Chapter 2

NGS-based screening of known causative genes in CH with *gland-in-situ*

2.1 Introduction

2.1.1 What is congenital hypothyroidism?

Congenital hypothyroidism (CH) is a rare condition of thyroid hormone deficiency, occurring in 1 of 3000-4000 newborns [36] due to a complete or partial failure of thyroid gland development or thyroid hormone production.

Thyroid hormones, triiodothyronine (T3) and thyroxine (T4), are tyrosine and iodine-based hormones produced by the thyroid gland; they are responsible for regulating vital metabolic processes for normal growth and development and are particularly important for the correct myelination and maturation of the brain, a process that starts in utero but that extends into postnatal life [69, 339]. Consequently, severe hormonal deficiency can cause irreversible cognitive impairment and neurological damage if not promptly treated. In the 1970s, CH was the most common neonatal endocrine disorder and also the leading preventable cause of intellectual disability [36]. The introduction of neonatal screening programs in most developed countries in late 1970s/early 1980s enabled early detection of the disease and initiation of thyroid hormone replacement therapy (Levothyroxine) [124]. This decision transformed the outlook for children with CH so that severe growth and mental retardation as a consequence of CH is now rarely seen.

Routine screening includes serum thyroid function tests, such as measurement of thyroid hormone (T4) and Thyroid-Stimulating Hormone (TSH) levels. Further investigations may, if needed, include thyroid imaging and anti-thyroid antibody determinations, to rule out autoimmune thyroid disease [414]. Biochemical diagnosis of CH is confirmed by demonstrating reduced circulating levels of T4 in response to elevated levels of TSH, the pituitary hormone that stimulates the thyroid gland to produce T4.

Most newborns with CH have no or only subtle, non-specific symptoms at birth, including feeding difficulty, lethargy and constipation. However, in severe cases, suspicious signs at birth include an enlarged neck (goitre), excessive intrauterine growth, and prolonged jaundice [414]. In almost all cases, the thyroid phenotype is isolated, however, it may also be seen alongside other congenital abnormalities, resulting in distinct clinical phenotypes. Examples of co-morbid features include sensorineural hearing loss, cardiac defects, spiky hair, cleft palate, neurologic abnormalities and genitourinary malformations [414]. I refer to these CH manifestations as "syndromic CH", and will cover them in greater detail in the following chapter. CH can also be classified into permanent or transient CH, depending whether or not there is a persistent deficiency of thyroid hormone that requires life-long treatment.

Historically, thyroid developmental defects were thought to account for approximately 85% of CH cases [171, 373], with the remaining resulting from impaired hormone production within a structurally normal gland or *gland-in-situ* (GIS). However, recent observational studies have reported a doubling in CH incidence, reaching 1 in 1,500 live births [99, 194], predominantly driven by an increase in CH with GIS, which accounted for almost two-thirds of recently diagnosed cases in a region of Italy [99]. Decreased TSH cutoffs upon screening may be the major drive for this increase in diagnosis, although changes in the demographic composition of the screened population, increased multiple and premature births, misclassification of transient forms of CH as permanent and variable iodine status, very likely contribute [376, 385].

2.1.2 The known genetics of CH with *gland-in-situ*

The molecular basis of CH with GIS remains poorly understood [229, 409]. Genetic defects in eight genes involved in thyroid hormone biosynthesis (*TG*, *TPO*, *DUOX2*, *DUOXA2*, *IYD*, *SLC5A5*, *SLC26A4* and *TSHR*) are known to mediate some cases. **Figure** 2.1 illustrates the role of the proteins encoded by these genes within the thyroid hormone production pathway. Disease-causing mutations in these loci are

usually biallelic and are thus inherited in an autosomal recessive manner, with the exception of monoallelic *DUOX2*, *IYD* and *TSHR* mutations, which may also result in CH (**Table** 2.1). Biallelic mutations in *SLC26A4* result in a syndromic CH phenotype (Pendred syndrome, OMIM: 274600) clinically defined by goiter and congenital bilateral sensorineural hearing loss [266], because in addition to being expressed in the thyroid tissue, the gene is also expressed in the inner ear [178].

Similar to many other Mendelian diseases [50, 166, 292, 421, 430], there is considerable inter- and intra-familial phenotypic variability in CH cases harboring the same causative mutation [179, 342], suggesting that both mono and polygenic factors, as well as environmental modulators, may play a role in determining disease severity [26]. While there have been occasional reports of digenic mutations, involving *TSHR* and either *DUOX2* [229, 409] or *TPO* [460], the role of oligogenicity in disease development and modulation of disease penetrance remains unclear, with no evidence for an additive effect of digenic mutations in one large published kindred [460].

2.1.3 Previous genetic studies of CH with *gland-in-situ*

Genetic characterization of CH with GIS has been limited by the cost and labour implications of Sanger sequencing multiple exons: collectively, these eight genes encode a total of 148 exons. Therefore, previous studies have generally focused on either a small number of genes (e.g. *TG*, *TPO*, *TSHR* and *DUOX2* in 43 Korean cases) [229], specific phenotypic subsets of cases [342, 409], or multiple genes in a small cohort of patients [345]. Recently, large-scale multiplex genetic screening of *TPO*, *TSHR*, *DUOX2*, *DUOXA2*, *SLC5A5* and *PAX8*, a transcription factor involved in thyroid gland development [372], was conducted for the first time in a cohort of 170 CH patients from Korea. However, *TG*, *IYD* and *SLC26A4* were not included in the sequencing panel of that study, and the patients were not selected on the basis of thyroid morphology, meaning some may have been incorrectly defined as *gland-in-situ* patients.

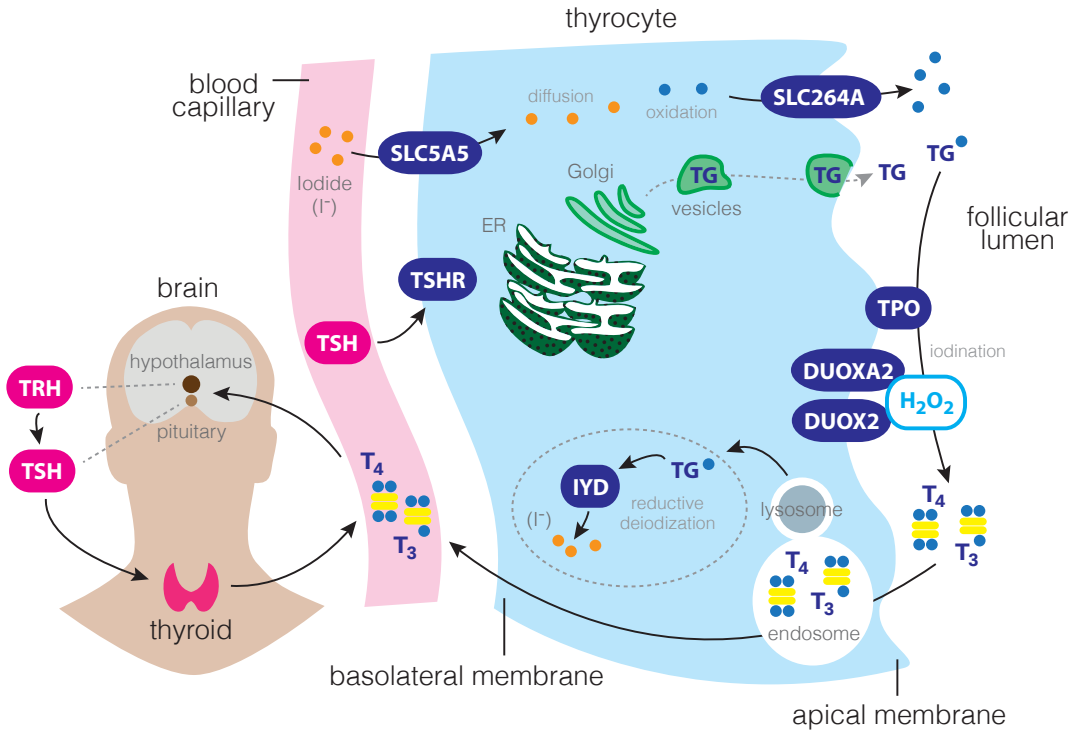


Figure 2.1 Key steps and players involved in thyroid hormone synthesis and regulation.

1) Thyroid hormones are secreted from the thyroid gland under the tight regulation of the hypothalamic-pituitary-thyroid axis, which ensures a negative feedback control dependent on the concentration of blood-circulating thyroid hormones [120]. Thyrotropin-releasing hormone (TRH), secreted from the hypothalamus, acts upon the pituitary gland to induce thyroid-stimulating hormone (TSH) synthesis and secretion; 2) TSH binds to the TSH receptor (TSHR) located on the basolateral membrane of thyrocytes; 3) ingested iodide (I-), the rate-limiting substrate for hormone synthesis, is transported in the plasma to the thyrocytes, where it is actively transported by the sodium-iodide symporter (*SLC5A5*) and concentrated into the thyrocyte cytoplasm; 4) intracellular iodide is then transported by pendrin (*SLC26A4*) into the follicular lumen; 5) the thyrocyte endoplasmic reticulum (ER) synthesizes TG and TPO proteins, which are transferred to the apical surface via exocytotic vesicles; 6) on the luminal side, TPO oxidises iodide using H_2O_2 produced by *DUOX2* and *DUOXA2* and attaches it to tyrosyl residues of the intrafollicular TG (a process known as iodination or organification); 7) after a variable period of storage in the follicles and when thyroid hormone is needed, iodinated TG is retrieved by phagocytosis and is subject to proteolysis in lysosomes to generate T3 and T4 hormones; 8) IYD subsequently recycles iodide and tyrosine to be reutilised in subsequent hormone synthesis; 9) lastly, T3 and T4 are secreted into the circulation and carried to target tissues via thyroid binding globulins.

Gene	Protein	Affected process	Mutation type	Characteristic features	Biochemical phenotype	CH duration	Diagnostic test
*TPO**	thyroid peroxidase	iodination	Biallelic	TIOD/PIOD	Severe/mild CH [171, 423]	P	ClO_4- discharge test
*TG***	thyroglobulin	substate for hormone synthesis	Biallelic	Absent/very low serum TG levels	Severe/mild CH [471]	P	Serum TG levels
DUOX2	dual oxidase 2	H_2O_2 production	Monoallelic	PIOD	Severe/mild CH [178, 342, 502]	P/T	
			Biallelic	TIOD	Severe/mild CH [178, 342, 389, 502]	P/T	ClO_4- discharge test
DUOXA2	DUOX maturation factor 2	H_2O_2 production	Biallelic	PIOD	Mild CH [178, 537]	P	ClO_4- discharge test
SLC5A5	sodium iodide symporter	iodide uptake into the thyrocyte	Biallelic	Reduced thyroidal iodide uptake	Severe/mild CH [459]	P	Saliva/plasma RAI ratio
SLC26A4	pendrin	iodide efflux into follicular lumen	Biallelic	Sensorineural hearing loss***	Very mild CH [178]	P	MRI/CT of temporal bones
TSHR	TSH receptor	initiation of hormone synthesis cascade	Biallelic	TSH resistance	Severe CH [171]	P	
			Monoallelic		Mild CH [178]	P	
IYD	iodotyrosine deiodinase	intrathyroidal iodide recycling	Monoallelic	Excessive urinary excretion of iodine	Mild CH	P	
			Biallelic		Severe CH [63]	P	Rapid thyroidal loss of iodine

TIOD: total iodine organification defect; PIOD: partial iodine organification defect; P: permanent CH; T: transient CH; ClO_4-: percholrate; RAI: radioactive iodide.*Most frequent cause of CH with GIS [179]; **Second most frequent cause of CH with GIS [179].***Sensorineural hearing loss (Pendred syndrome, OMIM 274600), because the gene is also expressed in the inner ear [178]. *TPO*, *DUOX2* and *DUOXA2* defects are characterise by discharge of substantial percentage of radio labeled iodide from the thyroid after administration of perchlorate (perchlorate discharge test). This discharge indicates a defect in converting accumulated iodide to thyrosine-bound iodide. The discharge may be incomplete or complete, thus defining partial (PIOD) or total defects (TIOD). PIODs are characterise by release of <50% of the accumulated radioiodine, whereas TIODs are characterise by release of >90% of the accumulated radioiodine [439].

Table 2.1 Known gene defects causing CH with *gland-in-situ*

2.2 Aims

The aim of the project reported in this chapter was to conduct, for the first time, a comprehensive next-generation sequencing-based screening of all eight known CH-GIS genes (*TG*, *TPO*, *DUOX2*, *DUOXA2*, *IYD*, *SLC5A5*, *SLC26A4* and *TSHR*) in an ethnically and biochemically heterogeneous CH cohort with GIS. Further, my collaborators and I, aimed to investigate the associated clinical phenotypes of mutation-positive and negative patients and to investigate potential digenic causes of CH involving these eight known genes.

2.3 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, Adeline Nicholas, Martin Howard and Dr Eric Schoenmakers. Some parts of this work have been published at The Journal of Clinical Endocrinology & Metabolism (JCEM).

2.4 Methods

2.4.1 Patients

All investigations conducted in this work were part of an ethically approved protocol and/or clinically indicated, being undertaken with the consent from patients and/or next of kin. Dr Nadia Schoenmakers and Professor Krishna Chatterjee recruited a cohort of 49 cases from 34 families, of which 14 families constituted multi-affected siblings and the rest were singleton cases. All patients were referred from centres in the UK, Oman, Saudi Arabia, UAE and Turkey on the basis of newborn screening and/or raised venous TSH levels. Inclusion criteria required clinical evidence of goitre, or radiological evidence of a normally-sited thyroid gland in the proband (or in one affected family member) and a diagnosis of overt or subclinical primary CH. Thyroid

biochemistry was measured using local analysers in the referring hospitals. None of the patients have been previously screened for mutations in the eight known CH genes.

2.4.2 Next-generation DNA sequencing

This study employed three NGS-based strategies: HiSeq whole-exome sequencing (WES), HiSeq targeted-sequencing (HiSeq-TS) and MiSeq targeted-sequencing (MiSeq-TS) (**Table** 2.2). The first two experiments were performed at the WTSI as part of the UK10K project (www.uk10k.org) and the last was performed either at the University of Cambridge Metabolic Research Laboratories or the Department of Medical Genetics of the University of Cambridge. Cost constraints precluded the use of WES in all samples.

NGS protocol	Samples
Whole-exome sequencing (N = 17)	F3a,b
	F6a,b
	F7a,b
	F8a,b
	F9a,b
	F10
	F13
	F15a,b,c
	F33a,b
HiSeq targeted sequencing (N = 11)	F2a,b
	F11
	F12a,b
	F17
	F26
	F28
	F29a,b
	F34
MiSeq targeted sequencing (N = 21)	F1a,b
	F4
	F5a,b
	F14a,b
	F16
	F18
	F19a,b
	F20
	F21
	F22
	F23
	F24
	F25
	F27
	F30
	F31
	F32

Table 2.2 Summary of samples sequenced for each NGS protocol. Indexes *a*, *b* and *c* refer to siblings.

HiSeq exome sequencing (WES)

Sample processing and sequencing was performed by the Sanger Institute Core Sequencing pipeline. Genomic DNA (1-3µg) was extracted from blood and was sheared to 100-400bp using a Covaris E210 or LE220 (Covaris, Woburn, Massachusetts, USA). Sheared DNA was subjected to Illumina paired-end DNA library preparation and enriched for target sequenced (Agilent Technologies, Santa Clara, CA, USA; Human All Exon 50Mb – ELID S02972011) according to the manufacture's recommendations (Agilent Technologies, Santa Clara, CA, USA; SureSelectXT Automated Target Enrichment for Illumina Paired-End Multiplexed Sequencing). Enriched libraries were sequenced (eight samples over two lanes) using the HiSeq 2000 platform (Illumina) as paired-end 75 base reads according to the manufacturer's protocol.

The Human Genetics Informatics team at Sanger performed the alignment of the raw sequencing data to human reference genome build UCSC hg19/Grch37 using the Burrows-Wheeler Aligner [281]. Tarjinder Singh from the Medical Genomics team at Sanger performed the variant calling. Variants were first called at the single sample level using GATK Haplotype Caller (version 3.2-2-gec30cee) [116] and then joint-called using GATK CombineVCFs and GenotypeVCFs at default settings.

For variant QC, I applied Variant Quality Score Recalibrator (VQSR) with the recommended training sets (see Appendix **Table** A.1 for more details). VQSR uses annotation metrics such as quality by depth, mapping quality, variant position within reads and strand bias, based on "true" sites provided as input, i.e. high confidence, validated gold standard variants, to generate an adaptive error model. VQSR then applies this model to the remaining variants called to calculate a probability (the Variant Quality Score Log Odds Ratio score, VQSLOD) that each variant is a true genetic variant versus a sequencing or data processing artefact. Using this recalibrated quality score, one can filter low quality variants rather than relying on multiple hard filters. Following current GATK best practices [116], I applied VQSR separately to SNVs and indels, and variants within the 99.9% truth sensitivity threshold were considered of sufficient quality. However, recent studies have shown that poor quality variants remain in datasets following GATK's best practices [71]. To mitigate this, I set genotypes to missing when the genotype quality (i.e. the probability of the genotype being real, GQ) was below 20 or the depth (DP) was below eight. These combinatory thresholds filter out genotypes with $\leq$ 99% likelihood [71] and are recommended because VQSR does not explicitly filter genotypes, allowing low quality genotypes generated at variant sites that passed the VQSLOD filter to persist in the dataset and to contribute to a

major source of errors in sequencing studies [71]. **Figure** 2.2 illustrates the beneficial effect of these extra filters on the overall exome data quality, measured in the form of mean GQ and mean ratio of transitions to transversions (Ts/Tv) per sample at variant sites. The Ts/Tv metric is used in almost all sequencing studies as a parameter for checking overall SNV quality and is computed as the number of transition SNVs (A←→G, C←→T) divided by the number of transversions SNVs (A←→T, G←→C, A←→C, G←→T). High quality exome datasets are expected to have Ts/Tv ratios between 2.7 and 3.0 [71, 116], as higher Ts/Tv ratios are associated with lower false positives.

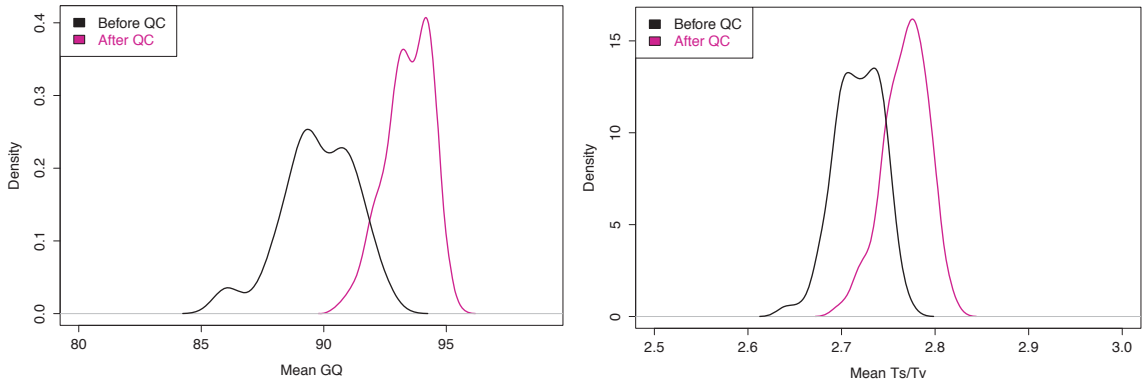


Figure 2.2 Distribution of mean genotype quality (GQ, left panel) and ratio of transitions to transversions (Ts/Tv, right panel) pre- and post- extra genotype-QC.

Only variants passing the VQSR and the extra genotype-QC thresholds and located within target regions were considered in downstream genetic analyses.

HiSeq targeted-sequencing (HiSeq-TS)

Sample processing and sequencing was performed by the Sanger Institute Core Sequencing pipeline. The GenomiPhi V2 DNA Amplification kit (GE Healthcare) was used for whole-genome amplification of 1ng/µg template DNA prior to pull-down. Target enrichment and amplification were performed with the HaploPlex Target Enrichment kit (Agilent Technologies) according to the manufacture's protocol, and sequenced using the HiSeq 2000 platform (Illumina).

Sequencing alignment and variant calling was conducted by Dr Shane McCarthy from the UK10K production team. Raw alignment BAMs were realigned around known indels (1000 Genomes pilot data [402]), base quality scores were recalibrated using

GATK [116] and base alignment quality tags were added using SAMtools calmd (version 0.1.19-3-g4b70907) [281]. BAMs were then merged to sample level and duplicate reads marked using Picard (http://broadinstitute.github.io/picard). SNPs and indels were called on each sample individually with both SAMtools mpileup [281] and GATK UnifiedGenotyper (version 2.4-9-g532efad) [116].

For variant QC, Dr Shane McCarthy added standard variant quality filters (see Appendix **Table** A.2) to each call set separately using vcf-annotate. Similar filters have been used in many research studies [160, 183, 272, 394, 522]. Calls were then merged, giving precedence to GATK information, when possible.

Again, I only took forward for downstream analysis those variants that passed all standard QC thresholds and variants that were located within target regions.

MiSeq targeted-sequencing (MiSeq-TS)

This experiment was performed by Adeline Nicholas and Martin Howard. Primers to amplify the full coding sequences of all genes were designed using Primer3. Primer uniqueness and the presence of SNPs in the primer binding sites were checked using SNPCheck3 (National Genetics Reference Laboratory, Manchester, UK). PCRs were performed using SequalPrep Long PCR Kit (Thermo Fisher Scientific), with amplicons ranging in size from 1 to 7.6 kb. PCR products were purified using the Agencourt AMPure XP system (Beckman), and products for all genes were pooled for each patient.

Illumina paired-end DNA libraries were prepared using the Nextera XT DNA sample preparation kit, from 1 ng of pooled amplicons. Libraries were normalized and pooled according to the manufacturer's recommendations, then diluted in water and quantified by qPCR on a Roche LightCycler 480, using the KAPA Library Quantification Kit (KAPA Biosystems, MA. USA). Libraries were sequenced on an Illumina MiSeq as paired end 150bp according to the manufacturer's protocol.

The MiSeq protocol was validated by re-sequencing the 21 patient DNAs for 25 known variants. All 25 alleles were successfully detected at >20x coverage, giving a sensitivity of 100% for this sequencing depth.

2.4.3 Sequencing efficiency of WES and HiSeq-TS experiments

To evaluate the coverage levels of each gene within the WES and HiSeq-TS experiments, I ascertained the read depth at each nucleotide (within exonic sequences of each gene) on a per-BAM level using SAMtools mpileup. The read depth was then averaged per-coordinate across samples to produce an average capture per position, as well as a median coverage per gene. Because expressing gene coverage as a median read depth does not imply that all bases within that gene are covered at the same depth, I also calculated the proportion of each gene covered at various depths.

2.4.4 Variant annotation

After conducting variant calling and quality control in all three datasets, I annotated these data against a large number of resources, including dbSNP v137 rsIDs and allele frequencies computed from several datasets such as: 1000 Genomes Phase I (1KG, N=2,818) [402], NHLBI GO Exome Sequencing Project 6,500I (ESP, N=6,500) [476], UK10K low-coverage study (N=3,781) [507], other UK10K whole-exome sequencing studies (N=4,975) [507] and Exome Aggregation Consortium r0.3 (ExAC) (N=60,706) [135].

Functional annotations were then added using Ensembl Variant Effect Predictor (VEP, version 75) to annotate all variants according to Gencode v19 coding transcripts, keeping the most severe consequence for the gene [322]. Next, I used Sorting Intolerant From Tolerant (SIFT) [349] and Polyphen-2 [4] to predict missense deleteriousness scores, and Genomic Evolutionary Rate Profiling (GERP) [107] to assess whether variants affected evolutionary conserved amino acid sites.

2.4.5 Identifying likely damaging variants per sample

After annotation, I filtered for rare and functional variants in the eight genes in each sample. Rare variants were defined as those that were absent or with AFs <1% in all of the above population datasets. Functional variants were defined as changes that affected the protein coding sequence with the following consequences: transcript ablation, stop gained/lost, stop retained, splice donor/acceptor/region, frameshift, inframe insertion/deletion, initiator codon and missense variants (see **Figure** 1.9 for definitions of splice sites).

Likely damaging variants were defined here as LoF variants (i.e. nonsense, frameshift and splice acceptor/donor variants) and as missense variants with a Polyphen-2 or SIFT pathogenicity prediction of 'possibly damaging/deleterious' or above, or if demonstrated to disrupt the protein structure via *in silico* mutation modelling.

The structural modelling of missense mutations was conducted by Dr Erik Schoenmakers using Phyre2 (Protein Homology/analogy Recognition Engine 2) [241]. Briefly, this software works by scanning the user protein sequence via a Hidden Markov model against a large database of approximately 10 million known sequences of proteins, to detect evolutionary relationships to other protein sequences (i.e. homologies) and high confidence similarities. This scanning procedure generates an alignment between our sequence of interest and sequences of known structure, which then permits the generation of a tri-dimensional (3D) model for our protein of interest and the investigation of specific amino acid changes.

Novel variants were defined as those that were absent from HGMD Professional and were classified, by Dr Nadia Schoenmakers, according to the standards described by the American College of Medical Genetics [419].

2.4.6 Capillary sequencing for variant validation

Adeline Nicholas validated all variants identified in this study via Sanger sequencing. Where possible, DNA obtained from family members was also sequenced to verify inheritance of variants and segregation with phenotype. All compound heterozygous mutations were confirmed by sequencing the probands' parents. Briefly, 50ng of genomic DNA was amplified using Illustra Genomiphi V3 ready-to-go kit (GE Healthcare Life Sciences, Buckinghamshire, UK) according to the manufacturer's instructions. PCR products were sequenced using the BigDye Terminator v3.1 Cycle Sequencing kit (Applied Biosystems, Foster City, USA) and 3730 DNA Analyzer (Applied Biosystems) according to the manufacturer's instructions.

2.5 Results

2.5.1 Sequencing data quality

In the samples sequenced using WES or the HiSeq targeted sequencing panel, optimal median coverage (>30x) [260] was achieved for all genes except *DUOXA2* and *SLC5A5* in the eleven samples screened by HiSeq targeted sequencing, which displayed a median coverage of 5x and 24x, respectively (**Figure** 2.3). Exons sequenced using the MiSeq targeted sequencing panel either achieved >20x coverage, or were repeated by Sanger sequencing (Dr Nadia Schoenmakers, personal communication).

In the WES and HiSeq protocols, in common with previous studies employing similar NGS techniques [246, 304], although median coverage was generally high, coverage was non-uniform across individual genes (**Figure** 2.3). This was most marked with the HiSeq targeted sequencing panel in which specific exons exhibited <10x coverage, below which detection of heterozygous SNPs is severely compromised [84]. This affected the following genes and exons: *DUOXA2* (exons 1, 2, 4, 5 and 6), *SLC5A5* (exons 1-3, 5, 6, 11, 12 and 15), *DUOX2* (exons 2, 5, 6, 8, 15 and 34), *TG* (exons 13, 15, and 16), *TPO* (exons 3, 7, 8, and 16), *SLC26A4* (exon 21) and *IYD* (exon 6) (**Figure** 2.4).

Comparison of the WES and TS approaches in greater detail revealed the HiSeq targeted-sequencing experiment showed considerably greater variability in coverage between genes, while the WES experiment suffered from higher inter-sample variability (data not shown), again findings that have been observed elsewhere [246, 304].

Figure 2.3 Proportion of gene sequence covered at various depth thresholds for the **A)** WES experiment and **B)** Hi-Seq targeted-sequencing experiment.

SAMtools mpileup was used to calculate the depth at each base within every exonic region of every gene for all samples. The median coverage across samples per gene (at exonic sequences only) is represented on top of each bar. Numbers at the top of the bars represent the median coverage.

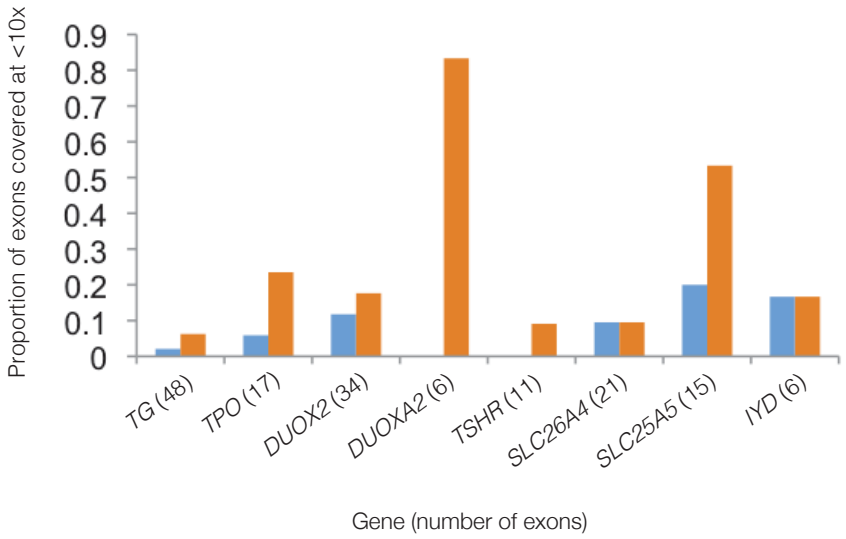


Figure 2.4 Proportion of exons with mean depth coverage <10x in the samples sequenced by WES (blue) and HiSeq targeted sequencing (orange).

2.5.2 Genetic diagnostic yield

Forty-nine cases from 34 families of European, Asian, Middle Eastern and Afro-Caribbean descent were investigated in this study, and a total of 39 likely damaging variants were detected across patients (**Table** 2.3).

Twenty-nine cases (20 families, 59%) were considered 'solved' following identification of a decisive link between genotype and phenotype (**Figure** 2.5); these patients harbored likely damaging variants with genotypes that were consistent with the known mode of inheritance of the gene in which they occurred. The causative variants identified comprised known pathogenic (38%) or novel mutations (62%) predicted to be damaging to the encoded protein, i.e. predicted to compromise the normal levels or biochemical function of the gene or gene product. **Figure** 2.6 illustrates the impact of these variants at the protein level, with missense variants being the most frequently observed functional class, followed by frameshift, stop gained and splice region variants.

In a further 11 cases (7 families) where mutations were identified, the ascertained genotype could plausibly be contributing to the phenotype, but the evidence to support a causal link was weaker than in the 'solved' group (**Figure** 2.5). In this case, the observed genotypes were inconsistent with the known mode of inheritance of the gene (i.e. patients harbored monoallelic variants in genes known to express disease recessively). These cases were therefore classified as 'ambiguous'.

Finally, nine cases (7 families) were considered 'unsolved', as they carried no mutations in any of the screened genes (**Figure** 2.5).

Detailed genetic and phenotype data for all samples is supplied in the Appendix **Tables** A.3 and A.4.

Gene	Protein change	Nucleotide change	Family ID	Known pathogenic mutation?	Mutation location	SIFT	PolyPhen	GERP	Allele frequency
TG	R159X	.	F2a,b	.	type 1 repeat	.	.	2.42	0.000042 (ExAC)
	C160S	.	F5a,b	.	type 1 repeat	T	PD	5.84	0.00013 (UK10K cohorts)
	R296X	.	F5a,b	yes	type 1 repeat	.	.	5.62	0.000362 (ExAC)
	R451X	.	F1a,b	yes		.	.	1.1	.
	S528X	.	F3a,b	.		.	.	1.7	.
	.	c.638+5G>A	F7a,b	.		.	.	5.36	0.000025 (ExAC)
	C726Y	.	F6a,b	.	type 1 repeat	D	PD	5.66	.
	Q771X	.	F13	.		.	.	4.61	0.0002 (UK10K)
	Y759C	.	F14a,b	.	type 1 repeat	D	PD	5.81	0.000017 (ExAC)
	Q870H	.	F12a,b	yes	type 1 repeat	D	PD	3.15	0.00325 (ExAC)
	W1050L	.	F6a,b	.	type 1 repeat	D	PD	4.91	0.001 (1KG)
	C1493Y	.	F8a,b	.	type 2 repeat	D	PD	5.52	.
	Q1644E	.	F11	.		D	B	5.3	.
	R1691C	.	F10a	.		D	B	0.561	0.000881 (ExAC)
	S2121AfsX32	.	F4	.	type 3 repeat	.	.	4.83	.
	L2547Q	.	F10a	.	ACHE domain	D	PD	4.83	0.0006 (UK10K)
	W2685L	.	F9a,b	.	ACHE domain	D	PD	4.84	.
	.	c.3433+3_3433+6delGAGT	F15a,b,c	.		.	.	5.52	.
TPO	E17DfsX77	.	F21	yes	SP cleavage site	.	.	.	0.0003 (UK10K)
	R291H	.	F18	.		D	PD	-0.174	.
	G331V	.	F18	.		D	PD	-0.431	.
	A397PfsX76	.	F16	.		.	.	.	.
	Y453D	.	F21	yes		D	PD	5.3	0.0003 (ESP)
	R491H	.	F11, F16	yes		D	PD	5.3	0.002762 (1KG)
	E510AfsX14	.	F22	.		.	.	.	.
	R584Q	.	F19a,b	.		D	PD	4.78	0.00023(ESP)
	R665Q	.	F17	yes		D	PD	4.84	0.000025 (ExAC)
	C808AfsX24	.	F20	.		.	PD	-2.18	0.00934 (ESP)
DUOX2	R354W	.	F9a,b	.	NADPH oxidase domain	D	PD	4.82	0.00014 (ExAC)
	Q570L	.	F10a	yes		T	PD	4.99	0.002866 (ExAC)
	Q686X	.	F8a,b, F6b	yes		.	.	5.51	.
	R764W	.	F25	.		D	PD	3.67	0.004 (1KG)
	F966SfsX29	.	F23	yes		.	.	.	0.0031 (UK10K)
	F966SfsX29	.	F23	yes		.	.	.	0.0031 (UK10K)
TSHR	P68S	.	F26	yes		D	PD	0	0.0041 (1KG)
SLC26A4	N324Y	.	F19a	yes		D	PD	5.62	0.00023 (ESP)
	E384G	.	F21	yes		D	PD	5.92	0.0001 (UK10K)
	I713M	.	F19b	.		D	PD	-1.72	0.0028 (ESP)
DUOXA2	.	c.555-5G>A	F27	.		.	.	1.08	.

Table 2.3 Known and novel mutations detected in the CH cohort with GIS. Variants identified in solved and in ambiguous cases are listed. Known pathogenic mutation refers to whether the variant is present in HGMD Professional database. Mutation location refers to the domains of the protein. The allele frequency column is annotated with the maximum alternative allele frequency observed across all the AF datasets used in the annotation. T: tolerated by SIFT; B: benign by Polyphen-2; D: damaging by SIFT; PD: possibly damaging or probably damaging by Polyphen-2. Table is sorted by amino acid position within each gene.

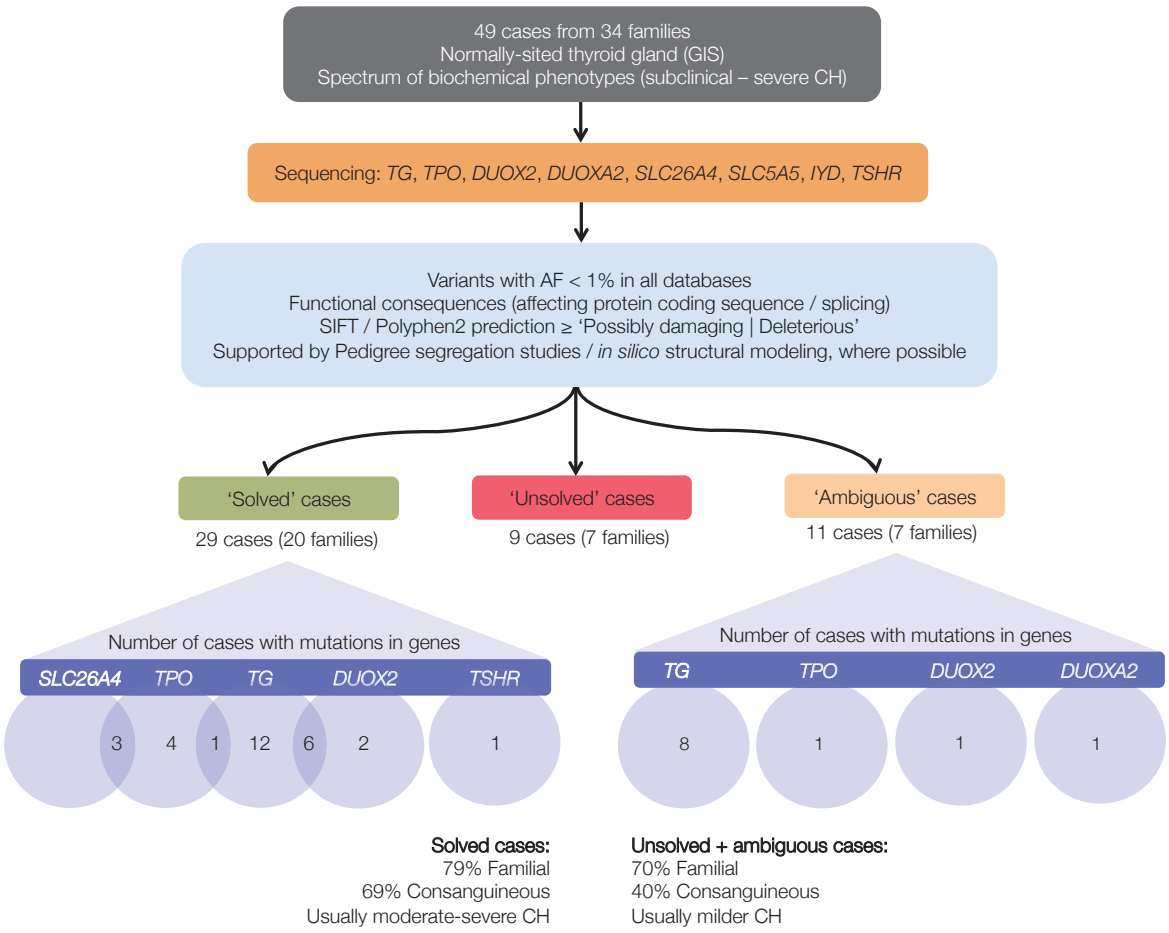


Figure 2.5 Schematic illustrating case selection, variant filtering and distribution of mutations in the GIS cohort studied.

'Solved' cases refers to those in who a clear link between genotype and CH phenotype was established: they carried likely damaging variant with a genotype that was consistent with the known mode of inheritance of the gene. Solved cases harbored mutations in a single gene or in two genes, and will be explained separately in the main text. 'Ambiguous' cases refers to samples for who the variants identified did not conclusively explain their CH phenotype: they carried likely damaging variants, but the observed genotype was inconsistent with the known mode of inheritance of the gene. 'Unsolved' cases did not harbor any likely damaging variants in the screened genes. The number of cases harboring monoallelic or biallelic mutations in each gene are listed beneath the corresponding gene name for the 'solved' and 'ambiguous' cases. Numbers in the intersect between circles denote triallelic cases harboring mutations in both genes (one biallelic; one monoallelic). In the 'ambiguous' cases, all mutations except *DUOXA2* were monoallelic. Solved and ambiguous+unsolved cases were equally likely to be familial, yet CH was generally more severe in the solved category (mean TSH 100mU/L vs 36mU/L at diagnosis, P=0.02, Welch's t-test). Splice refers to a splice region variant (see **Figure** 1.9 for definitions).

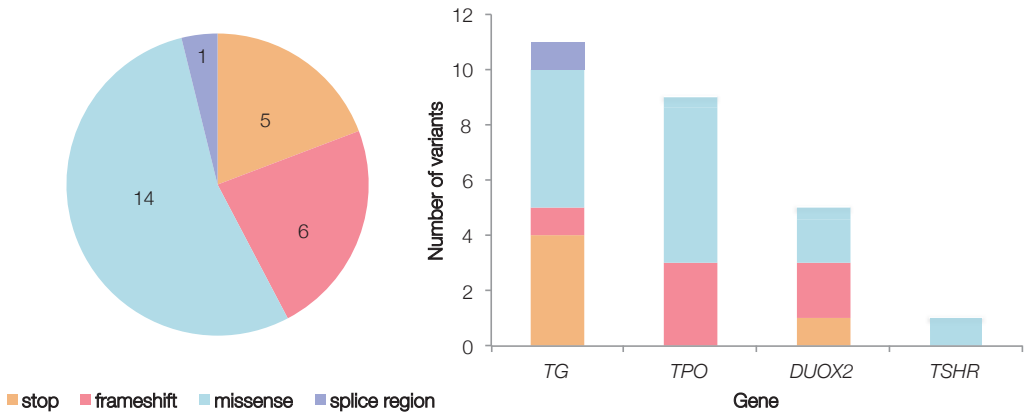


Figure 2.6 Causative variants identified in CH cohort with GIS.

A) Pie chart with distribution of consequence classes. **B)** Distribution of consequence classes per gene. Total of 26 variants identified in the 29 solved cases (20 families). Only variants assumed to contribute to the phenotype are depicted, i.e. mutations observed in 'ambiguous' cases are not included in the image.

Dr Nadia Schoenmakers classified CH severity according to the European Society for Paediatric Endocrinology (ESPA) criteria, on the basis of serum free-T4 levels (the active form of T4, fT4): severe <5, moderate 5 to <10 and mild >10pmol/L, respectively [277]. This analysis suggested CH was more severe biochemically in solved cases than in unsolved or ambiguous cases (mean TSH 100mU/L vs 36mU/L at diagnosis, P=0.02, Welch's t-test). Solved cases were also more frequently from consanguineous backgrounds (69% cases vs. 40% cases), which likely reflects the increased incidence of recessive disease in the presence of consanguinity, since CH-associated mutations in five of the eight targeted genes (*TG*, *TPO*, *DUOXA2*, *SLC5A5* and *SLC26A4*) are usually biallelic. Cases with affected siblings were common in both solved and unsolved or ambiguous categories (79% vs. 70% cases, **Figure** 2.5, and Appendix **Tables** A.3 and A.4.).

2.5.3 'Solved' families with mutations in one gene (monogenic families)

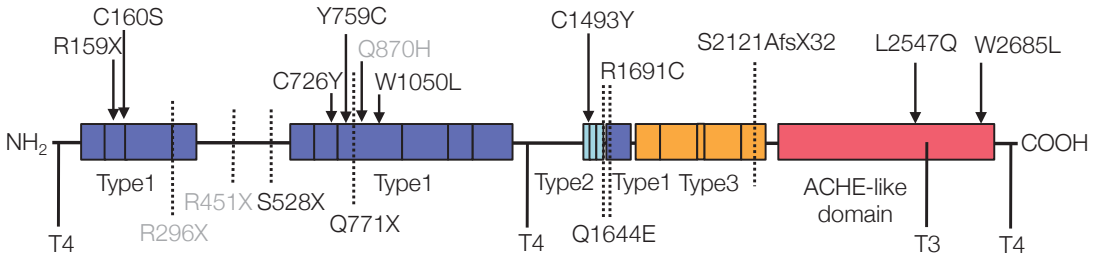
Nineteen of the 29 solved cases had a monogenic basis of disease, most commonly involving biallelic mutations in *TG* (12 cases), followed by *TPO* (four cases), *DUOX2* (one monoallelic and one biallelic mutation) and *TSHR* (one case) (**Figure** 2.5). I did not identify cases of CH attributable to monogenic mutations in *IYD*, *SLC5A5* and *SLC26A4*.

TG mutations

TG is the secretory protein upon which thyroid hormone is synthesized and is the most abundantly expressed protein in the thyroid gland [472]. The 12 cases that harbored monogenic *TG* mutations predominantly exhibited moderate to severe CH (**Figure** 2.7). One known and three novel homozygous nonsense or frameshift mutations were identified that truncate TG before the carboxy-terminal acetyl cholinesterase (ACHE)-like domain (F1, 2, 3, 4). This region has been shown to function as an intramolecular chaperone and is essential for normal conformational maturation and efficient intracellular trafficking of TG from the ER to the Golgi and the follicular lumen [273].

Two siblings (F5) were compound heterozygotes for a novel, maternally inherited mutation (C160S) and a known, paternally inherited stop mutation (R296X). Even though C160S is predicted to be benign by SIFT, it is highly conserved (GERP score 5.84, with the maximum being 6), extremely rare (AF <0.1% in UK10K cohorts) and estimated to be damaging by Polyphen-2 (**Table** 2.3). Cysteine residues within repetitive domains (type 1, 2 and 3) in TG form intramolecular disulphide bonds needed for protein folding, thus p.C160S may be deleterious to TG by affecting the tertiary structure and by preventing the availability of hormonogenic sites for thyroid hormone production [471].

Two siblings (F7) harbored the same homozygous splice region variant (c.638+5 G>A) inherited from heterozygous parents. Because the amino acid change is located within the 3-8 bases of the intron, and not at the 5' or 3' end of the intron known as the splice donor or acceptor sites, respectively, it is difficult to ascertain the pathogenicity of this variant *in silico*. Yet, the fact it is unique to the affected siblings, and adjacent to a known pathogenic mutation (c.638+1G>A [15]) supports causality, albeit in association with a milder CH phenotype.



ID	CH Severity	Protein change Allele 1	Protein change Allele 2	Likely Consequence	Oligogenicity
Monogenic families					
F1a, b*	Severe	R451X	R451X	Known pathogenic	
F2a, b	Mild, Moderate	R159X	R159X	Pathogenic	
F3a, b	Moderate	S528X	S528X	Pathogenic	
F4	TSH>100	S2121AfsX32	S2121AfsX32	Pathogenic	
F5a, b	Moderate, Mild	R296X	WT	Known pathogenic	
		WT	C160S	Likely pathogenic	
F6 a	Moderate	C726Y	C726Y	VUS	
		W1050L	W1050L	Likely pathogenic	
F7a, b	Mild, bs. TSH 16	c.638+5G>A	c.638+5G>A	VUS (Splice-region)	
Digenic families					
F6b	Moderate	C726Y	C726Y	VUS	*DUOX2* Q686X (het)
		W1050L	W1050L	VUS	
F8a*, b	Severe	C1493Y	C1493Y	VUS	*DUOX2* Q686X (het)
F9a, b	Severe	W2685L	W2685L	VUS	*DUOX2* R354W (het)
F10a	TSH>150 (v)	R1691C	L2547Q	VUS, VUS	*DUOX2* Q570L (hom)
F11	Severe	Q1644E	WT	VUS	*TPO* R491H (het)
Ambiguous families					
F12a, b	Mild	Q870H	WT	Known pathogenic	
F13	Mild	Q771X	WT	Pathogenic	
F14a, b	.	Y759C	WT	VUS	
F15a*, b*, c	Mild	c.3433+3_3433+6delGAGT	WT	VUS (splice region)	

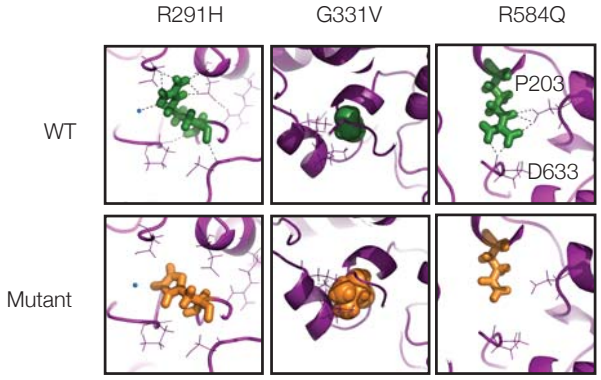
Figure 2.7 Mutations identified in *TG*.

Summary of *TG* mutations identified in solved and ambiguous case categories and associated biochemical phenotypes. CH severity was classified according to ESPE criteria on the basis of serum fT4 levels; severe, <5, moderate 5 to <10, and mild >10 pmol/l, respectively [277] and pathogenicity is predicted according to ACMG guidelines [419]. VUS: variant of uncertain significance. Mutation position of the variants is illustrated using a schematic representation of the thyroglobulin protein and its key structural domains, including the repetitive type 1, 2 and 3 cysteine-rich regions, acetylcholinesterase homology (ACHE-like) domain and hormonogenic domains. T4 synthesis occurs at amino acid position 5 (exon 2), 1291 (exon 18) and 2747 (exon 48) and T3 synthesis at position 2554 (exon 48). Known mutations are shown in grey, novel mutations in black. *: cases for which complete biochemical data at diagnosis is not available and CH severity refers to sibling. bs: blood spot.

TPO mutations

TPO is the hemeprotein peroxidase that catalyzes the final steps of thyroid hormone synthesis [423]. Biallelic mutations were identified in four monogenic kindreds, two of which were compound heterozygotes (F16 and F18). The variants identified across the four families included two known pathogenic missense mutations (F16; p.R491H, F17; p.R665Q), two novel frameshifts (F20; p.C808Afs*24, F16; p.A397Pfs*76) and two novel missense variants (F18; p.R291H, p.G331V) (**Table** 2.3).

Structural modelling of the novel *TPO* missense mutations revealed the p.R291H variant is predicted to disrupt a hydrogen bond network close to the TPO heme group, the electron source for catalytic reactions [423], and is thus predicted to destabilize the TPO catalytic domain. G331 is located close to the substrate binding domain, and mutation to the larger valine amino acid will likely cause steric hindrance impeding substrate binding (**Figure** 2.8).



ID	CH Severity	Protein Change Allele 1	Protein Change Allele 2	Likely Consequence	Oligogenicity
Monogenic families					
F16	.	R491H	A397PfsX76	Known pathogenic, Pathogenic	
F17	TSH 27	R665Q	R665Q	Known pathogenic	
F18	Severe	R291H	G331V	Likely pathogenic, Likely pathogenic	
F20	.	C808Afs*24	C808Afs*24	Pathogenic	
Digenic families					
F11	Severe	R491H	R491H	Known pathogenic	*TG* Q1644E (het)
F19a	Severe	R584Q	R584Q	Likely pathogenic	*SLC26A4* N324Y (het)
F19b	Severe	R584Q	R584Q	Likely pathogenic	*SLC26A4* I713M (het)
F21	Severe	E17DfsX77	Y453D	Pathogenic, Known pathogenic	*SLC26A4* E384G (het)
Ambiguous families					
F22	Subclinical	E510AfsX14	WT	Pathogenic	

Figure 2.8 Mutations identified in *TPO*.

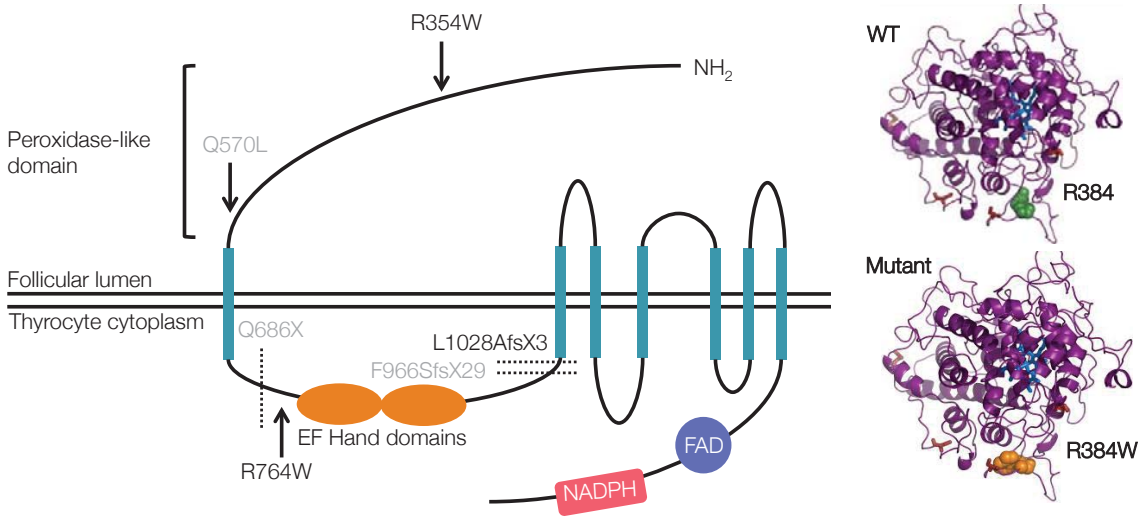
Summary of TPO mutations identified in all case categories and associated biochemical phenotypes. CH severity was classified according to ESPE criteria on the basis of serum fT4 levels; severe, <5, moderate 5 to <10, and mild >10 pmol/l, respectively [277] and pathogenicity is predicted according to ACMG guidelines [419]. VUS: variant of uncertain significance. The effect of the novel missense mutations was investigated by Dr Eric Schoenmakers. The protein structure was modelled using the phyre2-server and the image was generated using the MacPyMOL Molecular Graphics System (Schrödinger LLC). Figures in the top row show the wild-type (WT) model, with amino acids of interest in green; figures on the bottom row show the model with the mutant amino acid (orange); local polar contacts are shown with black dashed lines. The R291H and R584Q mutations affect amino acids contributing to an intensive network of H-bond contacts close to the catalytic domain involving the heme-group. R291 makes polar contacts with R585 and R582, interacting directly with the heme-group and R584 makes direct polar contacts with the heme-group itself, as well as with P203 and D633. The mutations R291H (increased hydrophobicity) and R584Q (resulting in a smaller polar group) are likely to disrupt polar contacts affecting local structure and are predicted to affect catalytic activity. The G331V mutation affects local space filling with the larger valine predicted to impair substrate binding by displacement of the nearby helix and/or disruption of polar contacts (orange amino acids, H_2O molecules in blue), affecting the local structure of TPO.

DUOX2 mutations

DUOX2 is the NADPH oxidase that generates H_2O_2 required for thyroid hormone synthesis [537]. Two solved cases with monogenic *DUOX2* mutations were identified (**Figure** 2.9), including one known heterozygous mutation (F23; p.F966Sfs*29) and one novel homozygous variant (F24; p.L1028Afs*3, **Table** 2.3). Both of these variants are predicted to abrogate protein function as they truncate DUOX2 prematurely, resulting in a shorter protein without the C-terminal NADPH oxidase domain (**Figure** 2.9), which is required for electron transfer. Affected cases generally had a milder or transient (F23) CH phenotype compared with cases harboring monogenic *TG* and *TPO* mutations.

TSHR mutations

A single individual from the UAE with mild CH harbored a known pathogenic heterozygous *TSHR* mutation (F26; p.P68S) (**Table** 2.3). Parental samples were not available to determine whether the variant consituted a *de novo* event, however, the mild CH phenotype was consistent with previously reported biochemistry associated with this mutation [475].



ID	CH Severity	Protein Change Allele 1	Protein Change Allele 2	Likely Consequence	Oligogenicity
Monogenic families					
F23	Mild	F966SfsX29	WT	Known pathogenic	
F24	TSH 55	L1028AfsX3	L1028AfsX3	Pathogenic	
Digenic families					
F10a	TSH>150	Q570L	Q570L	Known pathogenic	*TG* 1691C, *TG* 2547Q (het)
F6b	.	Q686X	WT	Known pathogenic	*TG* C726Y, *TG* W1050L (hom)
F8a*, b	Severe	Q686X	WT	Known pathogenic	*TG* C1493Y (hom)
F9a, b	Severe	R354W	WT	VUS	*TG* W2685L (hom)
Ambiguous families					
F25	Moderate	R764W	WT	VUS	

Figure 2.9 Mutations identified in *DUOX2*.

Summary of DUOX2 mutations identified in all case categories and associated biochemical phenotypes. CH severity was classified according to ESPE criteria on the basis of serum fT4 levels; severe, <5, moderate 5 to <10, and mild >10 pmol/l, respectively [277] and pathogenicity is predicted according to ACMG guidelines [419]. VUS: variant of uncertain significance. Mutation position is illustrated using a schematic representation of the domain structure of the DUOX2 protein. The protein contains seven transmembrane domains (blue) and a C-terminal cytosolic domain containing flavin adenine dinucleotide (FAD) and NADPH-binding sites (to provide electron transfer needed for DUOX2 function). Known mutations are shown in grey and novel mutations in black. Structural modelling of the novel missense mutation (p.R354W), performed by Dr Eric Schoenmakers, suggests that R354 is part of an intensive hydrogen network. The novel missense mutation R354W replaces the hydrophilic arginine by the hydrophobic tryptophan disrupting this network and also leading to a possible repositioning of the loop containing R354 and C351, which mediates interactions between the peroxidase domain and extracellular loops obligatory for DUOX2 function. The protein structure was modelled using the phyre2-server and the image was generated using the MacPyMOL Molecular Graphics System (Schrödinger LLC).

2.5.4 'Solved' families with mutations in two genes (digenic families)

Ten solved cases from seven families harbored digenic pathogenic variants, which constitute the simplest form of oligogenic inheritance. These variants were predominantly triallelic and most commonly involved a biallelic variant in one gene, in association with a monoallelic variant in the other locus, and affected the following gene pairs: *TG* and *DUOX2* (6 cases), *SLC26A4* and *TPO* (3 cases) and *TPO* and *TG* (1 case, **Figure** 2.5).

TG and *DUOX2*

TG and *DUOX2* digenic mutations were detected in consanguineous Turkish families F6, 8 and 9 (**Figure** 2.10). In all these families, although defined as variants of uncertain significance by ACMG criteria, the biallelic *TG* mutations were rare (AF<0.1% in 1KG Europeans and absent in all other population datasets, including the ∼61,000 ExAC samples) or private to patients, affected conserved amino acids, and were predicted to be pathogenic by both Polyphen-2 and SIFT.

Two siblings with CH in F6 (a, b) were both homozygous for two novel *TG* mutations (W1050L and C726Y), but one sibling (F6b) harbored an additional, maternally-inherited *DUOX2* mutation (p.Q686X), previously reported in association with transient CH [334]. Biochemistry at diagnosis could not be retrieved from F6b for comparison with F6a, however both presented with neonatal goitre and had similar treatment requirements (Dr Nadia Schoenmakers, personal communication). Their mother exhibited adult-onset hypothyroidism of unknown etiology.

Two unrelated sibling pairs also harbored homozygous *TG* mutations in association with a heterozygous *DUOX2* mutation: *TG* p.1493Y and *DUOX2* p.Q686X in F8 a, b and *TG* p.W2685L and *DUOX2* R354W (predicted to perturb the DUOX2 peroxidase-like domain) in F9a, b (**Figure** 2.9). There was also a strong history of goitre (mother and maternal aunt) in F8 but maternal DNA was not available to confirm the *DUOX2* genotype.

In all three kindreds (F6, 8, 9) the most severe phenotype was observed in individuals harboring biallelic *TG* or triallelic (biallelic *TG* plus monoallelic *DUOX2*) mutations, however it was impossible to disentangle the relative contribution of each mutation to the phenotype reliably in these small pedigrees with limited subphenotype data.

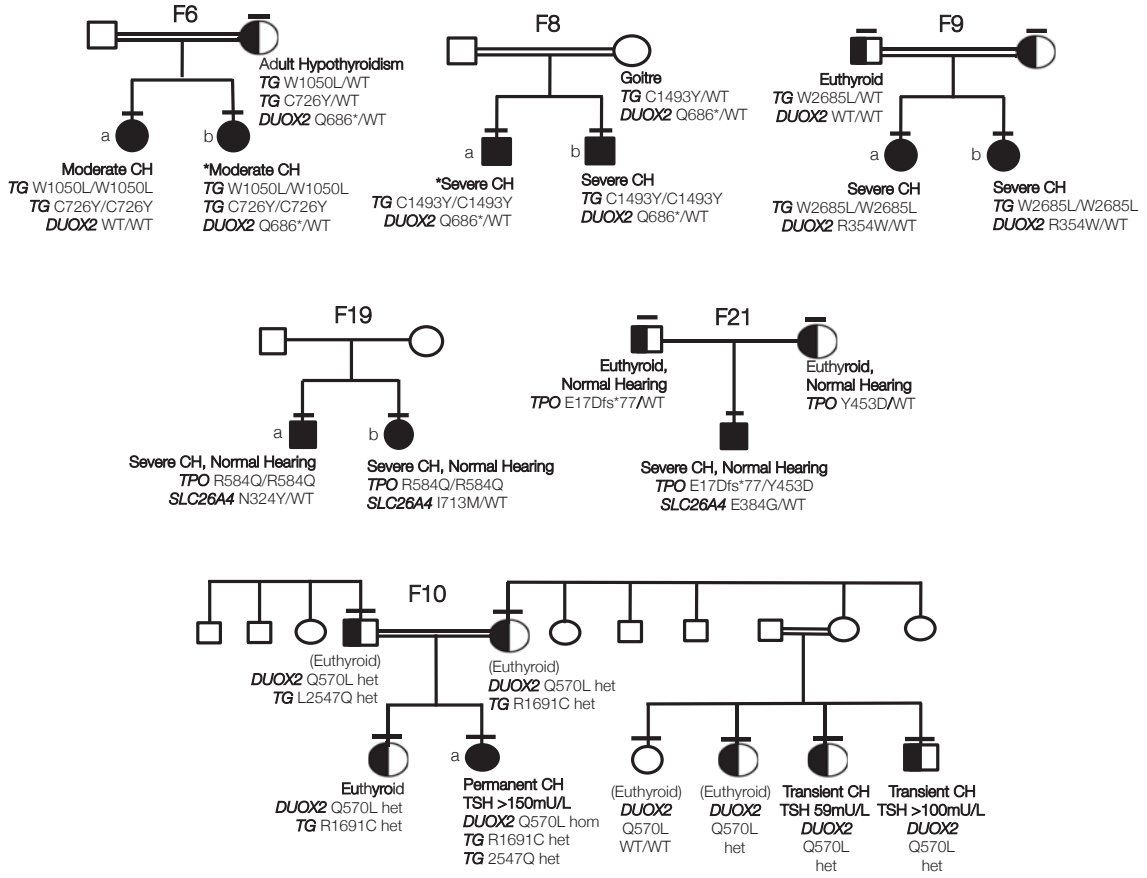


Figure 2.10 Genotype-phenotype segregation in six families with oligogenic variants.

Horizontal bars denote individuals who have been genotyped. Black shading denotes homozygous individuals and half-black shading denotes heterozygotes for *TG* mutations (F9, F6, F8), *TPO* mutations (F19, F21) and *DUOX2* mutations (F10). Potential oligogenic modulators are included by aligning genotype and phenotype data with the individual to whom they refer in the pedigree. *; cases for whom complete biochemical data at diagnosis is not available (F6b, F8a) and CH severity refers to sibling. In F10, black, half-black and white shading denote the *DUOX2* genotype (Q570L homozygous, heterozygous or wild-type respectively). The pedigree is annotated with *TG* genotype in those cases harboring variants (L2547Q, R1691C), and phenotype (euthyroid, transient or permanent CH) with venous screening TSH results for CH cases. Cases annotated (euthyroid) were born in Pakistan and although euthyroid in adulthood, the fact that they were not screened neonatally for CH may have precluded detection of transient CH.

Since monogenic, heterozygous *DUOX2* mutations (including p.Q686X) are frequently associated and sufficient to cause CH, I hypothesized that an additive phenotypic contribution of all three mutations was plausible. To understand whether the heterozygous *DUOX2* mutations are contributing to the CH phenotype in these families, in addition to their *TG* genotypes, one needs to test if the observed number of *TG* carriers with

a *DUOX2* monoallelic variant differs from what would be expected under the null. To minimise the impact of stratification, the null expectation should be calculated using a large control population matched as closely as possible to patients in terms of ancestry. However, since no Turkish exomes were publicly available, I used the ExAC population with the largest number of *DUOX2* variants as controls (N=8,654 East Asian chromosomes). The frequency of rare predicted damaging heterozygous variants in *DUOX2* in ExAC East Asians was 0.06. Therefore under the null, one would expect to see 0.66 *TG* family carriers also carrying a *DUOX2* mutation by chance (11 x 0.06 = 0.66), however, the observed value was three *TG* families, which is significantly higher than the expectation (P=0.03, Fisher's exact one-tail). This finding supports a potential phenotypic contribution of the *DUOX2* mutation in these individuals, however a much larger cohort of sequenced CH cases will be required to assess the phenotypic consequences of digenicity in CH thoroughly.

TPO and *SLC26A4* / *TG*

Biallelic mutations in *TPO* were identified in two European families in addition to heterozygous known *SLC26A4* mutations, previously associated with Pendred syndrome (OMIM: 274600) when homozygous: F19a: *TPO* p.R584Q (homozygous) and *SLC26A4* p.N324Y (heterozygous); F19b: *TPO* p.R584Q (homozygous) and *SLC26A4* p.I713M (heterozygous); F21: *TPO* p.[E17DfsX77 + Y453D] (compound heterozygous) and *SLC26A4* p.E384G (heterozygous) (**Figure** 2.10). The novel *TPO* p.R584Q missense variant is predicted to perturb polar contacts possibly affecting the catalytic domain (**Figure** 2.8). The occurrence of Pendred syndrome usually mandates biallelic *SLC26A4* mutations, and manifests universally with congenital or postnatal progressive sensorineural hearing loss, whereas thyroid dysfunction is usually mild or absent [266]. In both these kindreds (F19, 21), only the biallelic *TPO* mutations segregated with CH, which was severe whereas the hearing was normal, suggesting the *SLC26A4* mutations do not play a role in the CH phenotype of these patients.

In F11, a known homozygous pathogenic *TPO* mutation (p.R491H) was inherited together with a heterozygous *TG* variant (p.Q1644E). Again, since biallelic inheritance is also usually required for CH due to *TG* mutations, this observations suggest the *TPO* mutations are the predominant drivers of the CH phenotype in this family as well.

Genotype-phenotype correlation analysis in family F10

Detailed investigation of the contribution of oligogenicity to genotype-phenotype variability requires the study of large pedigrees, with a spectrum of genotypes, e.g. F10 (**Figure** 2.10). In this large, consanguineous Pakistani family, the proband (F10a) harbored a known pathogenic *DUOX2* mutation, p.Q570L, previously published in [342]. Homozygosity for this mutation segregated with permanent CH in the proband, whereas his parents, sister and cousins, who were all heterozygotes, presented with either euthyroidism (i.e. normal thyroid state) or transient CH. Two novel, rare *TG* variants (p.L2547Q, predicted to be pathogenic by PolyPhen and SIFT, and p.R1691C, of less certain significance) were also identified in this kindred, yet neither of these variants segregated with transient CH in the *DUOX2* p.Q570L heterozygotes, suggesting digenic mutations in *TG* and *DUOX2* do not explain the phenotypic variability seen in this kindred.

2.5.5 'Ambiguous' and 'unsolved' families

The ambiguous category included two cases harboring heterozygous pathogenic *TG* variants: a novel nonsense mutation in F13 (p.Q771X) and a previously described missense mutation in F12 (p.Q870H, **Table** 2.3, **Figure** 2.7). An additional case was heterozygous for a frameshift mutation in *TPO* (p.E510AfsX14, F22). Previous reports of CH due to *TG* and *TPO* mutations most commonly involve biallelic mutations, therefore it is unclear whether the mild or subclinical hypothyroidism observed in these patients is attributable to the monoallelic mutation or whether they harbored a second 'hit' not detected by the exome and targeted-sequencing methods employed here. Other cases in this category harbored novel heterozygous *TG* missense (p.Y759C, F14) or splice region (c.3433+3_3433+6delGAGT, F15) variants, a novel heterozygous *DUOX2* variant (p.R764W, F25) inherited from a healthy parent, and a homozygous *DUOXA2* splice site (c.555–5G>A) variant for which *in silico* predictions were inconclusive (F27).

Overall, nine cases from seven families remained completely unsolved, with no likely disease-causing variants identified in the eight genes screened.

Exome and Hi-Seq targeted-sequencing samples from both of these categories (i.e. 14 out of 20 cases) were subject to further genetic analyses to investigate whether inherited or *de novo* variation, including copy-number defects, outside the known *gland-in-situ* CH genes contribute to their phenotype. This work will be presented in the following chapter.

2.6 Discussion

In this study, whole-exome and targeted-sequencing strategies enabled the efficient screening of eight known genes associated with CH and GIS in 49 cases from the UK, Turkey, Middle East and Asia, and with a spectrum of biochemical phenotypes. In addition to single-gene mutations, the contribution of oligogenic variants was assessed. Mutations in the screened genes collectively explained 59% of the cases. Previous genetic analyses in *gland-in-situ* CH cohorts have been less comprehensive, screening smaller numbers of genes or fewer cases with specific ethnicities [229, 310, 345, 508]. The only large-scale multiplex study in CH did not select cases on the basis of thyroid morphology and excluded *TG*, *SLC26A4* and *IYD* from its sequencing panel [372]. Direct sequencing of *DUOX2*, *TG*, *TPO* and *TSHR* has been undertaken in 43 Korean CH cases with GIS [229] and, in common with our study, only around 50% of cases harbored pathogenic variants in one or more genes.

The frequency of mutations in known CH causative genes depends on the selection criteria and the ethnic origin of the cohort [28, 229]. The cohort studied here included individuals of diverse ethnicities, in whom biochemical diagnosis of CH was achieved using different, country-specific, screening protocols, or following neonatal or early childhood presentation with clinical hypothyroidism. This variation precludes a detailed comparison of relative mutation frequencies with other population studies with a uniform ethnicity or biochemical diagnostic approach. However, the spectrum of TSH levels at diagnosis in this cohort would almost all be regarded as positive on the UK neonatal screening programme (Dr Nadia Schoenmakers, personal communication). Further, the mixed ethnicity of our cohort removes bias from founder mutations in specific genes, and reflects the ethnic heterogeneity of real clinic populations in some regions of the UK, meaning the findings presented here have broader relevance to CH with GIS even though they cannot be easily compared to previous studies.

In this cohort of mixed ethnicities, mutations were most frequently found in *TG*, followed by *TPO*. *DUOX2* mutations were relatively infrequent compared with findings by Jin *et al*, who reported mutations in ∼35% of their East Asian cases [229]. This finding probably reflects the higher prevalence of *DUOX2* mutations in individuals of East Asian ethnicity (which is corroborated by the ExAC data), who were poorly represented in our study, rather than incomplete or unsuccessful sequencing of *DUOX2* in our cohort, as I have demonstrated. No convincing pathogenic mutations were found in *DUOXA2*, *IYD* and *SLC5A5*, which is in line with previous reports suggesting these are rarer genetic causes of dyshormonogenesis [345, 372, 459]. The paucity of *TSHR*

mutations in a CH cohort with GIS is surprising [229]; however, the high incidence of consanguinity predicts occurrence of biallelic mutations that, in the case of *TSHR*, normally causes thyroid hypoplasia, i.e, the incomplete development of the thyroid gland [386], which would have been excluded by the selection requirement for normal-sized or goitrous gland. Despite unselected recruitment of either sporadic or familial cases, this CH cohort was greatly enriched for familial CH (76% cases), which may have increased the percentage of cases harboring an underlying genetic etiology. In a standard clinic population with a greater proportion of sporadic cases, the proportion of mutation-negative cases could indeed be higher.

2.6.1 The significance of the causative variants identified

Interpretation of novel genetic variants requires *in vitro* or *in vivo* functional studies in order to confirm pathogenicity. For the case of *TG* mutations, *in vivo* measurement of serum Tg levels could be conducted to confirm the genetic defect (**Table** 2.1) [471]. For the case of *TPO* and *DUOX2*, the enzymatic activity of patients could be evaluated using a radioiodine uptake and perchlorate (CLO_{4^-}) discharge test (**Table** 2.1). Because CLO_{4^-} competitively interferes with iodide trapping in the thyroid, the test would measure the amount of tyrosyl-unbound radioiodine that would be lost from the gland after perchlorate administration; for a wild-type (WT) *TPO* individual, the discharge should be no more than 10% [439]. Alternatively, for *DUOX2* defects, site-directed mutagenesis on the WT cDNA, followed by transfection in cells and measurement of H_2O_2 production, could be conducted to evaluate the degree of functional impairment.

Although such investigations were not undertaken, the novel mutations identified herein are mostly private to a given family or extremely rare (after screening of more than ∼81,000 population samples from diverse ethnic background), segregate with the phenotype within families, and have strong *in silico* (bioinformatic or structural) predictions of pathogenicity, supporting a causal role. Approximately 38% of the novel variants identified are loss-of-function, such as nonsense and frameshifts variants, which are known to be extremely rare in the general human population, with only one *TG* and two *DUOX2* nonsense heterozygous carriers in 60,706 ExAC individuals and no homozygous samples. Moreover, the location of the novel variants in *TPO* (heme- or substrate-binding regions) and *DUOX2* (peroxidase-like domain) matches that of previously described pathogenic mutations [178, 423]. The analysis of novel variants in *TG* is hindered by an incomplete knowledge of its functional domains and crystal structure [472], but the variants that were identified affect similar regions to

previously documented mutations, which are normally located in N-terminal cysteine-rich repetitive elements or in the C-terminal ACHE-like domain, which also supports causality [342, 423, 472].

2.6.2 Clinical phenotypes of mutation carriers

The associated clinical phenotypes in our mutation-positive patients were similar to published cases. *TG* mutations may result in mild or severe hypothyroidism [472], and monoallelic and biallelic *DUOX2* mutations may cause both permanent or transient CH with significant inter- and intrafamilial phenotypically variability [310, 312, 334, 342, 508]. Even *TPO* mutations, although classically associated with total iodide organification defects, can cause milder phenotypes [423]. In our cohort, biallelic *TG* mutations were predominantly associated with moderate to severe CH. In cases harboring *DUOX2* mutations, a spectrum of phenotypes were observed, ranging from transient to permanent CH with intrafamilial variability noted specially in association with monoallelic *DUOX2* mutations.

Solved cases usually had a more severe phenotype than unsolved or ambiguous cases, however the latter group included four cases of subclinical or mild CH harboring heterozygous mutations in *TPO* or *TG*. Such monoallelic mutations have previously been described in association with CH, but are usually assumed to coexist with an additional undetected CNV, intronic or regulatory mutation in the other chromosome [28, 82, 157]. This may be the case in these patients as well, however, the sequencing techniques employed here would not have detected mutations in non-coding regions of the genome, and CNVs were not called in this cohort (but are investigated in the exome-sequenced samples in the following chapter).

2.6.3 The role of digenicity in disease development

Oligogenicity has often been proposed to underlie the intrafamilial variability seen in known genetic causes of CH, especially in association with *DUOX2* mutations [342]. Despite reports of digenic GIS cases in the literature, pedigree studies have either not been performed [229, 372] or have not confirmed a genotype-phenotype correlation [460]. In this study, we detected likely pathogenic variants in more than one CH-associated gene, especially in consanguineous Turkish kindreds, most commonly involving *TG* and *DUOX2*. I have also demonstrated that the rate of this event is significantly

higher in our cohort when compared to the expected rate seen in the ExAC population harboring the largest number of *DUOX2* mutations (ExAC East Asians). Therefore, even though cases and controls were not appropriately matched in terms of ancestry, this analysis was conducted as conservatively as possible. Nevertheless, small pedigree sizes, poor information about mutation frequencies in populations matched exactly to the CH cases, and a paucity of subphenotype data preclude definitive statements regarding the relative aetiological contribution of digenicity in CH, which still remains inconclusive. In addition, it is also possible that our study is underestimating the frequency of oligogenicity in CH with GIS; the high percentage of consanguinity in our cohort facilitates the identification of potentially pathogenic variants in a disease model with recessive inheritance, but also increases the likelihood of detecting variants which are contributory to the CH phenotype but not causative, due to the occurrence of genomic regions with loss- of-heterozygosity involving CH-associated genes.

Further studies with large pedigrees and clear phenotypic variability are required to ascertain the role of polygenic modulators in CH with GIS. Alternative candidate genes involved in the same biological pathways as known *gland-in-situ* causative genes, may be implicated, and these may either exacerbate or play a compensatory role in the context of loss-of-function mutations. Examples include *DUOX1*, *DUOXA1*, and *NOX*, which are also involved in H_2O_2 production in thyrocytes, and whose expression may be upregulated in the context of *DUOX2* deficiency [220, 460].

2.6.4 Limitations

It is conceivable that despite adequate median coverage, non-uniform coverage of genes could have resulted in failure to detect variants. This is most likely to be significant for the eleven cases (eight families) that underwent HiSeq targeted-sequencing, and in which coverage of specific exons was <10-fold, predominantly affecting *DUOXA2* and *SLC5A5*. Suboptimal coverage of these regions raises the possibility of a type II error. However, undetected variants in these cases are unlikely to affect the conclusions of this work since five of these cases harbored mutations that explained their CH (F26, F2a, b, F11, F17), and two ambiguous cases harbored heterozygous *TG* variants (F12 a, b). Previous studies have also reported considerable variability in uniformity and depth of coverage across the exome [122, 246, 304], so this finding is not uncommon and represents a well known limitation of target enrichment and sequencing technologies, which may sometimes impact and limit variant identification.

2.6.5 Future work

The aetiology of CH with GIS remains elusive and factors other than known *gland-in-situ* associated genes must be implicated. The high familial component (57%) in the unsolved case category favors an etiological contribution of genetic factors rather than environmental modulators (e.g. iodine status). Future studies with exome or whole-genome sequencing in familial cases may identify novel genetic aetiologies for CH with GIS, elucidating novel pathways in thyroid development and physiology. Specifically, other genes involved in thyroid hormone synthesis, but expressed outside the thyroid follicular unit, may play a role in disease, as well as genes that have been recently postulated, by GWAS studies, to influence TSH and free-T_4 levels [474]. Such hypotheses will be explored, via exome-sequencing analyses of these and additional CH samples, in the following chapter.