Chapter 3

Exome and targeted-sequencing of families with congenital hypothyroidism

3.1 Introduction

As previously mentioned, congenital hypothyroidism (CH) is a rare condition of thyroid hormone deficiency caused by thyroid hormone production defects or by abnormal embryological development of the thyroid gland [144]. In the previous chapter, I described the eight genes that are known, thus far, to be associated with thyroid hormone production defects, and a cohort of CH patients with structurally normal glands (or *gland-in-situ*) was screened for likely causative mutations within those genes.

In this chapter, I present an exome and targeted-sequencing study of a phenotypically heterogeneous cohort of CH patients. The studied individuals comprise not only the *gland-in-situ* cases for who no causative mutations were identified in Chapter 2, but also patients that suffer from thyroid gland abnormalities or syndromic forms of CH seen in the context of other congenital malformations. I start this chapter by presenting what is currently known about thyroid developmental abnormalities and its genetic causes, and then explain why exome-sequencing analyses of CH phenotypes can be of value.

3.1.1 Thyroid developmental defects

Thyroid developmental malformations are collectively referred to as thyroid dysgenesis (TD), an umbrella term encompassing a spectrum of phenotypes, usually non-syndromic, that result in a gland that is either completely absent (agenesis), underdeveloped (hypoplasia), or located in an unusual position (ectopia) [144]. Ectopia is the commonest phenotype, with patients usually exhibiting sub-lingual glands due to a failure of the thyroid gland to migrate to its proper anatomical location [144].

While *gland-in-situ* CH is generally accepted to be a Mendelian condition, CH due to thyroid dysgenesis is historically thought to occur as a sporadic disorder, and to be caused by nongenetic mechanisms. This stemmed from early observations that ∼98% of cases appeared to be non-familial [74] and from the fact 92% of monozygotic twins were discordant for the phenotype [384]. Yet, germinal genetic defects have been identified during the last few years in around 5% of TD cases [74, 144]. Known defects include mutations in the TSH receptor (*TSHR*) [47, 67], and in all but one of the transcription factors (TFs) that control thyroid gland morphogenesis, including *NKX2-1* [257], *FOXE1* [88] and *PAX8* [301] (**Figure** 3.1). Because *TSHR* is expressed late in thyroid development (**Figure** 3.1), inactivating recessive mutations in this locus result in mild, non-syndromic thyroid hypoplasia [47, 67]. In contrast, mutations in the other three genes lead to the development of several clinically relevant conditions (**Table** 3.1), which represent multisystem phenotypes that are linked to the specific expression of these proteins in multiple tissues of the developing fetus. The thyroid phenotype characteristic of these syndromes is very heterogeneous and has a broad spectrum of expression (i.e. agenesis/hypoplasia/ectopia) even within families [92, 412, 486].

Fundamental insights into the mechanism by which mutations in these TFs affect thyroid organogenesis has been gained through the analysis of knockout mice with targeted disruption of such genes [171]. Earlier studies of *Nkx2.1* and *Pax8* null mice revealed TD pathologies result from the degeneration of thyroid tissue (possibly due to apoptotic mechanisms) following the specification of the gland precursors, or from a complete defective initiation process [137, 248, 328, 377]. Collectively, these and additional studies led to a total of 22 mouse genetic models, in which different types of thyroid malformations are reported alongside extrathyroidal features (**Table** 3.2) [140, 144]. Notably, many of these phenotypes result from inactivation of endodermic genes implicated in thyroid bud formation (*Hoxa3*, *Hoxb3*, *Hoxd3*, *Hoxa5*, *Shh*, *Hes1* and *Isl1*) [70, 305, 306, 327, 518], or of genes implicated in cardiac (*Nkx2.5*, *Hhex*, *Tbx1*, *Fbln1*, and *Chordin*) [115, 139] or musculoskeletal malformations (*Shh* and *Fgf10*) [96, 138].

3.1.2 Arguments for a genetic involvement in TD

The view of TD as a non-genetic disease is gradually changing [59, 373, 397], with several lines of evidence, in addition to the already mentioned genetic defects, indicating that genetic factors are involved in the pathogenesis of TD. First, there is a small (2%) but significant proportion of familial cases, an estimate that is 15-fold greater than the frequency expected based on chance alone [144]. Second, there is a significantly higher number of asymptomatic thyroid abnormalities, especially ectopia, in first-degree relatives of sporadic CH cases compared with the general population (8% vs. 1%) [311]. This observation suggests that severe forms of TD and these mild alterations could originate from the same genetic defects affecting thyroid organogenesis, albeit with incomplete penetrance, as strongly suggested by the observation that *Foxe1* null mice show either ectopy with a very small thyroid or no thyroid at all [143]. Third, in populations where consanguineous unions are common, the incidence of CH is increased [144]. Fourth, discordance of TD in MZ twins may be due to the presence of *de novo* events (SNVs or CNVs) or somatic mutations in one of the children, rather than just environmental effects. Lastly, the significantly higher frequency of extrathyroidal congenital malformations in CH cases than in the general population further point towards genetic (or epigenetic) factors that have yet to be discovered [144].

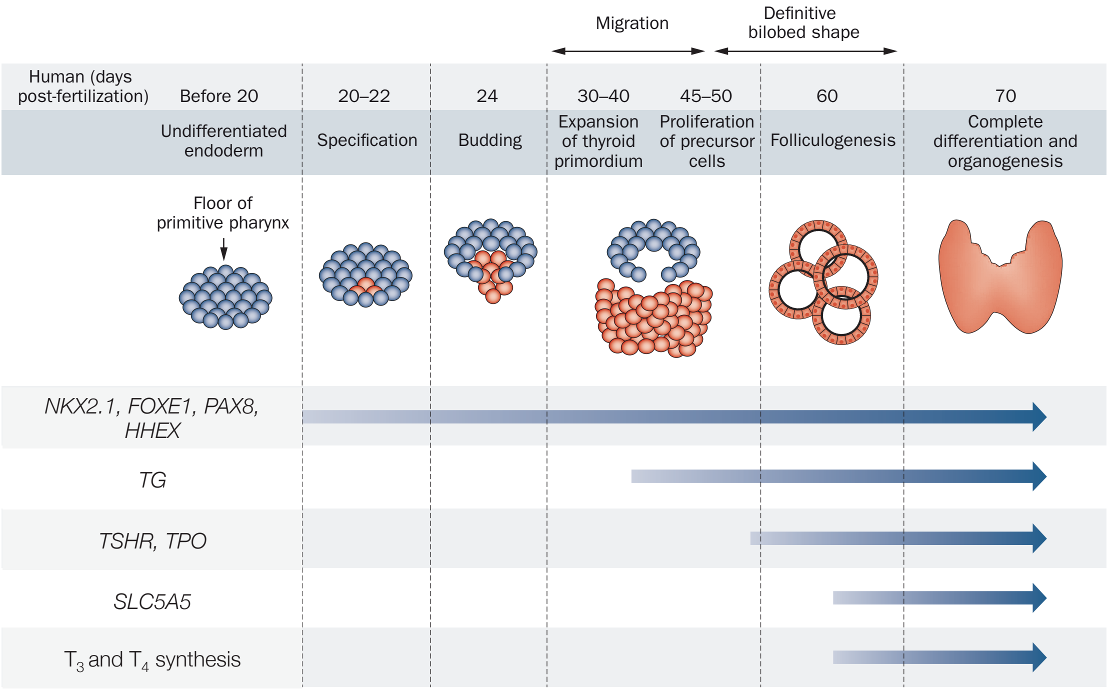
Migration
Definitive bilobed shape
Human (days post-fertilization)
Before 20
20–22
24
30–40
45–50
60
70
Undifferentiated endoderm
Specification
Budding
Expansion of thyroid primordium
Proliferation of precursor cells
Folliculogenesis
Complete differentiation and organogenesis
Floor of primitive pharynx
NKX2.1, FOXE1, PAX8, HHEX
TG
TSHR, TPO
SLC5A5
T3 and T4 synthesis

Figure 3.1 Gene expression during the stages of thyroid gland development in humans.

The thyroid gland is the first endocrine structure to differentiate during fetal development, at approximately 3 to 5 weeks of gestation. It develops initially as an endodermal thickening of the pharyngeal floor, whose cells invaginate and then migrate caudally to their final position in the trachea where they form the follicular cells of the thyroid. The first process of specification is determined by the combined expression of four transcription factors (*NKX2.1*, *FOXE1*, *PAX8* and *HHEX*) which, from that point on, will permanently represent the molecular hallmark of thyroid follicular cells and drive its commitment towards a differentiated thyroid fate. In humans, genetic defects that result in thyroid dysgenesis have been identified in *NKX2.1*, *FOXE1* and *PAX8* but not in *HHEX*. Half-way between the migration and bilobation processes, these four TTFs drive the expression of the genes that are necessary for thyroid hormone production, such as *TG*, *TSHR*, *TPO*, *SLC5A5* and *SLC26A4* (the latter not pictured). Thyroid hormone synthesis in the form of T_3 and T_4 starts only after the gland has been completely formed, at around day 70 post-fertilization. Image adapted from Fernández *et al* [147].

Gene	Location	Non-thyroid expression in adult	Phenotype	Mutation location	Mutation type
NKX2.1	14q13.3	Lung and nervous system.	Normal thyroid/agenesis/hypoplasia/single lobe and/or benign hereditary chorea (movement disorder) and respiratory distress. Brain-lung-thyroid syndrome, OMIM: 610978.	Homeobox or region encoding the transactivation domain.	Heterozygous *de novo* deletions, missense, nonsense and frameshift variants that most often result in haploinsufficiency, and only rarely have a dominant-negative effect on the NKX2.1 wild-type.
FOXE1	9q22.33	Tongue, palate, oesophagus and in ectoderm-derived organs: anterior pituitaty, choanae and hair follicles.	Thyroid agenesis, hypoplasia and/or cleft palate, choanal atresia, bifid epiglottis, spiky hair and tongue-tie. Bamforth-Lazarus syndrome, OMIM: 241850.	Forkhead box.	Homozygous missense mutations that partially or completely impact the capacity of FOXE1 to bind DNA and thus activate transcription.
PAX8	2q14.1	Kidney, excretory system, endometrium, ovary, fallopian tube, pancreatic islet cells and lymphoid cells.	Thyroid agenesis, hypoplasia, ectopia (sub-lingual most often) and rarely unilateral kidney and problems in urogenital tract. OMIM: 218700.	Paired box (binding domain), region encoding the transactivation domain or promoter region.	Heterozygous missense or nonsense mutations whose molecular mechanism of effect is still not elucidated and can be via dominant-negative effects, haploinsufficiency or monoallelic expression.

Table 3.1 Human phenotypes and syndromes associated with mutations in thyroid transcription factor genes. Table adapted from Fagman *et al* [140].

Gene in mouse (Mus musculus)	Gene in human (Homo sapiens)	Description	Thyroid phenotype	Extrathyroidal features
Shh	*SHH*	Sonic Hedgehog	Bilobation defect	Holoprosencephaly, cardiac outflow tract defects
Foxe1	***FOXE1***	Forkhead Box E1	Ectopia or agenesis	Cleft palate
Chrd	*CHRD*	Chordin	Hypoplasia	Cardiac outflow tract defects, hypoplasia of thymus, parathyroid
Edn1	*EDN1*	Endothelin 1		Craniofacial, cardiac and thymus defects
Eya1	*EYA1*	EYA Transcriptional Coactivator And Phosphatase 1		Aplasia of kidneys, thymus, parathyroid
Fbln1	*FBLN1*	Fibulin 1		Craniofacial, cardiac and thymus defects
Hes1	*HES1*	Hes Family BHLH Transcription Factor 1		Craniofacial, cardiac and thymus defects
Hoxa5	*HOXA5*	Homeobox A5		Cardiovascular and skeletal defects
Isl1	*ISL1*	ISL LIM Homeobox 1		Heart, pancreas and neural defects
Nkx2-5	***NKX2-5***	NK2 Homeobox 5		Cardiac defects
Frs2	*FRS2*	Fibroblast Growth Factor Receptor Substrate 2	Hypoplasia plus bilobation defects	Thymus and parathyroid defects
Hoxa3	*HOXA3*	Homeobox A3		Cardiovascular and skeletal defects
Hoxb3	*HOXB3*	Homeobox B3		Cardiovascular and skeletal defects
Hoxd3	*HOXD3*	Homeobox D3		Thymus and parathyroid defects
Pax3	*PAX3*	Paired Box 3		Cardiac outflow tract defects, hypoplasia of thymus, parathyroid
Tbx1	*TBX1*	T-Box 1		Cardiac outflow tract defects, hypoplasia of thymus, parathyroid
Fgfr2	*FGFR2*	Fibroblast Growth Factor Receptor 2	Agenesis	Atresia of the lungs
Fgf10	*FGF10*	Fibroblast Growth Factor 10		Aplasia of limbs, lungs, pituitary, salivary glands
Hhex	*HHEX*	Hematopoietically Expressed Homeobox		Forebrain truncations, liver aplasia, cardiac defects
Nkx2-1	***NKX2-1***	NK2 Homeobox 1		Pulmonary atresia, neural defects
Pax8	***PAX8***	Paired Box 8		Reproductive tract defects
Twsg1	*TWSG1*	Twisted Gastrulation BMP Signaling Modulator 1		Vertebral defects, spectrum of midline defects, agnathia

Table 3.2 Mouse models of thyroid dysgenesis. Human genes marked in bold denote those for which defects have been identified in human patients. Table adapted from Fagman *et al* [140].

3.1.3 Genetic studies of thyroid dysgenesis

Few genetic investigations of TD phenotypes have been conducted to date; the studies that have been reported generally focused on screening cohorts of patients for mutations in *TSHR* and in the three TFs [9, 61, 83, 92, 209, 226]. More recently, one of the genes uncovered via TD mouse models (*NKX2.5*) was postulated to also underlie a fraction of human TD cases, after observations of four heterozygous probands in a cohort of 241 TD patients [115]. *NKX2.5* encodes a homeodomain-containing transcription factor that is expressed in thyroid morphogenesis [141], but is mostly known to play a pivotal role in heart development [34]. Indeed, mutations in *NKX2.5* are a well established cause of several dominantly-inherited congenital heart diseases (CHDs) including atrial septal defects (OMIM: 108900), tetralogy of Fallot (OMIM: 187500) and ventricular septal defects (OMIM: 614432) [438, 516]. Because CHD is overrepresented among children with TD [362, 414], a developmental association between the two systems had been suggested. Yet, the possible involvement of *NKX2.5* in TD pathogenesis is now thought to be ambiguous, after a study that examined the literature evidence and functional impact of the reported mutations concluded there was a lack of clear evidence of pathogenicity of *NKX2.5* mutations in an isolated TD context [499].

Besides point mutations, additional genetic research in TD has focused on identifying copy-number-variants (CNVs) that could potentially explain the apparent sporadic nature of TD. These investigations, based either on fluorescence *in situ* hybridization (FISH) [494], array comparative genomic hybridization (aCGH) [485] or SNP genotyping [363], identified rare, non-recurrent CNVs encompassing several genes in thyroid agenesis and hypoplasia patients. Yet, none of those variants have been linked to or put into context of thyroid disease, being majorly non-informative, with the exception of one duplication in a single agenesis patient that overlapped with *TBX1* in the DiGeorge critical region 22q11. *TBX1*, enconding the T-box 1 protein, is another example of a TD candidate gene identified through mouse experiments (**Table** 3.2). *Tbx1*-null mice exhibit hypoplastic phenotypes due to delayed expression of *Nkx2.1* in thyroid progenitor cells [139]. Although thyroid abnormalities have been reported sporadically in patients with a 22q11.2 deletion [513], the majority of DiGeorge and 22q11 duplication patients do not have CH [367, 515], which has made this CNV finding difficult to interpret in the context of an isolated thyroid abnormality.

3.1.4 Why exome-sequence CH cases

Despite being the most common congenital endocrine disorder [99], the pathogenesis of CH remains elusive in the vast majority of patients. Cumulatively, known genetic defects linked to CH with *gland-in-situ* and TD phenotypes account for less than 20% of all CH cases [373]. TD probands represent an exceptionally small fraction of that percentage, and are mostly syndromic. Besides *TSHR*, genetic causes underlying non-syndromic TD are lacking. Extensive searches for mutations in *NKX2-1*, *FOXE1* or *PAX8* [83, 209, 226], have explained only a handful of TD cases, and linkage analyses have excluded these genes in several multiplex families with TD [75]. Because many CH patients suffer from congenital malformations adjacent to the thyroid gland, other factors that govern multiorgan development (such as cardiac, lung or musculoskeletal) may be equally involved, but none have yet been discovered.

The majority (57%) of *gland-in-situ* patients for whom no causative variants were observed in the previous chapter, represented familial cases with multiple affected siblings. This finding has also been observed in another study [66], and suggests genetic factors of CH with *gland-in-situ* have yet to be discovered. Other genes involved in thyroid hormone synthesis, but outside the follicular unit, may well be implicated [179], as well as genes that are known to modulate TSH and free-T_4 levels [398, 474].

Exome-sequencing has previously only been employed once, with the aim of identifying additional genetic defects causative of CH, but it was conducted in a single consanguineous TD family [263]. Exome-sequencing of large CH cohorts has not yet been performed, and such a strategy is therefore warranted to discover novel genetic factors [66].

3.2 Aims

The aim of the research present in this chapter was to identify novel genetic aetiologies associated with CH phenotypes. This was addressed by means of a whole-exome and targeted-sequencing study of a cohort of CH families that previously screened negative for known genetic causes. The cohort was phenotypically heterogeneous and consisted of non-syndromic TD cases, syndromic patients with a multiplicity of phenotypes alongside CH, and the *gland-in-situ* CH patients for whom no convincing causative mutations in known GIS-CH genes were identified in the previous chapter.

3.3 Colleagues

All the work presented in this chapter is my own work, unless otherwise stated. This research was carried out under the supervision and guidance of Dr Carl Anderson at the Wellcome Trust Sanger Institute (WTSI) and Dr Nadia Schoenmakers at the Institute of Metabolic Science (IMS), Cambridge, UK. This work was done in close collaboration with other colleagues at the IMS, namely Professor Krishna Chatterjee and Adeline Nicholas.

3.4 Methods

3.4.1 Patients

A clinical team consisting of Dr Nadia Schoenmakers and Professor Krishna Chatterjee recruited a cohort of CH. Adeline Nicholas collected DNA from these cases and from unaffected and/or affected relatives, whenever possible. A total of 75 samples (48 affecteds and 27 unaffecteds) from 27 families were whole-exome sequenced, some of which as part of the UK10K rare-disease project (www.uk10k.org). Additionally, 33 samples (25 affecteds and 8 unaffecteds) from 21 families were sequenced for a panel of selected genes, as part of the UK10K targeted sequencing experiment. **Table** 3.3 lists all the pedigree structures available in this study and **Figure** 3.2 illustrates the different phenotype categories across patients. All investigations conducted in this work were part of an ethically approved protocol, being undertaken with the consent from patients and/or next of kin.

Pedigree type	Description	Number of families	
		WES	TS
Trio (with or without unaffected relatives)	Unaffected parents and proband	13	1
Affected siblings (with or without unaffected relatives)	At least two affected siblings	9	3
Multiplex family (with or without unaffected relatives)	One affected parent	3	.
Extended family	Cousins affected	1	1
Unaffected parent-proband duo	Single unaffected parent and proband	.	1
Singletons	No family relative sequenced	1	15
Total		27	21

Table 3.3 Pedigree structures available in this CH cohort. WES: whole-exome sequencing; TS: targeted-sequencing.

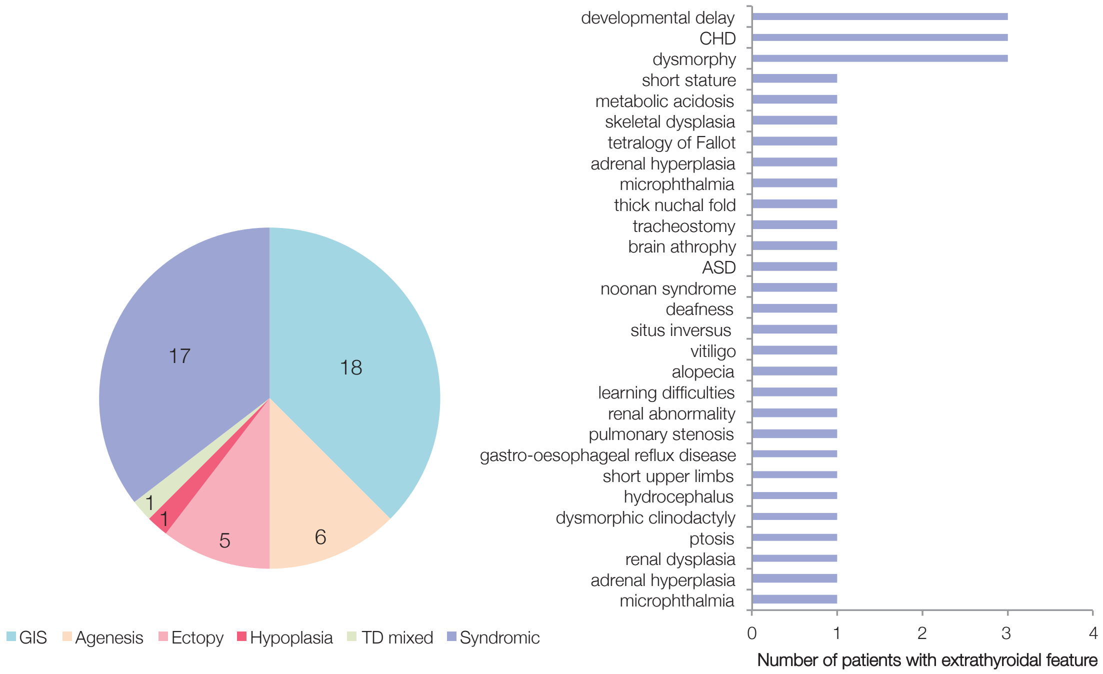

Figure 3.2 Phenotype categories within the CH cohort.
A) GIS: *gland-in-situ* CH; TD mixed: multiple dysgenesis phenotypes in different affected relatives of the same family (agenesis in one sib, ectopy in the other, for example); Syndromic: any case that presents with CH and additional extrathyroidal symptoms, displayed in bar chart on the right. Numbers in pie chart represent the number of families with the given phenotype. **B)** List of extrathyroidal features recorded for the 17 syndromic CH patients. CHD: congenital heart disease. ASD: autism spectrum disorder.

3.4.2 Sequencing

The whole-exome sequencing (WES) and Hi-Seq targeted-sequencing (HiSeq-TS) of the CH samples included in this study were performed exactly as described in the previous chapter. The genes that were included in the targeted panel were selected prior to my PhD studies by Dr Nadia Schoenmakers, and Dr James Floyd designed the Agilent SureSelect pull-down array. This panel was a subset of a large targeted-sequencing study of seven rare diseases, comprising a total of 2,812 individuals, which was carried out within the UK10K study (www.uk10k.org). Overall, the target, was composed of 3.4Mb of sequence from the coding exons (UCSC hg19/Grch37 human reference genome build) of 1,188 genes, of which only 20 were candidates for either *gland-in-situ* CH or thyroid dysgenesis phenotypes (**Table** 3.4). Most of these loci were selected based on representing constituents of thyroid hormone biological pathways, GWAS hits for thyroid function levels or mouse/zebrafish knockouts with evidence of CH (Dr Nadia Schoenmakers, personal communication).

Gene	Evidence
DUOX1	Thyroid synthesis biological pathway
DUOXA1	
KCNQ1	
KCNE2	
LHX3	GWAS hits for TSH and free T_4 levels
PDE8B	
SLC26A7	Mouse knockout exhibits CH
TBX1	Established mouse models
ISL1	
NKX2.5	
EDN1	
FBLN1	
HOXA3	
FGF8	
HAND2	
DICER1	Involved in thyroid carcinoma
TPST2	Mouse knockouts exhibits CH
MCHR1	
GLIS3	Neonatal diabetes with CH, OMIM 610199
WWTR1	Zebrafish knockout exhibits thyroid follicles abnormalities

Table 3.4 Candidate GIS and TD genes selected for the targeted-sequencing experiment. Blue rows mark the genes that are candidates for CH with *gland-in-situ* while rows in orange mark the genes that are candidates for thyroid dysgenesis phenotypes.

3.4.3 Data quality control

Before embarking on downstream genetic analyses, I conducted a series of quality control (QC) assessments on the called VCF files to make sure the sequencing data were of high quality. Specifically, I worked to detect whether there was evidence of poorly sequenced samples, or samples that were outliers for several population genetics expectations, including the ratio of heterozygous to alternative homozygous variants (Het/Alt ratio), the ratio of transtitions to transversions (Ts/Tv ratio) and the number of variants called at various allele frequencies and functional categories. Before interpretation of those results however, I ran two analyses: one, to infer genetically the ancestry of the samples and the second, to infer genetically their consanguinity status. These two analyses are important because factors such as ethnicity and consanguinity can influence how a sample behaves across several population-based metrics. For instance, African samples will tend to have a higher number of variants called when compared to European samples, not due to a sequencing error, but because of the higher genetic diversity across African genomes [207]. Consanguinity status matters because offspring of related relatives will display a higher number of homozygous alternative calls because a higher proportion of their genome is autozygous [282], i.e. it contains a larger proportion of alleles that represent physical copies of each other or physical copies of an ancestral allele, known as identical-by-descent (IBD) alleles [482].

Inferring ancestry origin

I evaluated the ethnicity of the WES samples via a principal component analysis (PCA) of 2,504 samples from the 1KG phase 3, followed by projecting our samples onto the first (PC1) and second principal components (PC2). Sites from the CH exomes were restricted to autosomal, biallelic SNVs with minor allele frequency $\geq 5\%$ that did not deviate significantly from the Hardy-Weinberg equilibrium (HWE) <10^{-5} and that had a call rate > 90% across all samples. I then took the overlap of SNVs between the CH exomes and 1KG and pruned the markers for LD with the command –indep 100 1 0.1 in PLINK [406]. This is a useful step to increase computational efficiency by making sure only independent SNVs are used to create the PCA. This step left a total of 13,850 SNVs available for analysis. The PCA calculation and projection was carried out with the EIGENSTRAT package [400]. The ancestry analysis in the targeted-sequencing dataset was performed following the same protocol, but using 1,192 HapMap3 samples instead of 1KG, and a total of 1,520 SNVs.

Both of these PCA analyses revealed that 61% and 21% of the WES and targeted-sequencing samples, respectively, were not of European ancestry (**Figure** 3.3). Non-European ethnicity mostly included Pakistanis and Bangladeshis (19 individuals), followed by individuals from Turkey (6 individuals), Saudi Arabia (4 individuals), Iraq (3 individuals), Africa (3 individuals) and South Africa (1 individual), as subsequently reported by Dr Nadia Schoenmakers.

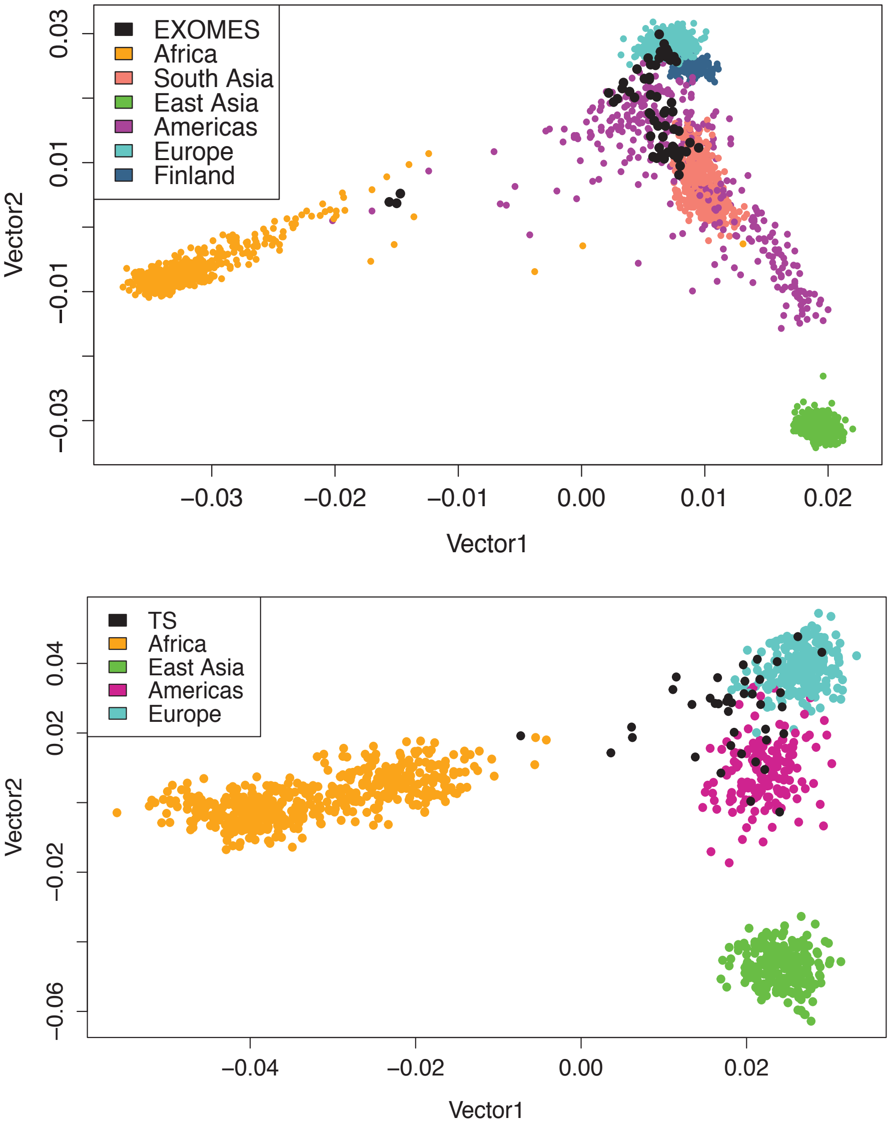

Figure 3.3 Principal component analysis of WES and targeted-sequencing CH samples. **A)** 75 WES samples plotted against 2,504 1KG Phase 3 samples. **B)** 33 targeted-sequencing (TS) samples plotted against 1,192 HapMap3 samples.

Inferring consanguinity status

A consanguineous union is commonly defined as an union between a couple related as second cousins or closer, equivalent to a coefficient of inbreeding (F) in their offspring of $F \geq 0.0156$ [48]. One way of calculating F is by identifying runs-of-homozygosity (ROHs), which are uninterrupted stretches of homozygous variants in the genome. Such long segments of homozygosity can represent large deletions, loss-of-heterozygosity (LOH), segmental uniparental disomy, or autozygosity regions [282], as defined above. The proportion of the genome that is autozygous is the closest estimation of the real F [282]. I used BCFtools to estimate the proportion of the genome that is autozygous (F) in every CH case from the WES experiment. BCFtools implements a statistical framework that takes into account genotype likelihoods and the recombination rate along the genome to provide a probability of autozygosity for every site along the exome [344]. A statistical model is helpful because it is crucial to accurately distinguish truly autozygous ROHs from the larger pool of often non-autozygous ROHs [215, 344]. As an additional line of evidence, I also calculated the ratio of heterozygous to alternative homozygous variants (Het/Alt ratio), a metric that is inversely correlated with F when F is well estimated.

From comparing F and the Het/Alt ratio, I confirmed that all samples reported to be consanguineous upon sample recruitment indeed had $F \geq 0.0156$ and their Het/Alt ratio was also lower, as expected (**Figure** 3.4). More importantly, this analysis identified four samples (two sib-pairs) that, contrary to what was reported, appeared to be from consanguineous unions.

Assessing sequencing quality

Overall, the WES and the targeted-sequencing datasets were of high quality. The two datasets had a mean depth (DP) of 76x and 53x, respectively. Because high-depth regions that result from unspecific binding of the capture regions can easily affect mean depth [187], I also calculated the median DP achieved in the two different datasets, which were 59x and 43x, respectively. These values are much higher than the minimum 30x estimated to be required for accurate detection of heterozygous variants for Mendelian disease studies [260].

The mean number of SNVs and indels detected per sample were 36,480 and 1,612, respectively (**Figure** 3.5), both of which are within the expected range seen in other exome studies that used the same technology [31, 122, 169]. The number of high quality

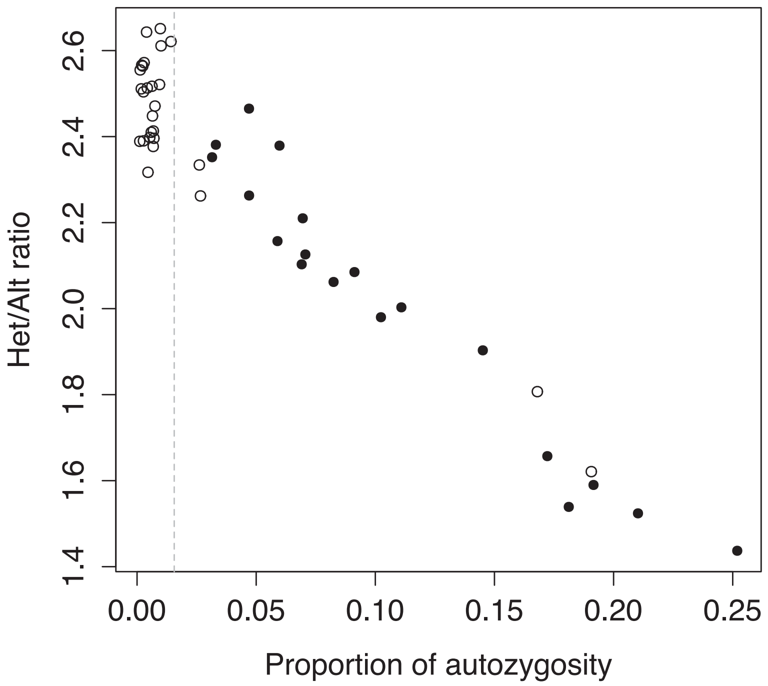

Figure 3.4 Estimation of consanguinity status for CH cases. HetAlt ratio: the ratio of heterozygous to homozygous SNVs. Filled circles denote the individuals whichs were reported to be consanguineous upon sample recruitment whereas empty circles represent samples that were reported to be non-consanguineous. Gray dashed line denotes the inbreeding coefficient (F) of 0.0156 [48].

SNVs and indels were consistent across samples, with a small variation dependent on the ancestry of the samples: for example, three African samples exhibited a larger number of both SNVs and indels exome-wide. However, no sample exhibited SNV or indel counts significantly outside the boundary marked by ±3 standard deviations (SD) from the mean. The Ts/Tv ratio (mean = 2.8) was also consistent across samples and was within the expected values (between 2.7 and 3.0) for exome datasets [71, 116]. The same was true for the Het/Alt ratio (mean = 2.4), which was also close to the expected value of 2.5 [187].

Also consistent with expectation, ∼94% of the SNVs were common in the population ($\geq$ 5% in 1KG Phase 1), 2% were rare ($\leq$ 1% in 1KG Phase 1) and around 2.7% were novel, i.e. absent from both in 1KG Phase 1 and dbSNP137 (**Figure** 3.6). The rest of the variants had population frequencies between 1-5% (data not shown). Similar proportions have been documented in other exome studies [11, 72].

Of all SNVs, the majority were intronic and located in untranslated regions (UTRs), with the rest representing either functional or silent variants, both of which occurred at a similar rate (**Figure** 3.6). In terms of missense variants detected, the majority were predicted to be benign by both SIFT and Polyphen-2, with only ∼16% considered to be damaging by both tools and a further ∼5% for which predictions were uncertain or unknown. Finally, of the loss-of-function (LoF) SNVs (i.e. nonsense, frameshift and splice acceptor/donor variants), only 8% were novel while the rest had been previously observed in the 1KG Phase 1 data.

The targeted-sequencing experiment behaved similarly to the WES data across the 1,188 targeted genes included in the UK10K experiment, with a Ts/Tv mean ratio of 2.9. There were no outliers for the quality metrics or population genetics expectations described above (data not shown).

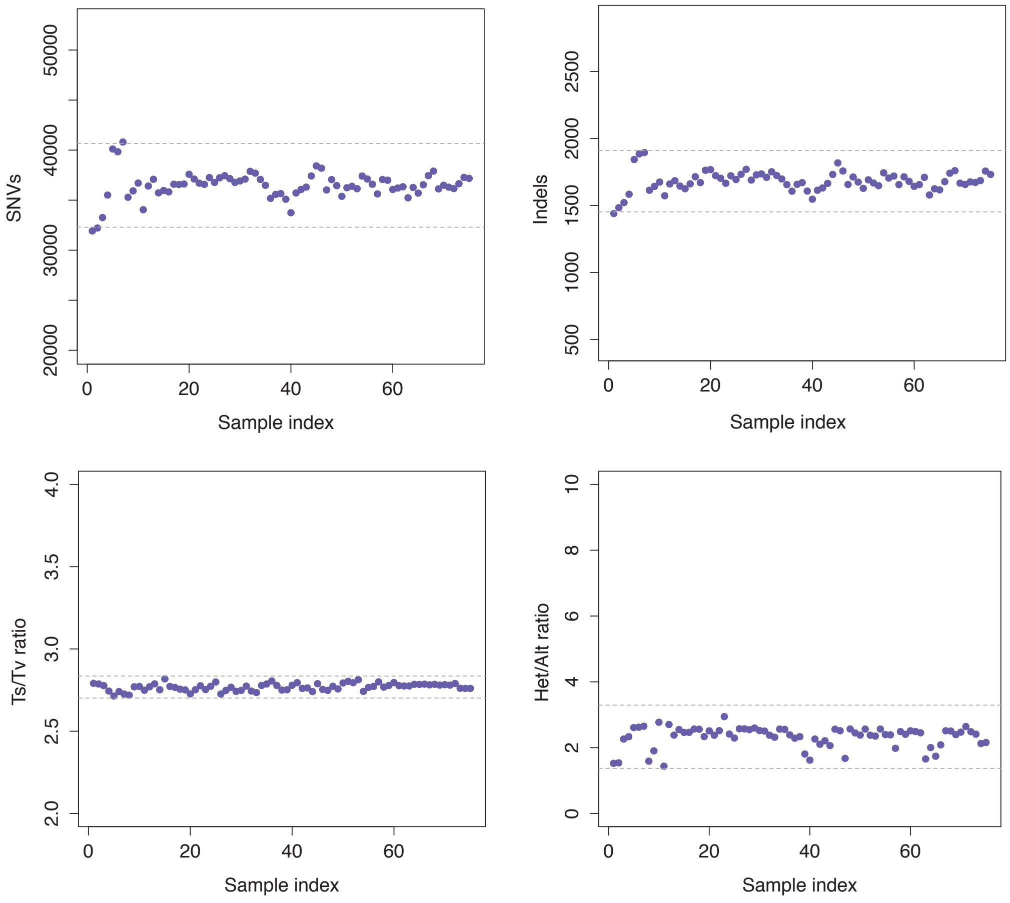

Figure 3.5 Quality control metrics for the WES experiment. **A)** Number of SNVs called and passing the variant QC per sample (see previous chapter for details). The cluster of three samples with higher numbers correspond to individuals of African ethnicity; **B)** Number of indels called and passing the variant QC per sample (see previous chapter for details); **C)** Transitions to transversions ratio (Ts/Tv) per sample and **D)** Heterozygous to homozygous (altervative allele) ratio per sample. Samples with lower Het/Alt ratio correspond to individuals with some degree of consanguinity.

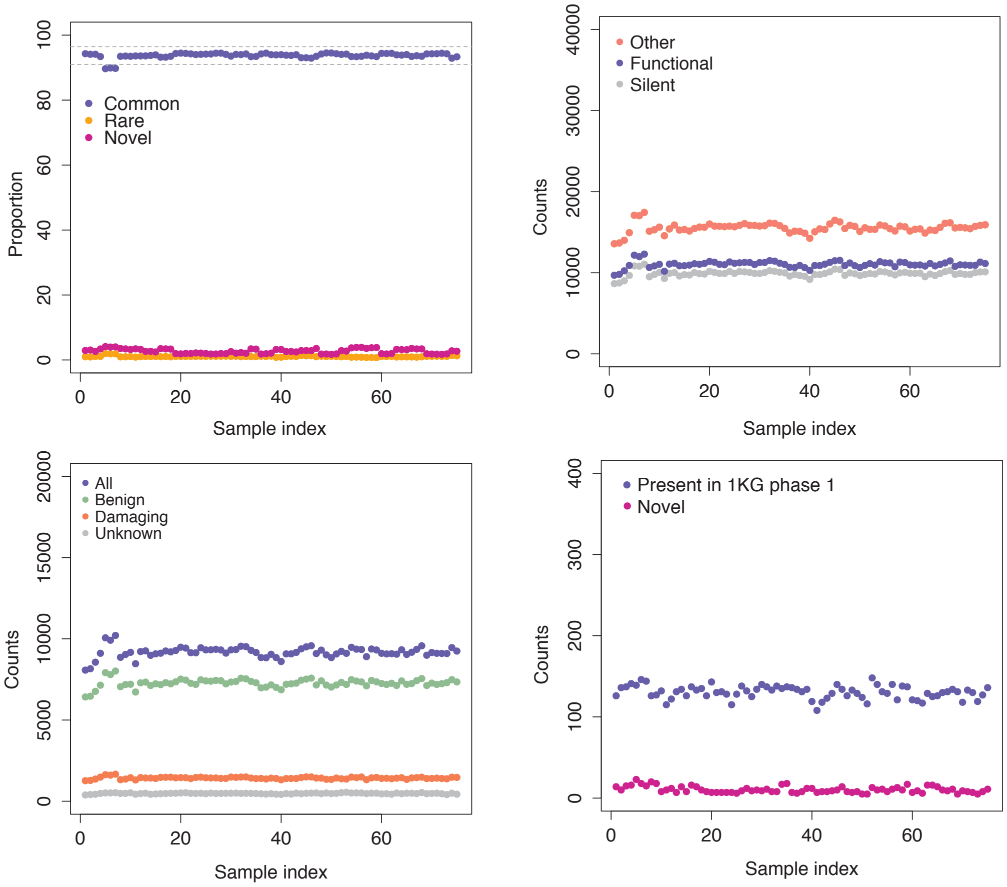

Figure 3.6 Population genetics metrics for the WES experiment. **A)** Proportion of SNVs that are common ($\geq$ 5% in 1KG Phase 1), rare ($\leq$ 1% in 1KG Phase 1) and novel (absent from both in 1KG Phase 1 and dbSNP137) per sample. The cluster of three samples with lower common variants correspond to samples of African ethnicity; **B)** Number of SNVs that are functional, silent or 'other' per sample; protein consequences given by Ensembl Variant Effect Predictor v75. Functional variants include stop gained, splice donor, splice acceptor, frameshift, missense variant, inframe deletion, inframe insertion, initiator codon variant and splice region variant. Silent variants include synonymous varians and 'other' include the rest of the variants that are non-coding such as intronic variants and variants in 5' or 3' untranslated regions, UTRs; **C)** Number of missense variants per sample broken down by their deleteriousness prediction provided by SIFT and Polyphen-2; **D)** Number of loss-of-function (LoF) SNVs that are present in 1KG Phase 1 and that are novel (absent from both in 1KG Phase 1 and dbSNP137) per sample. LoF variants were defined as those with consequences given by Ensembl Variant Effect Predictor v75 (detailed in following section) of: stop gained, splice donor, splice acceptor and frameshift (see **Figure** 1.9 for definitions of splice sites).

3.4.4 Gene mapping within CH families

Inherited variation

Post-QC, I designed a variant filtering pipeline for the identification of rare and functional variants segregating within CH families (**Figure** 3.7). The pipeline started by merging the VCF files of all members of each family into a multi-sample, family VCF file. Next, only variants that met the high-quality thresholds (see variant-QC in previous chapter) and contained within the baits/targeted regions were kept. The following step used the Ensembl Variant Effect Predictor (VEP) version 75 to annotate the functional consequences of all variants according to Gencode v19 coding transcripts, keeping the most severe consequence for the gene [322]. Functional variants were defined as any variant with an impact at the protein level that fell in the following consequence classes: transcript ablation, stop gained, splice donor variant, splice acceptor variant, frameshift variant, inframe insertion, initiator codon variant, splice region variant, stop lost, missense, variant, inframe deletion, stop retained variant. All variants were then annotated for Genomic Evolutionary Rate Profiling (GERP) conservation scores [107], and missense variants were annotated with deleteriousness prediction scores generated from Sorting Intolerant From Tolerant (SIFT) [349] and PolyPhen-2 [4].

Next, variants were annotated with allele frequencies (AFs) that I computed from several population datasets including 1000 Genomes Phase 1 integrated callset 2012-07-19 (1KG, N=2,504) [402], NHLBI Exome Sequencing Project 6,500I (ESP, N=6,500) [476] and Exome Aggregation Consortium r0.3 (ExAC) (N=60,706) [135], as well as from a set of control exomes sequenced at the WTSI, including UK10K whole-genome sequenced cohorts (N=3,781), other UK10K exomes (N=4,818) and other UK10K targeted sequenced samples (N=2,634) [507]. Collectively, these data constituted ∼81,000 control sequences, some of which were disease-cases but, in principle, unrelated to thyroid disease, and free from severe paediatric samples in the ExAC dataset (Daniel McArthur personal communication). Similarly as others have demonstrated [474], including the AFs from other projects conducted internally at the WTSI (i.e. UK10K) was crucial to increase the specificity of the filtering, because it allowed for the exclusion of systematic technical artefacts that were specific to the WTSI sequencing pipeline. Rare variants were defined as variants that were absent or with AFs <1% in all of the above population and internal control datasets.

The different pedigree structures available in the study (**Table** 3.3) meant that the pipeline had to be flexible and allow for specific downstream filtering of variants under different Mendelian models, assuming 100% penetrance. **Table** 3.5 summarises the different segregation rules assumed for each pedigree structure.

Finally, the genotypes of segregating variants were then cross-checked with genotypes of other UK10K exomes and UK10K targeted-sequencing samples. As recommended by Macarthur *et al* [300], variants were removed if UK10K samples harbored the same genotype as to that seen in CH cases, or if any UK10K sample was homozygous for heterozygous or compound heterozygous sites observed in CH patients.

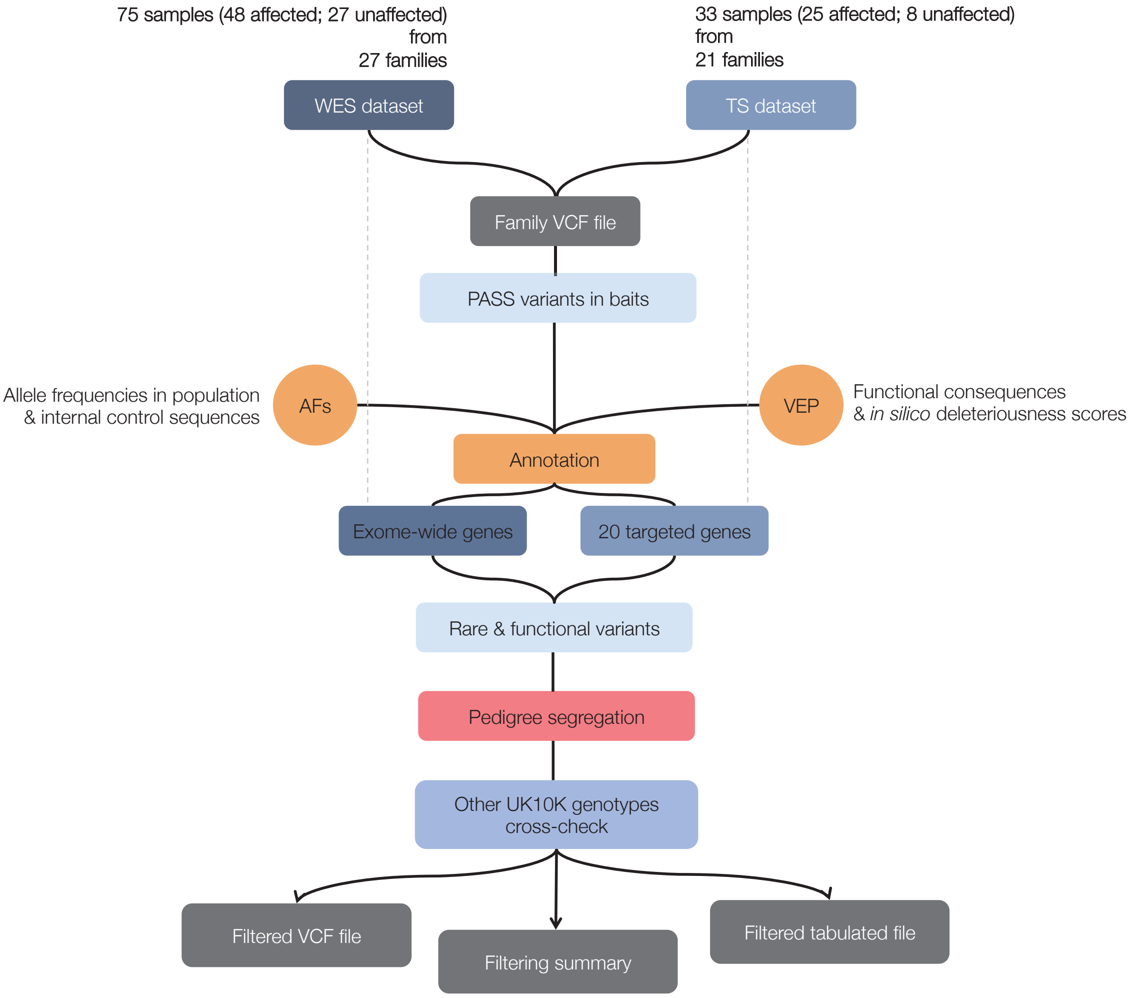

Figure 3.7 Variant filtering pipeline.

WES: whole-exome sequencing; TS: targeted-sequencing; AFs: alternate allele frequency; VEP: Ensembl Variant Effect Predictor v75. Rare variants were defined as variants that were absent or with AFs <1% in all of the population and internal control datasets (see main text for details). Functional variants were defined as any variant with an impact at the protein level that fell in the following consequence classes: transcript ablation, stop gained, splice donor variant, splice acceptor variant, frameshift variant, inframe insertion, initiator codon variant, splice region variant, stop lost, missense, variant, inframe deletion and stop retained variant. Deleteriousness scores included SIFT and Polyphen-2 and conservations scores were provided by GERP. For pedigree segregation rules see **Table** 3.5 on the following page.

Pedigree structure	Description	*De novo* variants (putative only)	Inherited variants		
			Compound heterozygotes	Homozygous	Heterozygotes
Trios	Unaffected parents and proband	1,0,0/2,1,0	a+b, a, b	2,1,1	.
Affected sibs	At least two affected siblings	.	.	2,2	.
Multiplex families	One affected parent affected	.	.	.	1,1,0
Extended families	Cousins affected	.	.	2,2	1,1
Unaffected parent-proband	Single unaffected parent-proband duo	.	.	2,1	1,0
Singletons	No familial data available	.	.	2	1

Table 3.5 Pedigree segregation rules for different family structures.

In the case of **trios**, the pipeline looked for genotype inconsistencies between proband and parents i.e. putative *de novo* variants (to be later cross-checked with the output from a proper *de novo* caller), compound heterozygote variants and homozygous variants. Putative *de novo* variants represented either heterozygote variants in child that were reference in both mother and father (i.e. 1,0,0) or homozygous variants in the child that were heterozygous in one parent and reference in the other (i.e. 2,1,0). Compound heterozygous variants in a given gene were both present in the child ($a + b$) but each one came from exactly one of the parents (i.e, a in mother; b in father). Homozygous variants in child were considered if both parents were obligate carriers (i.e. 2,1,1). For X-chromosome variants, I assumed an X-linked model of inheritance in male probands only. In **affected siblings**, both siblings had to share the same homozygous variant (i.e. 2,2), as dominant mutation in multiple siblings is extremely unlikely unless the father exhibits germline mosaicism. In **multiplex families**, I kept heterozygote variants shared by the proband and affected parent and that were reference in the unaffected relative (i.e, 1,1,0). In **extended families**, homozygous and heterozygous variants were kept as long as they were shared by both cousins. In the **unaffected parent-proband duo**, I considered all heterozygous variants in the child that were reference in the available parent (i.e. 2,1, with the caveat that a substantial proportion of these will still represent inherited variation rather than *de novo* events) and homozygous variants in the child that were heterozygous in the parent (i.e. 2,1). As no familial DNA was available in **singletons**, both heterozygous and homozygous variants were considered.

De novo variation

In a basic approach, *de novo* variation (DNV) can be detected by simply identifying genotype inconsistencies between parents and offspring, as included in **Table** 3.5 and as implemented in my variant filtering pipeline described above. Although this is straightforward, a great proportion of these inconsistencies will turn out to be false positives that result from failure to call the corresponding germline variants in one of the parents. Indeed, a study recently demonstrated that the ability to accurately distinguish *de novo* from familial inherited variants is more limited by high false-negative rates in the parents than by high false-positive rates in the child [364].

To identify DNVs in trios with increased sensitivity compared to the simple filtering approach offered by the pipeline, I used DeNovoGear [413]. This program calculates a posterior probability for observing a polymorphism or a real DNV at any given site in the genome by taking into account individual genotype likelihoods (from parents and child) and a prior mutation rate, which together, increase the accuracy of the calls [413]. Moreover, DeNovoGear uses a beta-binomial distribution fit, instead of the binomial distribution typically used by genotype calling algorithms, to handle the over dispersion in the distribution of alternate and reference read frequencies that is typical of exome sequencing data [198]. Ultimately, this approach has been shown to reduce the false positive rate associated with DNVs discovery by 50%, with no loss of power compared to other genotype calling algorithms such as SAMtools and GATK [413].

Conservatively, I decided to only focus on DeNovoGear DNV calls with posterior probabilities greater than 80%, as recommended by Ramu *et al* [413]. However, even though these calls were identified with high confidence through the statistical framework of DeNovoGear, they can still be enriched for false positives. To mitigate this, I used several metrics computed by the program to filter the output. First, variants were removed if located in tandem repeats or segmental duplication sites, as false positive calls are frequently observed in these regions [31]. This is because these regions are highly unstable and known to mutate at higher rates than those of point mutations in repeat-free sequence [264, 297]. Second, variants were also removed when >10% of the reads in either parent supported the alternative allele, as the variant would be more likely inherited from a parent than a true *de novo* event. Thirdly, I focused on functional and rare variants (as defined above) and excluded variants not called by the independent and ordinary variant caller (HaplotypeCaller in WES dataset; SAMTools and/or GATK in the targeted-sequencing dataset). Finally, to verify variants were not associated with reads that were incorrectly mapped, I visually inspected all DNVs using the Integrative Genomics Viewer (IGV) [484].

Copy-number-variants

CNVs in the whole-exome sequended CH samples were detected using CoNVex (ftp://ftp.sanger.ac.uk/pub/users/pv1/CoNVex/). This software applies a comparative read depth approach that compares, at a given genomic location, the read depth of a given sample with the read depth of a set of control exomes. This is a desirable approach since it corrects for technical variation between samples, which normally arise due to poor read mappability, GC bias, and also batch effects between sequencing experiments [278, 477]. CoNVex is capable of detecting deletions and duplications of the WES targeted sequences (probes) from a few hundred base pairs in size to a few mega bases or more, i.e. high resolution events. CoNVex was calibrated to call approximately 200 CNVs per sample, a number very close to the expected number of common CNVs (>1% frequency) present in a human exome when taking into account the number of probes available for the CNV calling procedure (Dr Vijay Parthiban, personal communication).

Dr Vijay Parthiban ran CoNVex on the whole-exome sequended CH cohort. CNV calling was not conducted for the targeted-sequenced CH samples due to the very low number of genes sequenced and because the CNV boundaries, if any, would be difficult to ascertain. Moreover, the whole genome amplification process performed prior to sequencing in this dataset is known to compromise CNV calling [405].

To filter the output produced by CoNVex, I considered only calls with confidence scores ≥ 10, as recommended by Dr Vijay Parthiban. I then annotated the CNV calls against published datasets, including 2,026 clinically well-characterised healthy individuals [442] and results from a whole genome screen for CNVs at 500-bp resolution [94] to filter out common CNV calls. Similarly as in Carss *et al* [72], CNVs were considered identical if their sequences overlaped by 50% and were excluded if this was the case. To further filter for rarer CNVs, I considered only those calls that were absent in any of the other UK10K samples and calls that overlapped with at least one protein-coding gene and that were covered by more than one probe. Finally, I inspected plots of regional $\log_2$ ratios of the exome read depth in each proband and in the available family members. I filtered out variants that did not properly segregate with disease status and Mendelian rules of inheritance within families, or variants that constituted likely technical artefacts.

3.4.5 Predicting the impact of splice donor mutations

Recent guidelines for evaluating the impact of splice-disrupting variants in human disease recommend the use of *in silico* prediction tools, similarly as routinely conducted when judging missense variants [419]. These tools essentially examine whether a variant observed in a 5' or 3' splicing consensus region is likely to disrupt the exon-intron boundaries of the protein and affect RNA splicing [228].

In the present study, I used MaxEntScan [285] to predict the potential impact of rare splice donor mutations that I observed in CH cases. This tool has been shown to have the highest accuracy at predicting the effects of mutations at the 5' invariant splice sites [118] and, as an example, it has been successfully applied to the prediction of splicing mutations in the *ATM* gene responsible for the neurological disorder ataxia-telangiectasia, in which three mutations were correctly interpreted as disrupting normal splice sites [128]. Briefly, MaxEntScan provides a score as a numerical measure of the strength of the splicing signal. This basically represents the probability or the confidence of a site being a true splice site used during the splicing process. To evaluate the effects of nucleotide substitutions occurring at 5' splice donor sites, I generated MaxEntScores for both the wild-type (WT) and mutant 5' sequences and compared the difference between the two, as recommended by Houdayer *et al* [213]. This difference in scores is thus a reflection of the deleteriousness of the variant at that splice site [228].

3.5 Results

3.5.1 Inherited variants in CH families

Targeted-sequenced families

DNA samples from a total of 21 families were put through targeted sequencing across a customised panel of 20 genes. The variant filtering pipeline identified four rare functional variants in three singleton samples (**Table** 3.6). Three of these variants are novel, i.e. have not been previously observed in ~81,000 population controls, and are predicted to be damaging and to affect conserved aminoacid sites (GERP>2 [98]).

Phenotype	Family	Gene	GT	CQ	AA change	Exon	Intron	GERP	1KG AF	ESP AF	EXAC AF	UK10K cohort AF	UK10K exomes AF
Ectopy (submandibular gland)	R8	***GLIS3***	**0/1**	**MI**	**H837R**	**10/11**	**.**	**5.93**	**.**	**.**	**.**	**.**	**.**
Gland-in-situ CH	R13	*DUOX1*	0/1	MI	K653N	18/35	.	0.74	.	.	5.00E-05	.	.
		TBX1	**0/1**	**SD**	**.**	**.**	**4/8**	**5.08**	**.**	**.**	**.**	**.**	**.**
Syndromic (CH and congenital heart disease)	R22	***NKX2-5***	**0/1**	**MI**	**G206R**	**2/2**	**.**	**4.58**	**.**	**.**	**.**	**.**	**.**

Table 3.6 Rare functional variants identified in three targeted-sequenced CH patients.

GT: genotype; CQ: consequences at the protein level given by Ensembl Variant Effect Predictor v75; AA change: amino acid change; GERP: conservation score (ranging from -12 to 6); 1KG AF: 1000 Genomes allele frequencies, ESP AF: NHLBI ESP project allele frequencies. STOP: stop gained variant; MI: missense variant; SD: splice donor variant. Variants highlighted in bold represent those that were predicted to be damaging by Polyphen-2 and SIFT or that represented LoF variants. All patients were singleton cases, i.e. no DNA from family relatives was available for analysis.

A *GLIS3* missense mutation (H837R) predicted to be damaging and affecting a conserved amino acid was discovered in a patient (R8) presenting with an ectopic, submandibular gland. This gene was included in the targeted sequencing experiment because mutations in this transcription factor are associated with a rare syndrome characterised by CH and neonatal diabetes (Neonatal Diabetes with Hypothyroidism, NDH, OMIM: 610199). Patients with NDH exhibit hypoinsulinemia, hyperglycaemia, reduced levels of T_3 and T_3, and elevated levels of TSH and TG [441], symptoms that can also be accompanied by glaucoma, polycystic kidney disease, hepatic fibrosis, osteopenia and mild mental retardation, depending on the nature of the mutation [234]. *Glis3* is highly expressed in the kidney, thyroid gland, endocrine pancreas, thymus, testis and uterus, and claimed to play a critical role in the maintenance and proliferation of endocrine progenitor cells [237, 511]. Thyroid ultrasound investigations of patients with NDH caused by *GLIS3* mutations revealed thyroid developmental abnormalities such as agenesis or hypoplasia phenotypes [234]. While the ectopic thyroid phenotype seen in patient R8 is a thyroid abnormality itself, there is no record of the patient suffering from neonatal diabetes which, in the case of NDH, necessarily develops within the first few weeks of life [6]. This coexistence of CH and diabetes manifestations in NDH is not only limited to *GLIS3*-positive humans patients, with *Glis3* mouse knockout models also consistently developing both clinical entities [234]. A possible explanation for the lack of a diabetes phenotype in our patient would be if the H837R amino acid change affected a transcript that is selectively expressed in the thyroid. This was not the case however, as the affected transcript (ENST00000324333) is the major protein coding sequence and the one with the highest expression levels across all tissues available in GTEx, the Genotype-Tissue Expression database (GTEx) (www.gtexportal.org). This finding, together with the fact the mutation is monoallelic rather than homozygous, and located in the penultimate exon of the gene, suggests this variant is unlikely to be causative of the CH phenotype observed in this patient.

The *gland-in-situ* CH patient R13 is an unsolved patient of the previous chapter. The subsequent sequencing of the additional 20 candidate genes and the variant filtering conducted here revealed two heterozygous variants in two separate candidates: *DUOX1* and *TBX1*. *DUOX1* is a long-standing candidate gene for dyshormonogenesis phenotypes after observations that a complete loss of *DUOX2* activity in homozygous patients does not completely revoke the ability to synthesize hormone [220], as some degree of oxidase activity is maintained by *DUOX1*, its paralogue. Similarly to *DUOX2*, this protein is also present at the apical membrane of thyrocytes, although at lower expression levels [179]. Given the compensatory mechanism between the two *DUOX* oxidases, the monoallelic amino acid change in *DUOX1* observed in this patient will probably not be sufficient to cause a phenotype, plus it is predicted to be benign and is located in a non-conserved amino acid site. The other candidate mutation found in this patient, resides within *TBX1*, a transcription factor involved in regulating the cell fate of organs and tissues derived from the pharyngeal apparatus, including the thyroid, and the adjacent secondary heart field from which the cardiac outflow tract derives [159]. Both reduced *Tbx1* dosage and dysregulation of *Tbx1* expression due to gain-of-function effects have been shown to affect pharyngeal and heart development in mice, in which a hypoplastic thyroid phenotype is usually developed [270, 504]. The R13 patient presents with a *gland-in-situ* but thyroid gland size was not quantitated formally for this patient, meaning mild thyroid hypoplasia could have been missed (Dr

Nadia Schoenmakers, personal communication). The splice donor variant observed in this patient was intriguing, not only because it disrupts the conserved T-box region of *TBX1*, a critical region for its function [127], but especially since no single LoF mutation has been recorded in more than 60,000 ExAC individuals, an observation that is in agreement with the known haploinsufficient nature of *TBX1* [30]. To better understand whether this variant has a potential deleterious effect on the protein via disruption of this 5' consensus sequence, I analysed both the wild-type and the mutant splice sequences using MaxEntScan and compared the two predicted splice scores, as is commonly conducted [118, 213]. My analysis revealed this substitution shifted the strength of the WT splice signal down by 8.5 units (from 8.72 to 0.22, **Figure** 3.8). To put these values in context, the strength of the wild-type sequence is comparable to that observed across 10,000 randomly selected splice donor sites occurring in the genome, while the mutant score is located at the lower tail of that distribution (**Figure** 3.8), suggesting it may well have a detrimental effect to the protein.

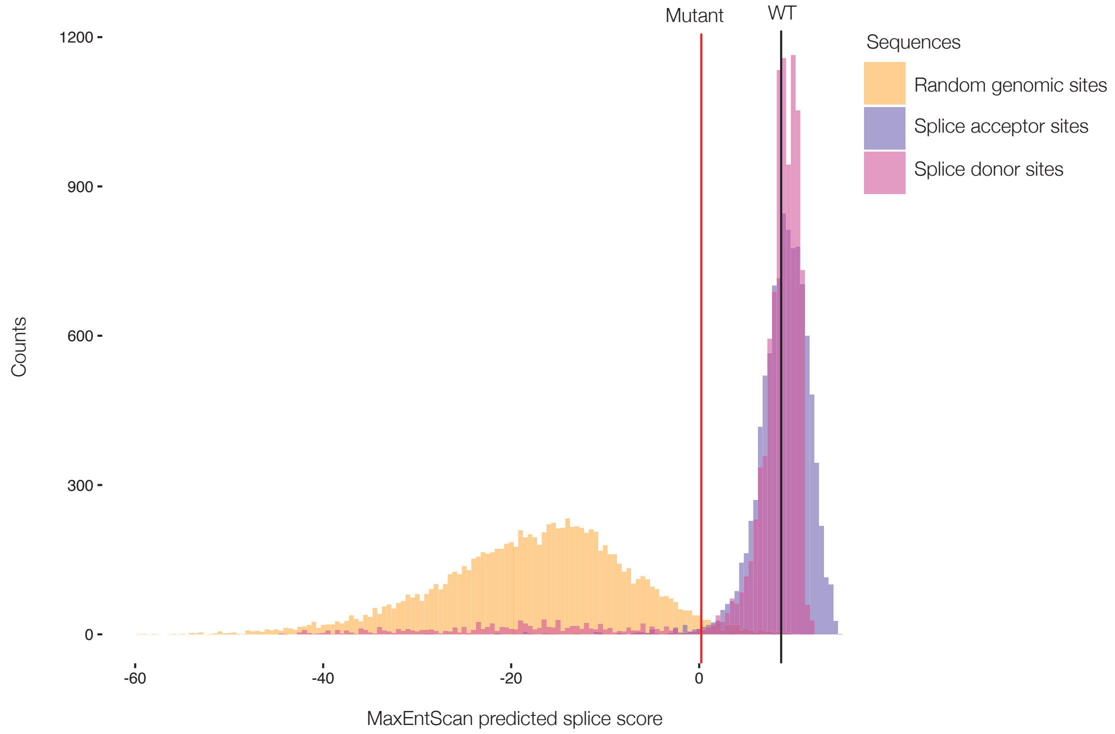

Figure 3.8 Predicted MaxEntScan scores for wild-type and mutant splice donor sequences of *TBX1* intron 4. To put the WT and mutant *TBX1* splice donor scores into context, MaxEntScan scores were also generated for a random set of 10,000 genomic regions, splice acceptor and splice donor regions ocurring in the genome. *TBX1* splice WT score: 8.72; *TBX1* splice mutant score: 0.22.

Even though this result looks encouraging, it is difficult to link this (apparently) deleterious mutation to the *gland-in-situ* CH phenotype of this patient when one takes into account the multiplicity of phenotypes *TBX1* has been linked to. This gene is considered to be the major genetic determinant of four genetic syndromes (conotruncal anomaly face syndrome (OMIM:21709), DiGeorge syndrome (OMIM:188400), Tetralogy of Fallot (OMIM:18750) and velocardiofacial syndrome (OMIM:192430) that result from deletions of the 22q11 critical region [532]. Together, these five clinical entities constitute a contiguous gene syndrome characterised by multiple, apparently unrelated clinical features that include cardiac outflow tract anomalies, abnormalities of the thymus and parathyroid glands, cleft palate and facial dysmorphism [532], with thyroid abnormalities only sporadically reported and always in combination with other severe clinical manifestations [513]. Importantly, apart from deletions encompassing *TBX1*, single point missense mutations in *TBX1* have also been identified in patients that did not carry 22q11 deletions but that nevertheless exhibited the major features of 22q11.2 deletion syndromes [175, 547]. I therefore concluded the variant observed herein is of uncertain significance.

Finally, a missense mutation (G206R) predicted damaging in *NKX2.5* was identified in a patient (R22) suffering from CH and congenital heart disease (CHD). *NKX2.5* is an established gene for several dominantly inherited non-syndromic CHDs [516]. A frameshift mutation segregating with disease in a CHD family has been previously observed to disrupt the same amino acid that is mutated here in our patient, a Glycine at position 206 [2]. Also, the missense mutation observed here is located within the homeodomain of the protein, where roughly one third of all CHD causative mutations have been observed [2]. Functional studies evaluating the impact of most of these missense, homeodomain alleles at the protein level have shown the mutated proteins have reduced RNA binding capacity or reduced transcriptional activity [238, 543]. Together, this suggests the mutation observed here may also equally compromise protein function and therefore contribute to the CHD phenotype of this patient, but functional demonstration is pending. Interestingly, apart from having a pivotal role in heart development, *NKX2.5* has been shown to be expressed during thyroid morphogenesis and to drive the transcriptional activation of *TG* and *TPO* in combination with *NKX2.1* [141]. That finding, plus the observation that *Nkx2-5*-null mice develop thyroid bud hypoplasia in addition to cardiac defects [46, 115, 298], initially suggested mutations in this gene could potentially underlie the pathology of both CHD and thyroid disease phenotypes. Indeed, as mentioned in the introduction, four missense mutations in *NKX2-5* were deemed causative of TD in four TD patients [115]. However, subsequent analyses have shown those published variants do not segregate with the phenotype of TD in any of the families, and functional investigations have further confirmed those variants behave similarly to the wild-type protein and thus have no discernible pathogenic role in TD [499]. Although incomplete penetrance cannot be totally excluded, there is no longer genetic evidence of a clear pathogenic effect of *NKX2.5* mutations in thyroid disease. In addition, heterozygous *Nkx2-5* knockout mice are viable and are not reported to have TD [46], indeed suggesting that the loss of one *Nkx2-5* allele is tolerated, perhaps by compensation during development by paralogue genes such as *Nkx2.1*. All in all, given the lack of clear evidence of pathogenicity of the previously reported *NKX2-5* mutations, the high number of patients with TD without *NKX2-5* mutations [9, 61, 67, 346, 378, 499], and the absence of thyroid abnormalities in *NKX2-5* mutation carriers [416], it seems unlikely that this *NKX2.5* substitution contributes to the CH phenotype of the patient, but the the co-occurrance of a thyroid phenotype alongside the CHD defect is nevertheless intriguing, and a role of *NKX2.5* as a genetic modifier cannot be excluded.

WES families

In addition to the 21 families that underwent targeted-sequencing, 27 CH families were whole-exome sequenced in this study. The variant filtering pipeline identified a total of 800 rare functional variants that segregated with disease status within families, with an average of ∼26 inherited variants per family (range = 0 - 198). Unsurprisingly, most variants (∼85%) were missense and very few (∼4%) represented LoF variants (**Figure** 3.9). In total, the 800 variants identified were distributed across 678 unique genes, the majority of which (∼51%) fit the dominant model of inheritance.

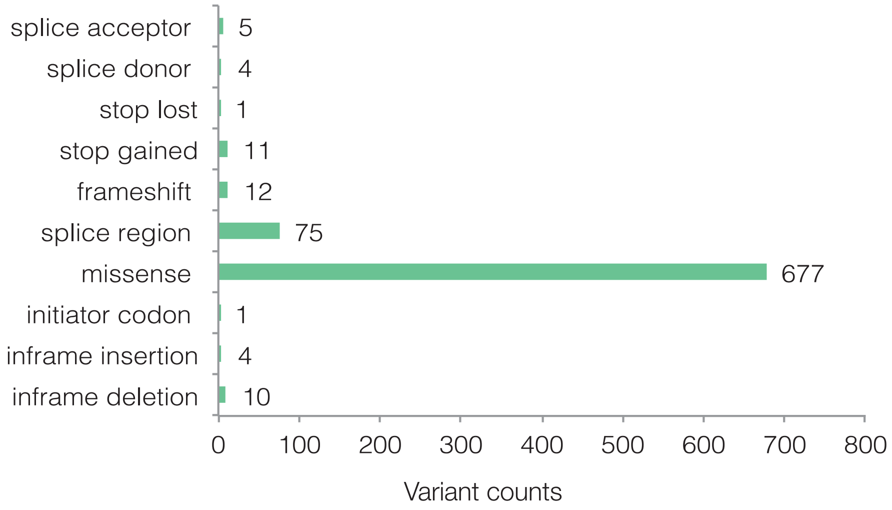

Figure 3.9 Functional consequences of the 800 inherited variants identified in 27 WES CH families.

Figure 3.10 illustrates how these 800 variants are distributed across the families and their allelic status, i.e. whether variants represent heterozygous, homozygous or compound heterozygote alleles. The diversity of pedigree structures available in the study resulted in the large variation observed in the number of variants identified per family (SD = 38.8). As expected, the singleton sample harbored the largest number of putative causal variants (N=198), whereas the smaller search space for causative variants was observed in affected siblings, who shared an average of only 8 homozygous variants. The variation seen in different multiplex pedigrees (families 28, 34 and B8) is related to the number of affected cases sequenced, while the variation seen in the trio families is mainly driven by the significantly higher number of homozygous candidates in consanguineous trios than in non-consanguineous trio families (Student *t-test* $P =$ 0.0145), as expected. The variation seen across different affected-sib families is related to the number of affected siblings sequenced (two or three), and whether data from unaffected siblings were also available for the filtering process.

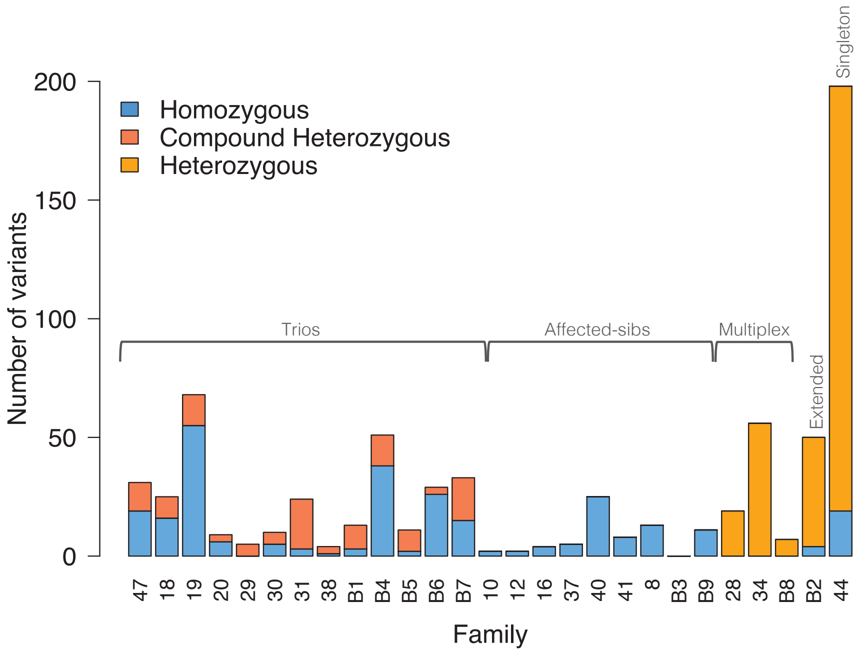

Figure 3.10 Distribution of inherited rare functional variants across families. Homozygous: variants that are homozygous for the non-reference allele; Compound heterozygous variants: in genes that have at least two variants, each of which inherited from exactly one parent. Trios: unaffected parents and affected child; Affected-sibs: two or three affected siblings; Multiplex: at least one parent equally affected; Extended family: more distant relatives equally affected (cousins); Singleton: unique index case.

3.5.2 *De novo* variation in CH trios

De novo SNV and indel events in trio families were identified using a dedicated caller, DeNovoGear. The software called around 249 variants of varying genotype posterior probabilities per trio, ranging from 155 to 402. Around 17% of the DNVs had posterior probabilities greater than 80% (**Figure** 3.11), the recommended cutoff for selecting true positive events [413].

The fraction of *de novo* mutations that were synonymous was 27%, which is very close to the expected percentage (28.6%) of synonymous DNVs based on mutation probabilities [259], suggesting the overall proportion of *de novo* events predicted to have a functional impact at the protein level in these CH trios is not significantly enriched over what would be expected by chance ($P_{binomial} = 0.49$).

After filtering low quality calls post DNV discovery (**Table** 3.7), I identified a total of nine candidate *de novo* variants (mean ~ 0.7 events per trio, range = 0-3), of which eight were SNVs and one an indel (**Table** 3.8). The average exome is estimated to contain only 0-3 DNVs [1, 365, 503], thus this result is within the expected range from known germline mutation rate and consistent with results from NGS of other disease cohorts [25, 111, 415]. Of note, no gene harbored DNVs in multiple independent trios and only four of the nine DNVs were predicted to have a detrimental effect on protein sequence by both SIFT and Polyphen-2.

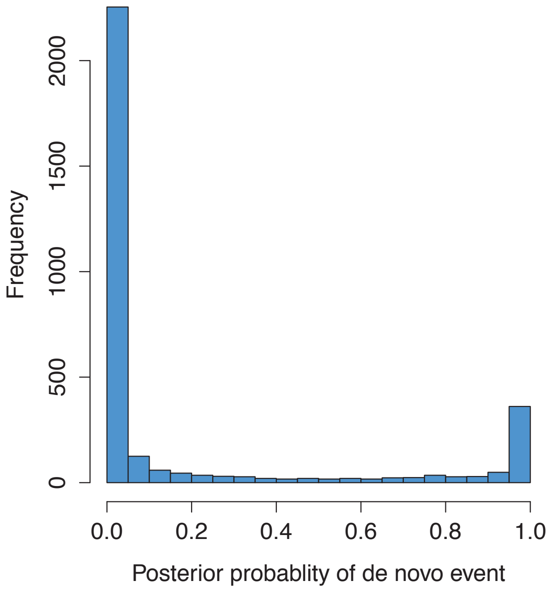

Figure 3.11 Histogram of posterior probabilities of *de novo* events called by DeNovoGear.

Phenotype	Family	Raw calls	PP_DNV > 0.8	TR/SegDup regions	>10 ALT reads parents	Rare DNVs	Functional DNVs	Independently called
Syndromic (unspecified)	31	247	50	9	5	5	3	3
Syndromic (agenesis, developmental delay)	B7	337	34	7	4	4	0	0
Syndromic (ectopy, pulmonary stenosis, renal abnormality, dysmorphic)	30	248	54	16	4	3	2	2
Syndromic (agenesis, tetralogy of Fallot)	B1	402	84	8	4	4	1	1
Syndromic (agenesis, developmental delay, hydrocephalus, short upper limbs, gastro-esophageal reflux disease)	29	240	49	14	5	3	1	1
Agenesis	20	278	37	5	1	1	1	1
Agenesis	B6	194	15	4	4	4	0	0
Agenesis	19	155	16	2	2	2	1	1
Ectopy	18	251	20	8	4	1	0	0
Syndromic (*gland-in-situ CH*, skeletal dysplasia)	B4	239	34	5	4	2	0	0
Syndromic (*gland-in-situ CH*, developmental delay, dysmorphic, metabolic acidosis)	B5	268	33	2	2	1	0	0
Syndromic (*gland-in-situ CH*, *situs inversus*, CHD)	47	203	22	5	3	2	0	0
Gland-in-situ CH	38	174	19	2	1	1	0	0
	Sum	3236	467	87	43	33	9	9
	Mean	248.9	35.9	6.7	3.3	2.5	0.7	0.7
	Median	247.5	34	6	4	2.5	0.5	0.5
	Min	155	15	2	1	1	0	0
	Max	402	84	16	5	5	3	3

Table 3.7 Summary of *de novo* calls per family along each filtering step.

PP_DNV: posterior probability given by DeNovoGear; TRSegDup regions: tandem repeats or segmental duplication regions; ALT: alternative; Independently called: whether the DNV was also called by the independent, ordinary variant caller (HaplotypeCaller in WES dataset; SAMTools and GATK in the targeted-sequencing dataset).

Phenotype	Family	Gene	Alleles	PP_DNV	CQ	AA change	Exon	Intron	GERP	1KG AF	ESP AF	EXAC AF	UK10K cohort AF	UK10K exomes AF
Agenesis	**19**	***ATXN2L***	**T>C**	**1**	**MI**	**V130A**	**3/24**	**.**	**4.88**	**.**	**.**	**.**	**.**	**.**
Agenesis	20	*RYR1*	G>A	0.806	MI	V2280I	42/106	.	5.04	.	.	0.000034	.	0.000105
Syndromic (agenesis, developmental delay, hydrocephalus, short upper limbs, gastro-esophageal reflux disease)	**29**	***HNRNPD***	**G>T**	**1**	**STOP**	**Y244X**	**5/9**	**.**	**1.83**	**.**	**.**	**.**	**.**	**.**
Syndromic (ectopy, pulmonary stenosis, renal abnormality, dysmorphic)	**30**	***SSH1***	**C>T**	**0.996**	**MI**	**R453Q**	**14/15**	**.**	**5.46**	**.**	**.**	**.**	**.**	**.**
Syndromic (ectopy, pulmonary stenosis, renal abnormality, dysmorphic)	30	*LAMB1*	TGGGG >TGG	0.999	SR	-/1786	.	33/33	.	.	.	.	.	.
Syndromic (unspecified)	31	*SEC23IP*	C>T	1	MI	T893M	16/19	.	5.31	.	.	0.000008	.	0.000105
Syndromic (unspecified)	31	*NYNRIN*	C>A	1	MI	L1385M	9/9	.	5.03	.	.	.	.	.
Syndromic (unspecified)	31	*FUK*	C>T	1	MI	R820C	19/24	.	2.2	.	0.0079	0.000017	0.000411	.
Syndromic (agenesis, tetralogy of Fallot)	**B1**	***ZRANB2***	**C>A**	**1**	**MI**	**C71F**	**3/10**	**.**	**5.53**	**.**	**.**	**.**	**.**	**.**

Table 3.8 *De novo* mutations identified in nine trios.

PP_DNV: posterior probability given by DeNovoGear; CQ: consequences at the protein level given by Ensembl Variant Effect Predictor v75; AA change: amino acid change; GERP: conservation score (ranging from -12 to 6); 1KG AF: 1000 Genomes allele frequencies, ESP AF: NHLBI ESP project allele frequencies. STOP: stop gained variant; MI: missense variant; SR: splice region variant (amino acid change is located within the 3-8 bases of the intron, and not at the two base region at the start of the intron known as the splice acceptor site). Variants highlighted in bold represent those that were predicted to be damaging by Polyphen-2 and SIFT. All variants were heterozygous.

To assess the biological candidacy of the nine DNVs, I gathered functional and biological information for the nine gene using PubMed (www.ncbi.nlm.nih.gov/pubmed), the GeneCards database (www.genecards.org), and DAVID, a functional annotation tool (https://david.ncifcrf.gov) [216]. Collectively, this included information on GO terms (biological processes, molecular function and location within the cell), KEGG and REACTOME pathways, OMIM (Online Mendelian Inheritance in Man), BioGPS (Biology Gene Portal System), NHGRI (National Human Genome Research Institute), the GWAS catalogue and model organisms (ZFIN, Zebrafish Model Organism Database; IKMC, International Knockout Mouse Consortium). For all genes, except one (*HNRNPD*), there was no clear biological link (i.e. relevant biological function or process) either to thyroid biology or the specific syndromic phenotype of a given patient, neither were genes involved in other overlapping phenotypes (i.e. affecting a relevant system) in human or in model organisms.

HNRNPD warrants discussion due to a possible link to thyroid biology and the patient's specific phenotype (**Table** 3.8). This variant has been confirmed through capillary sequencing to be a true *de novo* event (Adeline Nicholas, personal communication). HNRNPD (Heterogeneous Nuclear Ribonucleoprotein D AU-Rich Element RNA Binding Protein 1) is a protein involved in mRNA stabilization through binding to adenylate-uridylate-rich element motifs (AREs), which are present in many genes related to growth regulation, such as proto-oncogenes, growth factors, cytokines and cell cycle-regulatory genes [490]. Through literature review, I found evidence relating *HNRNPD* to the thyroid gland. Firstly, many thyroid related genes contain ARE-motifs and can be regulated at the transcriptional level. These include mRNAs related to thyroid development (*NKX2.1*, *PAX8*, *HHEX*, *EYA1*, *HOXA3*, *HOXA5*, *PAX9*), thyroid function (*SLC5A5*, *SLC26A4*, *TPO*, *DUOX1*, *TSHR*) and response to thyroid hormones and thyroid pathology [491]. It is still unclear however, whether these mRNAs are targets of *HNRNPD*. Secondly, investigations on malignant thyroid tissues revealed the expression of *HNRNPD* was increased when compared with benign thyroid tissues, and further knockdown of *HNRNPD* in thyroid cancer cell lines decreased thyroid cell proliferation [490]. Finally, the nonsense mutation found in this study (Y244X) affects a conserved amino acid located in the RNA recognition domain. This specific protein truncation has been shown to decrease the ability of *HNRNPD* to destabilize vascular endothelial growth factor RNA [145], a regulator of vascularization in development and a key growth factor in tissue repair.

The statistical significance of *de novo* findings, when multiple DNVs in unrelated probands hit the same gene, is normally assessed using a one-sided binomial test [72, 324].

Such a test incorporates the known exome mutation rate of 1.5×10^{-8} per base per generation [347], the proportion of *de novo* mutations that are expected to be functional (71.4%) or nonsense (3.4%) [259] (depending on which were identified), the sample size of the cohort under study, and the length of the coding sequence of the gene of interest. In this case, given a single DNV hit, I evaluated the level of background nonsense variation in *HNRNPD* in the ExAC database (N=60,706). This revealed a single heterozygous nonsense variant (4:83280743, T>A) private to a single sample, suggesting LoF mutations are infrequent in this locus. Further supporting this observation is the fact that *HNRNPD* is in the top 16% of genes in the genome more intolerant to LoF mutation (**Figure** 3.12) [431], with a pLI score (probability of loss-of-function intolerance) of 0.96, slightly above the pLI threshold of 0.9 that defines extremely LoF intolerant genes [135]. These findings point towards the truncation of this protein being pathogenic, possibly interfering with downstream HNRNPD-target interactions.

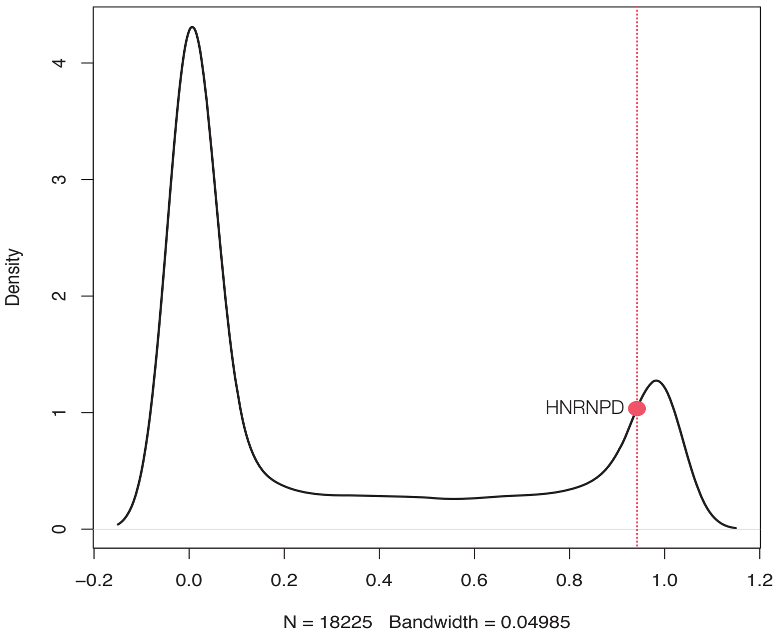

Figure 3.12 Distribution of intolerance to loss-of-function (LoF) mutation (ExAC pLI score) for all genes in the genome (N=18,225). The red dot represents *HNRNPD*, with a score of 0.96. This graph made use of data taken from Samocha *et al* [431].

The mutation identified in *HNRNPD* (Y244X) was seen in a syndromic patient (F29) aged 10 who presented with thyroid agenesis and extrathyroidal features, including marked body disproportion, short arms and legs, hydrocephalus, gastro-esophageal reflux disease, bilateral sensorineural hearing loss, nasal bridge, and global developmental delay (delayed speech and behavioral difficulties). Since the patient was heterozygous for a likely pathogenic nonsense mutation, Dr Nadia Schoenmakers searched the DECIPHER database [149] and the literature for deletions involving the gene. There were 14 patients with chromosomal deletions containing *HNRNPD* and additional genes in DECIPHER, and additional nine cases of 4p21 *HNRNPD*-containing microdeletions, of various sizes, in the literature [44, 51]. The patient in our study and the 23 deletion cases shared several phenotypic features, in particular the skeletal phenotype (short hands and feet, nasal bridge, macrocephaly) and the intellectual impairment.

The difficulty in further interpreting the *HNRNPD* Y244X finding in the context of the F29 patient phenotype is that, although there is good variant-level evidence to suggest Y244X may be a pathogenic mutation (i.e. it occurs in a gene known to be intolerant to LoF variation) and likely to account for some of the extrathyroidal phenotypes seen in this patient (i.e. the skeletal phenotype), none of the DECIPHER or published haploinsufficiency cases were reported to have thyroid abnormalities. One explanation for this discrepancy could be that the impact of a stop mutation is different to that of a deletion. A nonsense HNRNPD mutant can potentially still bind RNA to some degree and act as a dominant-negative mutant, which could result in the more severe CH phenotype observed in the F29 patient. This hypothesis remains to be experimentally validated.

3.5.3 Copy-number-variants in the WES dataset

CNVs in WES CH families were discovered using CoNVex, which called an average of 187 CNVs per sample. Of these, around 60% (110/887), on average, passed the post-discovery QC per sample, with the majority (∼97%) overlapping with common (>1% AF) population variants. There was an average of 0.5 rare (<1% AF) CNVs encompassing protein coding sequences with more than one probe per sample (range from 0 to 3). After inspection of regional $\log_2$ plots a total of 10 rare CNVs remained in the WES cohort, with sizes ranging from 920 bp to 65.7Kb.

Importantly, none of these rare CNVs encompassed known CH genes, nor did they overlap with genes contained in previously reported CNV regions [363, 485, 494]. **Figure** 3.13 illustrates the 10 CNVs identified in families (four duplications and six deletions). None of these CNVs fit the model of inheritance expected (*de novo* or recessive inheritance) in each of these families, as variants were either present in one of the unaffected parents in trio families (**Figure** 3.13 A, B, D and E) or absent in one of the multi affected siblings (**Figure** 3.13 C).

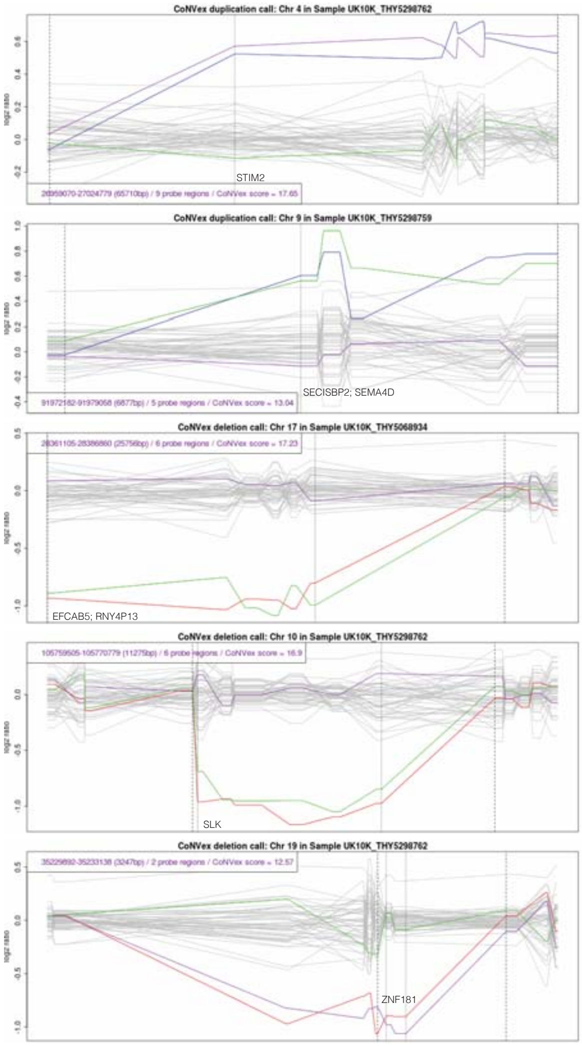

Figure 3.13 Log$_2$ ratios of rare (<1% AF) CNVs identified by CoNVex. (A) F31, (B) F30, (C) F12, (D) F31 and (E) F31.

The x-axis indicates the genomic coordinates. The y-axis indicates the normalised log$_2$ ratio of the exome read depth, compared to a group of controls. The red line shows the log$_2$ ratio of the proband, where the variant is a deletion, and the blue line shows the log2 ratio of the proband where the variant is a duplication. The purple line shows the log$_2$ ratio of the mother, and the green line shows the log$_2$ ratio of the father. The grey lines show the log$_2$ ratio of control samples. The vertical small dashed lines show the minimum deleted/duplicated region and the vertical wide dashed lines show the maximum deleted/duplicated region. The protein-coding genes present in each region are also represented.

3.5.4 Searching for novel genetic causes of CH across families

After identifying both inherited and *de novo* variants segregating with disease in each CH family, I determined whether any genes showed such variation recurring across families. The vast majority of candidate genes (∼96%) harboring *de novo* (N=9) or inherited variants (N=678) in the combined WES and targeted-sequence datasets were mutated in only a single family. Of the remaining 26 loci, 22 harbored variants in two independent families, three genes (*AHNAK2*, *MUC16* and *MUC4*) were detected in three families and *TTN* was recurrently mutated in four families. Interestingly, *AHNAK2*, *MUC16* and *TTN* are genes that are systematically mutated in multiple exome analyses [223, 347, 352, 366, 432] due to their high tolerance to mutation and/or their exceptionally large sequences, where many variants fall by chance [448].

To assess the statistical significance of the recurrent observations in each of the 26 genes, I conducted a case-control analysis. To do so, I used a control dataset comprised of 2,120 unrelated healthy individuals from the INTERVAL study (www.intervalstudy.org.uk) that had also been exome-sequenced at the WTSI. This cohort of healthy individuals was assembled by the NHS Blood and Transplant England (www.nhsbt.nhs.uk), and consists of more than 50,000 individuals that donated blood at NHS blood donation centres across England. These data were generated by the Human Genetics Informatics team using the same alignment and variant calling protocol that was previously used to call the WES CH dataset. To ensure, as much as possible, that cases and controls harbored broadly similar quality metrics, I applied the same variant- and sample-QC steps that I used in the WES CH dataset. This resulted in globally similar profiles in both datasets for Ts/Tv ratios (median $\sim$ 2.7), genotype qualities (median $\sim$ 94) and Het/Alt ratios (median $\sim$ 2.4). It is important to note however, that this control dataset is not perfectly matched to the case cohort in terms of depth of sequencing and ethnic composition. In the case of depth, controls were sequenced at a median of 45x while the WES CH cases had a median coverage of 59x. In terms of ethnicity, all controls were of European descent whereas 49% of all CH cases were non-European individuals encompassing an array of diverse ethnicities, as highlighted by my PCA analysis. The combination of these two factors meant that the case samples harbored a larger number of rare functional variants exome-wide than controls ($\text{mean}_{\text{cases}}$ = 1285 vs $\text{mean}_{\text{controls}}$ = 1124), so any statistically significant finding resulting from this analysis would need to be carefully inspected further to examine its real validity.

I tested the enrichment, per gene, of rare functional variants in cases versus controls using a one-tailed Fisher's exact test. The one-tail test here basically assumes that rare variation implicated in Mendelian disease is damaging rather than protective, so the alternative hypothesis being tested is that cases have a higher number of such variants than controls. Variants that were shared by relatives of the same family were counted only once, so the effective number of "cases" used here was the total number of families that were sequenced at that gene (note that for the majority of genes, only the WES families were informative, since the targeted-sequenced samples were only sequenced for 20 genes).

None of the 26 genes reached the conservative, yet standard, exome-wide significance threshold of 1.7×10^{-6} [300] (**Table** 3.9). Of all 26 genes, only one (*DUOX1*) represented an interesting biological candidate, in which one exome sample (F44) harbored another variant in *DUOX1* in addition to the targeted-sequenced sample (R8) that was previously discussed. Both of these samples, who were singletons, shared the same *gland-in-situ* CH phenotype, but the likelihood of this finding happening by chance is high ($P_{\text{uncorrected}}$=0.1967), according to this analysis. For reference, three genes (*CACNA1A*, *FLIP1* and *OBSCN*) were recurrently mutated in families sharing exactly the same phenotype (i.e. agenesis or *gland-in-situ* CH), and *CBFA2T2* and *KLHDC4* recurred in TD families with varying TD defects. None of those variants however had consistent mode of inheritance across the different families. Finally, amongst the list of recurring genes, only one (*TLN2*) contained LoFs or missense variants predicted to be damaging, which highlights the small search space for likely causative variants offered by this CH cohort.

Despite the negative results, this analysis illustrates the importance of using control exomes in evaluating and interpreting rates of mutation in genes. For instance, *OBSCN* and *TTN*, which are the largest genes in the exome, displayed P-values close to 1, meaning there is indeed no difference between groups other than due to random variation, i.e. healthy individuals carry a large number of rare functional variants in these two genes by chance so observing such variants at this rate in a case cohort of this size is not surprising.

	Cases		Controls			
Gene	Carriers	Non-carriers	Carriers	Non-carriers	*P*-value	OR
TRBV6-9	2	25	1	2119	4.54E-04	164.92
GGT1	2	25	4	2116	2.22E-03	41.80
ZNF623	2	25	4	2116	2.22E-03	41.80
CBFA2T2	**2**	**25**	**8**	**2112**	**6.44E-03**	**20.97**
MUC4	3	24	41	2079	0.0167	6.32
FILIP1	2	25	16	2104	0.0206	10.48
KLHDC4	2	25	17	2103	0.0228	9.86
ATP2B3	2	25	18	2102	0.0252	9.31
SCNN1A	2	25	18	2102	0.0252	9.31
ADAMTS15	2	25	19	2101	0.0276	8.82
DMBT1	2	25	20	2100	0.0301	8.37
AHNAK2	3	24	79	2041	0.0813	3.23
PTPRB	2	25	37	2083	0.0849	4.50
TLN2	2	25	40	2080	0.0964	4.15
CACNA1A	2	25	52	2068	0.1464	3.18
MUC17	2	25	59	2061	0.1776	2.79
DUOX1	**2**	**46**	**35**	**2085**	**0.1967**	**2.59**
DCHS2	2	25	67	2053	0.2145	2.45
LRP1B	2	25	69	2051	0.2239	2.38
SPTBN5	2	25	72	2048	0.2380	2.27
LRP2	**2**	**25**	**85**	**2035**	**0.2996**	**1.91**
HSPG2	2	25	102	2018	0.3792	1.58
FSIP2	2	25	117	2003	0.4467	1.37
MUC16	3	24	237	1883	0.5957	0.99
OBSCN	2	25	228	1892	0.8029	0.66
TTN	4	23	505	1615	0.9134	0.56

Table 3.9 Case-control burden analysis for 26 genes recurrently mutated in independent CH families.

Carriers: number of families (if case cohort) or samples (if control cohort) with at least one rare functional variant in a given gene. Non-carriers: number of families (if case cohort) or samples (if control cohort) without a rare functional variant in a given gene. Note if gene was sequenced in both the WES and targeted-sequencing experiments, then the total of CH families used in the test is 48. If the gene was only exome-sequenced and not included in the custom array, the total of CH families used in the test is 27. The total of control samples used in the test is always 2,120. P-values taken from the Fisher's exact test one-tail, which assumes rare functional variation in rare Mendelian diseases is damaging and not protective, so only assumes one direction of effect. P-values are uncorrected for multiple testing but none reach the standard exome-wide significance threshold of 1.7×10^{-6} [300]. OR: odds ratio given by the test. Table is sorted by P-value.

3.5.5 Searching for likely damaging variants in candidate genes

To leverage the data generated by this project, and since no gene was recurrently mutated across CH families at a higher rate than expected, I conducted a candidate-gene approach where I searched for likely damaging variants in long standing CH candidate genes. The motivation behind this analysis was that it could potentially reveal disrupted biologically meaningful loci, which could then be screened in additional CH cohorts and in interrogated in future CH-mapping efforts.

To define candidate genes, I collected information from several biological sources (**Table** 3.10). Candidates most relevant for TD phenotypes included loci that have been directly implicated in thyroid development based on mouse or zebrafish knockout studies [140]. Another relevant source for thyroid abnormalities was a recent microarray study that defined gene expression profiles in the mouse thyroid and lung primordia at embryonic day 10.5 [137]. The output of this work is relevant here because the thyroid and lungs originate as neighbouring bud shaped outgrowths, and it is possible that genes affecting both systems may be implicated in thyroid development defects. My list also included putative novel targets of *FOXE1* and *PAX8* that were identified through transcriptomic analysis of thyroid follicular cells after separate knockdown of each gene in mice [119, 146]. These knockdown candidates are relevant for TD because such abnormalities could result from defects in transcription factors/mediators acting downstream of *FOXE1*/*PAX8* and that cooperate in maintaining key cellular processes for thyrocyte biology.

Candidate gene list	Number of genes
Mouse models of TD	22
Zebrafish models of TD	4
Genes enriched in thyroid bud at E10.5	42
Genes enriched in thyroid bud and lung at E10.5	39
Targets of *FOXE1*	52
Targets of *PAX8*	13
GWAS loci associated with TSH and T_4 levels	17
Other candidates from collaborators	70
DDG2P genes	1952

Table 3.10 CH candidate genes compiled from different sources.

Candidate genes most relevant for *gland-in-situ* CH phenotypes included loci that were found to influence physiological TSH and free-T_4 levels in a recent whole genome sequencing study [474]. The hypothesis here was that hormone production defects could result from mutations in genes that are not necessarily located within the follicular unit but that are nevertheless involved in hormone biology.

Finally, because 35% of our CH families (i.e, 17 out of 48) were composed of patients with syndromic forms of CH, my candidate gene list also included 1,952 genes that have been linked to developmental disorders. This list was extracted from the Development Disorder Genotype-Phenotype database (DDG2P), which is a curated list of genes compiled by clinicians as part of the of the Deciphering Developmental Disorders (DDD) study. This DDG2P list is categorised into the level of certainty that the gene causes developmental disease (confirmed or probable), the consequence of a mutation (loss-of function, activating, etc) and the allelic status associated with disease (monoallelic, biallelic, etc), information which was taken into account in my analysis. The full list of candidates, excluding the DDG2P genes which can be found at DECIPHER (https://decipher.sanger.ac.uk/), is included in Appendix **Tables** A.5 and A.6.

To focus on alleles that are more likely to be impactful, I restricted the candidate-gene analysis to the list of *de novo* and inherited variants that were LoF or missense predicted to be damaging by both SIFT and Polyphen-2. This meant only around 28% of the genes and just over 25% of their variants were considered here. Two biallelic variants in two separate genes (*HSPG2* and *SLC26A7*) were identified in two independent families.

A biallelic variant (g.1:22178283 C>T) in *HSPG2* was identified in a consanguineous patient presenting with *gland-in-situ* CH and skeletal displasia. Both parents were obligate carriers. *HSPG2* is a DDG2P gene that is responsible for skeletal displasia phenotypes such as the Schwartz-Jampel (OMIM: 255800) and the dyssegmental dysplasia Silverman-Handmaker type syndromes (OMIM: 224410). The variant identified in the patient, a substitution in the splice donor site of the 54^{th} intron (out of 96) of *HSPG2*, is consistent with the general mechanism of disease of *HSPG2* defects, i.e. biallelic LoF mutations. My MaxEntScan analysis predicted this mutation to be deleterious, as it shifted the strength of the splice signal from the 60% down to the 9% percentile (WT score: 9.22; mutant score: 1.22). Overall, this suggests this mutation is highly likely to explain the skeletal phenotype of this case but since congenital hypothyroidism is not a feature of *HSPG2* defects, the CH phenotype of our patient is unlikely to be linked to the *HSPG2* defect identified herein. A closer look at the variants segregating with disease in this trio, revealed a biallelic splice region variant in *DUOXA2*, where both parents were also carriers. Even though this variant affects the 3-8 bases of the intron, and not the splice acceptor region, it is possible that it may be contributing to the CH phenotype of the patient. Functional studies are ongoing to try and understand if this is the case.

A biallelic stop mutation (R277X) disrupting *SLC26A7* was observed in two consanguineous siblings (F16) suffering from *gland-in-situ* CH. For reference, no homozygous LoF variants in *SLC26A7* were observed in ExAC individuals. *SLC26A7* encodes a sulfate/anion transporter transmembrane protein thought to be mainly expressed in the stomach and kidney [125, 388], and it belongs to the same family of *SLC26A4*, the iodide transporter in the thyroid follicular cells. Both genes are multifunctional anion exchangers, sharing the same biological REACTOME pathways of transmembrane transport of small molecules and transport of inorganic cations/anions and amino acids/oligopeptides. Several lines of evidence support *SLC26A7* as a strong candidate for thyroid hormone defects. First, mouse *Slc26a7* has been shown to be capable of both chloride and iodide transport [247] and, in a recent study, *Slc26a7* knockout mice (N=7) exhibited hypothyroidism, with serum T_4 levels reduced by 87% in males (P<0.001) and by 47% in females (P=0.003) [58]. Further histological observations showed hyperplastic (i.e. enlarged) thyrotrophs in male mice [58]. In addition, down regulation of *Slc26a7* was observed in another study that conducted microarray measurement of thyroid RNA levels at embryonic day 18 in double *Nkx2.1+Pax8* null mouse [16]. Finally, RNAseq of thyroid tissues (N=112) as part of the GTEx resource, revealed a clear overexpression of *SLC26A7* in the thyroid over kidney and stomach tissues, suggesting this gene may indeed have a more prominent role in thyroid biology. This latter finding is important because it results from a much more comprehensive and robust assay than initial *SLC26A7* expression studies [125, 388], as it assayed more tissues and more samples, respectively. A collaborator of Dr Nadia Schoenmakers subsequently screened other cases of CH with GIS in whom linkage had not identified a likely known genetic cause, and found the same mutation in two different families, involving two different haplotypes.

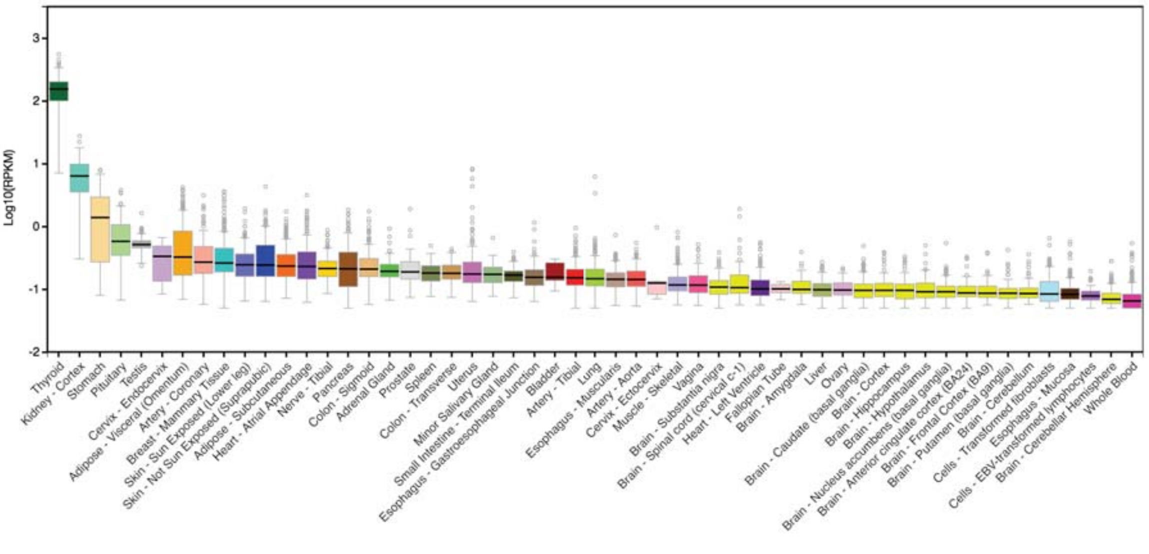

Figure 3.14 *SLC26A7* expression in GTEx tissues (www.gtexportal.org.)

3.6 Discussion

Even though CH is easily circumvented by hormonal therapy, such that patients are able to have growth and mental development that is as close as possible to their genetic potential, the etiology of CH remains a long lasting enigma in the pathophysiology of human endocrine diseases [59].

To try and elucidate novel genetic mechanisms contributing to CH, my collaborators and I conducted whole-exome and targeted-sequencing of 48 CH families for whom no causative mutations in known genes have been identified to date. This study was the first to employ next-generation sequencing technologies in a cohort of CH families, and also the first one to comprehensively screen for the presence of likely pathogenic variants in long-standing CH candidate genes. To narrow the search space for causative variants, I developed and implemented several variant filtering pipelines to identify rare inherited variation segregating with disease within CH families, as well as protein-altering *de novo* and CNVs events. No gene carrying such types of variation recurred significantly mutated across families, meaning this study was unable to identify a novel CH-associated gene, which is unsurprising given the high phenotypic variability and small sample size of our case cohort.

3.6.1 A putative causative gene for CH with *gland-in-situ*

The candidate-gene approach leveraging the data produced in this study, identified a novel LoF mutation in *SLC26A7*. This gene represents a putative causative gene for CH with *gland-in-situ* that is related to the classical *SLC26A4* gene, leads to CH when deleted in mice [58] and is overexpressed in thyroid tissue, as I demonstrated. The same mutation was subsequently identified in two additional unrelated *gland-in-situ* CH families external to this study, and no homozygous LoF alleles were found in this gene in the ExAC dataset. The present hypothesis is that this anion transporter contributes to iodide uptake in thyrocytes and that recessive mutations in the gene lead to hypothyroidism in humans. Functional experiments are currently being performed by Dr Nadia Schoenmakers and collaborators to understand the mechanism by which these mutations contribute to disease. Specifically, they are investigating thyroid follicular cell localisation of this molecule, together with *in vitro* assessment of iodide transport in transfection studies, and detailed assessment of thyroid physiology in the affected patients and in *Slc26A7* knockout mice.

3.6.2 *De novo* and CNVs in TD and syndromic CH

The mapping of *de novo* and CNVs events in this study was particularly relevant for thyroid dysgenesis and syndromic CH phenotypes, as *de novo* and CNV variants represent prime candidates to explain the general lack of clear familial transmission that is typical of TD phenotypes [97, 101, 114, 136, 541]. Most TD and syndromic families (14/20) included in this study were non-familial, in agreement with what is generally seen. Of those that were exome-sequenced (N=7), the read-depth analyses conducted here suggested CNVs do not contribute to disease in those families, as all rare or novel CNVs present in patients were also observed in unaffected relatives. Of those that were trios, four harbored rare and predicted damaging functional *de novo* variants, but only one family harbored a *de novo* event (in *HNRNPD*) that could potentially be relevant to thyroid biology and the specific syndromic phenotype of the patient. The lack of biological candidacy for most of the *de novo* variants identified here is not limited to our study and, in fact, the clinical significance of most *de novos* detected in the plethora of trio studies published so far remains unclear [210]. The lack of genes with recurrent *de novos* in independent families is also unsurprising, given the small number of trios available for study and the heterogeneity of phenotypes of those patients. As an example, yet perhaps more extreme than the case of CH, given the (presumed) higher locus heterogeneity and the different genetic architecture [258, 365], a total of 238 ASD trios were needed to identify a recurrently mutated gene in two unrelated families [432]. Future studies aimed at elucidating whether *de novo* variation contributes to TD and syndromic CH phenotypes will certainly need to recruit substantially larger cohorts.

3.6.3 Limitations

Sample size limitation is something that is not limited solely to the present study. Recruitment of large patient cohorts of any rare human condition is especially challenging and several other Mendelian disease studies have reached the same conclusion [160, 199]. To increase the sample size of the current study, additional 50 CH patients have already been recruited by Dr Nadia Schoenmakers and are currently undergoing exome-sequencing. While this number is still modest, it certainly represents a step forward in the gene-mapping path of CH phenotypes, and repeating the analyses presented here in the expanded patient collection may prove fruitful.

Apart from sample size, I have identified several other factors that may have hindered the success of this project and that should be pondered over carefully when designing future genetic studies of CH. The main factor is perhaps the heterogeneity of the phenotypes collected, as previously mentioned, which definitely influences the likelihood of mutational recurrence in a given gene in unrelated families, since this is known to be inversely correlated with genetic heterogeneity [210]. Approximately 30% of the families included in this study displayed extrathyroidal features affecting an array of different systems such as the brain, skeleton, kidneys, ears and lungs. Including syndromic patients, in addition to isolated cases, in genetic studies of CH is advantageous because the prior probability of detecting a genetic defect is higher. However, including such patients can also compromise the ability to statistically implicate novel genes when the total cohort size is small, as observed here. Further, it may also be difficult to discern *a priori* whether the CH phenotype observed in syndromic-CH cases is directly linked to the extrathyroidal phenotype of the patient, or whether it represents a parallel and coincidental manifestation. This issue is well illustrated in my results, with the identification of likely causative variants in well established disease genes (*NKX2.5* and *HSPG2*) that very likely explain the extra clinical phenotypes observed in the patients (congenital heart disease and skeletal dysplasia, respectively) but that are unlikely, on the other hand, to play any role in the aetiology of their thyroid hormone deficiency, which remains unsolved. Similarly, evidence of "blended" and often complicated phenotypes resulting from multiple-gene defects have been documented in two recent clinical exome-sequencing studies: Yang *et al* [536] reported that, of 504 patients with a molecular diagnosis, 23 (4.6%) had a phenotype resulting from two single-gene defects, and Retterer *et al* [417] identified 25 patients (out of 3,040 probands) that had two concurrent genetic diagnoses and three with three distinct genetic diagnoses.

Finally, the diverse ethnicities of the cases meant the case-control analysis performed here did not use appropriate control data. Sequencing of healthy population individuals has been mainly conducted for European populations and appropriate control sequences for other ethnicities are still lacking. Ultimately, this did not represent an issue for this study because no gene was significantly enriched for variation more than expected by chance. However, if the opposite had been the case, one would certainly need to further interrogate control sequences matched on ancestry to ensure the finding was not driven by population stratification.

3.6.4 Future work

The genetic hypothesis explored in this study was that fully penetrant single-gene defects cause CH. Future studies going forward should explore the role of more complex genetic aetiologies. One possibility is a digenic mode of inheritance, where the variant genotypes at two loci, each transmitted from different parents, affect two independent genes that interact in a way to manifest the phenotype [434]. Such a model could account for the apparent sporadic nature of TD and also explain the incomplete penetrance and variable expressivity that is often observed in familial TD cases [92, 97, 412, 486]. The *Pax8/Nkx2-1* murine model exemplifies the role of digenicity in thyroid dysgenesis, since only mice doubly heterozygous for the two null alleles manifest a phenotype [16]. In humans, evidence of a digenic inheritance came from a single dysgenesis patient who carried heterozygous mutations in *NKX2.5* and *PAX8* [201]. It is challenging to investigate digenic causes of disease in an exome-wide manner, since it is not always clear whether a given digenic observation represents a true digenic case or simply a co-inheritance of two mutations by chance. Future investigations could, for example, focus on recruiting large numbers of both affected and unaffected trio families, and then look for rare coding variants in gene pairs that are transmitted to the affected offspring more often than in the offspring of controls. However, assuming there are ∼21,000 protein-coding genes in the exome, the search space for gene-pairs would be huge ($2.1\text{x}10^8$ unique gene pairs) and, in addition, there would be scant biological evidence to support the vast majority of potential interactions. A more fruitful alternative may be to only consider genes that have proven protein-protein interactions [434]. This would also facilitate the development and interpretation of downstream experimental studies aiming to convincingly implicate a mutated gene-pair in disease.

In sufficiently large datasets, future studies will also be able to test formally for an incomplete penetrance model using, for example, a modified version of a Transmission Disequilibrium Test (TDT) [10]. TDT tests comprise a group of family-based association tests to detect the distortion in transmission of alleles from a heterozygous healthy parent to their affected offspring [458]. A simple modification to this test would accept the collapsed counts, per gene, of transmitted and non-transmitted rare functional alleles across cases and control trios. By using the transmission of silent alleles as internal control, one would be able to detect whether there is significantly over-transmission of rare protein-altering alleles to the affected offspring for a given gene. A similar approach was used successfully in an autism study, where a significant maternal transmission bias of private truncating SNVs in conserved genes were observed in probands in comparison to unaffected siblings [258].

In a more complicated scenario, the apparent sporadic nature of TD can result from a two-hit model combining a germinal mutational hit (consistent with the rare occurrence of familial cases [74]) and a somatic mutation in the thyroid tissue. A much less common congenital endocrine disorder, focal hyperinsulinism, has been shown to result from such a model: in the pancreatic lesions found in these patients, a paternally inherited mutation in the *SUR1* gene is found together with the loss of a maternal 11p15 allele (loss-of-heterozygosity, LH), a locus which contains many imprinted genes. In this case, the LH event is a somatic event restricted to the pancreas, which explains why focal congenital hyperinsulinism is a sporadic disease with a genetic aetiology. The same model, affecting haploinsufficient genes specific for thyroidal morphogenesis, could be involved in TD. To accelerate these discoveries, studies investing in the creation of animal models with a thyroid-specific conditional inactivation of a given gene should be initiated.

Finally, new mutations that contribute to thyroid dysgenesis should be sought in introns and regulatory regions, such as the 3' and 5' UTRs, where microRNAs bind, in addition to the coding regions of genes. Nonclassical mechanisms of disease involving epigenetic changes should also not be forgotten, as these could account for differences in the phenotypic expression and incomplete penetrance of CH [97, 147, 445]. Epigenetic mechanisms, particularly DNA methylation, have been shown to contribute to the development of several endocrine and metabolic diseases [163] including Beckwith–Wiedemann syndrome [113], pseudohypoparathyroidism type IA [268]), as well as thyroid cancers [254], suggesting it may indeed represent a potentially relevant pathogenic mechanism involved in congenital hypothyroidism.