of Amish, Greek and Turkish origins) [178]. In a single Amish family comprising three affected siblings, whole genome SNP genotyping identified a single notable homozygous region of 12.2Mb common to all affected siblings. Despite the relatively large size of the homozygous genomic region identified, it fortuitously contained only five protein coding genes, in which dideoxy sequencing identified a single nonsense variant (c.1240C>T; p.Q414X) in one; exon 2 of *SLITRK6*, that appropriately co-segregated in the family. Our studies of *SLITRK6* -/- mice showed delayed synaptogenesis of the retina in postnatal development [178] and HEK293 cell transfection studies showed impaired cell surface localisation of the mutant protein. The SNHL in these individuals was bilateral, moderate to severe and of prelingual onset, requiring hearing aids. Congenital myopia was severe from –6 to –11 diopters refractive error. The heterozygous parents of the affected siblings both had a low degree of myopia, perhaps indicative of an intermediate or milder phenotype in heterozygotes, although neither had HL and it is impossible to conclusively determine whether the myopia may relate to the SLITRK6 variant. Subsequently other Amish families from Pennsylvania were identified with the condition [159] in whom the same homozygous c.1240C>T mutation was identified. This group undertook control studies to estimate a 4.7% carrier frequency in an Old Order Amish population (n=571). All affected individuals display a similar phenotype of high myopia and SNHL, which is progressive and severe to profound by early adulthood. Audiological testing in all affected individuals showed absent distortion product otoacoustic emissions (DPOAEs) and auditory brainstem responses (ABRs)

were dys-synchronised bilaterally. Four affected individuals had absent ipsilateral middle ear muscle reflexes (MEMRs).

HARS

The *HARS* (histidyl-tRNA synthetase) gene is located on chromosome 5q31 and comprises 13 exons [347]. *HARS* encodes a molecule with a role in protein biosynthesis, which catalyses ligation of histidine to tRNA. Puffenburger et al. identified three patients from Amish families in Pennsylvania who displayed features suggestive of Usher syndrome type 3 [179]. Usher syndrome is classified into three major types, distinguished by severity of HL, age of onset, and the presence of vestibular abnormalities. Usher syndrome shows genetic heterogeneity and is further subdivided by genetic basis, with 15 genes currently associated with the condition. *HARS* mutation as a novel cause of Usher syndrome is classified as type 3B [348]. Clinical features of individuals in whom the *HARS* variant was identified include progressive visual impairment in childhood with horizontal nystagmus, optic pallor, photosensitivity, bull's eye macula and pigmentary changes consistent with retinitis pigmentosa (RP), progressive SNHL, delay in gross motor development, lower limb brisk reflexes and ataxia with normal intellect. Other features described include visual hallucination in response to fever (Charles-Bonnet syndrome), acute psychosis with catatonia, and sudden unexplained death in one child. No newborn auditory screening data was available, although for affected individuals all evoked auditory waveforms were absent by 5 years.

Homozygosity mapping performed using samples from the three affected individuals determined that the gene most likely resides within an 8.4Mb region of homozygosity of chromosome 5q31, containing 187 genes. WES of two

individuals with filtering for homozygous variants within the mapped region identified a single novel variant (c.1361A>C;p.Tyr454Ser) in *HARS*. Screening of 406 Old Order Amish control chromosomes identified 7 heterozygous carriers of the variant, with an allele frequency of 1.72%. The group identified a further individual of Amish descent in Canada with an identical phenotype who was also homozygous for the c.1361A>C variant.

YARS

The *YARS* gene is located on chromosome 1p35.1 and encodes tyrosyl-tRNA synthetase, an enzyme that catalyses the aminoacylation of tRNA. Dominant variants in YARS have been associated with Charcot-Marie-Tooth (CMT) disease type C dominant intermediate [349]. More recently, a recessive multisystem disorder characterised by developmental delay, small stature, spasticity, areflexia, hypertriglyceridaemia, liver dysfunction, lung cysts and abnormal subcortical white matter was identified in two siblings of Polish origin [350]. Expanding on this phenotype Williams and Demczko [180] have presented unpublished work describing an Amish family with distantly related common ancestors and four individuals (three siblings and one cousin) affected with bilateral SNHL, nystagmus, visual impairment, developmental delay, pancreatic insufficiency, cholestatic liver disease, hypoglycaemia and white matter abnormalities on MRI brain, who were identified through whole exome sequencing to be homozygous for a variant in *YARS* (c.499C>A; p.Pro167Thr). Further work is ongoing to confirm this association, including functional studies. Variants in other aminoacyl-tRNA synthetases including *HARS* (see above) and *KARS* have been associated with both SNHL and dominant CMT disease [350].

ST3GAL5

The ST3 beta-galactoside alpha-2,3-sialyltransferase 5 (*ST3GAL5*) gene encodes GM3 synthase; a sialyltransferase enzyme that synthesises GM3 ganglioside from lactosylceramide, the first step in synthesis of complex gangliosides [351]. The *ST3GAL5* gene is resident on chromosome 2p11.2 and contains seven exons [352]. Our group first identified an infantile onset epilepsy syndrome caused by mutation in *ST3GAL5* resulting in GM3 synthase deficiency [30]. In a large Amish family, eight affected individuals from two interlinking branches were identified to have infantile onset epilepsy, with developmental delay and blindness. Seizures arise in the first year, with associated developmental regression, with several seizure types, which are refractory to treatment. Other features include generalised irritability, poor feeding requiring gastrostomy in two children, dystonic arm movements, hypotonia and visual deterioration of likely cortical origin. Homozygosity mapping identified a single candidate genomic region in which a homozygous nonsense mutation (c.862C>T; p.R288X) was identified in *ST3GAL5*, with appropriate segregation within the family. GM3 synthase deficiency has also been termed Salt and Pepper Developmental Regression Syndrome and has been associated with further features within the Amish population, including SNHL (Wang, Wang et al. 2016), pigmentary changes of the skin [353] and optic nerve defects [354]. Auditory brainstem responses (ABRs) showed absent cochlear microphonics and abnormal thresholds were recorded in all eight Amish children homozygous for the c.862C>T variant, with waveform phase reversal in most. Cortical auditory-evoked potentials showed abnormal morphology in seven of these individuals [355]. GM3 synthase deficiency has subsequently been described by several further groups in other populations worldwide [356-358].

LONP1

The lon peptidase 1 (*LONP1*) gene encodes a homohexameric enzyme of the AAA+ superfamily of ATPases. Lon peptidase 1 is multifunctional; it regulates quality control processes for protein synthesis, assembles protein complexes within the respiratory chain and regulates mitochondrial gene expression [359-361]. *LONP1* consists of 19 exons at cytogenetic location 19p13.3 [362]. Mutation in *LONP1* has been associated with cerebral, ocular, dental, auricular, and skeletal anomalies (CODAS) syndrome. CODAS syndrome was first described in an endogamous Mennonite community [183] characterised by clinical features of developmental delay, dysmorphic facial features including ptosis, median nasal groove and malformed ears, bilateral cataracts, dental anomalies including delayed tooth eruption and anomalous cusp morphology, SNHL and skeletal features of short stature, delayed epiphyseal ossification, metaphyseal hip dysplasia and coronal clefts of the vertebrae. A further case of CODAS in a child of Mennonite ancestry was subsequently described and further characterised the phenotype [363], suggesting autosomal recessive inheritance of the condition. Strauss et al. identified ten further individuals with CODAS syndrome from Amish, Mennonite and mixed European backgrounds [364]. WES or dideoxy sequencing identified four *LONP1* mutations; among the Pennsylvania Amish, eight individuals were identified to be homozygous for a founder mutation (c.2161C>G; p.Arg721Gly) in *LONP1*. In an Amish control group a high population allele frequency of 5.9% was identified (compared with 0.00073% in GnomAD). The group sequenced *LONP1* in the Mennonite individual originally described by Shebib et al. [183] who was found to be homozygous for a different variant (c.2026C>T; p.Pro676Ser) and a further individual of mixed European ancestry was found to be compound heterozygous for two further mutations c.1892C>A

(p.Ser631Tyr) and c.2171C>T (p.Ala724Val). The affected Amish individuals showed a mixed picture of hearing loss; two had low frequency conductive hearing loss (CHL) with impaired tympanic membrane mobility. The two further individuals had both CHL with mild to moderate SNHL at medium to high frequency. Further details of the CODAS phenotype from these individuals highlight the severity of the condition with several children dying from laryngeal obstruction in the first few days of life and usually have swallowing difficulties requiring gastrostomy feeding. Other features include hypotonia, scoliosis, imperforate anus, omphalocele, rectovaginal fistula, cryptorchidism and tongue hemiatrophy. Due to the high allele frequency of c.2161C>G it would be expected that more cases of CODAS syndrome would be present in the Amish population, although this discrepancy may be explained in part by allele frequency differences in different Amish communities, as well as the high rate of neonatal mortality associated with gene mutation and many cases remaining undiagnosed.

7.6.2 Conductive hearing loss in the Amish community

HYAL2

The *HYAL2* gene encodes hyaluronidase 2, an enzyme with weak activity to degrade hyaluronan, an extracellular matrix glycosaminoglycan that is expressed during development [365]. *HYAL2* is located on chromosome 3p21.31 and consists of four exons [366]. We identified *HYAL2* as a novel gene associated with SHL in an extended Amish family of three distantly related nuclear families with five affected individuals, as well as a further two affected siblings with a similar phenotype in a Saudi Arabian family [184]. Mutation in *HYAL2* causes principle clinical features of cleft lip or palate (CLP), which can be unilateral or bilateral, and facial dysmorphism with characteristic features of frontal bossing,

hypertelorism, flattened and wide nasal bridge and tip, cupped ears with thickened helices and micrognathia. Other more variable features include congenital cardiac abnormalities, including cor triatriatum, predominantly CHL (although one individual had SNHL), pectus excavatum, single palmar creases, 2-3 toe syndactyly, myopia, staphyloma and cataract. A combination of homozygosity mapping of the five affected individuals with WES of a single affected individual identified a 10.18Mb region of homozygosity shared between affected individuals within the extended Amish family. Only a single candidate deleterious variant was identified, located in the *HYAL2* gene (c.443A>G; p.K148R). In the Saudi Arabian family similar methods identified a missense variant (c.749C>T; p.P250L) in *HYAL2*. Mice *HYAL2*-/- studies demonstrated comparable phenotype with cleft palate, facial dysmorphism and variable features of cor triatriatum sinister in 50%, and hearing loss in 100% of *HYAL2*-/- mice. Functional studies demonstrated that both reported mutations significantly reduced the level of hyaluronidase-2 expression. In these families, CHL was described in four individuals and was variable, from mild to moderate impairment, unilateral or bilateral and pre or postlingual. A further individual had severe to profound SNHL that was prelingual.

7.6.3 Mixed hearing loss in the Amish community

COL1A2

Collagen type I alpha 2 chain (*COL1A2*) is a large gene consisting of 52 exons located at 7q21.3.The *COL1A2* gene encodes the pro-alpha 2 chain of type 1 collagen, which is comprised of two alpha-1 chains and one alpha-2 chain that form a triple helix. Mutations in *COL1A2* and *COL1A1* have been identified as a cause of osteogenesis imperfecta (OI) types I II, III and IV. The four different types

of OI are inherited in an AD manner and have overlapping phenotypes and genotypes, varying in severity from a perinatal lethal form to a mild form with blue sclera and occasional fractures. Variability is often seen between individuals within the same family. Clinical features include fractures with minimal or no trauma, bone deformity, short stature, progressive mixed hearing loss, dentinogenesis imperfecta and connective tissue abnormality. Hearing loss in OI occurs in the majority of individuals by adulthood [367]. Often this starts with CHL caused by fracture of middle ear bones, with later SNHL developing with age. Usually onset is progressive and postpubertal, although hearing loss starting in childhood can occur.

Mcbride et al. [186] reported a *COL1A2* mutation (c.2098G>T; p.Gly610Cys) identified by dideoxy sequencing in several interrelated Amish families with 64 individuals subsequently identified with the *COL1A2* gene variant, displaying a variable phenotype [185]. A subsequent study used linkage analysis in these individuals to identify further candidate linkage loci that may contain further modifier genes that influence the severity of OI. This identified a candidate modifier locus on chromosome 1q, which the investigators suggested may involve the *PTGS2* gene which is involved in regulating bone formation [368].

7.7 APPENDIX G - Expression studies of Slc15a5 inner ear mouse tissue

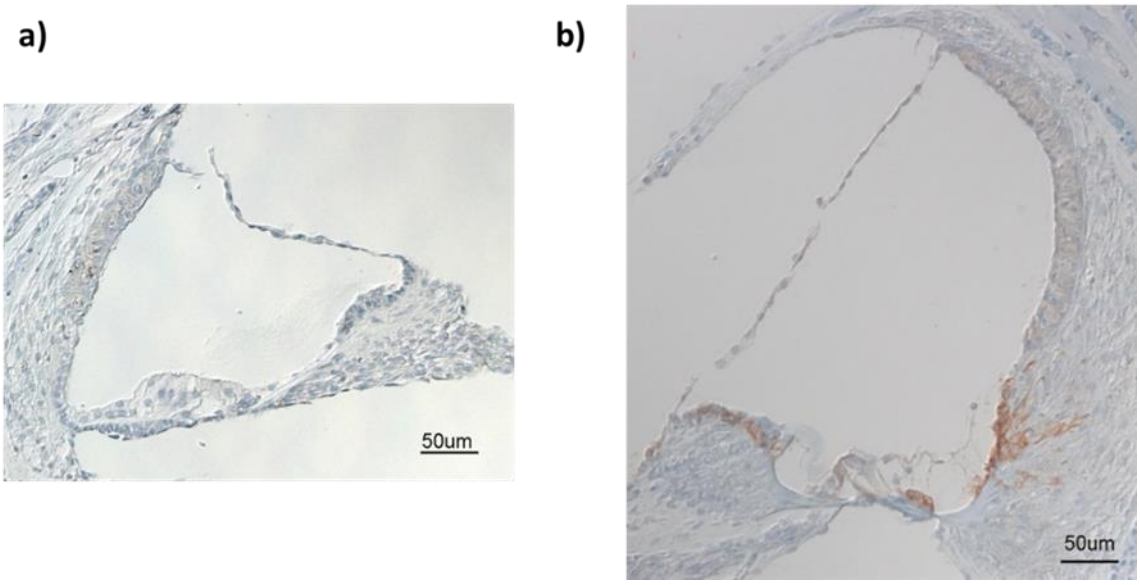


Figure 7.1: Antibody staining of the inner mouse ear (E16.5) a) with the antibody raised against slc15a5 and b) with antibody raised against odf2. Brown staining indicates positive staining.

7.8 APPENDIX H - SYT1-associated neurodevelopment disorder HPO terms

Phenotypic features reported for at least one case of *SYT1* mutation

System	HPO term	Feature
Eye		
	HP:0000565	Esotropia
	HP:0000540	Hypermetropia
	HP:0000486	Strabismus
	HP:0000639	Nystagmus
Cutaneous		
	HP:0025247	Dermoid cyst
Cardiovascular		
	HP:0001631	Atrial septal defect
Respiratory		
	HP:0010536	Sleep apnea
	HP:0001601	Laryngomalacia
	HP:0002883	Hyperventilation
Musculoskeletal		
	HP:0002650	Scoliosis
	HP:0008081	Pes valgus
	HP:0002938	Lumbar hyperlordosis
	HP:0001883	Talipes
	HP:0008081	Pes valgus
Gastrointestinal		
	HP:0002020	Gastroesophageal reflux
	HP:0012450	Chronic constipation
	HP:0011968	Feeding difficulties
Neurological		
	HP:0000733	Stereotypy
	HP:0012169	Self-biting
	HP:0000742	Self-mutilation
	HP:0100716	Self-injurious behavior
	HP:0012168	Head-banging

	HP:0100710	Impulsivity
	HP:0002353	EEG abnormality
	HP:0002451	Limb dystonia
	HP:0007098	Paroxysmal choreoathetosis
	HP:0001344	Absent speech
	HP:0002465	Poor speech
	HP:0001263	Global developmental delay
	HP:0002457	Abnormal head movements
	HP:0100022	Abnormality of movement
	HP:0002828	Multiple joint contractures
	HP:0100021	Cerebral palsy
	HP:0002072	Chorea
Connective Tissue		
	HP:0002828	Multiple joint contractures

7.9 APPENDIX I - Amish Genome Project additional tables

7.9.1 PubMed literature review genes – Nonsense

Gene	Phenotype	PubMed ref
APOL5	Schizophrenia	18571626
B4GALNT3	Neuroblastoma	21741930
C10orf53	Meningiomas	25981829
CCDC17	Keratoacanthoma	27788211
CCDC66	Human retinal dystrophy	28369829
CCDC83	Colon cancer	22923163
CES1	Carboxylesterase 1 deficiency	18485328
CFAP61	Obesity	28224759
CLDN17	Gastric cancer	24325792
CLEC4G	Inflammatory Response	26943817
COQ7	Coenzyme Q10 deficiency, primary, 8	26084283
DNAH6	Heterotaxy and ciliary dysfunction	26918822
DNALI1	Immotile cilia syndrome	19944400
EFS	Prostate cancer	25296736
EPHA10	Breast cancer	27566654
FAM109B	Meningiomas	25981829
FAT2	Spinal meningioma	27900010
FLT1	Eclampsia	14764923
FOLR3	Meningomyelocele	20683905
GPT	Glutamate pyruvate transaminase polymorphism	9119391
HAGH	Glyoxalase II deficiency	7424909
HLA-DRB4	Encephalitis	28026046
HLTF	Colorectal cancer, associated	21479407
HOXC5	Cervical cancer	10208853
HSD17B8	Colorectal cancer	20049862
IGSF10	Hypogonadotropic hypogonadism	27137492
KCNC4	Spinocerebellar Ataxia 13	20712895
KMT2C	Intellectual disability	22726846
KRT23	Colon cancer cells Periodontitis	24039993 15081423
KRT37	Celiac disease	21627641
MAPK13	Gynecological cancer	26969274
MED12	Lujan-Fryns syndrome, Ohdo syndrome, X-linked, Opitz-Kaveggia syndrome	16711603 17369503 28279489 317334363
MUC19	Sjogren's Syndrome	18184611
MUC3A	Inflammatory bowel disease Hypertrichotic Osteochondrodysplasia Cap Polyposis	11289722 21303913

MUC6	Pancreatic Ductal Carcinoma Signet Ring Cell Adenocarcinoma Gastric cancer	127165582 223573307 327298226
PADI2	Alzheimer Disease squamous cell carcinoma	20002008 28331341
PIFO	Laterality defects	120643351
PRSS2	Pancreatitis, chronic, protection against	16699518
PYHIN1	Diffuse Scleroderma and Limited Scleroderma, Asthma (risk)	21804549
RERGL	Familial colorectal cancer	24127187
RPS19	Diamond-Blackfan anemia 1	9988267 10598818 12586610 17517689
RXFP2	Cryptorchidism	20636340
SECISBP2	Thyroid hormone metabolism, abnormal	16228000
SLC22A11	Gout Persistent Fetal Circulation Syndrome Hereditary hypouricaemia	27225847 27103454
SRR	Streptococcal Meningitis, Serine Deficiency	22990841
TCF20	Vertebrobasilar Insufficiency and Breast Sarcoma Autism spectrum disorder	25228304
THBS3	Gaucher Disease, Type I Cancer	7558000 9608355 27197191
TIMP4	Open angle glaucoma patients Focal epilepsy Kawasaki disease	26539028 25595263 19048177
TMEM78	Nasopharyngeal carcinoma	22260379
TMPRSS5	Autosomal recessive hearing loss	17918732
TNNI3K	Cardiac conduction disease with or without dilated cardiomyopathy	24925317
TOP1MT	Non-small-cell lung cancer	28355294
TRAP1	Cakut With Or Without Vacterl & Brain Glioblastoma Multiforme Colorectal cancer	28088229
TTC21B	Nephronophthisis 12 Short-rib thoracic dysplasia 4 with or without polydactyly	21258341 21258341
TTC39B	Endometriosis Steatohepatitis and atherosclerosis	27453397 27383786
ULK4	Schizophrenia Neurogeneis and Brain function	27670918 28596978
USP54	Cancer	28129647
VPS13C	Parkinson disease 23, autosomal recessive, early onset	26942284
XIRP1	Brachial Plexus Lesion and Myopathy, Myofibrillar, 5	24725425
ZNF599	Hypospadias	22378287
ZNF675	Chronic periodontitis	25056994

7.9.2 PubMed literature review genes – Frameshift

Gene	Phenotype	PubMed ref
ABHD14B	neuroendocrine tumors	21681495
ACSM3	Hypertension, essential	7907320
ACTA1	Nemaline myopathy 3, autosomal dominant or recessive	10508519 11333380 9185179 16427282 19553116 22442437 23650303
ACTA1	Myopathy, congenital, with fiber-type disproportion 1	15468086
ACTA1	Myopathy, actin, congenital, with cores	15520409
ACTA1	Myopathy, scapulohumeroperoneal	4952447 25938801
ACTA1	Myopathy, actin, congenital, with excess of thin myofilaments	9185179 10508519
AK2	Reticular dysgenesis	19043417 19043416
ANO9	Colorectal carcinoma (stage II & III)	26317553
AP1G2	Cardiac Arrest Long Qt Syndrome 1	21658281
APOBR	Obesity	25955518
ARMC2	Lung function	21946350
ATP5J	Colorectal cancer	24124598
ATP5J	Alzheimer's disease and severe cerebral amyloid angiopathy	22008262
ATXN3	Machado-Joseph disease (Spinocerebellar ataxia type 3)	7874163
ATXN3	Machado-Joseph disease (Spinocerebellar ataxia type 3)	7874163
AVPR2	Diabetes insipidus, nephrogenic	1356229 1303271 8479490 8479491 8401502 8078903 7714087 9369448 9329382 9711877 10770218 11232028
AVPR2	Nephrogenic syndrome of inappropriate antidiuresis	15872203
BCL7C	Lymphoma	9931421
BRMS1	Breast Cancer and Malignant Melanoma, Somatic Breast cancer	18841483 23771732
BTNL9	Bipolar disorder and schizophrenia	25243493

BUB1B-PAK6	Colon and Prostate Cancer	25426562 24946957
CACNA1G	Childhood Absence Epilepsy	17397049
CACNA1G	Spinocerebellar Ataxia Type 42	26456284
CCDC40	Ciliary Dyskinesia, Primary, 15 Primary Ciliary Dyskinesia15: Ccdc40-Related Primary Ciliary Dyskinesia	21131974
CCDC7	Colorectal cancer	22024937
CCDC88B	Sarcoidosis	22837380
CEP164	Nephronophthisis 15	22863007
CHD7	CHARGE syndrome	15300250 16400610 16155193 17334995 18074359
CHD7	Hypogonadotropic hypogonadism 5 with or without anosmia	17661815 18074359 18834967 17937444 18978652
CLTCL1	Chromosome 22Q11.2 Microduplication Syndrome	26549885
CLTCL1	Congenital Insensitivity To Pain With Severe Intellectual Disability	26549885
CPNE8	Prion disease	19795140
CRIPAK	Non-small cell lung cancer	25444907
CSPG4	Chordoma	26689475
CSPG4	Chondrosarcoma	27292772
CWH43	Colorectal cancer	24959000
DAGLB	Hallucinogen Dependence	26595473
DLEC1	Lung cancer	27287342
DNAH14	Embryonic lethal genes	26036949
DNAH14	Endometrial cancer	28339086
ECHDC1	Breast and ovarian cancer	19517271
EGFR	Inflammatory skin and bowel disease, neonatal, 2	24691054
EGFR	Adenocarcinoma of lung, response to tyrosine kinase inhibitor in	15118125 15118073 15728811
EPPK1	Vater/Vacterl Association	23549274
ERV3-1	Choriocarcinoma (and other cancers)	25016529 24043713
FAM46A	Retinitis Pigmentosa 25	17803723
FAM46A	Skeletal dysplasia	26803617
FAM81B	Breast cancer	27491861
FBXW8	Intellectual disability	25626716
FCGBP	Endometriosis	27817035
GJC3	Hearing Loss	26074771 23179405 19876648

GPR27	Developmental delay and congenital anomalies (craniofacial dysmorphism, including a cleft lip)	12836054
GPR27	Speech delay, contractures, hypertonia and blepharophimosis	19332160
HELQ	Natural menopause	28118297
HELQ	Breast cancer	27792995 26351136
HELQ	Ovarian cancer	28101207 26351136
HLA-B	Ankylosing spondylitis, susceptibility to (synovitis, chronic, suscpetibility to)	8053961
HLA-B	Severe cutaneous adverse reaction, susceptibility to Stevens-johnson syndrome, susceptibility to, included Toxic epidermal necrolysis, susceptibility to, included	15057820
HLA-B	Severe cutaneous adverse reaction, susceptibility to Stevens-johnson syndrome, susceptibility to, included Toxic epidermal necrolysis, susceptibility to, included	15743917
HLA-B	Abacavir hypersensitivity, susceptibility to Drug-induced liver injury due to flucloxacillin, included	15247624 12462283
HLA-DPA1	Inflammatory Bowel Disease 3	12073072
HLA-DPA1	Kikuchi Disease	28613580
HLA-DPA1	Posner-Schlossman Syndrome	25863099
HLA-DRB1	Sarcoidosis, susceptibility to, 1	14508706
HLA-DRB1	Pemphigoid, susceptibility to	23502333
HLA-DRB1	Rheumatoid arthritis, susceptibility to	28711139
HLA-DRB1	Multiple sclerosis, susceptibility to, 1	21833088 28676141
IL36B	Psoriasis	21881584
KBTBD13	Nemaline myopathy 6, autosomal dominant	21109227 21104864 12805120
KCNJ12	Hirschsprung's disease	28399120
KCNJ18	Thyrotoxic periodic paralysis, susceptibility to, 2	20074522 28131627 27178871
LMO7	Emery-dreifuss muscular dystrophy	24825363
LY9	Systemic Lupus Erythematosus	23956418 18216865
MFF	Encephalopathy due to defective mitochondrial and peroxisomal fission 2	22499341 26783368
MOCOS	Xanthinuria, type II	11302742 25967871 14624414 17368066
MS4A12	colon cancer	27881006
MTMR11	breast cancer	20413845
MTX1	Gaucher's Disease	15024629
MUC21	Colorectal Cancer 1	28575854

MUC3A	Ulcerative colitis and Crohn's disease.	11289722
NDOR1	bladder carcinoma	26722457
NDUFA8	Mitochondrial Complex I Deficiency	15576045 9860297
NETO1	autism	20499253
PAK6	Prostate Cancer	26459798 23931236 18642328
PDE11A	Pigmented nodular adrenocortical disease, primary, 2	16767104
PDGFRL	Colorectal cancer, somatic	7898930
PDGFRL	Hepatocellular cancer, somatic	7898930
PDGFRL	Behçet disease	22926996
PGK2	Ovary and ovarian cancers	19333399
PLAC4	Early onset preeclampsia	23312075
PNPLA7	Susceptibility to menstrual disorder	25867316
POLQ	Breast cancer, ovarian cancer and other cancer types	27264557 25409685
PRMT6	HIV-1	26611710
PRPF3	Retinitis pigmentosa 18	11773002 17932117 11773002 18412284
RGL4	Associated with chemosensitivity	20224928
RP1L1	Occult macular dystrophy	20826268 23281133 20826268 23281133 22605915 27623337
SCAPER	Parkinson's disease	25294124
SCAPER	Atopic dermatitis (recalcitrant)	25935106
SCGB3A2	Asthma, susceptibility to	11813133
SCGB3A2	Graves' disease	23934357 21170691
SH2B3	Myelofibrosis, somatic	20404132
SH2B3	Thrombocythemia, somatic	20404132
SH2B3	Erythrocytosis, somatic	20843259
SHROOM4	Stocco dos Santos X-linked mental retardation syndrome	12673656 16249884 26740508
SLC22A3	Coronary Artery disease	27417586 27621937
SLC22A3	Oesophageal cancer	28533408 28743982
SLC25A5	Non-Syndromic Intellectual Disability	23783460
SLC6A18	Myocardial infarction	21420947

SLC6A18	Iminoglycinuria, Digenic	20377526 19033659
SP7	Osteogenesis Imperfecta, Type XII	20579626
SPZ1	Non-small cell lung cancer	23463593
TCHH	Uncombable hair syndrome 3	27866708 27487801 26414620 27866708
TEX9	Nevoid basal cell carcinoma syndrome	9388465
TMEM242	Developmental delay	26391891
TRPM1	Night blindness, congenital stationary (complete), 1C, autosomal recessive	19878917 19896113 19896109 19436059 20300565
ULK4	Schizophrenia	24284070 27670918
ULK4	Hypertension	25519392 25249183 27980663
VN1R2	Gliomas	23451178
ZNF443	Wilms tumor cells	23267699
ZNF510	Oral squamous cell carcinoma	21497587
ZNF677	Non-small cell lung cancers	25504438

AN AMISH FOUNDER VARIANT CONSOLIDATES DISRUPTION OF *CEPP5* AS A CAUSE OF HYDRANENCEPHALY AND RENAL DYSPLASIA

MNS1 MUATATION ASSOCIATED WITH SITUS INVERUS AND MALE INFERILITY

8 REFERENCES

1. Alkuraya, F.S., *The application of next-generation sequencing in the autozygosity mapping of human recessive diseases.* Hum Genet, 2013. **132**(11): p. 1197-211.
2. Muensterer, O.J., et al., *Ellis-van Creveld syndrome: its history.* Pediatr Radiol, 2013. **43**(8): p. 1030-6.
3. Angelman, H., *'Puppet' Children A Report on Three Cases.* Developmental Medicine & Child Neurology, 1965. **7**(6): p. 681-688.
4. MENKES, J.H., P.L. HURST, and J.M. CRAIG, *A NEW SYNDROME: PROGRESSIVE FAMILIAL INFANTILE CEREBRAL DYSFUNCTION ASSOCIATED WITH AN UNUSUAL URINARY SUBSTANCE.* Pediatrics, 1954. **14**(5): p. 462-467.
5. Blau, N., F.J. van Spronsen, and H.L. Levy, *Phenylketonuria.* The Lancet, 2010. **376**(9750): p. 1417-1427.
6. Studies, Y.C.f.A.a.P. *Amish Population Change 2009-2018*. 2018; Available from: https://groups.etown.edu/amishstudies/files/2018/08/Population_Change_2009-2018.pdf.
7. Slatkin, M., *A Population-Genetic Test of Founder Effects and Implications*
for Ashkenazi Jewish Diseases. Am. J. Hum. Genet., 2004. **75**: p. 282-293.
8. Norio, R., *Finnish Disease Heritage I: characteristics, causes, background.* Hum Genet, 2003. **112**(5-6): p. 441-56.
9. de la Chapelle, A., *Disease gene mapping in isolated human populations: the example of Finland.* J Med Genet, 1993. **30**: p. 857-865.
10. de la Chapelle, A. and F.A. Wright, *Linkage disequilibrium mapping in isolated populations: The example of Finland revisited.* Proc. Natl. Acad. Sci., 1988. **95**: p. 12416–12423.
11. Palo, J.U., et al., *Genetic markers and population history: Finland revisited.* Eur J Hum Genet, 2009. **17**(10): p. 1336-46.
12. Tsai, H., et al., *Distribution of genome-wide linkage disequilibrium based on microsatellite loci in the Samoan population.* HUMAN GENOMICS, 2004. **1**(55): p. 327-334.
13. Hudjashov, G., et al., *Investigating the origins of eastern Polynesians using genome-wide data from the Leeward Society Isles.* Sci Rep, 2018. **8**(1): p. 1823.
14. Aberg, K., et al., *Susceptibility loci for adiposity phenotypes on 8p, 9p, and 16q in American Samoa and Samoa.* Obesity (Silver Spring), 2009. **17**(3): p. 518-24.
15. Hawley, N.L., et al., *Prevalence of adiposity and associated cardiometabolic risk factors in the Samoan genome-wide association study.* Am J Hum Biol, 2014. **26**(4): p. 491-501.
16. Minster, R.L., et al., *A thrifty variant in CREBRF strongly influences body mass index in Samoans.* Nat Genet, 2016. **48**(9): p. 1049-1054.
17. Naka, I., et al., *A missense variant, rs373863828-A (p.Arg457Gln), of CREBRF and body mass index in Oceanic populations.* Journal of Human Genetics volume, 2017. **62**: p. 847-849.

18. Neel, J.V., *Diabetes Mellitus: A "Thrifty" Genotype Rendered Detrimental by "Progress"?* Am J Hum Genet., 1962. **14**: p. 353-362.
19. McWhirter, R.E., et al., *Genome-wide homozygosity and multiple sclerosis in Orkney and Shetland Islanders.* Eur J Hum Genet, 2012. **20**(2): p. 198-202.
20. Relethford, J.H. and E.R. Brennan, *Temporal Trends in Isolation by Distance on Sanday, Orkney Island.* Human Biology, 1982. **54**: p. 315-327.
21. Sajedi, S.A. and F. Abdollahi, *Which Environmental Factor Is Correlated with Long-Term Multiple Sclerosis Incidence Trends: Ultraviolet B Radiation or Geomagnetic Disturbances?* Mult Scler Int, 2017. **2017**: p. 4960386.
22. Visser, E.M., et al., *A new prevalence study of multiple sclerosis in Orkney, Shetland and Aberdeen city.* Journal of Neurology, Neurosurgery & Psychiatry, 2012. **83**(7): p. 719-724.
23. Rosati, G., *The prevalence of multiple sclerosis in the world: an update.* Neurol Sci, 2001. **22**: p. 117-139.
24. Francomano, C.A., V.A. McKusick, and L.G. Biesecker, *Medical genetic studies in the Amish: historical perspective.* Am J Med Genet C Semin Med Genet, 2003. **121C**(1): p. 1-4.
25. Cross, H.E. and V.A. McKusick, *The Troyer syndrome. A recessive form of spastic paraplegia with distal muscle wasting.* Arch Neurol, 1967. **16**(5): p. 473-85.
26. Cross, H.E. and V.A. McKusick, *The mast syndrome. A recessively inherited form of presenile dementia with motor disturbances.* Arch Neurol, 1967. **16**(1): p. 1-13.
27. McKusick, V.A., *Mendelian Inheritance in Man and its online version, OMIM.* Am J Hum Genet, 2007. **80**(4): p. 588-604.
28. Ankeny, R.A., *Geneticization in MIM/OMIM(R)? Exploring Historic and Epistemic Drivers of Contemporary Understandings of Genetic Disease.* J Med Philos, 2017. **42**(4): p. 367-384.
29. Patel, H., et al., *SPG20 is mutated in Troyer syndrome, an hereditary spastic paraplegia.* Nat Genet, 2002. **31**(4): p. 347-8.
30. Simpson, M.A., et al., *Infantile-onset symptomatic epilepsy syndrome caused by a homozygous loss-of-function mutation of GM3 synthase.* Nat Genet, 2004. **36**(11): p. 1225-9.
31. Exeter, U.o. *Windows of Hope*. 2018; Available from: http://www.wohproject.org/.
32. Mardis, E.R., *Next-generation DNA sequencing methods.* Annu Rev Genomics Hum Genet, 2008. **9**: p. 387-402.
33. Erzurumluoglu, A.M., et al., *Importance of Genetic Studies in Consanguineous Populations for the Characterization of Novel Human Gene Functions.* Ann Hum Genet, 2016. **80**(3): p. 187-96.
34. Alkuraya, F.S., *Autozygome decoded.* Genet Med, 2010. **12**(12): p. 765-71.
35. Lander, E.S. and D. Botstein, *Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children.* Science, 1987. **236**(4808): p. 1567-70.
36. Vahidnezhad, H., et al., *Research Techniques Made Simple: Genome-Wide Homozygosity/Autozygosity Mapping Is a Powerful Tool for*

Identifying Candidate Genes in Autosomal Recessive Genetic Diseases. J Invest Dermatol, 2018. **138**(9): p. 1893-1900.
37. Carr, I.M., et al., *Autozygosity mapping with exome sequence data.* Hum Mutat, 2013. **34**(1): p. 50-6.
38. Gibbs, J.R. and A. Singleton, *Application of genome-wide single nucleotide polymorphism typing: simple association and beyond.* PLoS Genet, 2006. **2**(10): p. e150.
39. Woods, C.G., et al., *A new method for autozygosity mapping using single nucleotide polymorphisms (SNPs) and EXCLUDEAR.* J Med Genet, 2004. **41**(8): p. e101.
40. Watson, C.M., et al., *Rapid Detection of Rare Deleterious Variants by Next Generation Sequencing with Optional Microarray SNP Genotype Data.* Hum Mutat, 2015. **36**(9): p. 823-30.
41. Affymetrix. *The Affymetrix SNP Array 6.0: A solution for molecular cytogenetics*. 2010; Available from: http://tools.thermofisher.com/content/sfs/brochures/snp_6_array_sol_cytogenetics_brochure.pdf.
42. Ching-Wan, L., L. Kin-Chong, and T. Sui-Fan, *Chapter 1 - Microarrays for Personalized Genomic Medicine.* Advances in Clinical Chemistry, 2010. **52**: p. 1-18.
43. Zhang, F., C.M. Carvalho, and J.R. Lupski, *Complex human chromosomal and genomic rearrangements.* Trends Genet, 2009. **25**(7): p. 298-307.
44. Sebat, J., et al., *Strong association of de novo copy number mutations with autism.* Science, 2007. **316**(5823): p. 445-9.
45. Diskin, S.J., et al., *Copy number variation at 1q21.1 associated with neuroblastoma.* Nature, 2009. **459**(7249): p. 987-91.
46. MacDonald, J.R., et al., *The Database of Genomic Variants: a curated collection of structural variation in the human genome.* Nucleic Acids Res, 2014. **42**(Database issue): p. D986-92.
47. Nowakowska, B., *Clinical interpretation of copy number variants in the human genome.* J Appl Genet, 2017. **58**(4): p. 449-457.
48. Peiffer, D.A., et al., *High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping.* Genome Res, 2006. **16**(9): p. 1136-48.
49. Illumina. *Infinium™ HumanCytoSNP-12 v2.1 BeadChip*. 2016; Available from: Infinium™ HumanCytoSNP-12 v2.1 BeadChip.
50. Howrigan, D.P., M.A. Simonson, and M.C. Keller, *Detecting autozygosity through runs of homozygosity: a comparison of three autozygosity detection algorithms.* BMC Genomics, 2011. **12**: p. 460.
51. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* Am J Hum Genet, 2007. **81**(3): p. 559-75.
52. Gusev, A., et al., *Whole population, genome-wide mapping of hidden relatedness.* Genome Res, 2009. **19**(2): p. 318-26.
53. Zhang, L., et al., *cgaTOH: extended approach for identifying tracts of homozygosity.* PLoS One, 2013. **8**(3): p. e57772.
54. Ceballos, F.C., S. Hazelhurst, and M. Ramsay, *Assessing runs of Homozygosity: a comparison of SNP Array and whole genome sequence low coverage data.* BMC Genomics, 2018. **19**(1): p. 106.

55. Browning, B.L. and S.R. Browning, *Improving the accuracy and efficiency of identity-by-descent detection in population data.* Genetics, 2013. **194**(2): p. 459-71.
56. Vigeland, M.D., K.S. Gjotterud, and K.K. Selmer, *FILTUS: a desktop GUI for fast and efficient detection of disease-causing variants, including a novel autozygosity detector.* Bioinformatics, 2016. **32**(10): p. 1592-4.
57. Narasimhan, V., et al., *BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data.* Bioinformatics, 2016. **32**(11): p. 1749-51.
58. Szpiech, Z.A., A. Blant, and T.J. Pemberton, *GARLIC: Genomic Autozygosity Regions Likelihood-based Inference and Classification.* Bioinformatics, 2017. **33**(13): p. 2059-2062.
59. Eddy, S.R., *Profile hidden Markov models.* Bioinformatics, 1998. **14**: p. 755–763.
60. Ghahramani, Z., *An introduction to Hidden Markov Models and Bayesian Networks.* International journal of pattern recognition and Artificial Intelligence, 2001. **15**: p. 9-42.
61. Eddy, S.R., *Hidden Markov models.* Curr Opin Struc Biol, 1996. **6**: p. 361-365.
62. Eddy, S.R., *What is a hidden Markov model?* Nat Biotechnol, 2004. **22**: p. 1315-1316.
63. Robinson, M.A., . *Linkage Disequilibrium*. Encyclopedia of Immunology (Second Edition), ed. P.J. Delves. 1998.
64. Li, Y., et al., *Leveling the Playing Field in Homozygosity Mapping Using Map Distances.* Ann Hum Genet, 2015. **79**(5): p. 366-372.
65. Goode, E.L., *Linkage Disequilibrium*, in *Encyclopedia of Cancer*, M. Schwab, Editor. 2011, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 2043-2048.
66. Pengelly, R.J., et al., *Whole genome sequences are required to fully resolve the linkage disequilibrium structure of human populations.* BMC Genomics, 2015. **16**: p. 666.
67. Barrett, J.C., *Population Genetics and Linkage Disequilibrium*, in *Analysis of Complex Disease Association Studies*. 2011. p. 15-23.
68. Warr, A., et al., *Exome Sequencing: Current and Future Perspectives.* G3 (Bethesda), 2015. **5**(8): p. 1543-50.
69. Maxam, A.M. and W. Gilbert, *A new method for sequencing DNA.* Proc Natl Acad Sci U S A, 1977. **74**: p. 560-564.
70. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors.* Proc. Nati. Acad. Sci. USA, 1977. **74**: p. 5463–5467.
71. Smith, M., *DNA Sequence Analysis in Clinical Medicine, Proceeding Cautiously.* Front Mol Biosci, 2017. **4**: p. 24.
72. Ameur, A., W.P. Kloosterman, and M.S. Hestand, *Single-Molecule Sequencing: Towards Clinical Applications.* Trends Biotechnol, 2019. **37**(1): p. 72-85.
73. Silva, P.J., V.M. Schaibley, and K.S. Ramos, *Academic medical centers as innovation ecosystems to address population -omics challenges in precision medicine.* J Transl Med, 2018. **16**(1): p. 28.
74. Suwinski, P., et al., *Advancing Personalized Medicine Through the Application of Whole Exome Sequencing and Big Data Analytics.* Front Genet, 2019. **10**: p. 49.

75. Adams, D.R. and C.M. Eng, *Next-Generation Sequencing to Diagnose Suspected Genetic Disorders.* N Engl J Med, 2018. **379**(14): p. 1353-1362.
76. Zhang, G., et al., *Comparison and evaluation of two exome capture kits and sequencing platforms for variant calling.* BMC Genomics, 2015. **16**(1): p. 581.
77. Rabbani, B., M. Tekin, and N. Mahdieh, *The promise of whole-exome sequencing in medical genetics.* J Hum Genet, 2014. **59**(1): p. 5-15.
78. Choia, M., et al., *Genetic diagnosis by whole exome capture and massively parallel DNA sequencing.* PNAS, 2009. **106**: p. 19096-19101.
79. Albert, T.J., et al., *Direct selection of human genomic loci by microarray hybridization.* Nat Methods, 2007. **4**(11): p. 903-5.
80. Hodges, E., et al., *Genome-wide in situ exon capture for selective resequencing.* Nat Genet, 2007. **39**(12): p. 1522-7.
81. Gnirke, A., et al., *Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing.* Nat Biotechnol, 2009. **27**(2): p. 182-9.
82. Porreca, G.J., et al., *Multiplex amplification of large sets of human exons.* Nat Methods, 2007. **4**(11): p. 931-6.
83. Sulonen, A.M., et al., *Comparison of solution-based exome capture methods for next generation sequencing.* Genome Biol, 2011. **12**(9): p. R94.
84. Caspar, S.M., et al., *Clinical sequencing: From raw data to diagnosis with lifetime value.* Clin Genet, 2018. **93**(3): p. 508-519.
85. Sims, D., et al., *Sequencing depth and coverage: key considerations in genomic analyses.* Nat Rev Genet, 2014. **15**(2): p. 121-32.
86. Okou, D.T., et al., *Microarray-based genomic selection for high-throughput resequencing.* Nat Methods, 2007. **4**(11): p. 907-9.
87. Fokstuen, S., et al., *Experience of a multidisciplinary task force with exome sequencing for Mendelian disorders.* Hum Genomics, 2016. **10**(1): p. 24.
88. Liao, P., G.A. Satten, and Y.J. Hu, *PhredEM: a phred-score-informed genotype-calling approach for next-generation sequencing studies.* Genet Epidemiol, 2017. **41**(5): p. 375-387.
89. Karczewski, K.J., et al., *Variation across 141,456 human exomes and genomes reveals the spectrum of loss-offunction intolerance across human protein-coding genes.* 2019.
90. Popejoy, A.B. and S.M. Fullerton, *Genomics is failing on diversity.* Nature, 2016. **538**(7624): p. 161-164.
91. Need, A.C. and D.B. Goldstein, *Next generation disparities in human genomics: concerns and remedies.* Trends Genet, 2009. **25**(11): p. 489-94.
92. Ng, P.C., *SIFT: predicting amino acid changes that affect protein function.* Nucleic Acids Research, 2003. **31**(13): p. 3812-3814.
93. Adzhubei, I., D.M. Jordan, and S.R. Sunyaev, *Predicting functional effect of human missense mutations using PolyPhen-2.* Curr Protoc Hum Genet, 2013. **Chapter 7**: p. Unit7 20.
94. Schwarz, J.M., et al., *MutationTaster2: mutation prediction for the deep-sequencing age.* Nat Methods, 2014. **11**(4): p. 361-2.
95. Richards, S., et al., *Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American*

College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med, 2015. **17**(5): p. 405-24.
96. Shendure, J. and H. Ji, *Next-generation DNA sequencing.* Nat Biotechnol, 2008. **26**(10): p. 1135-45.
97. Schadt, E.E., S. Turner, and A. Kasarskis, *A window into third-generation sequencing.* Hum Mol Genet, 2010. **19**(R2): p. R227-40.
98. van Dijk, E.L., et al., *The Third Revolution in Sequencing Technology.* Trends Genet, 2018. **34**(9): p. 666-681.
99. Jain, M., et al., *Improved data analysis for the MinION nanopore sequencer.* Nat Methods, 2015. **12**(4): p. 351-6.
100. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.
101. Li, H. and R. Durbin, *Fast and accurate long-read alignment with Burrows-Wheeler transform.* Bioinformatics, 2010. **26**(5): p. 589-95.
102. Kayima, J., et al., *Association of genetic variation with blood pressure traits among East Africans.* Clin Genet, 2017. **92**(5): p. 487-494.
103. Dun, X.P., et al., *Drebrin controls neuronal migration through the formation and alignment of the leading process.* Mol Cell Neurosci, 2012. **49**(3): p. 341-50.
104. Dieffenbach, C.W., T.M.J. Lowe, and G.S. Dveksler, *General Concepts for PCR Primer Design.* PCR Methods and Applications 1993. **3**.
105. Thomas, P. and T.G. Smart, *HEK293 cell line: a vehicle for the expression of recombinant proteins.* J Pharmacol Toxicol Methods, 2005. **51**(3): p. 187-200.
106. Powles-Glover, N. and M. Maconochie, *Prenatal and postnatal development of the mammalian ear.* Birth Defects Res, 2018. **110**(3): p. 228-245.
107. Dror, A.A. and K.B. Avraham, *Hearing loss: mechanisms revealed by genetics and cell biology.* Annu Rev Genet, 2009. **43**: p. 411-37.
108. Krstic, R.V., *Human Microscopic Anatomy: An Atlas for Students of Medicine and Biology*. 1991: Springer.
109. Raphael, Y. and R.A. Altschuler, *Structure and innervation of the cochlea.* Brain Res Bull, 2003. **60**(5-6): p. 397-422.
110. Pepermans, E. and C. Petit, *The tip-link molecular complex of the auditory mechano-electrical transduction machinery.* Hear Res, 2015. **330**(Pt A): p. 10-7.
111. Stelma, F. and M.F. Bhutta, *Non-syndromic hereditary sensorineural hearing loss: review of the genes involved.* J Laryngol Otol, 2014. **128**(1): p. 13-21.
112. El-Amraoui, A. and C. Petit, *Usher I syndrome: unravelling the mechanisms that underlie the cohesion of the growing hair bundle in inner ear sensory cells.* J Cell Sci, 2005. **118**(Pt 20): p. 4593-603.
113. Barr-Gillespie, P.G., *Assembly of hair bundles, an amazing problem for cell biology.* Mol Biol Cell, 2015. **26**(15): p. 2727-32.
114. Trune, D.R., *Ion Homeostasis in the Ear: Mechanisms, Maladies, and Management.* Curr Opin Otolaryngol Head Neck Surg, 2010: p. 413-419.
115. Siemens, J., et al., *Cadherin 23 is a component of the tip link in hair-cell stereocilia.* Nature, 2004. **428**(6986): p. 945-50.
116. Fettiplace, R. and K.X. Kim, *The physiology of mechanoelectrical transduction channels in hearing.* Physiol Rev, 2014. **94**(3): p. 951-86.

117. Mittal, R., et al., *Indispensable Role of Ion Channels and Transporters in the Auditory System.* J Cell Physiol, 2017. **232**(4): p. 743-758.
118. Khimich, D., et al., *Hair cell synaptic ribbons are essential for synchronous auditory signalling.* Nature, 2005. **434**(7035): p. 886-9.
119. Dubyak, G.R., *Ion homeostasis, channels, and transporters: an update on cellular mechanisms.* Adv Physiol Educ, 2004. **28**(1-4): p. 143-54.
120. Zdebik, A.A., P. Wangemann, and T.J. Jentsch, *Potassium ion movement in the inner ear: insights from genetic disease and mouse models.* Physiology (Bethesda), 2009. **24**: p. 307-16.
121. Nin, F., et al., *The unique electrical properties in an extracellular fluid of the mammalian cochlea; their functional roles, homeostatic processes, and pathological significance.* Pflugers Arch, 2016. **468**(10): p. 1637-49.
122. Gagov, H., M. Chichova, and M. Mladenov, *PRE PRESS: Endolymph composition: paradigm or inevitability.* PRE PRESS ARTICLE, 2018.
123. Locher, H., et al., *Development of the stria vascularis and potassium regulation in the human fetal cochlea: Insights into hereditary sensorineural hearing loss.* Dev Neurobiol, 2015. **75**(11): p. 1219-40.
124. WHO, *Deafness and Hearing Loss: Factsheet*. 2018, World Health Organisation: Geneva:.
125. Vona, B., et al., *Non-syndromic hearing loss gene identification: A brief history and glimpse into the future.* Mol Cell Probes, 2015. **29**(5): p. 260-70.
126. Sloan-Heggen, C.M. and R.J. Smith, *Navigating genetic diagnostics in patients with hearing loss.* Curr Opin Pediatr, 2016. **28**(6): p. 705-712.
127. Chang, K.W., *Genetics of Hearing Loss--Nonsyndromic.* Otolaryngol Clin North Am, 2015. **48**(6): p. 1063-72.
128. Parker, M. and M. Bitner-Glindzicz, *Republished: Genetic investigations in childhood deafness.* Postgrad Med J, 2015. **91**(1077): p. 395-402.
129. Hoefsloot, L.H., et al., *Genotype phenotype correlations for hearing impairment: approaches to management.* Clin Genet, 2014. **85**(6): p. 514-23.
130. Smith, R.J., J.F. Bale, Jr., and K.R. White, *Sensorineural hearing loss in children.* Lancet, 2005. **365**(9462): p. 879-90.
131. Shearer, A.E., M.S. Hildebrand, and R.J.H. Smith, *Hereditary Hearing Loss and Deafness Overview*, in *GeneReviews((R))*, M.P. Adam, et al., Editors. 1993: Seattle (WA).
132. Jecmenica, J., A. Bajec-Opancina, and D. Jecmenica, *Genetic hearing impairment.* Childs Nerv Syst, 2015. **31**(4): p. 515-9.
133. Zhong, L.X., et al., *Non-Syndromic Hearing Loss and High-Throughput Strategies to Decipher Its Genetic Heterogeneity.* Journal of Otology, 2013. **8**(1): p. 6-24.
134. Powles-Glover, N. and M. Maconochie, *Prenatal and postnatal development of the mammalian ear.* Birth Defects Res, 2017.
135. Lancet, *Hearing loss: an important global health concern.* The Lancet, 2016. **387**(10036): p. 2351.
136. Keren, R., et al., *Projected cost-effectiveness of statewide universal newborn hearing screening.* Pediatrics, 2002. **110**(5): p. 855-64.
137. Schrijver, I., *Hereditary non-syndromic sensorineural hearing loss: transforming silence to sound.* J Mol Diagn, 2004. **6**(4): p. 275-84.
138. Egilmez, O.K. and M.T. Kalcioglu, *Genetics of Nonsyndromic Congenital Hearing Loss.* Scientifica (Cairo), 2016. **2016**: p. 7576064.

139. Moeller, M.P., et al., *Strategies for Educating Physicians about Newborn Hearing Screening.* J Acad Rehabil Audiol, 2006: p. 11–32.
140. England, P.H. *Newborn hearing screening: programme overview*. 2016 2 November 2016 [cited 2018 16/04/2018]; This overview of the NHS newborn hearing screening programme (NHSP) explains how a baby's hearing is tested, and the equipment used for the tests.]. Available from: https://www.gov.uk/guidance/newborn-hearing-screening-programme-overview.
141. Lenarz, T., *Cochlear implant – state of the art.* GMS Curr Top Otorhinolaryngol Head Neck Surg., 2018.
142. Lenarz, T., *Cochlear implant - state of the art.* GMS Curr Top Otorhinolaryngol Head Neck Surg, 2017. **16**: p. Doc04.
143. Shearer, A.E. and R.J. Smith, *Massively Parallel Sequencing for Genetic Diagnosis of Hearing Loss: The New Standard of Care.* Otolaryngol Head Neck Surg, 2015. **153**(2): p. 175-82.
144. Audiology, B.A.o.P.i. *British Association of Paediatricians in Audiology*. Available from: http://www.bapa.uk.com/.
145. Gao, X. and P. Dai, *Impact of next-generation sequencing on molecular diagnosis of inherited non-syndromic hearing loss.* Journal of Otology, 2014. **9**(3): p. 122-125.
146. Yawn, R., et al., *Cochlear implantation: a biomechanical prosthesis for hearing loss.* F1000Prime Rep, 2015. **7**: p. 45.
147. Zeng, F.G., et al., *Cochlear implants: system design, integration, and evaluation.* IEEE Rev Biomed Eng, 2008. **1**: p. 115-42.
148. Wilson, B.S., et al., *Cochlear implants matching the prosthesis to the brain and facilitating desired plastic changes in brain function.* Prog Brain Res, 2011. **194**: p. 117-29.
149. Chiossi, J.S.C. and M.A. Hyppolito, *Effects of residual hearing on cochlear implant outcomes in children: A systematic-review.* Int J Pediatr Otorhinolaryngol, 2017. **100**: p. 119-127.
150. Lustig, L.R. and O. Akil, *Cochlear gene therapy.* Curr Opin Neurol, 2012. **25**(1): p. 57-60.
151. Mittal, R., et al., *Recent Advancements in the Regeneration of Auditory Hair Cells and Hearing Restoration.* Front Mol Neurosci, 2017. **10**: p. 236.
152. Ahmed, H., O. Shubina-Oleinik, and J.R. Holt, *Emerging Gene Therapies for Genetic Hearing Loss.* J Assoc Res Otolaryngol, 2017. **18**(5): p. 649-670.
153. Carpena, N.T. and M.Y. Lee, *Genetic Hearing Loss and Gene Therapy.* Genomics Inform, 2018. **16**(4): p. e20.
154. Ginn, S.L., et al., *Gene therapy clinical trials worldwide to 2017: An update.* J Gene Med, 2018. **20**(5): p. e3015.
155. Kelsell, D.P., et al., *Connexin 26 mutations in hereditary non-syndromic sensorineural deafness.* Nature, 1997. **387**(6628): p. 80-3.
156. Goodenough, D.A. and D.L. Paul, *Beyond the gap: functions of unpaired connexon channels.* Nat Rev Mol Cell Biol, 2003. **4**(4): p. 285-94.
157. Berger, A.C., et al., *Mutations in Cx30 that are linked to skin disease and non-syndromic hearing loss exhibit several distinct cellular pathologies.* J Cell Sci, 2014. **127**(Pt 8): p. 1751-64.

158. Wingard, J.C. and H.B. Zhao, *Cellular and Deafness Mechanisms Underlying Connexin Mutation-Induced Hearing Loss - A Common Hereditary Deafness.* Front Cell Neurosci, 2015. **9**: p. 202.
159. Morlet, T., et al., *A homozygous SLITRK6 nonsense mutation is associated with progressive auditory neuropathy in humans.* Laryngoscope, 2014. **124**(3): p. E95-103.
160. Zhu, Y., et al., *Connexin26 (GJB2) deficiency reduces active cochlear amplification leading to late-onset hearing loss.* Neuroscience, 2015. **284**: p. 719-29.
161. Chen, J., et al., *Deafness induced by Connexin 26 (GJB2) deficiency is not determined by endocochlear potential (EP) reduction but is associated with cochlear developmental disorders.* Biochem Biophys Res Commun, 2014. **448**(1): p. 28-32.
162. Mei, L., et al., *A deafness mechanism of digenic Cx26 (GJB2) and Cx30 (GJB6) mutations: Reduction of endocochlear potential by impairment of heterogeneous gap junctional function in the cochlear lateral wall.* Neurobiol Dis, 2017. **108**: p. 195-203.
163. Snoeckx, R.L., et al., *GJB2 mutations and degree of hearing loss: a multicenter study.* Am J Hum Genet, 2005. **77**(6): p. 945-57.
164. Lebeko, K., et al., *Genetics of hearing loss in Africans: use of next generation sequencing is the best way forward.* Pan Afr Med J, 2015. **20**: p. 383.
165. Gasparini, P., et al., *High carrier frequency of the 35delG deafness mutation in European populations. Genetic Analysis Consortium of GJB2 35delG.* Eur J Hum Genet, 2000. **8**(1): p. 19-23.
166. Leclere, J.C., et al., *GJB2 mutations: Genotypic and phenotypic correlation in a cohort of 690 hearing-impaired patients, toward a new mutation?* Int J Pediatr Otorhinolaryngol, 2017. **102**: p. 80-85.
167. Rabionet, R., P. Gasparini, and X. Estivill, *Molecular genetics of hearing impairment due to mutations in gap junction genes encoding beta connexins.* Human Mutation, 2000. **16**(3): p. 190-202.
168. Norouzi, V., et al., *Did the GJB2 35delG mutation originate in Iran?* Am J Med Genet A, 2011. **155A**(10): p. 2453-8.
169. Rothrock, C.R., et al., *Connexin 26 35delG does not represent a mutational hotspot.* Hum Genet, 2003. **113**(1): p. 18-23.
170. Ohtsuka, A., et al., *GJB2 deafness gene shows a specific spectrum of mutations in Japan, including a frequent founder mutation.* Hum Genet, 2003. **112**(4): p. 329-33.
171. Morell, R.J., et al., *Mutations in the connexin 26 gene (GJB2) among Ashkenazi Jews with nonsyndromic recessive deafness.* N Engl J Med, 1998. **339**(21): p. 1500-5.
172. Chan, D.K. and K.W. Chang, *GJB2-associated hearing loss: systematic review of worldwide prevalence, genotype, and auditory phenotype.* Laryngoscope, 2014. **124**(2): p. E34-53.
173. Maheshwari, M., et al., *Screening of families with autosomal recessive non-syndromic hearing impairment (ARNSHI) for mutations in GJB2 gene: Indian scenario.* Am J Med Genet A, 2003. **120A**(2): p. 180-4.
174. Tekin, M., et al., *GJB2 mutations in Mongolia: complex alleles, low frequency, and reduced fitness of the deaf.* Ann Hum Genet, 2010. **74**(2): p. 155-64.

175. Bliznetz, E.A., et al., *Update of the GJB2/DFNB1 mutation spectrum in Russia: a founder Ingush mutation del(GJB2-D13S175) is the most frequent among other large deletions.* J Hum Genet, 2017. **62**(8): p. 789-795.
176. Cryns, K., et al., *A genotype-phenotype correlation for GJB2 (connexin 26) deafness.* J Med Genet, 2004. **41**(3): p. 147-54.
177. Baple, E.L., et al., *Hypomorphic PCNA mutation underlies a human DNA repair disorder.* J Clin Invest, 2014. **124**(7): p. 3137-46.
178. Tekin, M., et al., *SLITRK6 mutations cause myopia and deafness in humans and mice.* J Clin Invest, 2013. **123**(5): p. 2094-102.
179. Puffenberger, E.G., et al., *Genetic mapping and exome sequencing identify variants associated with five novel diseases.* PLoS One, 2012. **7**(1): p. e28936.
180. Williams, K. and M. Demczko, *Severe Multi-organ Dysfunction Due to a Novel YARS Mutation.* 2017, Clinic For Special Children: Lancaster County, Pennsylvania.
181. Chen, Q., et al., *Homozygous deletion in KVLQT1 associated with Jervell and Lange-Nielsen syndrome.* Circulation, 1999. **99**(10): p. 1344-7.
182. Yoshikawa, M., et al., *Ganglioside GM3 is essential for the structural integrity and function of cochlear hair cells.* Hum Mol Genet, 2015. **24**(10): p. 2796-807.
183. Shebib, S.M., et al., *Newly recognized syndrome of cerebral, ocular, dental, auricular, skeletal anomalies: CODAS syndrome--a case report.* Am J Med Genet, 1991. **40**(1): p. 88-93.
184. Muggenthaler, M.M., et al., *Mutations in HYAL2, Encoding Hyaluronidase 2, Cause a Syndrome of Orofacial Clefting and Cor Triatriatum Sinister in Humans and Mice.* PLoS Genet, 2017. **13**(1): p. e1006470.
185. Daley, E., et al., *Variable bone fragility associated with an Amish COL1A2 variant and a knock-in mouse model.* J Bone Miner Res, 2010. **25**(2): p. 247-61.
186. McBride, D.J., et al., *Variable expressivity of a COL1A2 gly-610-cys mutation in a large Amish pedigree.* American Journal of Human Genetics, 2002. **71**(4): p. 351-351.
187. Krumenacker, S., *Hearing Aid Dispensing Training Manual*. reprint ed. 2013: Plural Publishing.
188. Steinman, K.J., et al., *16p11.2 deletion and duplication: Characterizing neurologic phenotypes in a large clinically ascertained cohort.* Am J Med Genet A, 2016. **170**(11): p. 2943-2955.
189. Carrasquillo, M.M., et al., *Two different connexin 26 mutations in an inbred kindred segregating non-syndromic recessive deafness: implications for genetic studies in isolated populations.* Hum Mol Genet, 1997. **6**(12): p. 2163-72.
190. Verri, T., et al., *Di- and tripeptide transport in vertebrates: the contribution of teleost fish models.* J Comp Physiol B, 2017. **187**(3): p. 395-462.
191. Romano, A., et al., *High-affinity peptide transporter PEPT2 (SLC15A2) of the zebrafish Danio rerio: functional properties, genomic organization, and expression analysis.* Physiol Genomics, 2006. **24**(3): p. 207-17.
192. Jochems, P.G.M., et al., *Evaluating Human Intestinal Cell Lines for Studying Dietary Protein Absorption.* Nutrients, 2018. **10**(3).
193. Smith, D.E., B. Clemencon, and M.A. Hediger, *Proton-coupled oligopeptide transporter family SLC15: physiological, pharmacological*

and pathological implications. Mol Aspects Med, 2013. **34**(2-3): p. 323-36.
194. Sreedharan, S., et al., *Long evolutionary conservation and considerable tissue specificity of several atypical solute carrier transporters.* Gene, 2011. **478**(1-2): p. 11-8.
195. Bucci, C., et al., *Rab7: a key to lysosome biogenesis.* Mol Biol Cell, 2000. **11**(2): p. 467-80.
196. Kenneson, A., K. Van Naarden Braun, and C. Boyle, *GJB2 (connexin 26) variants and nonsyndromic sensorineural hearing loss: a HuGE review.* Genet Med, 2002. **4**(4): p. 258-74.
197. Indellicato, R., et al., *Total loss of GM3 synthase activity by a normally processed enzyme in a novel variant and in all ST3GAL5 variants reported to cause a distinct congenital disorder of glycosylation.* Glycobiology, 2018.
198. Najmabadi, H., et al., *Deep sequencing reveals 50 novel genes for recessive cognitive disorders.* Nature, 2011. **478**(7367): p. 57-63.
199. Riazuddin, S., et al., *Exome sequencing of Pakistani consanguineous families identifies 30 novel candidate genes for recessive intellectual disability.* Mol Psychiatry, 2017. **22**(11): p. 1604-1614.
200. Mithyantha, R., et al., *Current evidence-based recommendations on investigating children with global developmental delay.* Arch Dis Child, 2017. **102**(11): p. 1071-1076.
201. Eun, S.H. and S.H. Hahn, *Metabolic evaluation of children with global developmental delay.* Korean J Pediatr, 2015. **58**(4): p. 117-22.
202. Afroze, B. and B. Chaudhry, *Genetics of non-syndromic autosomal recessive mental retardation.* J Pak Med Assoc, 2013. **63**(1): p. 106-10.
203. Filges, I., et al., *Brain MRI abnormalities and spectrum of neurological and clinical findings in three patients with proximal 16p11.2 microduplication.* Am J Med Genet A, 2014. **164A**(8): p. 2003-12.
204. Miller, D.T., et al., *16p11.2 Recurrent Microdeletion*, in *GeneReviews((R))*, M.P. Adam, et al., Editors. 1993, University of Washington, Seattle
University of Washington, Seattle. GeneReviews is a registered trademark of the University of Washington, Seattle. All rights reserved.: Seattle (WA).
205. Berman, J.I., et al., *Abnormal auditory and language pathways in children with 16p11.2 deletion.* Neuroimage Clin, 2015. **9**: p. 50-7.
206. Chong, J.X., et al., *A population-based study of autosomal-recessive disease-causing mutations in a founder population.* Am J Hum Genet, 2012. **91**(4): p. 608-20.
207. Strauss, K.A. and E.G. Puffenberger, *Genetics, medicine, and the Plain people.* Annu Rev Genomics Hum Genet, 2009. **10**: p. 513-36.
208. Van Laer, L., et al., *A common founder for the 35delG GJB2 gene mutation in connexin 26 hearing impairment.* J Med Genet, 2001. **38**(8): p. 515-8.
209. Imtiaz, F., et al., *A comprehensive introduction to the genetic basis of non-syndromic hearing loss in the Saudi Arabian population.* BMC Med Genet, 2011. **12**: p. 91.
210. Tsukada, K., et al., *Ethnic-specific spectrum of GJB2 and SLC26A4 mutations: their origin and a literature review.* Ann Otol Rhinol Laryngol, 2015. **124 Suppl 1**: p. 61S-76S.

211. Scheffer, D.I., et al., *Gene Expression by Mouse Inner Ear Hair Cells during Development.* J Neurosci, 2015. **35**(16): p. 6366-80.
212. Hwang, J. and Y.K. Kim, *When a ribosome encounters a premature termination codon.* BMB Rep, 2013. **46**: p. 9-16.
213. Anderle, P., et al., *Genetic variants of the human dipeptide transporter PEPT1.* J Pharmacol Exp Ther, 2006. **316**(2): p. 636-46.
214. Zhou, Q., et al., *Evidence of genetic heterogeneity in Alberta Hutterites with Usher syndrome type I.* Mol Vis, 2012. **18**: p. 1379-83.
215. Bader, P.I., et al., *Infantile refsum disease in four Amish sibs.* Am J Med Genet, 2000. **90**(2): p. 110-4.
216. Ghaloul-Gonzalez, L., et al., *Mitochondrial respiratory chain disorders in the Old Order Amish population.* Mol Genet Metab, 2016. **118**(4): p. 296-303.
217. Alagramam, K.N., et al., *Mutations in the novel protocadherin PCDH15 cause Usher syndrome type 1F.* Hum Mol Genet, 2001. **10**(16): p. 1709-18.
218. Connolly, M.B., et al., *Hepatic dysfunction in Alstrom disease.* Am J Med Genet, 1991. **40**(4): p. 421-4.
219. Cruz-Aguilar, M., et al., *A Nonsense ALMS1 Mutation Underlies Alstrom Syndrome in an Extended Mennonite Kindred Settled in North Mexico.* Genet Test Mol Biomarkers, 2017. **21**(6): p. 397-401.
220. Puffenberger, E.G., et al., *A missense mutation of the endothelin-B receptor gene in multigenic Hirschsprung's disease.* Cell, 1994. **79**(7): p. 1257-66.
221. Vasudevan, P. and M. Suri, *A clinical approach to developmental delay and intellectual disability.* Clinical Medicine, 2017. **17**(6): p. 558-561.
222. Bradshaw, R.A. and E.A. Dennis, *Cell Signaling: Yesterday, Today, and Tomorrow*. Handbook of Cell Signaling, Three, ed. Elsevier. Vol. 2. 2010.
223. Derynck, R. and E.H. Budi, *Specificity, versatility, and control of TGF-b family signaling.* Sci. Signal. 1, 2019. **12**(570).
224. Horbelt, D., A. Denkis, and P. Knaus, *A portrait of Transforming Growth Factor beta superfamily signalling: Background matters.* Int J Biochem Cell Biol, 2012. **44**(3): p. 469-74.
225. Hinck, A.P., T.D. Mueller, and T.A. Springer, *Structural Biology and Evolution of the TGF-beta Family.* Cold Spring Harb Perspect Biol, 2016. **8**(12).
226. Rol, N., et al., *TGF-beta and BMPR2 Signaling in PAH: Two Black Sheep in One Family.* Int J Mol Sci, 2018. **19**(9).
227. Feng, X.H. and R. Derynck, *Specificity and versatility in tgf-beta signaling through Smads.* Annu Rev Cell Dev Biol, 2005. **21**: p. 659-93.
228. Rybska, M., et al., *Transforming growth factor (TGF) – is it a key protein in mammalian reproductive biology?* Medical Journal of Cell Biology, 2018. **6**(3): p. 125-130.
229. Drabsch, Y. and P. ten Dijke, *TGF-beta signalling and its role in cancer progression and metastasis.* Cancer Metastasis Rev, 2012. **31**(3-4): p. 553-68.
230. Kubiczkova, L., et al., *TGF-β – an excellent servant but a bad master.* Journal of Translational Medicine, 2012. **10**(183).
231. Kang, J.S., C. Liu, and R. Derynck, *New regulatory mechanisms of TGF-beta receptor function.* Trends Cell Biol, 2009. **19**(8): p. 385-94.

232. Massague, J., J. Seoane, and D. Wotton, *Smad transcription factors.* Genes Dev, 2005. **19**(23): p. 2783-810.
233. Attisano, L. and S.T. Lee-Hoeflich, *The Smads.* Genome Biology, 2001. **2**.
234. Samanta, D. and P.K. Datta, *Alterations in the Smad pathway in human cancers.* Front Biosci (Landmark Ed), 2012: p. 1281-93.
235. Jiang, W.G., et al., *Tissue invasion and metastasis: Molecular, biological and clinical perspectives.* Semin Cancer Biol, 2015. **35 Suppl**: p. S244-S275.
236. Attisano, L. and J.L. Wrana, *Signal transduction by the TFF-beta superfamily.* Science, 2002. **296**: p. 1646-1647.
237. Chen, L., et al., *Central role of dysregulation of TGF-beta/Smad in CKD progression and potential targets of its treatment.* Biomed Pharmacother, 2018. **101**: p. 670-681.
238. Kim, R.H., et al., *SNIP1 inhibits NF-kappa B signaling by competing for its binding to the C/H1 domain of CBP/p300 transcriptional co-activators.* J Biol Chem, 2001. **276**(49): p. 46297-304.
239. Liu, S., et al., *SUMO Modification Reverses Inhibitory Effects of Smad Nuclear Interacting Protein-1 in TGF-beta Responses.* J Biol Chem, 2016. **291**(47): p. 24418-24430.
240. Fujii, M., et al., *SNIP1 is a candidate modifier of the transcriptional activity of c-Myc on E box-dependent target genes.* Mol Cell, 2006. **24**(5): p. 771-783.
241. Kim, R.H., et al., *A novel smad nuclear interacting protein, SNIP1, suppresses p300-dependent TGF-beta signal transduction.* Genes Dev, 2000. **14**(13): p. 1605-16.
242. Roche, K.C., et al., *The FHA domain protein SNIP1 is a regulator of the cell cycle and cyclin D1 expression.* Oncogene, 2004. **23**(50): p. 8185-95.
243. Chen, E.Y., et al., *Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool.* BMC Bioinformatics, 2013. **14**.
244. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.* Nucleic Acids Res, 2009. **37**(1): p. 1-13.
245. Maia, G.H., et al., *Serotonin depletion increases seizure susceptibility and worsens neuropathological outcomes in kainate model of epilepsy.* Brain Res Bull, 2017. **134**: p. 109-120.
246. Galovic, M. and M. Koepp, *Advances of Molecular Imaging in Epilepsy.* Curr Neurol Neurosci Rep, 2016. **16**(6): p. 58.
247. Takenouchi, T., et al., *1p34.3 deletion involving GRIK3: Further clinical implication of GRIK family glutamate receptors in the pathogenesis of developmental delay.* Am J Med Genet A, 2014. **164A**(2): p. 456-60.
248. Boroveck, F., et al., *Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease.* PNAS, 2005. **102**: p. 11023–11028.
249. Tylee, D.S., D.M. Kawaguchi, and S.J. Glatt, *On the outside, looking in: a review and evaluation of the comparability of blood and brain "-omes".* Am J Med Genet B Neuropsychiatr Genet, 2013. **162B**(7): p. 595-603.
250. Consortium, G.T., *The Genotype-Tissue Expression (GTEx) project.* Nat Genet, 2013. **45**(6): p. 580-5.

251. Wagner, G.P., K. Kin, and V.J. Lynch, *A model based criterion for gene expression calls using RNA-seq data.* Theory in Biosciences, 2013. **132**(3): p. 159-164.
252. Johnson, E.L., *Seizures and Epilepsy.* Med Clin North Am, 2019. **103**(2): p. 309-324.
253. Chang, R.S., et al., *Classifications of seizures and epilepsies, where are we? - A brief historical review and update.* J Formos Med Assoc, 2017. **116**(10): p. 736-741.
254. Giovedi, S., et al., *Involvement of synaptic genes in the pathogenesis of autism spectrum disorders: the case of synapsins.* Front Pediatr, 2014. **2**: p. 94.
255. Hosaka, M. and T.C. Sudhof, *Synapsin III, a Novel Synapsin with an Unusual Regulation by Ca2□*.* THE JOURNAL OF BIOLOGICAL CHEMISTRY, 1998. **273**(22): p. 13371–13374.
256. Baker, K., et al., *SYT1-associated neurodevelopmental disorder: a case series.* Brain, 2018. **141**(9): p. 2576-2591.
257. Osimo, E.F., et al., *Synaptic loss in schizophrenia: a meta-analysis and systematic review of synaptic protein and mRNA measures.* Mol Psychiatry, 2019. **24**(4): p. 549-561.
258. Bornschein, G. and H. Schmidt, *Synaptotagmin Ca(2+) Sensors and Their Spatial Coupling to Presynaptic Cav Channels in Central Cortical Synapses.* Front Mol Neurosci, 2018. **11**: p. 494.
259. Lai, A.L., et al., *Synaptotagmin 1 and SNAREs form a complex that is structurally heterogeneous.* J Mol Biol, 2011. **405**(3): p. 696-706.
260. Brachya, G., C. Yanay, and M. Linial, *Synaptic proteins as multi-sensor devices of neurotransmission.* BMC Neurosci, 2006. **7 Suppl 1**: p. S4.
261. Zhou, H., et al., *Structural and Functional Analysis of the CAPS SNARE-Binding Domain Required for SNARE Complex Formation and Exocytosis.* Cell Rep, 2019. **26**(12): p. 3347-3359 e6.
262. Carr, C.M. and M. Munson, *Tag team action at the synapse.* EMBO Rep, 2007. **8**(9): p. 834-8.
263. Sudhof, T.C., *The presynaptic active zone.* Neuron, 2012. **75**(1): p. 11-25.
264. Fenner, B.J., M. Scannell, and J.H. Prehn, *Expanding the substantial interactome of NEMO using protein microarrays.* PLoS One, 2010. **5**(1): p. e8799.
265. Khvotchev, M.V. and J. Sun, *Synapsins*. Encyclopedia of Neuroscience, ed. L.R. Squire. 2009: Academic Press.
266. Mirza, F.J. and S. Zahid, *The Role of Synapsins in Neurological Disorders.* Neurosci Bull, 2018. **34**(2): p. 349-358.
267. Sudhof, T.C., et al., *Synapsins: Mosaics of Shared and Individual Domains in a Family of Synaptic Vesicle Phosphoproteins.* Science. , 1989. **245**: p. 1474-80.
268. Cesca, F., et al., *The synapsins: key actors of synapse function and plasticity.* Prog Neurobiol, 2010. **91**(4): p. 313-48.
269. Fassio, A., et al., *SYN1 loss-of-function mutations in autism and partial epilepsy cause impaired synaptic function.* Hum Mol Genet, 2011. **20**(12): p. 2297-307.
270. Prasad, D.K., et al., *Association of GABRA6 1519 T>C (rs3219151) and Synapsin II (rs37733634) gene polymorphisms with the development of idiopathic generalized epilepsy.* Epilepsy Res, 2014. **108**(8): p. 1267-73.

271. Oti, M., et al., *Predicting disease genes using protein-protein interactions.* J Med Genet, 2006. **43**(8): p. 691-8.
272. Zaltieri, M., et al., *alpha-synuclein and synapsin III cooperatively regulate synaptic function in dopamine neurons.* J Cell Sci, 2015. **128**(13): p. 2231-43.
273. Meurs, A., et al., *Seizure activity and changes in hippocampal extracellular glutamate, GABA, dopamine and serotonin.* Epilepsy Res, 2008. **78**(1): p. 50-9.
274. Richerson, G.B. and G.F. Buchanan, *The serotonin axis: Shared mechanisms in seizures, depression, and SUDEP.* Epilepsia, 2011. **52 Suppl 1**: p. 28-38.
275. Chen, C.P., et al., *Partial monosomy 3p (3p26.2 --> pter) and partial trisomy 5q (5q34 --> qter) in a girl with coarctation of the aorta, congenital heart defects, short stature, microcephaly and developmental delay.* Genet Couns, 2012. **23**(3): p. 405-13.
276. Tassano, E., et al., *Heterozygous deletion of CHL1 gene: detailed array-CGH and clinical characterization of a new case and review of the literature.* Eur J Med Genet, 2014. **57**(11-12): p. 626-9.
277. Cuoco, C., et al., *Microarray based analysis of an inherited terminal 3p26.3 deletion, containing only the CHL1 gene, from a normal father to his two affected children.* Orphanet J Rare Dis, 2011. **6**: p. 12.
278. Dias, A.T., et al., *Post-mortem cytogenomic investigations in patients with congenital malformations.* Exp Mol Pathol, 2016. **101**(1): p. 116-23.
279. Irintchev, A. and M. Schachner, *The injured and regenerating nervous system: immunoglobulin superfamily members as key players.* Neuroscientist, 2012. **18**(5): p. 452-66.
280. Hillenbrand, R., et al., *The close homologue of the neural adhesion molecule L1*
(CHL1): patterns of expression and promotion of neurite
outgrowth by heterophilic interactions. European Journal of Neuroscience, 1999. **11**: p. 813–826.
281. Kidd, T., et al., *Roundabout Controls Axon Crossing*
of the CNS Midline and Defines a Novel Subfamily
of Evolutionarily Conserved Guidance Receptors. Cell, 1998. **92**: p. 205-215.
282. Yadav, S.S. and G. Narayan, *Role of ROBO4 signalling in developmental and pathological angiogenesis.* Biomed Res Int, 2014. **2014**: p. 683025.
283. Ypsilanti, A.R., Y. Zagar, and A. Chedotal, *Moving away from the midline: new developments for Slit and Robo.* Development, 2010. **137**(12): p. 1939-52.
284. Fisher, S.E. and C. Francks, *Genes, cognition and dyslexia: learning to read the genome.* Trends Cogn Sci, 2006. **10**(6): p. 250-7.
285. Kruszka, P., et al., *Loss of function in ROBO1 is associated with tetralogy of Fallot and septal defects.* J Med Genet, 2017. **54**(12): p. 825-829.
286. Kodani, M., et al., *Application of TaqMan low-density arrays for simultaneous detection of multiple respiratory pathogens.* J Clin Microbiol, 2011. **49**(6): p. 2175-82.
287. Kitts, A. and S. Sherry, *The Single Nucleotide Polymorphism Database (dbSNP) of Nucleotide Sequence Variation.*, in *The NCBI Handbook [Internet]*, J. McEntyre and J. Ostell, Editors. 2002, National Center for Biotechnology Information (US): Bethesda (MD).

288. Hug, N., D. Longman, and J.F. Caceres, *Mechanism and regulation of the nonsense-mediated decay pathway.* Nucleic Acids Res, 2016. **44**(4): p. 1483-95.
289. Bahcall, O.G., *Genetic testing. ACMG guides on the interpretation of sequence variants.* Nat Rev Genet, 2015. **16**(5): p. 256-7.
290. Rivera-Munoz, E.A., et al., *ClinGen Variant Curation Expert Panel experiences and standardized processes for disease and gene-level specification of the ACMG/AMP guidelines for sequence variant interpretation.* Hum Mutat, 2018. **39**(11): p. 1614-1622.
291. Phillips, C., *Online resources for SNP analysis: a review and route map.* Mol Biotechnol, 2007. **35**(1): p. 65-97.
292. Sherry, S.T. and K. Sirotkin, *dbSNP—Database for Single Nucleotide Polymorphisms*
and Other Classes of Minor Genetic Variation. Genome Research, 1999. **9**: p. 667-679.
293. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation.* Nucleic Acids Research, 2001. **20**: p. 308–311.
294. Ismail, S. and M. Essawi, *Genetic polymorphism studies in humans.* Middle East Journal of Medical Genetics, 2012. **1**(2): p. 57-63.
295. Lek, M., et al., *Analysis of protein-coding genetic variation in 60,706 humans.* Nature, 2016. **536**(7616): p. 285-91.
296. Landrum, M.J., et al., *ClinVar: public archive of relationships among sequence variation and human phenotype.* Nucleic Acids Res, 2014. **42**(Database issue): p. D980-5.
297. Stenson, P.D., et al., *Human Gene Mutation Database (HGMD): 2003 update.* Hum Mutat, 2003. **21**(6): p. 577-81.
298. Stenson, P.D., et al., *The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine.* Hum Genet, 2014. **133**(1): p. 1-9.
299. Landrum, M.J., et al., *ClinVar: improving access to variant interpretations and supporting evidence.* Nucleic Acids Res, 2018. **46**(D1): p. D1062-D1067.
300. Landrum, M.J. and B.L. Kattman, *ClinVar at five years: Delivering on the promise.* Hum Mutat, 2018. **39**(11): p. 1623-1630.
301. Harrison, S.M., et al., *Using ClinVar as a Resource to Support Variant Interpretation.* Curr Protoc Hum Genet, 2016. **89**: p. 8 16 1-8 16 23.
302. Forbes, S.A., et al., *COSMIC: somatic cancer genetics at high-resolution.* Nucleic Acids Res, 2017. **45**(D1): p. D777-D783.
303. Stenson, P.D., et al., *The Human Gene Mutation Database: 2008 update.* Genome Med, 2009. **1**(1): p. 13.
304. Krawczak, M., et al., *Human gene mutation database-a biomedical information and research resource.* Hum Mutat, 2000. **15**(1): p. 45-51.
305. Stenson, P.D., et al., *The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies.* Hum Genet, 2017. **136**(6): p. 665-677.
306. Posey, J.E., et al., *Insights into genetics, human biology and disease gleaned from family based genomic studies.* Genet Med, 2019.
307. Philippakis, A.A., et al., *The Matchmaker Exchange: a platform for rare disease gene discovery.* Hum Mutat, 2015. **36**(10): p. 915-21.

308. Bruel, A.L., et al., *2.5 years' experience of GeneMatcher data-sharing: a powerful tool for identifying new genes responsible for rare diseases.* Genet Med, 2018.
309. Wahid, S., S. Aslam, and S. Minhas, *Ellis-Van Creveld Syndrome in a Neonate.* J Coll Physicians Surg Pak, 2018. **28**(3): p. S44-s45.
310. McKusick, V.A., *Ellis-van Creveld syndrome and the Amish.* Nature Genetics, 2000. **24**: p. 203-204.
311. Robinson, J.T., et al., *Integrative genomics viewer.* Nat Biotechnol, 2011. **29**(1): p. 24-6.
312. Hong, H.M., et al., *A novel mutation in the connexin 29 gene may contribute to nonsyndromic hearing loss.* Hum Genet, 2010. **127**(2): p. 191-9.
313. Su, C.C., et al., *Mechanism of two novel human GJC3 missense mutations in causing non-syndromic hearing loss.* Cell Biochem Biophys, 2013. **66**(2): p. 277-86.
314. Feroze, K.B. and B.C. Patel, *Buphthalmos [Updated 2019 Jan 12]*, S. Publishing, Editor. 2019, StatPearls [Internet].
315. Salviati, L., E. Trevisson, and M. Doimo, *Primary Coenzyme Q10 Deficiency*. GeneReviews [Internet], ed. H.H. Ardinger and R.A. Pagon. 2017, Seattle (WA): University of Washington, Seattle;.
316. Fabie, N.A.V., K.B. Pappas, and G.L. Feldman, *The Current State of Newborn Screening in the United States.* Pediatr Clin North Am, 2019. **66**(2): p. 369-386.
317. Kanungo, S., et al., *Newborn screening and changing face of inborn errors of metabolism in the United States.* Ann Transl Med, 2018. **6**(24): p. 468.
318. Sieren, S., et al., *Cross-Sectional Survey on Newborn Screening in Wisconsin Amish and Mennonite Communities.* J Community Health, 2016. **41**(2): p. 282-8.
319. Posey, J.E., et al., *Resolution of Disease Phenotypes Resulting from Multilocus Genomic Variation.* N Engl J Med, 2017. **376**(1): p. 21-31.
320. Armour, C.M., et al., *Syndrome disintegration: Exome sequencing reveals that Fitzsimmons syndrome is a co-occurrence of multiple events.* Am J Med Genet A, 2016. **170**(7): p. 1820-5.
321. Crosby, A.H., et al., *Defective mitochondrial mRNA maturation is associated with spastic ataxia.* Am J Hum Genet, 2010. **87**(5): p. 655-60.
322. Rawlins, L.E., et al., *An Amish founder variant consolidates disruption of CEP55 as a cause of hydranencephaly and renal dysplasia.* European Journal of Human Genetics, 2019.
323. Bondeson, M.L., et al., *A nonsense mutation in CEP55 defines a new locus for a Meckel-like syndrome, an autosomal recessive lethal fetal ciliopathy.* Clin Genet, 2017. **92**(5): p. 510-516.
324. Frosk, P., et al., *A truncating mutation in CEP55 is the likely cause of MARCH, a novel syndrome affecting neuronal mitosis.* J Med Genet, 2017. **54**(7): p. 490-501.
325. Leigh, M.W., *PRIMARY CILIARY DYSKINESIA*. 8th ed. Kendig & Chernick's Disorders of the Respiratory Tract in Children, ed. R.W. Wilmott, et al. 2012: Saunders, W. B.
326. Ta-Shma, A., et al., *Homozygous loss-of-function mutations in MNS1 cause laterality defects and likely male infertility.* PLoS Genet, 2018. **14**(8): p. e1007602.

327. Ridanpaa, M., et al., *The major mutation in the RMRP gene causing CHH among the Amish is the same as that found in most Finnish cases.* Am J Med Genet C Semin Med Genet, 2003. **121C**(1): p. 81-3.
328. Riley, P., Jr., et al., *Cartilage hair hypoplasia: characteristics and orthopaedic manifestations.* J Child Orthop, 2015. **9**(2): p. 145-52.
329. Taskinen, M. and O. Makitie, *[Cartilage-hair hypoplasia--much more than growth problem].* Duodecim, 2011. **127**(3): p. 273-9.
330. Polmar, S.H. and G.F. Pierce, *Cartilage hair hypoplasia: immunological aspects and their clinical implications.* Clin Immunol Immunopathol, 1986. **40**(1): p. 87-93.
331. Mäkitie, O. and S. Vakkilainen, *Cartilage-Hair Hypoplasia – Anauxetic Dysplasia Spectrum Disorders*. GeneReviews® [Internet], ed. M.P. Adam, H.H. Ardinger, and R.A. Pagon. 2018, Seattle (WA): University of Washington, Seattle.
332. IDF, *Newborn Screening: Chapter 26*. 5th ed. IDF Patient & Family Handbook or Primary Immunodeficiency Diseases,, ed. R.M. Blaese, et al. 2013, Towson, MD: Immune Deficiency Foundation, USa.
333. Grunert, S.C., et al., *Propionic acidemia: neonatal versus selective metabolic screening.* J Inherit Metab Dis, 2012. **35**(1): p. 41-9.
334. Wongkittichote, P., N. Ah Mew, and K.A. Chapman, *Propionyl-CoA carboxylase - A review.* Mol Genet Metab, 2017. **122**(4): p. 145-152.
335. Jurecki, E., et al., *Nutrition management guideline for propionic acidemia: An evidence- and consensus-based approach.* Mol Genet Metab, 2019.
336. Calore, C., et al., *A founder MYBPC3 mutation results in HCM with a high risk of sudden death after the fourth decade of life.* J Med Genet, 2015. **52**(5): p. 338-47.
337. Cirino, A.L. and C. Ho, *Hypertrophic Cardiomyopathy*. GeneReviews® [Internet], ed. M.P. Adam, H.H. Ardinger, and R.A. Pagon. 2008, Seattle (WA): University of Washington, Seattle.
338. Zahka, K., et al., *Homozygous mutation of MYBPC3 associated with severe infantile hypertrophic cardiomyopathy at high frequency among the Amish.* Heart, 2008. **94**(10): p. 1326-30.
339. Maron, B.J. and M.S. Maron, *Hypertrophic cardiomyopathy.* The Lancet, 2013. **381**(9862): p. 242-255.
340. Vears, D.F. and S.A. Metcalfe, *Carrier testing in children and adolescents.* Eur J Med Genet, 2015. **58**(12): p. 659-67.
341. Raz, A.E. and Y. Vizner, *Carrier matching and collective socialization in community genetics: Dor Yeshorim and the reinforcement of stigma.* Soc Sci Med, 2008. **67**(9): p. 1361-9.
342. Travali, S., et al., *Structure of the human gene for the proliferating cell nuclear antigen.* J Biol Chem, 1989. **264**(13): p. 7466-72.
343. Shivji, K.K., M.K. Kenny, and R.D. Wood, *Proliferating cell nuclear antigen is required for DNA excision repair.* Cell, 1992. **69**(2): p. 367-74.
344. Moldovan, G.L., B. Pfander, and S. Jentsch, *PCNA, the maestro of the replication fork.* Cell, 2007. **129**(4): p. 665-79.
345. Wilson, R.H., et al., *PCNA dependent cellular activities tolerate dramatic perturbations in PCNA client interactions.* DNA Repair (Amst), 2017. **50**: p. 22-35.
346. Aruga, J., N. Yokota, and K. Mikoshiba, *Human SLITRK family genes: genomic organization and expression profiling in normal brain and brain tumor tissue.* Gene, 2003. **315**: p. 87-94.

347. O'Hanlon, T.P. and F.W. Miller, *Genomic organization, transcriptional mapping, and evolutionary implications of the human bi-directional histidyl-tRNA synthetase locus (HARS/HARSL).* Biochem Biophys Res Commun, 2002. **294**(3): p. 609-14.
348. O'Neill, M.J.F. *USHER SYNDROME, TYPE IIIB; USH3B*. OMIM 2012 [cited 2019; Available from: https://mirror.omim.org/entry/614504.
349. Jordanova, A., et al., *Disrupted function and axonal distribution of mutant tyrosyl-tRNA synthetase in dominant intermediate Charcot-Marie-Tooth neuropathy.* Nat Genet, 2006. **38**(2): p. 197-202.
350. Nowaczyk, M.J., et al., *A novel multisystem disease associated with recessive mutations in the tyrosyl-tRNA synthetase (YARS) gene.* Am J Med Genet A, 2017. **173**(1): p. 126-134.
351. Ishii, A., et al., *Expression cloning and functional characterization of human cDNA for ganglioside GM3 synthase.* J Biol Chem, 1998. **273**(48): p. 31652-5.
352. Zeng, G., et al., *Characterization of the 5'-flanking fragment of the human GM3-synthase gene.* Biochim Biophys Acta, 2003. **1625**(1): p. 30-5.
353. Wang, H., et al., *Cutaneous dyspigmentation in patients with ganglioside GM3 synthase deficiency.* Am J Med Genet A, 2013. **161A**(4): p. 875-9.
354. Farukhi, F., et al., *Etiology of vision loss in ganglioside GM3 synthase deficiency.* Ophthalmic Genet, 2006. **27**(3): p. 89-91.
355. Inokuchi, J.I., et al., *Gangliosides and hearing.* Biochim Biophys Acta Gen Subj, 2017. **1861**(10): p. 2485-2493.
356. Fragaki, K., et al., *Refractory epilepsy and mitochondrial dysfunction due to GM3 synthase deficiency.* Eur J Hum Genet, 2013. **21**(5): p. 528-34.
357. Boccuto, L., et al., *A mutation in a ganglioside biosynthetic enzyme, ST3GAL5, results in salt & pepper syndrome, a neurocutaneous disorder with altered glycolipid and glycoprotein glycosylation.* Hum Mol Genet, 2014. **23**(2): p. 418-33.
358. Lee, J.S., et al., *GM3 synthase deficiency due to ST3GAL5 variants in two Korean female siblings: Masquerading as Rett syndrome-like phenotype.* Am J Med Genet A, 2016. **170**(8): p. 2200-5.
359. Bota, D.A. and K.J. Davies, *Lon protease preferentially degrades oxidized mitochondrial aconitase by an ATP-stimulated mechanism.* Nat Cell Biol, 2002. **4**(9): p. 674-80.
360. Lu, B., et al., *Roles for the human ATP-dependent Lon protease in mitochondrial DNA maintenance.* J Biol Chem, 2007. **282**(24): p. 17363-74.
361. Wohlever, M.L., T.A. Baker, and R.T. Sauer, *Roles of the N domain of the AAA+ Lon protease in substrate recognition, allosteric regulation and chaperone activity.* Mol Microbiol, 2014. **91**(1): p. 66-78.
362. Korenberg, J.R., et al., *Toward a cDNA map of the human genome.* Genomics, 1995. **29**(2): p. 364-70.
363. Innes, A.M., et al., *Third case of cerebral, ocular, dental, auricular, skeletal anomalies (CODAS) syndrome, further delineating a new malformation syndrome: first report of an affected male and review of literature.* Am J Med Genet, 2001. **102**(1): p. 44-7.
364. Strauss, K.A., et al., *CODAS syndrome is associated with mutations of LONP1, encoding mitochondrial AAA+ Lon protease.* Am J Hum Genet, 2015. **96**(1): p. 121-35.

365. Lepperdinger, G., J. Mullegger, and G. Kreil, *Hyal2--less active, but more versatile?* Matrix Biol, 2001. **20**(8): p. 509-14.
366. Strobl, B., et al., *Structural organization and chromosomal localization of Hyal2, a gene encoding a lysosomal hyaluronidase.* Genomics, 1998. **53**(2): p. 214-9.
367. Steiner, R.D., *COL1A1/2 Osteogenesis Imperfecta.*, ed. GeneReviews®. 2005 [Updated 2019], Seattle (WA): University of Washington, Seattle; 1993-2019.
368. !!! INVALID CITATION !!! {}.