

Understanding human disease using high-throughput sequencing



Eva Gonçalves Serra

Wellcome Trust Sanger Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Declaration

The work presented in this dissertation was carried out at the Wellcome Trust Sanger Institute between October 2012 and August 2016. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the Colleagues section or the text. It does not exceed the word limit set out by the Degree Committee for the Faculty of Biology, and is not substantially the same as any work that has been, or is being, submitted to any other University for any degree, diploma or any other qualification.

This is a post-viva dissertation, containing some minor corrections to that submitted on August 2016. The corrections were suggested by Dr Helen Firth and Prof William Newman.

Eva Gonçalves Serra
November 2016

Hold the vision, trust the process.

Abstract

Next-generation sequencing (NGS) is revolutionising Mendelian and complex disease research by enabling variant information to single-base resolution in a high-throughput way, scalable to the size of the human genome. In this dissertation, I describe four distinct projects in which NGS technologies were employed, in combination with different study designs and analytical strategies, to identify genetic determinants, or modifiers, of diseases that have been poorly studied thus far.

In Chapter 1, I provide a historical background of our understanding of how genetic variation contributes to disease phenotypes, and the technological advances in the last twenty years that have led to the NGS-based gene-mapping studies of today.

In Chapter 2, I describe a NGS-based screening of genes that are known to cause thyroid hormone production defects in a congenital hypothyroidism (CH) cohort of patients with *gland-in-situ*. I show how a stringent variant filtering pipeline, combined with pedigree segregation analyses and *in silico* predictions of pathogenicity for candidate variants, led to the identification of likely causal mutations in 59% of the patients.

In Chapter 3, I describe a family-based exome and targeted-sequencing analysis to identify novel genetic causes of CH. I explore different variant filtering pipelines to map *de novo*, inherited and copy-number-variants segregating with disease within families. I find that no gene is recurrently mutated across families over what is expected by chance. I then explore how a candidate-gene screening approach, leveraging rare disruptive mutations mapped in families, can highlight novel genes potentially associated with CH or the extrathyroidal features of some patients.

In Chapter 4, I describe a series of analyses to better understand the genetic architecture of very-early-onset inflammatory bowel disease (VEO-IBD). This condition is currently viewed as a Mendelian form of inflammatory bowel disease (IBD), a complex disorder of adulthood onset. Using exome data, I identify likely causative defects in known primary-immunodeficiency genes and explore the broader contribution of rare variants to VEO-IBD through case-control enrichment analyses at the level of single genes,

genesets and biological pathways. Moving beyond rare alleles, I generate polygenic risk scores leveraging the set of known, adult-onset IBD-risk alleles discovered to date, and demonstrate a polygenic component operating in VEO-IBD.

In Chapter 5, I describe a meta-analysis combining low-coverage whole-genome sequencing data and three genome-wide-association studies to identify genetic modifiers of age at IBD diagnosis. Four loci were discovered associated at suggestive significance with Crohn's disease (CD), ulcerative colitis or both, one of which may have a pleiotropic effect, being associated with both the risk of CD and a decrease age at CD diagnosis.

Finally, in Chapter 6, I highlight the major lessons learnt with these projects, discuss some immediate impact some of these results had for patients, and look forward to the future NGS-based studies that will shape gene-mapping strategies over the next coming years.

Acknowledgements

I owe my gratitude to a great many people who helped me through this PhD journey. First and foremost, I thank my primary supervisor, Dr Carl Anderson, for giving me the opportunity to pursue this PhD and for his faith in me throughout these years. I remember knocking on your office door four years ago and asking for some sequencing data I could get to grips with – I was keen to learn more about genetics with you, and I absolutely did! You taught me the importance of critical thinking and statistical rigour and you helped me grow immensely as a research scientist. I am beyond grateful for that. This dissertation would not have been possible without your support and constant optimistic encouragement, especially during tough times.

I'd also like to give a heartfelt, special thanks to my secondary supervisor and clinical collaborator, Dr Nadia Schoenmakers. Thank you for the opportunity to work with you, in an exciting and fruitful collaboration. It has been an absolute privilege. Thank you for your guidance, infectious passion for biology, and for always putting our work in a bigger and fulfilling context.

I am especially indebted to Dr Inês Barroso, my first mentor at Sanger, for the confidence she has shown in me throughout the years. Thank you for taking me as a young and naive Master's student and for turning me into an excited human geneticist wannabe one year later. I still have much to learn, but I owe my first steps to you.

I'd also like to thank my other advisers, thesis committee members and first year examiners including: Professor Krishna Chatterjee, Dr Chris Tyler-Smith, Dr Matt Hurles and Dr Jeff Barrett. I am also grateful to Dr Annabel Smith, Christina Hedberg-Delouka and the Committee of Graduate Studies for keeping the Sanger PhD programme running smoothly.

The work presented in this thesis would not have been possible without the collaboration of many colleagues at Sanger, Cambridge, Oxford and other points of the world. Thank you to Professor Krishna Chatterjee, Professor Holm Uhlig, Adeline Nicholas, Dr Tobias Schwerd, Dr Erik Schoenmakers, Martin Howard, Dr Hakan Cangul, Dr Amir Babiker,

Dr Irfan Ullah, Dr Saif Alyarubi, Dr Asma Deeb, Dr Abdelhadi Habeb, Dr Justin Davies, Philip Murray, Dr Shenoy Savitha, Mehul Dattani, Dr Ruben Willemsen, Dr Ajay Thankamony, Dr Soo-Mi Park, Dr Ahmed Massoud, Dr John Gregory, Dr Vijaya Parthiban, Dr Shane McCarthy, Nicola Corton, Tarjinder Singh, Katie De Lange, Dr Yang Luo and Daniel Rice.

Thank you also to the legion of doctors, nurses, researchers and administrators of the UK10K and the UK IBD Genetics Consortia for their efforts in bringing large cohorts of patients around the world towards the noble goal of advancing disease research. I'd also like to extend a big thank you to all the colleagues at Sanger who work in the sample management and sequencing team pipelines, as well as the many other people who keep the laboratory and computational facilities running. I have been tremendously fortunate to have had your help throughout. Special mention goes to Martin Pollard, Irina Colgiu, Allan Daly and Colin Nolan from the Human Genetics Informatics team, for going far beyond the call of duty.

I'd like to express my sincere gratitude to the Wellcome Trust for generously funding the four years of my PhD, as well as the various sponsors of the above Consortia for making these projects possible.

I am forever grateful to all patients (and their families) who participated in these research studies, without whose generosity none of this work would have been possible.

To members of the Anderson Group (both past and present) – Sun-Gou Ji, Tejas Shah, Javier Archury, Loukas Moutsianas, Jimmy Liu, Jamie Floyd, Velislava Petrova and Carmen Diaz – you became family. Thank you for patiently listening (and provoking!) all my daily rants; you really made the difference.

All the other people who have made the Genome Campus what it is for me – Joanna Kaplanis, Chris Franklin, Electra Tapanari, Pedro Albuquerque, Luis Pureza, Scott Shooter, Manuela Menchi, Ricardo Antunes, Rui Pereira, Alina Farmaki, Luis De Figueiredo, Ricardo Miragaia, Maria Fernandez, Neneh Sallah, Pinky Langat, Michal Szpak, Mia Petljak, Paris Litterick and Carol Dunbar – you brightened up my days.

Last, but certainly not least, I thank my family, who has provided unwavering support and limitless pride. Particularly: my dad for setting the bar high; my mum for reminding me that life is so much more than work; my four younger sisters Marta, Maria, Sofia and Inês for always picking up the phone; and finally my endlessly supportive soon-to-be husband Stathi, whose cooking skills really did improve throughout the course of this PhD. I love you all.

Publications

From this dissertation

Nicholas, A. K.* , **Serra, E. G.*** *et al*, (2016). Comprehensive screening of eight known causative genes in congenital hypothyroidism with *gland-in-situ*. *The Journal of Clinical Endocrinology and Metabolism*, 101(12): 4521–4531.

Serra, E. G. *et al*. Whole-exome sequencing and genome-wide genotyping defines the genetic architecture of very-early-onset inflammatory bowel disease. *In preparation*.

Arising elsewhere

Boudellioua, I., Razali, R. B. M., Kulmanov, M., Vladimir, B. B, **Serra, E. G.**, Schoenmakers, N., Gkoutos, G., Schofield, P. and Hoehndorf, R. (2016). Genome-scale identification of causative variants involved in human disease. *Under review*.

Luo, Y., de Lange, K. M., Jostins, L., Moutsianas, L., Randall, J., Kennedy, N. A., Lamb, C. A., McCarthy, S., Ahmad, T., Edwards, C., **Serra, E. G.**, Hart, A., Hawkey, C., Mansfield, J. C., Mowat, C., Newman, W. G., Nichols, S., Pollard, M., Satsangi, J., Simmons, A., Tremelling, M., Uhlig, H., Wilson, D. W., Lee, J. C., Prescott, N. J., Lees, C. W., Mathew, C. G., Parkes, M., Barrett, J. C., and Anderson, C. A. (2016). Exploring the genetic architecture of inflammatory bowel disease by whole genome sequencing identifies association at *ADCY7*. *Under review*.

* Jointly contributing authors.

Table of contents

List of figures	xix
List of tables	xxi
1 Introduction and historical perspective	1
1.1 The genetic architecture of disease	2
1.2 Gene-mapping in human disease	4
1.3 The start of gene-mapping: linkage analysis	4
1.4 Genome-wide association studies	7
1.5 The next-generation sequencing revolution	11
1.6 A standard NGS workflow	16
1.6.1 Sequence generation	16
1.6.2 Alignment and variant calling	18
1.6.3 Data annotation	19
1.7 NGS genetic analyses in Mendelian diseases	20
1.8 NGS genetic analyses in complex diseases	25
1.9 Outline of dissertation	28
2 NGS-based screening of known causative genes in CH with <i>gland-in-situ</i>	31
2.1 Introduction	31
2.1.1 What is congenital hypothyroidism?	31
2.1.2 The known genetics of CH with <i>gland-in-situ</i>	32
2.1.3 Previous genetic studies of CH with <i>gland-in-situ</i>	33
2.2 Aims	36
2.3 Colleagues	36
2.4 Methods	36
2.4.1 Patients	36

2.4.2	Next-generation DNA sequencing	37
2.4.3	Sequencing efficiency of WES and HiSeq-TS experiments	41
2.4.4	Variant annotation	41
2.4.5	Identifying likely damaging variants per sample	41
2.4.6	Capillary sequencing for variant validation	42
2.5	Results	43
2.5.1	Sequencing data quality	43
2.5.2	Genetic diagnostic yield	46
2.5.3	‘Solved’ families with mutations in one gene (monogenic families)	50
2.5.4	‘Solved’ families with mutations in two genes (digenic families)	56
2.5.5	‘Ambiguous’ and ‘unsolved’ families	59
2.6	Discussion	60
2.6.1	The significance of the causative variants identified	61
2.6.2	Clinical phenotypes of mutation carriers	62
2.6.3	The role of digenicity in disease development	62
2.6.4	Limitations	63
2.6.5	Future work	64
3	Exome and targeted-sequencing of families with congenital hypothyroidism	65
3.1	Introduction	65
3.1.1	Thyroid developmental defects	66
3.1.2	Arguments for a genetic involvement in TD	67
3.1.3	Genetic studies of thyroid dysgenesis	71
3.1.4	Why exome-sequence CH cases	72
3.2	Aims	72
3.3	Colleagues	73
3.4	Methods	73
3.4.1	Patients	73
3.4.2	Sequencing	75
3.4.3	Data quality control	76
3.4.4	Gene mapping within CH families	83
3.4.5	Predicting the impact of splice donor mutations	88
3.5	Results	90
3.5.1	Inherited variants in CH families	90

3.5.2	<i>De novo</i> variation in CH trios	96
3.5.3	Copy-number-variants in the WES dataset	102
3.5.4	Searching for novel genetic causes of CH across families	104
3.5.5	Searching for likely damaging variants in candidate genes	106
3.6	Discussion	110
3.6.1	A putative causative gene for CH with <i>gland-in-situ</i>	110
3.6.2	<i>De novo</i> and CNVs in TD and syndromic CH	111
3.6.3	Limitations	111
3.6.4	Future work	113
4	The genetic architecture of <i>very-early-onset</i> inflammatory bowel disease	115
4.1	Introduction	115
4.1.1	What is inflammatory bowel disease?	115
4.1.2	The genetics of IBD	116
4.1.3	Paediatric-IBD	120
4.1.4	Very-early-onset IBD	120
4.1.5	The genetics of VEO-IBD: the rare-variant hypothesis	121
4.1.6	Another hypothesis for the aetiology of VEO-IBD	125
4.2	Aims	126
4.3	Colleagues	127
4.4	Methods	127
4.4.1	Patients	127
4.4.2	Controls	128
4.4.3	Exome sequencing and variant calling	128
4.4.4	Data quality control	128
4.4.5	Annotations	134
4.4.6	Screening of IBD-like inflammatory genes	136
4.4.7	Gene-based association analysis	136
4.4.8	Geneset and pathway enrichment analysis	138
4.4.9	Calculation of polygenic risk scores in VEO-IBD cases	139
4.5	Results	144
4.5.1	Identification of causative defects in IBD-like inflammatory genes	144
4.5.2	Gene-based association analyses	149
4.5.3	Genesets and pathway enrichment analyses	152
4.5.4	The polygenic component of VEO-IBD	154

4.6	Discussion	156
4.6.1	The importance of screening IBD-like genes in VEO-IBD cohorts	156
4.6.2	The role of rare variants in VEO-IBD	157
4.6.3	VEO-IBD has a polygenic component	159
4.6.4	Limitations	161
5	A meta-analysis to map loci associated with age at IBD diagnosis	163
5.1	Introduction	163
5.1.1	The role of genetic variation in the age at IBD diagnosis	163
5.2	Aims	164
5.3	Methods	165
5.3.1	Association analyses	165
5.3.2	Meta-analysis within CD and UC studies	167
5.3.3	Meta-analysis for IBD	167
5.3.4	Post meta-analysis quality control	168
5.3.5	Power to detect previous ImmunoChip signals	168
5.4	Results	169
5.4.1	Suggestive association for the age at CD diagnosis	173
5.4.2	Suggestive associations for the age at UC diagnosis	177
5.4.3	Suggestive association for the age at IBD diagnosis	178
5.4.4	Comparison with the previous ADD ImmunoChip study	180
5.5	Discussion	182
5.5.1	The advantage of imputation	182
5.5.2	The pitfall and advantage of my genome-wide analysis	182
5.5.3	The possible pleiotropy of <i>FOSL2</i>	183
6	Conclusions and future prospects	187
6.1	Summary of my research	187
6.2	NGS: from bench to bedside	189
6.3	Common themes emerging from my research	190
6.3.1	Sample size	190
6.3.2	Phenotypic heterogeneity	192
6.3.3	Diverse ethnic origin	193
6.4	Future studies of rare and complex diseases	195
6.5	From variant discovery to disease mechanisms	197
6.6	Translation	200

6.6.1	Novel drug targets	200
6.6.2	Personalised treatments	200
6.6.3	Genetic risk prediction	201
6.7	Concluding remarks	202
References		203
Appendix A Appendix		271

List of figures

1.1	Inheritance of Mendelian and complex disorders	3
1.2	Linkage analysis within a family	5
1.3	Rate of Mendelian disease gene discovery between 1988-2012	7
1.4	Pace of GWAS publications since 2005	8
1.5	Schematic representation of a case-control GWAS study	10
1.6	Locus discovery in IBD over the past 15 years	11
1.7	High-throughput sequencing technology	12
1.8	Computational steps involved in NGS data generation	17
1.9	The functional impact of genetic variants at the protein level	20
1.10	A NGS-based study design for complex disease studies	27
2.1	Thyroid hormone synthesis	34
2.3	Sequencing efficiency per gene	44
2.4	Proportion of exons poorly covered per gene	45
2.5	Summary and distribution of mutations observed in the CH-GIS cohort	48
2.6	Causative variants identified in CH cohort with GIS	49
2.7	Mutations identified in <i>TG</i>	51
2.8	Mutations identified in <i>TPO</i>	53
2.9	Mutations identified in <i>DUOX2</i>	55
2.10	Genotype-phenotype segregation in families with oligogenic variants. . .	57
3.1	Gene expression during thyroid gland development	68
3.2	Phenotype categories within the CH cohort	74
3.3	Principal component analysis of exome and targeted-sequencing CH samples	77
3.4	Consanguinity status for exome-sequenced CH cases	79
3.5	Quality control metrics for the WES experiment	81
3.6	Population genetics metrics for the WES experiment	82

3.7	Variant filtering pipeline to identify inherited variation within CH families	85
3.8	MaxEntScan scores for the wild-type and mutant splice donor sequences of <i>TBX1</i>	92
3.9	Functional impact of inherited alleles identified in WES CH families	95
3.10	Distribution of inherited rare functional variants across CH families	96
3.11	Posterior probabilities of <i>de novo</i> events called by DeNovoGear	97
3.12	Intolerance to loss-of-function mutations for <i>HNRNPD</i> vs. the exome	101
3.13	Copy-number-variants identified in exome-sequenced CH families	103
3.14	<i>SLC26A7</i> expression in GTEx tissues	109
4.1	Epidemiological and clinical features of CD and UC	116
4.2	The genetic architecture of CD and UC	117
4.3	Overview of intestinal immunity in health and disease	119
4.4	Monogenic defects associated with VEO-IBD	124
4.5	Contamination analysis	130
4.6	Principal component analysis of VEO-IBD cases and controls	131
4.7	Mean genotype quality and mean depth per sample	132
4.8	Number of singletons per sample	133
4.9	Pre- and post- variant QC metrics for North European cases and controls	135
4.10	Per-sample heterozygosity rate at autosomal sites	141
4.11	Mutation identified in <i>XIAP</i>	146
4.12	Sequencing coverage for IBD-like genes vs. the exome	148
4.13	Burden of rare functional variants	150
4.14	Burden of rare disruptive variants at four allele frequency thresholds	151
4.15	Geneset enrichment of disruptive variants stratified by allele frequency	152
4.16	Distribution of CD-based and UC-based risk scores in VEO-IBD, CD, UC cases and control individuals	155
5.1	Distribution of age at disease diagnosis across the different studies	168
5.2	QQ plots of the individual association studies	171
5.3	QQ plots of the meta-analysis results for CD, UC and IBD	172
5.4	Regional association plot for 2p28	175
5.5	Genotype cluster plot for a directly genotyped proxy of rs2879179	176
5.6	Regional association plots for the associations with age at UC diagnosis	179
5.7	Regional association plots for the association with age at IBD diagnosis	180
5.8	Effect size estimations for <i>NOD2</i> rs5743293 across the studies	181

List of tables

1.1	Family-study designs used in NGS-based studies of Mendelian diseases	22
2.1	Known gene defects causing CH with <i>gland-in-situ</i>	35
2.2	Summary of CH samples sequenced for each NGS protocol	37
2.3	Known and novel mutations detected in the CH-GIS cohort	47
3.1	Phenotypes associated with mutations in thyroid transcription factors	69
3.2	Mouse models of thyroid dysgenesis	70
3.3	Pedigree structures available in the studied CH cohort	73
3.4	Candidate GIS and TD genes selected for targeted-sequencing	75
3.5	Pedigree segregation rules for different pedigree structures	86
3.6	Rare functional variants identified in three targeted-sequenced CH patients	90
3.7	Summary of <i>de novo</i> calls per family along each filtering step	98
3.8	<i>De novo</i> mutations identified in nine CH trios	99
3.9	Case-control analysis for recurrently mutated genes across CH families	106
3.10	List of CH candidate genes used in the candidate-gene approach	107
4.1	Biological processes involved in the pathology of IBD	118
4.2	Subgroups of paediatric-IBD according to age	120
4.3	Disease-causative genes discovered in the first WES studies of VEO-IBD	123
4.4	Genetic defects associated with IBD-inflammatory phenotypes	137
4.5	Variant subsets used in case-control enrichment tests	138
4.6	Genesets tested in case-control burden analyses	139
4.7	Predicted damaging and conserved variants identified in IBD-like inflammatory genes in VEO-IBD cases	144
4.8	Enrichment of disruptive variants in KEGG pathways at $\alpha = 0.05$	153
5.1	UKIBDGC sample breakdown per contributing study	165
5.2	Number of high-quality SNPs tested in each UKIBDGC study	166

5.3	Genetic loci associated at suggestive significance ($P_{\text{META-value}} \leq 5 \times 10^{-7}$) with age at CD, UC or IBD diagnosis	170
5.4	Power to detect previous loci associated with age at CD and UC diagnosis	181
A.1	VQSR training sets used in WES variant QC	271
A.2	Targeted-sequencing QC filters	272
A.3	Genotype and phenotype information for solved CH cases	273
A.4	Genotype and phenotype information for ambiguous and unsolved CH cases	274
A.5	List of CH candidate genes, part 1	275
A.6	List of CH candidate genes, part 2	276
A.7	Disease phenotype and therapy characteristics for VEO-IBD cohort . .	277

Chapter 1

Introduction and historical perspective

Identifying the genetic factors that determine or modify disease susceptibility phenotypes has become a central goal of human genetics. Genetic studies of disease offer insights into disease biology and pathological mechanisms which can bring tremendous benefits to humanity. Understanding the genetic aetiology of disease can ultimately lead to earlier and improved disease diagnostics, to drugs targeted at the biochemical pathways underlying the disease symptoms, to prevention strategies that reduce the risk of disease and to guidelines for prescribing more effective treatments based on a person's genetic makeup.

1.1 The genetic architecture of disease

In an oversimplified but nevertheless practical dichotomy, human diseases can be separated into Mendelian or complex disorders, depending on the underlying genetic architecture. A trait's genetic architecture comprises of the number of distinct genes that underlie a given disease and, more importantly, the frequency and the effect sizes of their alleles (**Figure 1.1**).

A disease is termed to be Mendelian if the disease alleles segregate according to Mendel's laws of inheritance, usually dominant, recessive or X-linked. These disorders are usually caused by rare and highly penetrant mutations of large effects in a single or very few genes, hence why they are often referred to as "monogenic" or "oligogenic" conditions, respectively. Mutations causing Mendelian disease are rare (usually <1% frequency in the population) because they tend to be negatively selected from the population due to their highly deleterious effects, and are highly penetrant because almost all individuals carrying a particular mutation also express the associated phenotype. There are at least 7,000 Mendelian phenotypes in OMIM, the Online Mendelian Inheritance in Man database [191], a catalogue of human genes and associated disorders. However, this number is never static, with ~300 new phenotypes being added each year [77]. Individually these diseases are usually rare, occurring 1 in 2,000 - <1 in 100,000 individuals, but collectively they affect millions of people worldwide.

Nearly all diseases with prevalence greater than ~1 in 500 are complex diseases (or polygenic/multifactorial), which do not appear to follow a classic Mendelian pattern of inheritance. They do not have a single cause (genetic or otherwise) but have been known from twin and family studies to have a genetic component [315, 498]. These disorders, as well as other human traits where variation is continuous (e.g. body mass), are the product of multiple genes and mostly common frequency alleles (>5% frequency in the population) of small effects, acting in an additive manner in combination with the environment. Contrary to Mendelian diseases, the variants associated with polygenic disorders do not directly cause disease, but rather influence disease risk. All the genetic and environmental factors contributing to a complex disease in a given individual can be summarised in a quantitative measure called "liability", which can be described in a population level as a normally distributed and continuous trait [329].

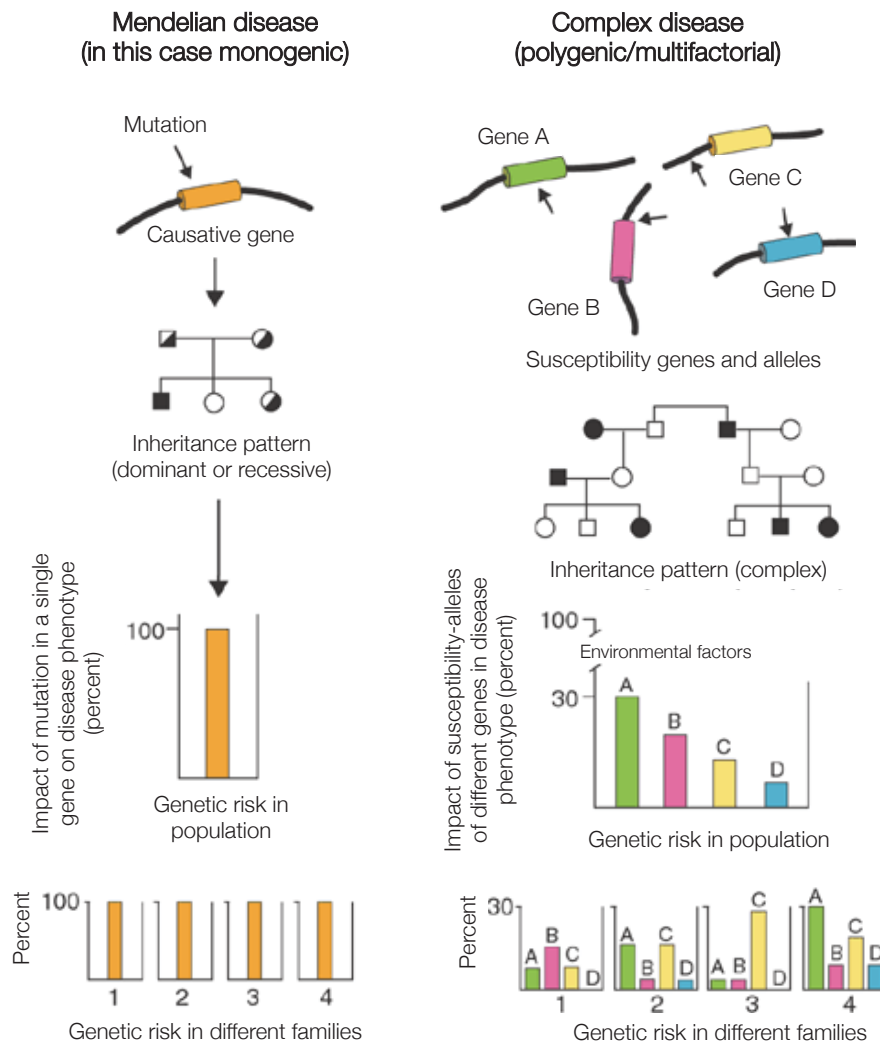


Figure 1.1 Inheritance of monogenic and complex disorders.

In Mendelian monogenic diseases, mutations in a single gene are both necessary and sufficient to produce the clinical phenotype and to cause disease. The genes and mutations involved in such diseases are termed to be “causative”. These mutations often have very high penetrance, meaning almost all affected individuals who carry a mutation also exhibit disease. The same mutation or different mutation in the same gene will be present in phenotypically-similar families, and their impact will be similar in all families. In complex disorders, several alleles in a number of genes result in a genetic predisposition to a clinical phenotype. Genes containing variation related to complex traits are thus referred to as “susceptibility genes”. Pedigrees reveal no clear Mendelian inheritance pattern, and variants are neither sufficient nor necessary to explain the disease phenotype. Environment and life-style factors are major contributors to the pathogenesis of these disorders. In a given population, epidemiological studies evaluate the relative impact of individual genes on the disease phenotype. In complex disorders, any single genetic or environmental factor is expected to explain only a very small fraction of disease risk in a population. Different people in a population may develop disease due to a combination of different genetic and/or environmental reasons. Image adapted from Peltonen *et al* [382].

1.2 Gene-mapping in human disease

With approximately 21,000 protein coding genes to choose from, assigning a specific gene, or group of genes, to a human disorder requires a methodological approach consisting of several steps, a process I refer to as "gene-mapping". Currently, there are many different technologies, study designs and analytical tools for gene-mapping in human disease, all of which have evolved over time and are a product of decades of technological advance in the field of human genetics. Collectively, they equip researchers with a truly diverse "genetic toolbox", where each component (technology/design/analysis) is chosen based on the known (or presumed) genetic architecture of the disease under study, the sample size collected and, of course, the available budget.

Much of this dissertation describes a collection of projects that used next-generation sequencing (NGS) technology, allied with different study designs and analytical strategies, to better understand the genetic basis of two poorly understood human conditions: congenital hypothyroidism and very-early-onset inflammatory bowel disease. The first disorder is considered to be Mendelian in nature, while the second is currently viewed as a Mendelian form, or extreme subtype, of a complex disease (inflammatory bowel disease). For the remainder of this chapter, I provide a brief history of the technological build-up to disease-mapping as we know it, including the techniques, tools and resources that have been developed throughout the years to aid gene-mapping efforts. I then describe the standard NGS data generation workflow that underlies any NGS-based study today, and describe the study designs and analytical approaches that are now commonly used in NGS-based gene-mapping studies of both Mendelian and complex disorders.

1.3 The start of gene-mapping: linkage analysis

Traditionally, linkage analysis was the standard and leading gene-mapping technique. This method identified regions of the genome underlying a given disease by testing a series of marker alleles for co-segregation, or linkage, with disease status within a family or across a number of families. Individuals were usually genotyped for restriction fragment length polymorphisms (RFLPs) [54] or repeat regions (microsatellites) [512] scattered throughout the genome. Markers that were close together on a chromosome were more likely to be co-inherited than would be expected by chance, as recombination was less likely to separate them (**Figure 1.2**).

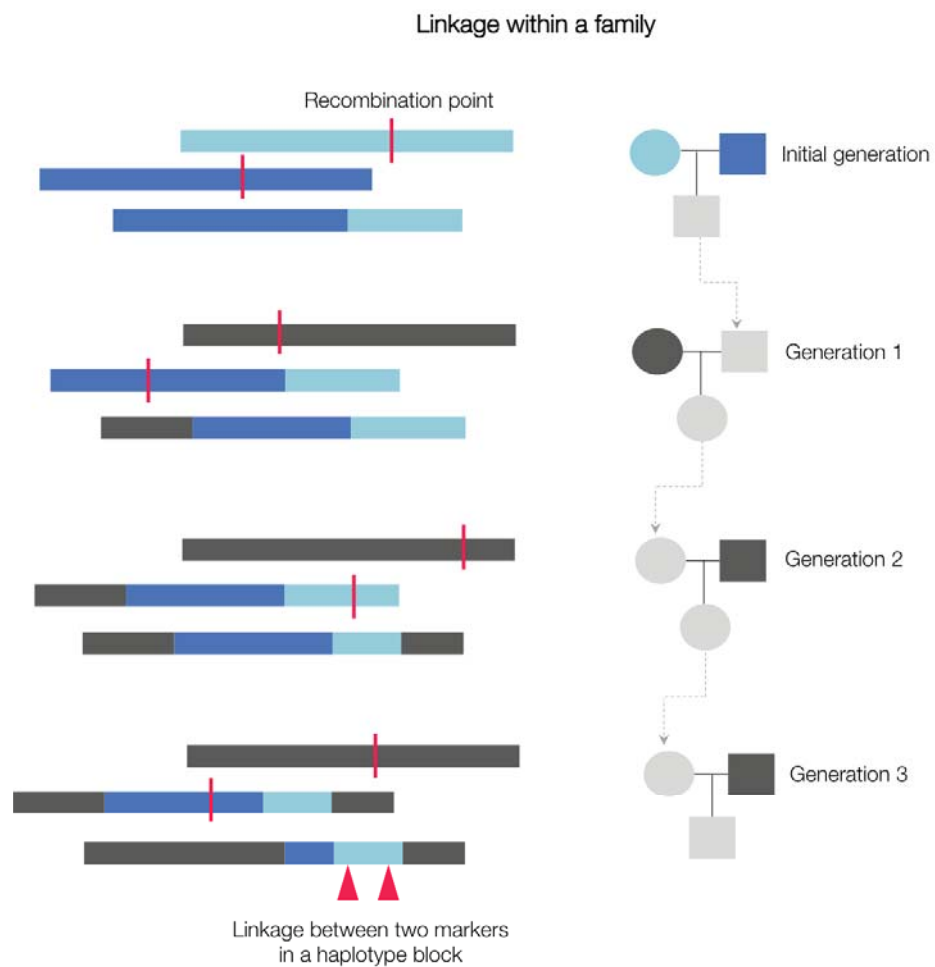


Figure 1.2 Linkage within a family

Within a family, linkage occurs when two genetic markers are co-inherited rather than being broken apart by recombination events during meiosis, shown as red lines. Co-inherited markers are said to be in linkage disequilibrium (LD) with each other and the region with such linked SNPs is called a "haplotype" block. Markers in LD are also termed to be correlated with each other and "tagged" by one another. Image adapted from Bush *et al* [64].

Most linkage studies used a sparse map of 300-400 markers evenly distributed, one every 10 cM, across the genome [131], and these were usually sufficient to capture the majority of the recombination events. The evidence for linkage in a region was measured statistically using a LOD score (logarithm of odds), which compared the likelihood that the genotyped marker and the hypothetical disease locus were inherited together in the observed data, to the likelihood of observing the co-segregation pattern simply by chance. This method would thus narrow down the chromosomal interval in which the disease gene was located, in relation to a known genetic marker, leading eventually to the gene being cloned, Sanger-sequenced and the genetic defects characterised (usually after a long, painstaking process). Even though it may now seem primitive and arduous by modern standards, linkage analysis contained many of the central principles of modern genetics: disease-genes were discovered through direct typing of genetic variants genome-wide, without any prior knowledge of disease biology, coupled with rigorous statistical analysis, careful design and sample ascertainment strategies.

By the mid 90's, linkage had proven to be an extremely effective approach for identifying highly penetrant and rare genetic defects underlying Mendelian diseases with simple genetic architectures, such as Huntington's [189] and cystic fibrosis [492]. More than 1,000 genes underlying Mendelian phenotypes were identified between 1987 and 1997, the decade since RFLP mapping became available [53]. An important lesson emerging from such studies was the notion that most disease-causing mutations cause major changes in the encoded proteins [13]. Linkage was also somewhat successful at identifying alleles with unusually large effects for some complex diseases that showed high familial aggregation. Notable well-replicated examples include *INS* and *CTLA4* in type 1 diabetes [27, 357] and *NOD2* in Crohn's disease [218, 219, 359]. Mendelian subtypes of complex disorders, such as obesity [86], type 2 diabetes [533], breast cancer [524] and Alzheimer's disease [461] were also discovered via linkage, highlighting how the boundaries between Mendelian and complex diseases can sometimes be blurred.

Despite extensive research efforts, linkage was largely unsuccessful at pinpointing the genetic factors involved in complex disorders. In retrospect, this failure was a result of the high locus heterogeneity and the low effect sizes characteristic of such diseases, which made it ill suited to study with this technique. Linkage was also underpowered to elucidate the genetic basis of some Mendelian disorders that were not as simple as initially thought. This was the case for conditions we now know have high levels of phenotypic and genetic heterogeneity, or diseases that occur sporadically due to *de novo* mutations, which were undetected by linkage as they were not transmitted across generations (due to substantially reduced reproductive fitness).

1.4 Genome-wide association studies

The sequencing of the reference genome, accomplished by the Human Genome Project (HGP) in 2003, marked a turning point in gene-mapping research. Knowing the precise location of genes within chromosomal regions enabled quicker progression from a linkage interval to a cloned disease-gene, which accelerated the identification of Mendelian disease genes (**Figure 1.3**). For complex disorders, instead of mapping disease genes by tracing transmission in families, the HGP allowed the creation of high-density polymorphism maps, which expedited population-based association testing at variant sites throughout the genome.

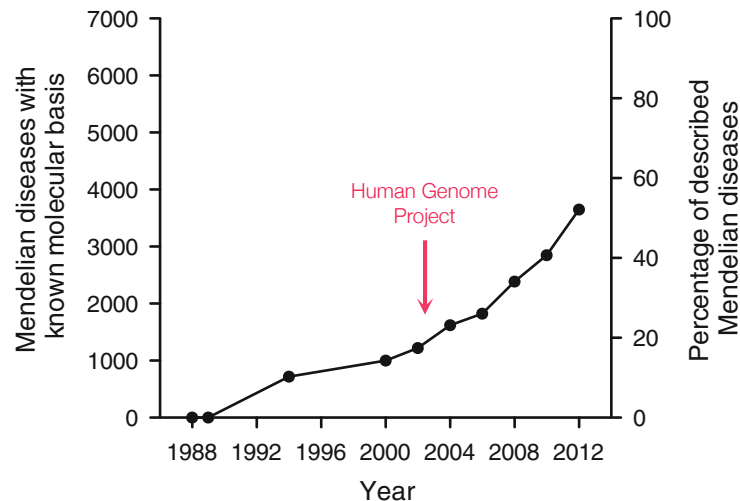


Figure 1.3 Mendelian disease genes of known molecular basis. The left-hand y-axis indicates the cumulative number of diseases for which a molecular basis is identified. The right-hand y-axis expresses that as a percentage of the $\sim 7,000$ Mendelian disorders that have been described and are present in OMIM. Following the release of the human reference genome in 2003, the rate of discovery of Mendelian disease genes increased greatly. Image adapted from Brunham *et al* [60].

In the early 2000s, along with the closing phases of the HGP, several initiatives such as the SNP Consortium and dbSNP were underway to discover and catalogue human genetic variation at the population level. Together, these two projects uncovered at least 1.4 million SNPs [446, 481] or single nucleotide polymorphisms with a population minor allele frequency (MAF) greater than 1%. It became clear that common-frequency SNPs in physical proximity tended to form LD blocks punctuated by recombination

hotspots occurring every 100-200 kb [325]. These correlated patterns (measured in terms of statistical r^2) were further characterised through the HapMap project, which by 2007 had identified a further ~ 3 million SNPs across 270 individuals from three ethnic populations (Europe, Asia and West Africa) [154]. Meanwhile, improvements in chip-based microarray technologies finally made possible the cost-effective and high-throughput genotyping of hundreds of thousands of SNPs in large number of individuals [468]. The newly discovered patterns of LD between SNPs meant that genotyping arrays could effectively survey the majority of common variants in a population by directly assaying only a fraction of the total number of SNPs in the genome. In the European population for example, ~ 5 million common SNPs can be almost entirely "tagged" by a selection of around 500,000 informative markers [32, 154]. Together, these achievements paved the way to the first genome-wide-association-studies (GWAS), a transformative step for the study of complex disorders. Over the last decade, the number of GWAS per year has increased linearly (**Figure 1.4**), with a total of 2,488 GWAS studies and 22,414 unique SNP associations currently reported in the latest release of the GWAS Catalogue [514], as of August 2016.

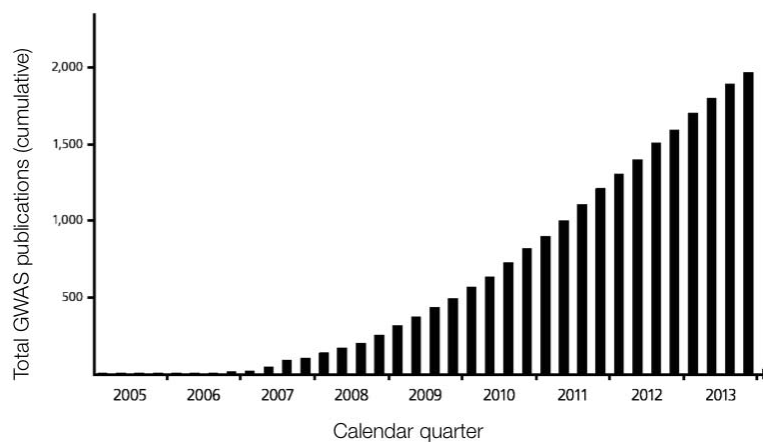


Figure 1.4 Number of genome-wide association studies published between 2005 and 2013. Image credit: Genome Research Limited.

In GWAS studies, allele (or genotype) frequencies at hundreds of thousands of SNPs are tested for association with disease status (**Figure 1.5**) or a quantitative trait value in thousands of individuals, usually under an additive genetic model. For quantitative traits (e.g. height), linear regression is used to test each SNP for association between trait value and genotype. For categorical traits (e.g. binary case/control or phenotypic

extremes), logistic regression is usually performed. The strength of the association is measured by the odds ratio (OR) or by the beta coefficient (β), depending whether the phenotype is binary or quantitative, respectively. The markers that show significant association with a disease or trait point to regions of the genome that are likely to harbour disease relevant genes. Because of LD however, associated SNPs do not represent causal variants *per se* and have yet to be dissected via subsequent fine-mapping strategies. These analyses aim to differentiate statistical signals at causal variants over their highly correlated neighbors, and usually involve a combination of statistical and functional analyses to narrow down the association signal to a single or very few variants [217, 456].

The first published GWAS, a study of age-related macular degeneration, identified a common variant association in the *CFH* locus that increased the risk of disease by a factor (OR) of ~ 7 [252]. Such large effects were soon recognised to be the exception rather than the rule. A landmark publication from the Wellcome Trust Case Control Consortium (WTCCC) in 2007 of a GWAS of 14,000 cases across seven diseases and 3000 shared controls [528] revealed most disease associations have in fact small effect sizes, typically between 1.1 and 1.4, such that the loci identified only explain a fraction of the estimated genetic component of disease risk [307].

Most of the quality control (QC) procedures that are now used in complex disease studies were also established by the WTCCC study, including several methods to identify poorly genotyped samples or markers, and protocols to deal with population stratification, a potential confounder in genetic studies that results from the fact SNP frequencies are variable across ethnic populations [18, 528]. The WTCCC also emphasised the importance of replicating association signals in an independent dataset and the use of stringent statistical criteria for declaring an association as genome-wide significant. The genome-wide significance threshold for association was set at $P < 5 \times 10^{-8}$ around this time. This roughly corresponds to a 5% type-I error rate when considering the number of independent SNPs tagged by common variants in the genome in individuals of European descent (~ 1 -2 million) [479].

To increase the overall sample size and statistical power of GWAS, many researchers subsequently embarked on large meta-analyses combining the results from individual studies. This approach essentially examines whether the observed effects at a given genomic region are consistent across studies, and whether the magnitude and direction of effects are also similar. Meta-analyses of GWAS studies, very often containing information from tens of thousands of individuals, were hugely successful at yielding novel

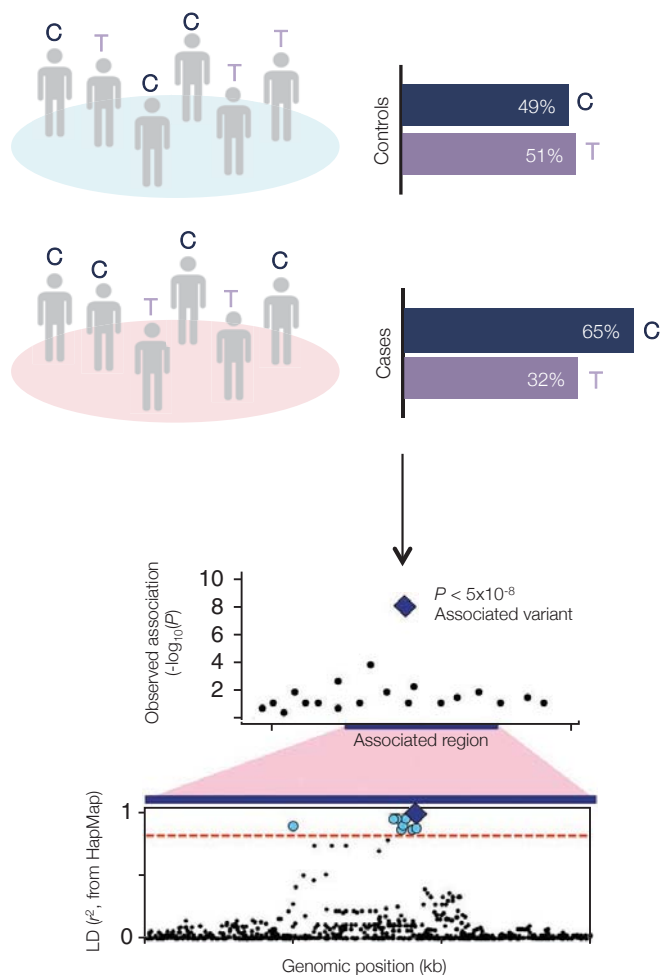


Figure 1.5 Schematic representation of a case-control (or binary) GWAS study.

In a case-control GWAS, a large cohort of diseased individuals (cases) and controls is genotyped for hundreds of thousands of SNPs spread throughout the genome. An associated region will often contain dozens of correlated SNPs in high LD with very similar association signals that, together, can span numerous genes. To narrow these multiple correlated signals down to a single or very few causal variants, researchers apply fine-mapping strategies. Such studies typically perform stepwise conditional analyses to identify independent signals within the associated regions. Statistical algorithms, in combination with functional genetic information (e.g. overlap with regulatory elements), can also be applied to assign posterior probabilities of causality to each candidate variant [217, 456].

disease-associations, and are still heavily used today. The story of inflammatory bowel disease (IBD, **Figure 1.6**) is a textbook example, where a total of four meta-analyses, conducted between 2008 to 2015, brought the number of loci from 21 (using 3,230 cases) to 231 (using 96,486 cases) [33, 153, 232, 290], ultimately yielding unprecedented insights into the biological mechanisms involved in IBD pathology (see Chapter 4).

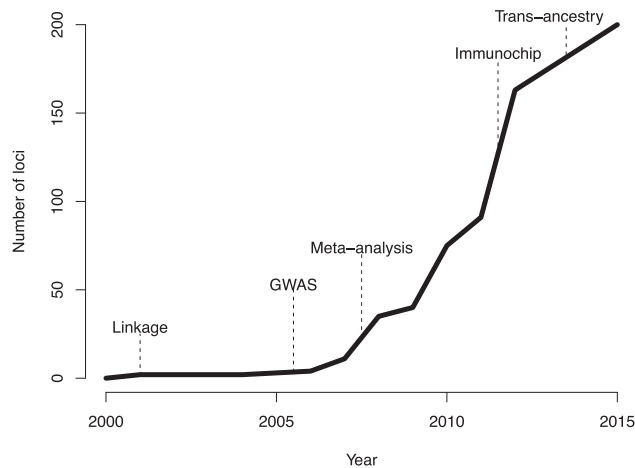


Figure 1.6 The number of IBD-associated loci identified using various study designs over the past fifteen years. Other than meta-analyses, IBD researchers also used a custom genotyping array (Immunochip) to aid replication and fine-mapping strategies, and to allow more cost-efficient genotyping in larger numbers of samples. The Immunochip contained a dense panel of 130,000 SNPs located in 186 regions known to be associated with one or more of 12 immune-related diseases, including IBD, autoimmune thyroid disease, ankylosing spondylitis, celiac disease, IgA deficiency, multiple sclerosis, primary biliary cirrhosis, psoriasis, rheumatoid arthritis, systemic lupus erythematosus and type 1 diabetes [375]. The latest GWAS meta-analysis, conducted by Liu *et al* in 2015 [289], also included individuals of non-European ancestry. Image taken from De Lange *et al* [109].

1.5 The next-generation sequencing revolution

The next big leap forward in human genetics was the arrival of massive parallel sequencing or "next-generation" technologies at the end of 2004. Before then, the sequencing field was dominated by Sanger sequencing, also known as "capillary sequencing" [221]. Also in 2004, the National Human Genome Research Institute (NHGRI) devised a 70 million dollar DNA sequencing initiative aimed at bringing the cost of sequencing

at a much lower cost. This gave birth to the terms we now routinely use of "targeted-sequencing" or "gene-panels" and "exome-sequencing". The "exome" comprises all the annotated protein-coding genes ($\sim 21,000$) and is equivalent to about 1% ($\sim 30\text{Mb}$) of the total genomic sequence [176].

Many personal genomes and exomes were fully sequenced by 2008 [79, 279, 296, 350, 381, 519], providing the first insights into the scale of variation within an individual's genome. Several lessons were learned with these studies, for example: 1) each individual differs from the reference genome at on average 3.5 million positions and contains ~ 1000 large ($>500\text{bp}$) copy-number-variants (CNVs) [176]; 2) most identified variants are common in the individual's population and are shared between continental populations [60]; and 3) individuals from older ethnic populations (e.g. Africa) show greater variation [321], consistent with the demographic history of the human species [29]. These and other subsequent studies [299, 531] also reported between 200-800 loss-of-function (LoF) variants (nonsense, frameshift and splice donor and acceptor sites) and many (13%) missense changes that were predicted to be damaging to proteins within one's genome, suggesting that healthy individuals do carry many gene-disrupting mutations despite not having disease. These observations have given us a glimpse of the likely complexity of the functional interpretation of sequencing data, and shaped many of the interpretation best-practices that we now follow in novel-gene discovery and in clinical diagnostics studies, i.e. the assessment of the background rate of a given class of variation in a particular gene in the general population.

Beyond personal genomes, the availability of sequencing technologies also meant that human variation of many types (single nucleotide variants (SNVs), small insertions and deletions (indels, below 50 base-pairs (bp)) and CNVs could also now be characterised in human populations. This was successfully accomplished by the 1000 Genomes Project (1KG), between 2007 and 2015, through low-coverage sequencing (2-4x) of 2,504 individuals from 26 populations [23]. This dataset is now considered the global reference for human variation, providing an unique insight into genetic variation at the population level. 1KG contains more than 38 million variants with $\geq 0.1\%$ frequency, which are now widely used in QC and variant filtering strategies in studies of Mendelian and complex diseases.

The first successful application of NGS for gene-mapping in a rare Mendelian disorder of unknown cause (Miller syndrome) was eventually published by Ng *et al* in 2010 [351]. The authors exome-sequenced four affected individuals from three independent kindreds and found compound heterozygous mutations in *DHODH* to be causal. This study

demonstrated that whole exome-sequencing (WES) is a powerful and cost-effective strategy to identify molecular defects underlying Mendelian diseases even without linkage or pedigree information, nor any biological information related to disease mechanism. Also importantly, this report showed that WES makes tractable those conditions that are too rare and in which appropriately sized families are not available for linkage, illustrating the power of this approach in situations where only small number of affected individuals are available for study.

Several other studies subsequently pioneered the application of NGS strategies (both exome and genome-sequencing) on a larger-scale by sequencing thousands of samples, and by focusing not only on Mendelian conditions but also on complex disorders and biomedically relevant quantitative traits. Two notable studies are the NIH Heart, Lung, Blood Institute GO Exome Sequencing (ESP) [476] and the UK10K [507] projects. The first study exome-sequenced 6,500 individuals to identify risk alleles associated with heart, lung and blood disorders. The latter study conducted low-coverage (7x) whole-genome sequencing (WGS) to assess the contribution of genetic variation to more than 50 cardiometabolic and anthropometric traits in 3,781 healthy individuals. In addition, the UK10K also embarked on high-depth ($\sim 80x$) WES and targeted-sequencing of specific genes, to identify causal mutations for $\sim 6,000$ individuals from three different collections (rare diseases, severe obesity and neurodevelopmental disorders). Some of the datasets analysed in Chapters 2 and 3 of this dissertation were generated within the rare-disease initiative of UK10K.

NGS technologies have enabled researchers to obtain variant information to the resolution of single-bases in a quick, high-throughput way, scalable to the size of the human genome. This has been revolutionary to both Mendelian and complex disorders for distinct reasons. For Mendelian diseases, NGS has finally enabled researchers to investigate conditions that were challenging to study before, such as sporadic and clinically heterogeneous disorders. Intellectual disability (ID) and neurodevelopmental disorders are examples of two broad category of heterogeneous conditions that have benefited tremendously from NGS [150, 170, 527], with more than 25 novel genes causative of ID discovered through exome-sequencing [408]. Combined with traditional genetic approaches including linkage, array comparative genomic hybridization and candidate gene-sequencing, WES and WGS have dramatically accelerated the pace at which novel genes are being linked to Mendelian phenotypes [77]. This has increased from a mean of ~ 166 per year between 2005-2009 to ~ 236 between 2010-2014, and this rate of progress shows no signs of abating as yet [77]. High-throughput sequencing now permits the genome or exome-wide identification of inherited, *de novo* and CNV

events within families and their subsequent joint analysis in a matter of weeks rather than years. Besides speeding up gene discoveries, NGS has been shown to dramatically decrease the length of the "diagnostic odyssey", i.e. the medical journey travelled by patients and their families from the onset of disease symptoms to a conclusive diagnosis. Multiple nation-wide and large-scale studies such as the FORGE (Finding of Rare Disease Genes) Canada Consortium [38], the DDD (Deciphering Developmental Disorders) [527], the UK10K [507] and many others [77, 535, 544], have demonstrated this benefit, with all studies providing genetic diagnoses in substantially less time than the usual time frame of around one decade [38].

For complex disorders, NGS has finally enabled researchers to search for low-frequency (1%-5%) and rare variants (<1%) underlying disease, rather than focusing solely on common-frequency alleles. It has long been hypothesised that rare variants are likely to play an important role in complex disease [401]. Loci that are associated with complex disease are enriched for rare variants that cause known Mendelian disorders, and it has been suggested that recessive variants confer risk to related complex diseases when the carrier is heterozygous [49]. Until recently, it had been unfeasible to explore the role of rare and low-frequency variation to complex disease genome-wide, because such variants were not represented in GWAS studies due to poor LD tagging by nearby SNPs [14]. NGS has now brought variants of all frequencies into view, meaning researchers can now more fully evaluate the spectrum of potential effects exerted by genetic variation. NGS-based studies of complex diseases have already yielded some fruitful results: studies such as the ESP, UK10K and many others [269, 403, 462, 476, 507] have already reported rare and low-frequency associations for many complex disorders and traits. Notable examples include *ADIPOQ* for adiponectin levels [507], *APOC3* for triglycerides and coronary heart disease [476], *PNPLA5* for low-density cholesterol [269] and *CCND2* for type 2 diabetes [462]. The most recent example [295] was the identification of a rare variant (0.6%) in *ADCY7* that doubles the risk of ulcerative colitis (UC). This association was detected after WGS of 4,280 cases and 3,652 population controls and is now the second strongest susceptibility-locus for UC after the *HLA*. One major benefit of detecting lower-frequency variants in complex disease is that fine-mapping may be easier, as such variants are correlated with fewer nearby SNPs. In addition, because rare alleles often have a direct functional impact at the protein level (if coding), they can be more straightforwardly transferred to cellular and animal models for mechanistic studies of disease [13], ultimately providing quicker insights into disease pathogenesis.

1.6 A standard NGS workflow

A standard NGS data-generation pipeline is composed of several steps that can be conceptualised as laboratory- or computational-based. Each one of the steps addresses a specific task that is needed to transform the raw sequencing data into meaningful information that can then be used by geneticists in downstream genetic analyses.

The laboratory steps start with genomic DNA being extracted from blood or saliva and then checked for high quality. The sequencing library is then created, i.e. the DNA is fragmented into smaller fragments of homogeneous length and linked to adaptors. Specific parts of the genome are then captured using predefined baits/probes of certain bp length, if conducting targeted- or exome-sequencing. Finally, this pulled-down library, or the whole-genome instead, is sequenced usually by indexing and pooling multiple samples over the same sequencing lane.

The several computational-steps that follow illustrate the complexity of the NGS data (**Figure 1.8**). This has meant that, in parallel to the development of the technology itself, the field of bioinformatics has become central and an invaluable discipline to NGS-based studies. It has developed multiple solutions and tools to store, process, maintain and to aid in the interpretation of the massive amount of data generated by the sequencing machines [278, 361]. Many of these tools (e.g. SAMtools [281], VCFtools [105]) had just finished being developed when I started my PhD studies back in 2012, others (e.g. VQSR, HaplotypeCaller [116]) were subsequently developed in the following years.

1.6.1 Sequence generation

The first computational-step entails the conversion of the raw data (fluorescent signal) into nucleotide bases with corresponding quality scores, and then the conversion into short sequencing reads. This process is termed as "base calling" and occurs on-board the sequencing machine, with the output being stored in a "FASTQ" file format. The base quality scores are useful to optimise downstream read-mapping and variant calling.

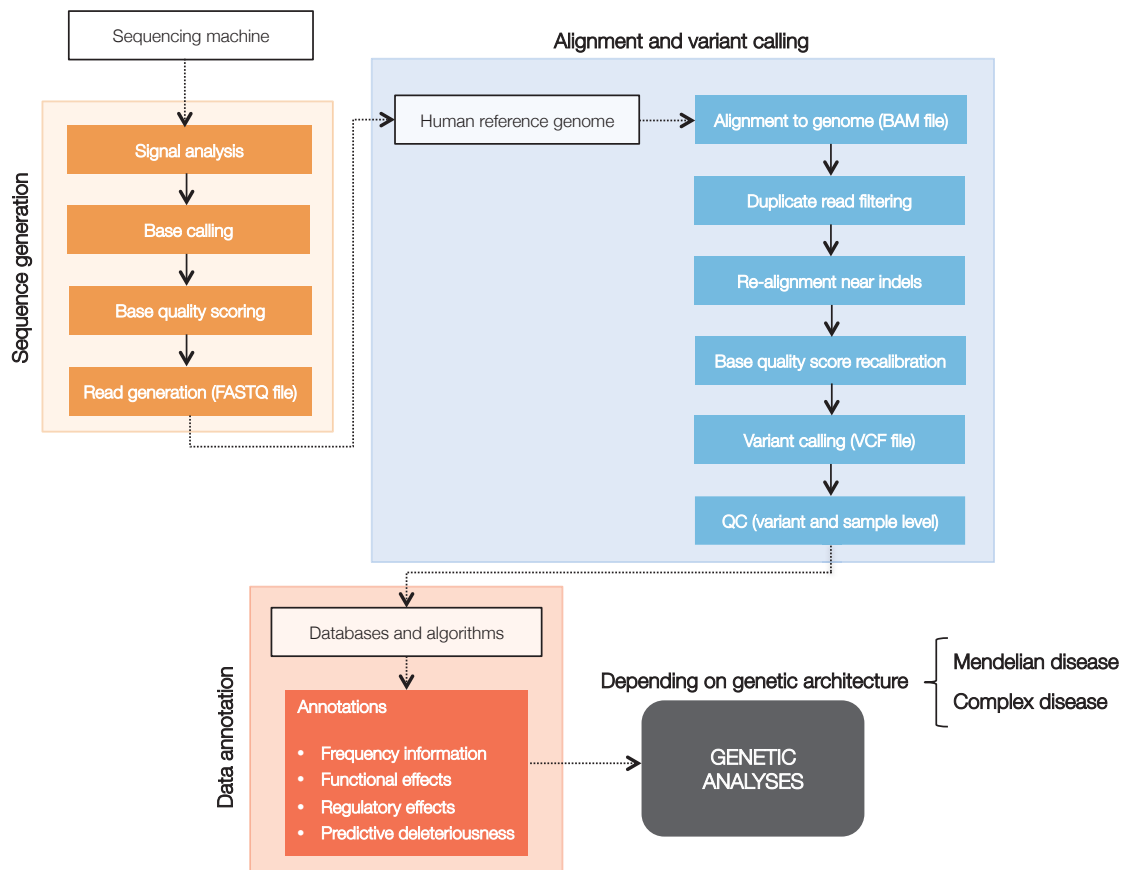


Figure 1.8 Flow diagram of the major computational steps involved in NGS data generation. The first step occurs inside the sequencing machine and involves the conversion of the raw imaging signal into sequencing reads. The second step is the alignment of the reads to the human genome, followed by several quality control procedures and variant calling. The third and final before downstream genetic analysis entails the annotation of the variant calls against allele frequency databases, functional and regulatory annotations, and predictive deleteriousness tools (e.g. PolyPhen2 [4], SIFT [349], GWAVA [424] and CADD [251]). All of these annotations are crucial for further genetic analyses, which vary depending on the genetic architecture of the disease under study. Image adapter from Oliver *et al* [361].

1.6.2 Alignment and variant calling

The next step is the alignment of reads to a reference genome (e.g. GRCh37, Genome Reference Consortium human build 37) and there are many tools to achieve this, with BWA being the most common [280]. Once the reads have been aligned, refinement steps are often performed, including the removal of duplicate reads (likely to be PCR artefacts), the re-alignment of reads around putative indels (to mitigate wrong alignments at the end of reads) and re-calibration of base quality scores (to correct for over- or under-estimated base quality scores). After alignment, reads are stored in BAM files, which can be the input to many read visualisation tools (e.g. Integrative Genomics Viewer [484]) for further judgement of putative variants directly from their reads.

Variant calling is then performed by identifying (or "calling") the positions (or "variants") of the sequenced reads that differ from the reference sequence. Depending on the application, this is done at the level of the genome, exome or specific genes, with all variants being stored in an easily accessible and readable VCF file. The calling itself depends heavily on accurate mapping to the reference genome and is accomplished using statistical modelling techniques that have been refined throughout the years to better distinguish genuine variation from sequencing errors [355]. One of such improvements was the incorporation of the degree of uncertainty when calling a genotype at a given position, rather than simply determining the genotype based on the effective counts of the alternative allele, i.e. the allele that did not match the one recorded in the reference. There are more than 60 different callers available to date (reviewed in [369]); which caller to use depends on the type of variation one aims to detect, i.e. SNVs/indels/*de novo*/CNVs. SAMtools [281] and GATK HaplotypeCaller [116] are the best established tools for SNV and indel calling. *De novo* and CNVs each have dedicated callers (see Chapter 3).

NGS provides a large amount of data with associated error rates ($\sim 0.1\text{-}15\%$) that are higher than those of traditional Sanger sequencing machines [177]. Moreover, there are many more sources of artefact and technical variation in NGS than in genotyping technologies, given the multiple preparation steps involved in a sequencing run. This problem is usually attenuated by sequencing at high depth, by performing variant-calling across all study samples [76], and by investing considerable amounts of time in downstream QC of variants and samples. Variant-QC steps can be performed either by using empirical thresholds derived from visualising the patterns of the data, by applying specific thresholds recommended by the variant calling software, or by using more

sophisticated statistical approaches (e.g. VQSR) [116]. The definition and the rationale for using many of these QC procedures are described within each of my thesis chapters. Also importantly, NGS technologies suffer from platform-specific error profiles [343]. If available, further analyses should take control sequences generated by the same lab into account, to successfully identify and remove systematic sequencing errors [474].

1.6.3 Data annotation

The number of variants identified through NGS strategies varies depending on many factors, such as the size of the sequenced regions, i.e. gene-panels/exome/genome, the ethnicity of the samples, the depth of sequencing coverage, etc [1]. In general, the number can range from 10,000-50,000 variants to four million variants in deep whole-genome sequences [158, 476, 507]. While these numbers certainly represent a challenge in interpretation, they are necessary to allow us to extract statistically robust and meaningful biological information from the data itself, and to engage in "data-driven" genetic hypotheses. Several biological annotations are normally added at this stage to facilitate downstream genetic analyses.

The first level of annotations is population-based allele frequencies for each alternative allele. Sources of frequency-based annotation include the HapMap [154], the 1KG [23], the ESP [476], the UK10K [507] and, more recently, the ExAC dataset [135]. The latter, only released two years ago, is the largest of all these datasets, consisting of variant calls from 60,706 exomes of different ethnicities, and has been especially developed to help prioritise variants in Mendelian diseases.

Functional-based annotations then assign the effect of a variant on the transcript(s) and encoded protein(s), based on the resulting amino acid change, and the effect is normally categorised into well-defined terms (**Figure 1.9**). Two tools commonly used for this purpose are the Ensembl VEP [322] and SnpEff [80]. Annotation of non-coding variants can be done using data from the ENCODE [478], Roadmap Epigenomics [425] and FANTOM5 [151] projects, all of which used applications of NGS such as ChIP-sequencing (chromatin immunoprecipitation assays), DNase I hypersensitive site mapping and CAGE (cap analysis of gene expression) to identify gene regulatory regions such as promoters, enhancers and transcription factor-binding sites in a variety of human cell and tissue types.

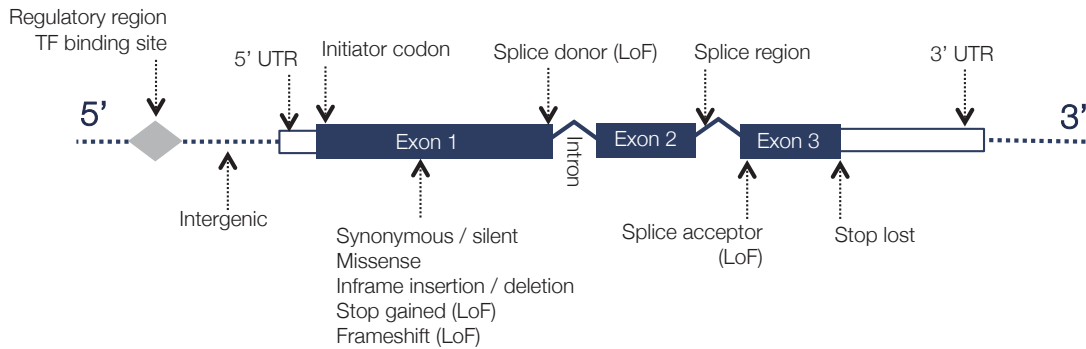


Figure 1.9 The impact of variants at the protein level. The diagram illustrates the set of functional consequence terms given by the Ensembl Variant Effect Predictor (VEP) tool [322]. A splice donor is splice variant that changes the invariable 2-base region at the 5' end of an intron. A splice acceptor is a splice variant that changes the invariable 2-base region at the 3' end of an intron. A splice region is a sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron, but not at the donor/acceptor splice sites. For a detailed description of each term see http://www.ensembl.org/info/genome/variation/predicted_data.html.

The final step before embarking on downstream genetic analyses is the use of prediction-based annotations which are added to infer the deleteriousness of missense changes on the resulting protein. This is done using computational tools that take into account the nucleotide and/or amino acid changes in combination with either: 1) sequence conservation within homologous sequences (e.g. SIFT [349] and GERP [107]), or 2) structural properties, such as the impact on the tri-dimensional protein model (e.g. PolyPhen2 [4]) [326]. The impact of splice donor and acceptor variants can be assessed using MaxEntScan [285], for example. Prediction for non-coding variants can also be done using recently developed tools such as GWAVA [424] or CADD [251], both of which use machine-learning algorithms trained with annotations from multiple sources of genomic, regulatory, functional and conservation data.

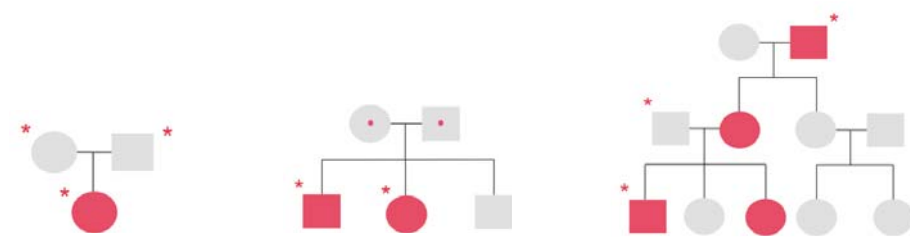
1.7 NGS genetic analyses in Mendelian diseases

WES at high coverage (60x-80x) is currently the most popular NGS approach for discovering genes underlying Mendelian diseases in research settings. Examining only the exonic portion of the genome is justified on the basis that the vast majority of Mendelian disease-associated mutations identified by linkage strategies result in the disruption of the protein-coding sequence [13].

Genetic studies of Mendelian diseases generally use family-based designs. A range of different pedigree structures can be used including, trios, affected sib-pairs or more distant relatives (e.g. cousins) or even larger pedigrees with multiple affected individuals. The design that is most useful depends on several factors including the known (or presumed) mode of inheritance of the disease under study, whether the disease is inherited or predicted to occur sporadically (i.e. parents often not affected) and also on the number of patients that can be sequenced in the study. Each pedigree structure has its advantages and disadvantages, both in terms of the feasibility of sample collection and the types of analytical approaches they allow to be explored (**Table 1.1**). The trio design is especially useful for sporadic diseases and when a dominant mode of inheritance and/or locus heterogeneity are suspected [169, 396]. In any case, the use of biological relatives is very valuable in the interpretation of genetic variation because it helps to identify neutral alleles, substantially narrowing down the search space for causative genes segregating within families [31].

The following assumptions are generally made when searching for causative mutations underlying simple, monogenic, Mendelian diseases: 1) a single mutation is sufficient to cause disease, 2) the mutation is coding and affects the function of the protein, 3) the allele is rare and probably private to the affected individual or family, 4) every carrier of the putative causative variant has the phenotype (complete penetrance), 5) every affected individual will carry the putative causative variants (complete detectance) and 6) the mutation is present in the same gene as in other unrelated affected individuals (genetic homogeneity) [464]. As such, when sifting through the data, researchers disregard variants located outside coding regions, silent amino acid changes and variants that are present in public variation datasets (e.g. 1KG, HapMap, ESP, UK10K, ExAC) and in internal control sequences at greater frequency than the expected carrier frequency [278]. Researchers then focus on variants that segregate with disease status within the pedigree and normally prioritise impactful variants (e.g. LoF and missense predicted to be damaging by *in silico* prediction tools) that occur in genes whose function is relevant for the disease [300].

Functional follow-up approaches of identified variants are then often conducted to confirm experimentally that the putative variant is detrimental to gene function. Examples of such approaches include *in silico* experiments such as computational modelling of the effect of a variant on the structure of a protein [65], *in vitro* investigation of the effect of the variant in patient cells [43], and *in vivo* investigations such as recapitulation of aspects of patient's phenotypes in animal models [483], which can ultimately inform about the biological mechanisms underlying disease pathogenesis.



Pedigree structure	TRIOS	AFFECTED SIB-PAIRS	MULTIPLEX FAMILIES
Well suited for	Autosomal dominant disorders	Autosomal recessive disorders	Autosomal dominant, recessive and X-linked disorders
Advantages	<i>De novo</i> and compound heterozygote variants can be identified	Few co-segregating rare homozygous variants shared by all affected sibs	Combine the power of both trios and affected sib-pairs designs Very small search space for causative variants
Disadvantages	Fewer patients sequenced if budget is limited	Further segregation analysis in parents and unaffected sibs needed Compound heterozygous variants cannot be identified	Difficult to collect Difficult to analyse if affected members have heterogeneous phenotypes
Analytical approaches	Identify <i>de novo</i> events (SNVs and CNVs): more likely in sporadic disorders Identify compound heterozygous: more likely in non consanguineous background Identify homozygous variants: more likely in consanguineous background Transmission-disequilibrium test (TDT)	Identify homozygous variants or putative compound heterozygotes shared by affected sibs Runs-of-homozygosity analysis Identical-by-descent analysis	Identify heterozygous or homozygous variants shared by affected relatives Linkage analysis (if pedigree is large enough)
Examples	Weaver syndrome (EZH2)	Postaxial polydactyly type 4 (ZNF141)	Familial diarrhea syndrome (GUCY2C)

Table 1.1 Overview of three possible family-based study designs used in NGS-based studies of Mendelian conditions. The table lists the advantages and disadvantages of each pedigree structure and provides examples of monogenic conditions that were successfully investigated using the corresponding study design. The analytical approaches to narrow down the search space for causative variants in NGS studies are also provided. If desired, traditional gene-mapping techniques (in pink) can also be used in combination with the NGS data, which can greatly increase power. Asterisks represent sequenced individuals.

Given the dramatic increase in novel-gene discoveries since NGS became available, there has been much discussion surrounding the exact extent and nature of the evidence that is required in order to state that a given gene is indeed causative, or associated, with a Mendelian disorder. Keeping with the history of the field of human genetics, the importance of a consistent and rigorous approach has been increasingly recognised, and a set of guidelines for this purpose was published in 2014 [300]. It is now clear that the identification of a single variant (even if LoF) segregating with disease in a single family is not on its own sufficient evidence that the allele is causative of disease. Therefore, observations in the same gene in additional individuals or families with similar phenotype should be accumulated and, more importantly, statistical support for the findings should be demonstrated. There is no one rule as to the number of independent individuals or families that are required to statistically demonstrate that the occurrence of a particular number of variants in a given gene is highly unlikely to have occurred by chance. Instead, the number required depends on several factors such as the size of the gene, its mutation rate, and how tolerant the gene is to the observed class of variation (e.g. missense or LoF) [300, 425]. A commonly used statistical approach to derive significance is to compare the number of cases that carry variants in a particular gene with that observed for a large cohort of controls using the Fisher's exact test [11, 160]. In principle, a novel gene can then be declared causative if its P -value surpasses the exome-wide significance level of 1.7×10^{-6} [300], corresponding to the Bonferroni corrected P -value for performing tests on $\sim 21,000$ protein-coding genes and $\sim 9,000$ long non-coding RNA genes [117, 195]. Such statistical analyses were made possible with the increasing availability of large-scale sequencing data that can be used as control sequences. This also now allows genome-scale approaches to gene discovery, in which the distribution of rare, predicted-damaging variants in cases is systematically compared to population controls to identify genes with an excess of potentially pathogenic variants for functional follow-up.

One should be mindful of potential technical differences existing between the two groups when performing case-control analyses, because any baseline differences can yield false-positive association signals that are not due to a biological reason but to technical artefact. Two possible confounders are population stratification and sequencing depth, both of which are usually correlated with the number of variants called within a sample, and even more so at rare or private sites [300, 314]. As such, the appropriate control group to use in such tests should be drawn from the same (or close) ethnicity as cases, its data should have been generated and analysed in similar fashion and QC checks should be conducted to ensure there are no discrepancies between the two groups.

Other than family-based designs, case-control enrichment strategies are increasingly being used in disease studies as they can often provide important insights into the aetiology of disease, especially when genetic heterogeneity is expected [407]. In such an approach, a cohort of unrelated cases is sequenced along with a large cohort of controls. Rare variants are then identified in both groups and a statistical test is applied to test the hypothesis that the cases have an excess of a defined category of variants (e.g. LoF) compared to controls. This can be performed at various testing units including, for example, assembled lists of candidate or biologically related genes (termed as "genesets"), biological pathways or even the whole exome. Ultimately, this approach is useful because it can highlight whether a specific category of variants and particular genes are important to disease pathogenesis, therefore providing insights into the genetic architecture of disease without necessarily assigning causality to individual alleles and genes [210, 407]. This can be viewed as a "top-down" approach, where one focus on identifying the overall rates of mutation, before proceeding to map particular disease-associated genes. Importantly, these enrichment analyses make fewer assumptions about causative variants than classical family-based approaches, and therefore take into account non-classical contributors to disease such as variants with incomplete penetrance, and variants that contribute to a phenotype in an oligogenic manner [335].

Several distinct statistical tests have been developed for use in rare-variant case-control enrichment analyses (reviewed in [276]), all of which evaluate the aggregate effects of multiple genetic variants in a testing unit. Four of the most commonly used tests are the cohort allelic sums test (CAST) [335], the BURDEN test [407], the weighted sum statistic [302], and the sequence kernel association test (SKAT) [529]. All of these tests have been developed with complex disease in mind, but the first two are often used in rare and Mendelian studies as well [103, 184, 407] since their underlying assumptions are appropriate: they both consider that all rare variants have the same direction of effect (e.g. all variants are disease-causing) and that the effects of the rare variants are all similar (e.g. all alleles exert large effects on the phenotype). The main difference between CAST and BURDEN is that the first one counts how many cases and controls have at least one alternative allele in a given region, while the second counts the exact number of alternative alleles per individual in a given region, summed for all cases and controls [276]. There is no one rule as to which category of variants to test in such analyses, therefore, researchers normally run tests for a series of increasingly rare allele frequency thresholds and also for different classes of mutations, e.g. all functional variants or just LoF [184, 407].

Several studies demonstrate the utility of case-control enrichment analysis in providing important insights into possible disease pathological mechanisms. In an early example, Purcell *et al* used the BURDEN test in an exome analysis of 2,536 schizophrenia cases and 2,543 controls and detected an enrichment of rare disruptive mutations in calcium channels and in components of the postsynaptic activity-regulated cytoskeleton (ARC) complex, emphasising their importance in the aetiology of schizophrenia [407]. Another study used the CAST test in a cohort of 986 individuals with ID and 903 controls that were targeted-sequenced for a panel of 565 known and candidate genes. Apart from an enrichment of LoF variants in known ID-associated genes, the authors also observed an enrichment in candidate genes, suggesting some of these may indeed be real causative genes but that have yet to be definitively proved as such [184]. D'Alessandro *et al* [103] exome-sequenced 81 patients with atrioventricular septal defects (AVSD) and used the 6,500 ESP exomes as controls. Using the CAST method, the authors reported a significant enrichment of rare missense damaging variants in 112 genes with biological associations to AVSD. Some of these genes included syndrome-associated genes, suggesting these can contribute to AVSD even in patients with isolated heart defects. On a different perspective, a targeted-sequencing of 44 candidate genes in 2,446 autism patients identified one *de novo* LoF mutation in *ADNP*, a candidate gene for autism [426]. Because this gene was part of a protein-protein interaction pathway that previously showed enrichment for *de novo* variants in autism in an earlier study [366], the authors embarked on further targeted resequencing experiments and identified several more cases with *de novo* mutations in *ADNP* [200]. This example illustrates how case-control enrichment analyses can also inform and drive novel gene discoveries.

1.8 NGS genetic analyses in complex diseases

Next-generation sequencing makes possible to study the low frequency and rare variants not covered by the GWAS approach. However, despite rapidly decreasing costs, it is still prohibitively expensive to deploy NGS on a scale similar to existing GWAS. The most important determinant of GWAS success has been the ability to analyse tens of thousands of individuals, and detecting rare variant associations will require even larger sample sizes, because the minor allele of a given rare variant is observed so infrequently [546]. The fundamental question that therefore arises when designing a NGS-based study for a complex disease is how to most efficiently distribute sequencing

reads across the genome and across individuals [295]. To maximise the number of individuals that can be sequenced, some researchers use exome-sequencing, which is relatively low cost [123, 245]. However, a major disadvantage of WES is that it only surveys coding variation, and results from GWAS have shown that the substantially majority ($\sim 92\%$) of complex disease associated variants lie in non-coding, presumed regulatory, regions of the genome [13, 288, 514]. An alternative approach is to use low coverage ($<10\times$) WGS, which captures this important non-coding variation and is cheap enough to enable thousands of individuals to be sequenced. This approach has already proven valuable in exploring rarer variants than those accessible in GWAS studies [95, 106, 123]. In addition, low-coverage WGS has been shown to maximise both cost and statistical power when budget is limited [283], meaning sequencing more individuals at lower depth is preferable to sequencing fewer samples at higher coverage.

Low-coverage WGS studies can be boosted further by using the dense genotype panel achieved with the low-coverage WGS as a reference panel to impute (or "predict" statistically) the genotypes of additional individuals genotyped in parallel on GWAS arrays (**Figure 1.10**). Briefly, imputation methods identify stretches of haplotypes that are shared between the study individuals (in this case the genotyped samples) and the haplotypes of a reference panel, and use those matching haplotypes to impute the missing alleles in study individuals [309]. Because the imputation of low-frequency and rare variants is more challenging compared with common alleles, the further use of very large-scale reference datasets (e.g. 1KG and UK10K) as reference panels, can greatly improve imputation performance at those sites [371]. This study design therefore allows researchers to infer genotypes in enough samples to test lower frequency variants genome-wide at approximately the same cost of WES. This approach has been successfully used in IBD [295], type 2 diabetes [156, 462], sick sinus syndrome [211] and in the UK10K study [507].

Downstream genetic analyses will often include single-point analysis, similarly to a standard GWAS. In this case, researchers often include variants with a lower bound frequency of 0.1%, 0.5% or 1%, depending on sample size, below which single-variant analysis is no longer well powered [295, 507]. The effect of rarer alleles, including those that are "private" to single individuals, can be tested in aggregate using collapsing tests such as the weighted sum method and SKAT. These two tests differ in the way variants are weighted and whether they incorporate alleles with opposite direction of effects, i.e. risk increasing/decreasing. Such enrichment analyses can be done at the level of genes, regulatory regions (e.g. promoters and enhancers) or even within genome-wide windows, therefore elucidating the aggregate impact of rare variation.

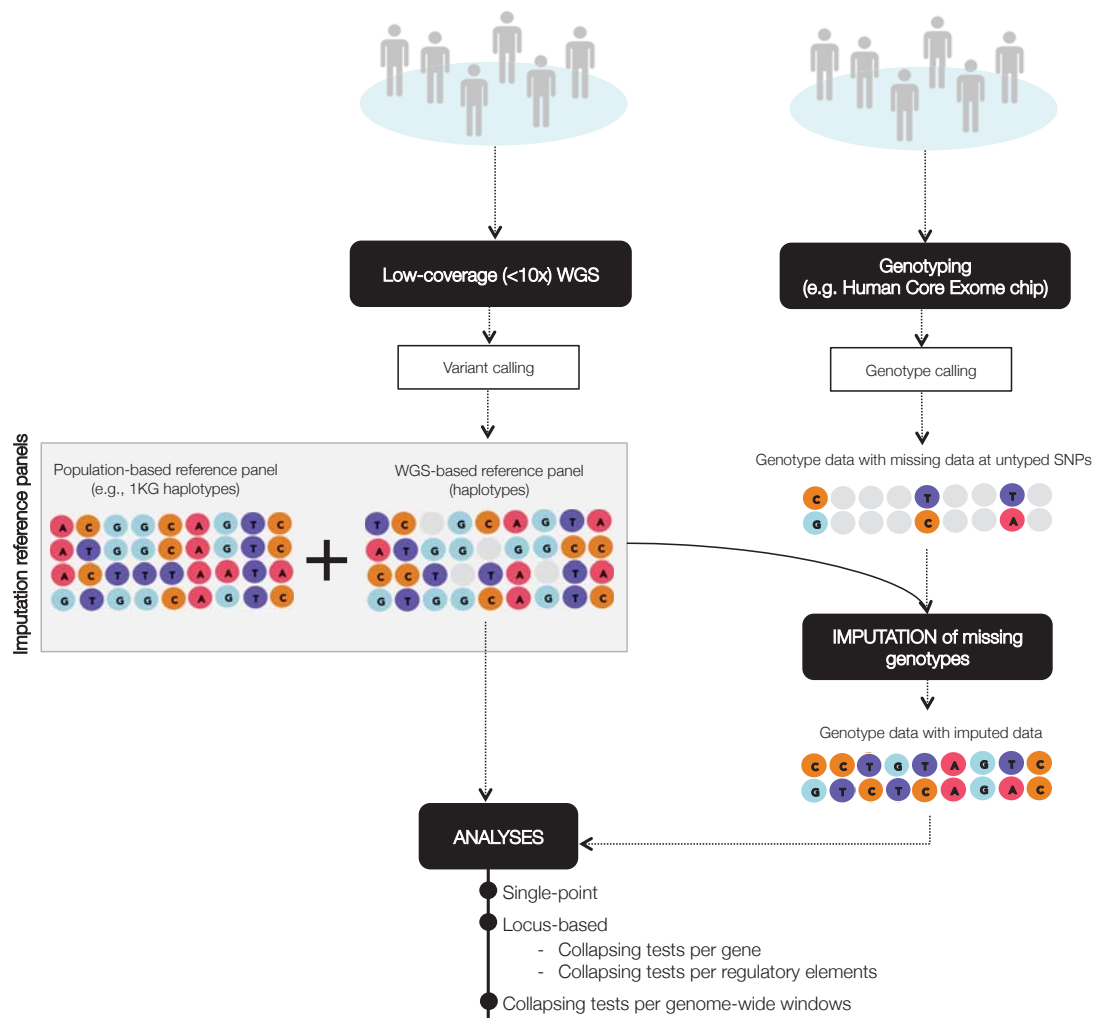


Figure 1.10 Diagram illustrating a popular study design now used in NGS-based complex disease studies. Low-coverage whole-exome sequencing is performed in as many cases and controls as possible. In parallel, additional cases and controls are genotyped for an ordinary GWAS array. The low-coverage sequences can then be combined with additional population-based reference panels and haplotypes can be generated. Through an imputation process, those haplotypes can be used to predict the genotype status of the genotyped samples at many sites that were not included in the genotyping array. The low-coverage sequencing and the boosted genotyping data can then be used together in downstream genetic analyses.

1.9 Outline of dissertation

In this dissertation I describe four distinct projects in which NGS technologies were employed, in combination with different study designs and analytical strategies, to identify genetic determinants, or modifiers, of human diseases that have not been extensively studied thus far. Because the projects are distinct, and encompass different phenotypes, the following four chapters are self-contained, and additional introductory material is located within each chapter.

The phenotype investigated in **Chapters 2 and 3** is congenital hypothyroidism (CH), a rare heterogeneous disease often caused by single-gene molecular defects that impair thyroid hormone production in a structurally normal thyroid gland ("*gland-in-situ*"), or that result in thyroid gland developmental abnormalities. The phenotype investigated in **Chapter 4** is very-early-onset inflammatory-bowel-disease (VEO-IBD), currently viewed as a Mendelian form of inflammatory bowel disease (IBD), a complex disorder of adulthood onset. In **Chapter 5**, I move beyond clinical disease *per se*, and use the age at IBD diagnosis as a quantitative phenotype.

The aim of the project described in **Chapter 2** was to conduct, for the first time, a comprehensive NGS-based screening of all genes that are currently known to cause thyroid hormone production defects in a CH cohort with *gland-in-situ* (N=49 cases from 34 families). Genetic screening of such patients has been traditionally limited by the cost and labour implications of Sanger-sequencing multiple exons, meaning many cases still await an exact genetic diagnosis. I show how a stringent variant filtering pipeline, combined with pedigree segregation analyses and *in silico* (bioinformatic and structural) predictions of pathogenicity for candidate variants, led to the identification of likely causal mutations in 59% of the patients.

In **Chapter 3**, I describe a family-based NGS study in which exome and targeted-sequencing were used, for the first time, with the aim of identifying novel genetic causes of CH in a phenotypically heterogeneous CH cohort comprised of 48 families. Historically, this condition has been refractory to traditional gene-mapping techniques, meaning it is still poorly understood. I describe the strategies I applied to map *de novo*, inherited and CNV variation segregating with disease within CH pedigrees, and the statistical analysis conducted to conclude no gene was recurrently mutated in multiple families over what was expected by chance. I will then show how a candidate-focused approach successfully uncovered a putative novel CH-associated gene and identified

further defects that very likely account for the extrathyroidal abnormalities seen in two CH patients.

In **Chapter 4**, I describe an exome-sequencing analysis of 145 VEO-IBD cases and 3,969 controls. The overall aim of this project was to investigate the contribution of rare variants, as well as known, common-frequency IBD-risk alleles, to the pathogenesis of VEO-IBD. I describe the analysis that led to the identification of likely causal defects in primary immunodeficiency-associated genes in four patients, and show several case-control enrichment analyses I also performed, at the level of single-genes, genesets and pathways, to more fully investigate the burden of rare disrupting alleles operating in VEO-IBD. I then demonstrate how the use of polygenic risk scores, leveraging the set of IBD GWAS associations discovered to date, can provide further unprecedented insights into the genetic architecture of this disease.

In **Chapter 5**, I present a meta-analysis study in which low-coverage whole-genome sequencing data was combined, with three previously imputed GWAS studies, to identifying genetic modifiers of age at IBD diagnosis. Much is already known about the factors that contribute to IBD-risk, but our understanding of the genetic factors modifying the onset of disease lags behind.

Lastly, in **Chapter 6**, I highlight the major lessons learnt with these projects, discuss some immediate impact some of these results had for patients, and look forward to the future developments and the types of studies that will shape gene-mapping strategies over the next coming years.

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.

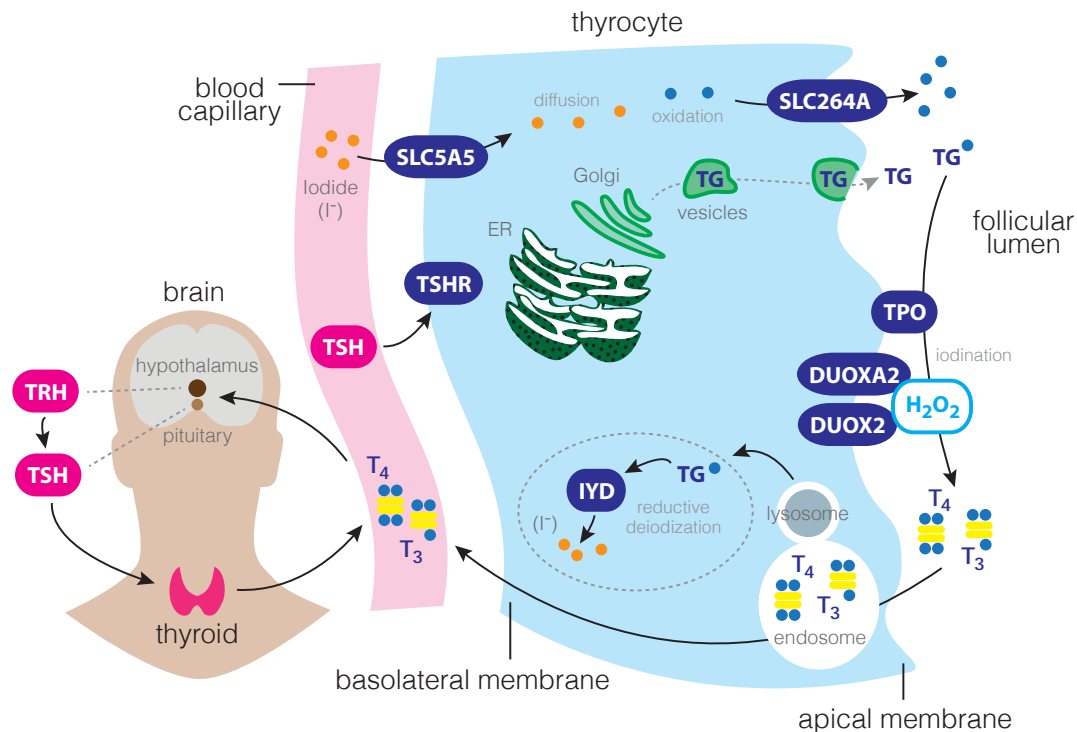


Figure 2.1 Key steps and players involved in thyroid hormone synthesis and regulation.

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 H₂O₂ produced by *DUOX2* and *DUOX2* 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
<i>TPO</i> *	thyroid peroxidase	iodination substrate for hormone synthesis	Biallelic	TIOD/PIOD	Severe/mild CH [171, 423]	P	CLO ₄ - discharge test
<i>TG</i> **	thyroglobulin	hormone synthesis	Biallelic	Absent/very low serum TG levels	Severe/mild CH [471]	P	Serum TG levels
<i>DUOX2</i>	dual oxidase 2	H ₂ O ₂ production	Monoallelic	PIOD	Severe/mild CH [178, 342, 502]	P/T	
<i>DUOX2</i>	dual oxidase 2	H ₂ O ₂ production	Biallelic	TIOD	Severe/mild CH [178, 342, 389, 502]	P/T	CLO ₄ - discharge test
<i>DUOX2</i>	DUOX maturation factor 2	H ₂ O ₂ production	Biallelic	PIOD	Mild CH [178, 537]	P	CLO ₄ - discharge test
<i>SLC5A5</i>	sodium iodide symporter	iodide uptake into the thyrocyte	Biallelic	Reduced thyroidal iodide uptake	Severe/mild CH [459]	P	Saliva/plasma RAI ratio
<i>SLC26A4</i>	pendrin	iodide efflux into follicular lumen	Biallelic	Sensorineural hearing loss***	Very mild CH [178]	P	MRI/CT of temporal bones
<i>TSHR</i>	TSH receptor	initiation of hormone synthesis cascade	Biallelic	TSH resistance	Severe CH [171]	P	
<i>IYD</i>	iodotyrosine deiodinase	intrathyroidal iodide recycling	Monoallelic	Excessive urinary excretion of iodine	Mild CH [178]	P	
			Biallelic		Mild CH	P	
					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; CLO₄:- perchlorate; 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 *DUOX2* 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 tyrosine-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 WTTSI 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		
MiSeq targeted sequencing (N = 21)	F34	
	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 (1-3 μ g) 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 manufacturer’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 ($A \leftrightarrow G$, $C \leftrightarrow T$) divided by the number of transversions SNVs ($A \leftrightarrow T$, $G \leftrightarrow C$, $A \leftrightarrow C$, $G \leftrightarrow T$). 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.

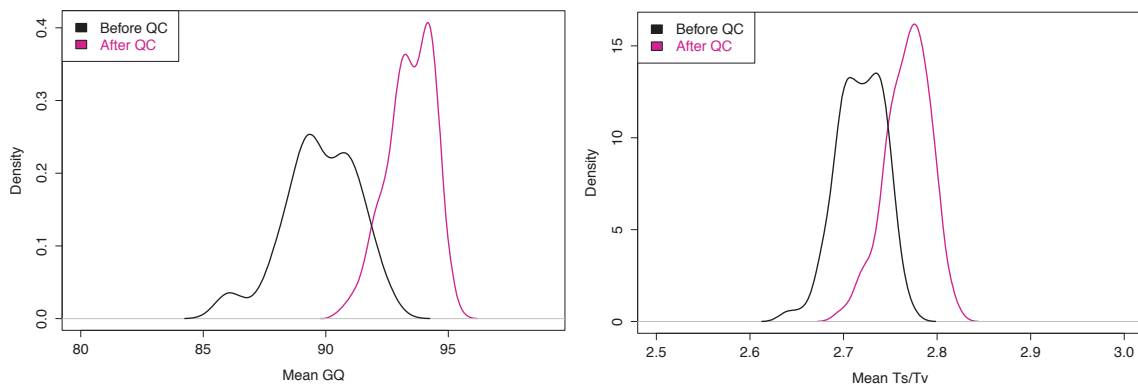


Figure 2.2 Distribution of mean genotype quality (GQ, left panel) and ratio of transitions to transversions (Ts/Tv, right panel) pre- and post- extra genotype-QC.

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/ μ g 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 SequelPrep 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].

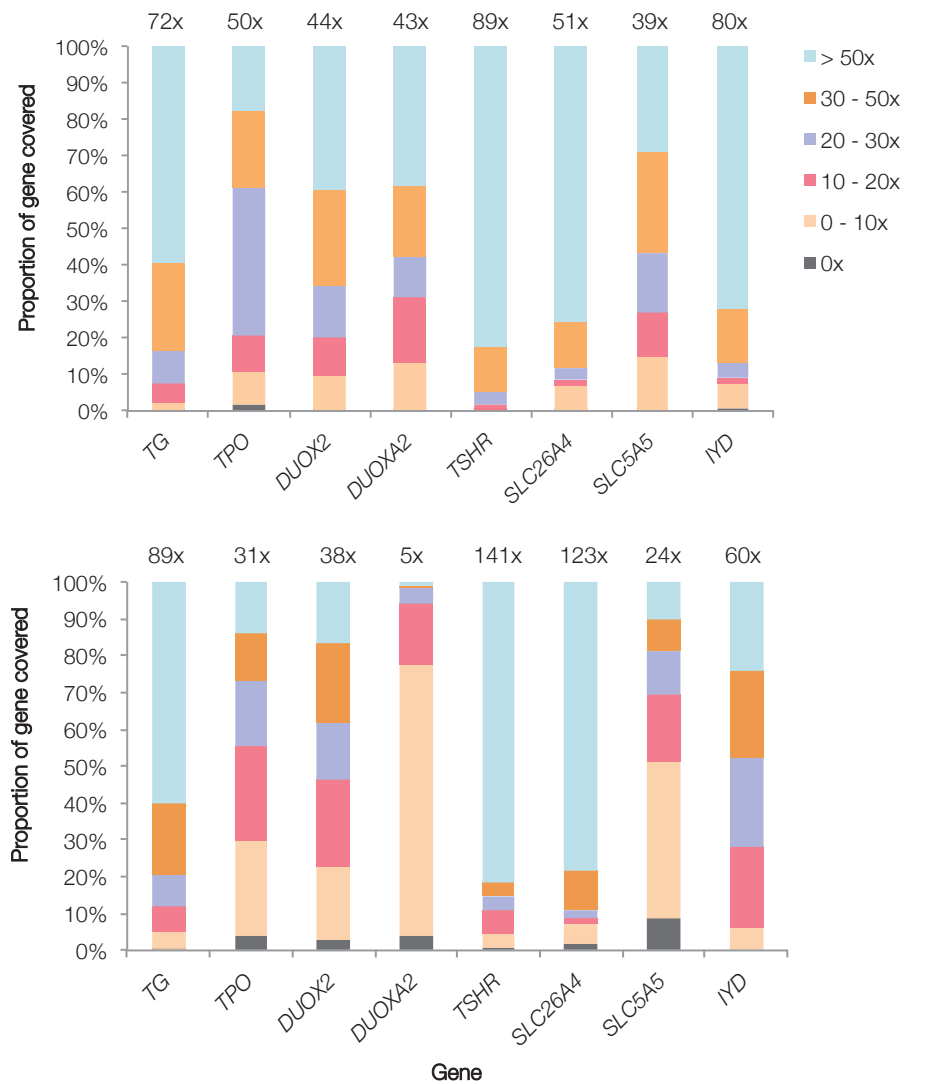


Figure 2.3 Proportion of gene sequence covered at various depth thresholds for the **A)** WES experiment and **B)** Hi-Seq targeted-sequencing experiment.

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.

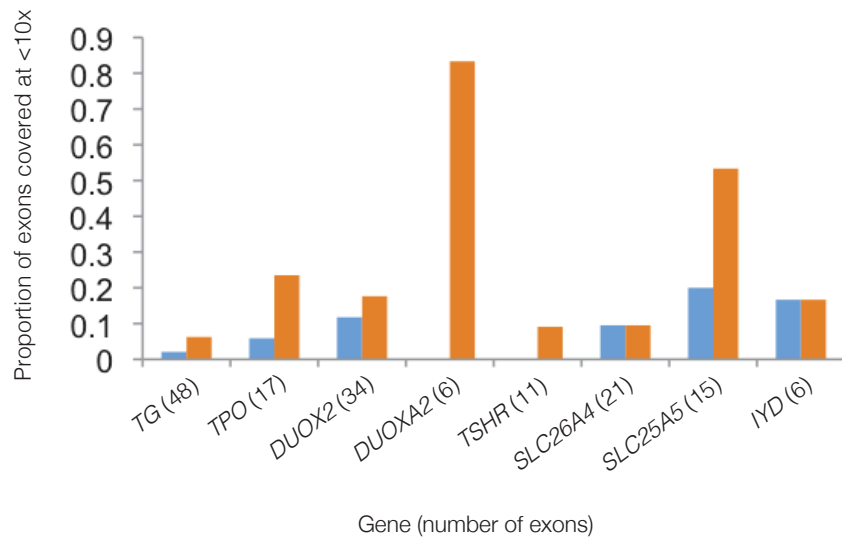


Figure 2.4 Proportion of exons with mean depth coverage <10x in the samples sequenced by WES (blue) and HiSeq targeted sequencing (orange).

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	.	.	5.84	0.000113 (UK10K cohorts)
	R296X	.	F5a,b	yes	type 1 repeat	T	PD	5.62	0.000362 (ExAC)
	R451X	.	F1a,b	yes	type 1 repeat	.	.	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)
	Y79C	.	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.3453+3_3453+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
	R684Q	.	F19a,b	.	.	D	PD	4.78	0.00023(ESP)
	R665Q	.	F17	yes	.	D	PD	4.84	0.000025 (ExAC)
	C808AfsX24	.	F20	.	.	D	PD	-2.18	0.00934 (ESP)
	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)
P68S	.	F26	yes	.	D	PD	0	0.0041 (1KG)	
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)	
.	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.

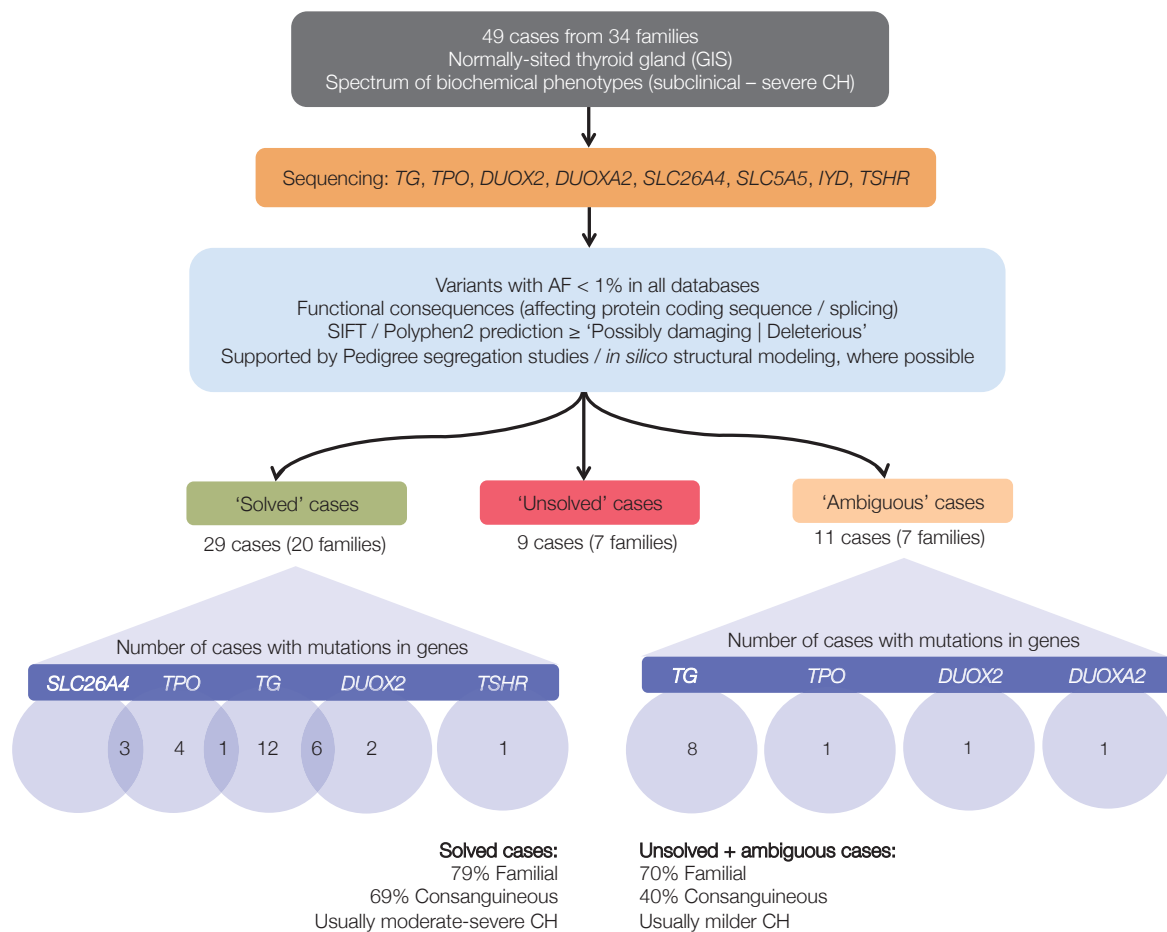


Figure 2.5 Schematic illustrating case selection, variant filtering and distribution of mutations in the GIS cohort studied.

'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).

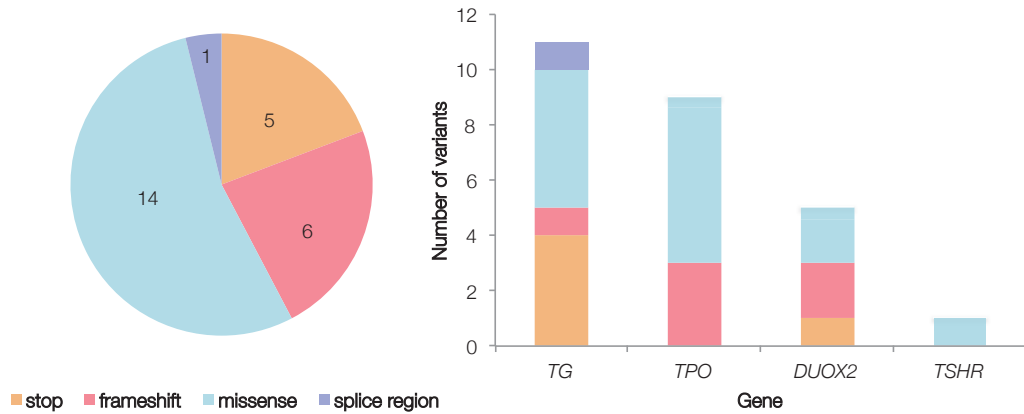


Figure 2.6 Causative variants identified in CH cohort with GIS.

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*, *DUOX2*, *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 homonogenic 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.

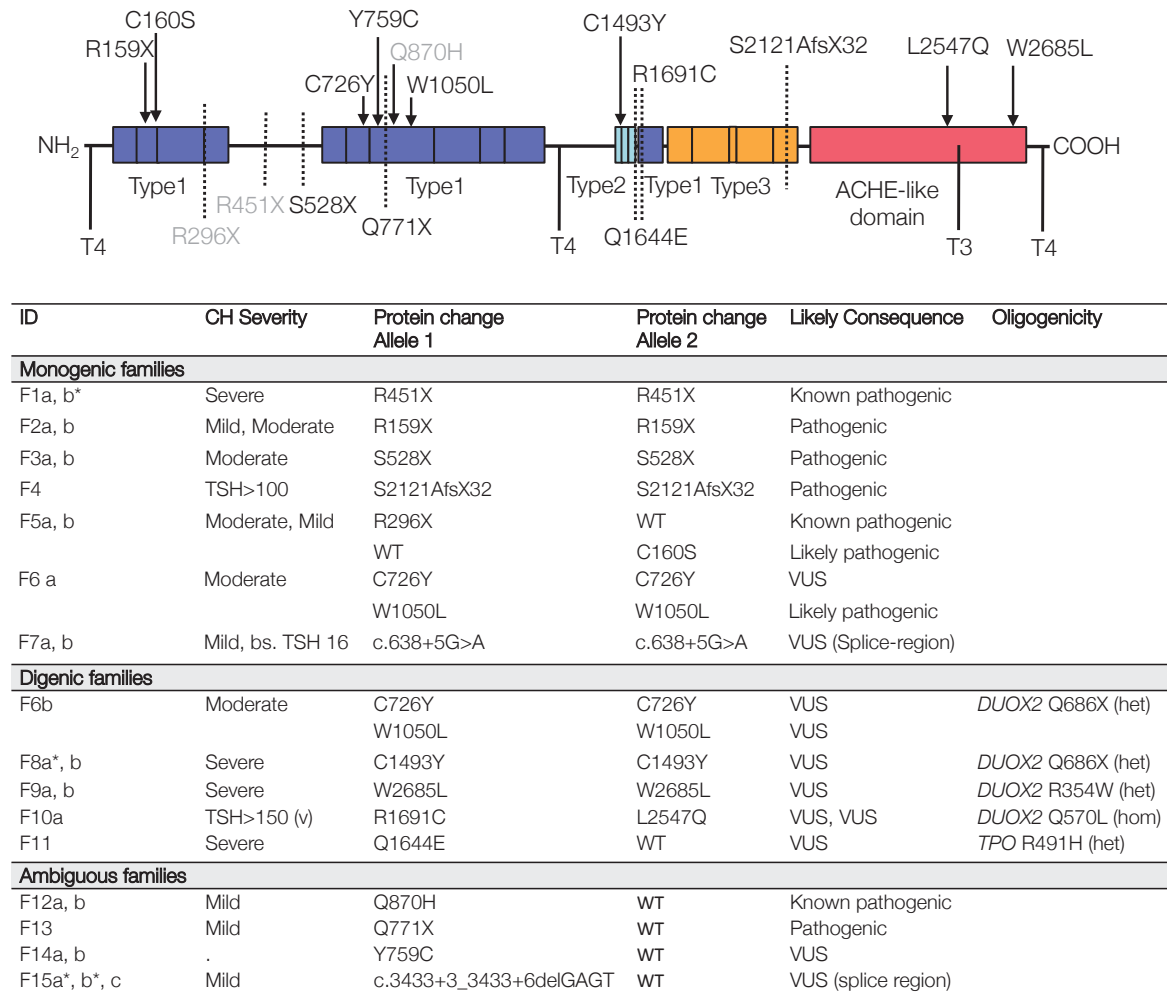


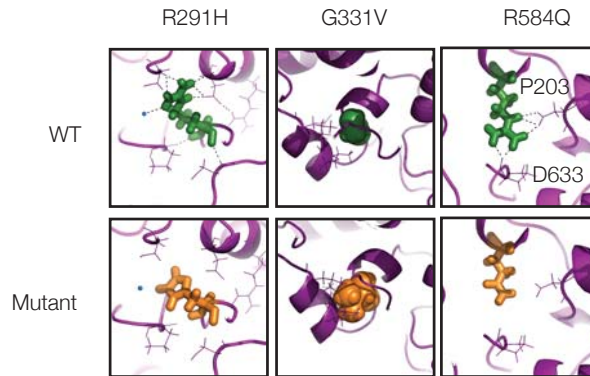
Figure 2.7 Mutations identified in *TG*.

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*.

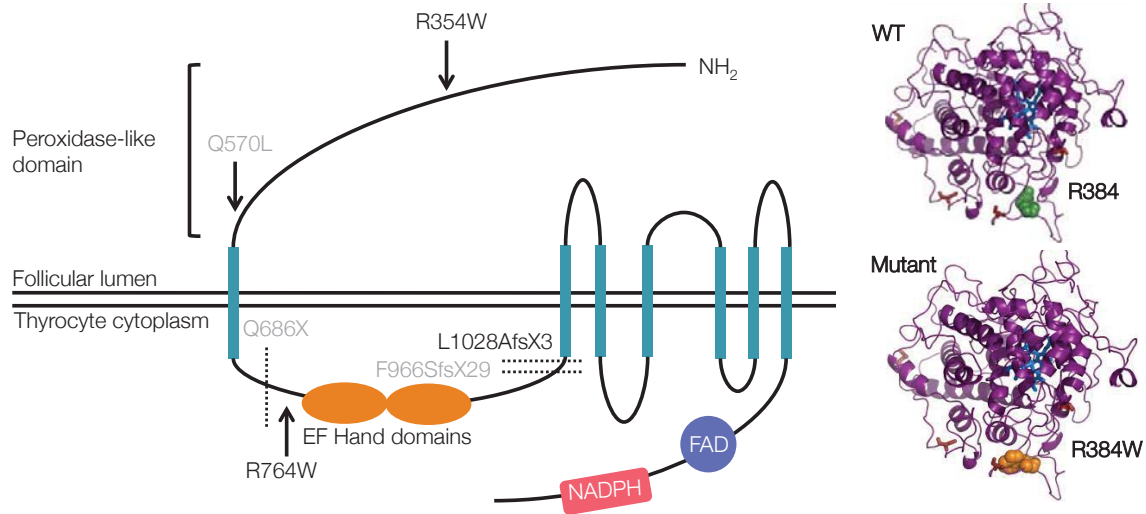
Summary of *TPO* mutations identified in all case categories and associated biochemical phenotypes. CH severity was classified according to ESPE criteria on the basis of serum fT4 levels; severe, <5, moderate 5 to <10, and mild >10 pmol/l, respectively [277] and pathogenicity is predicted according to 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, H₂O 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 constituted 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 ft_4 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.

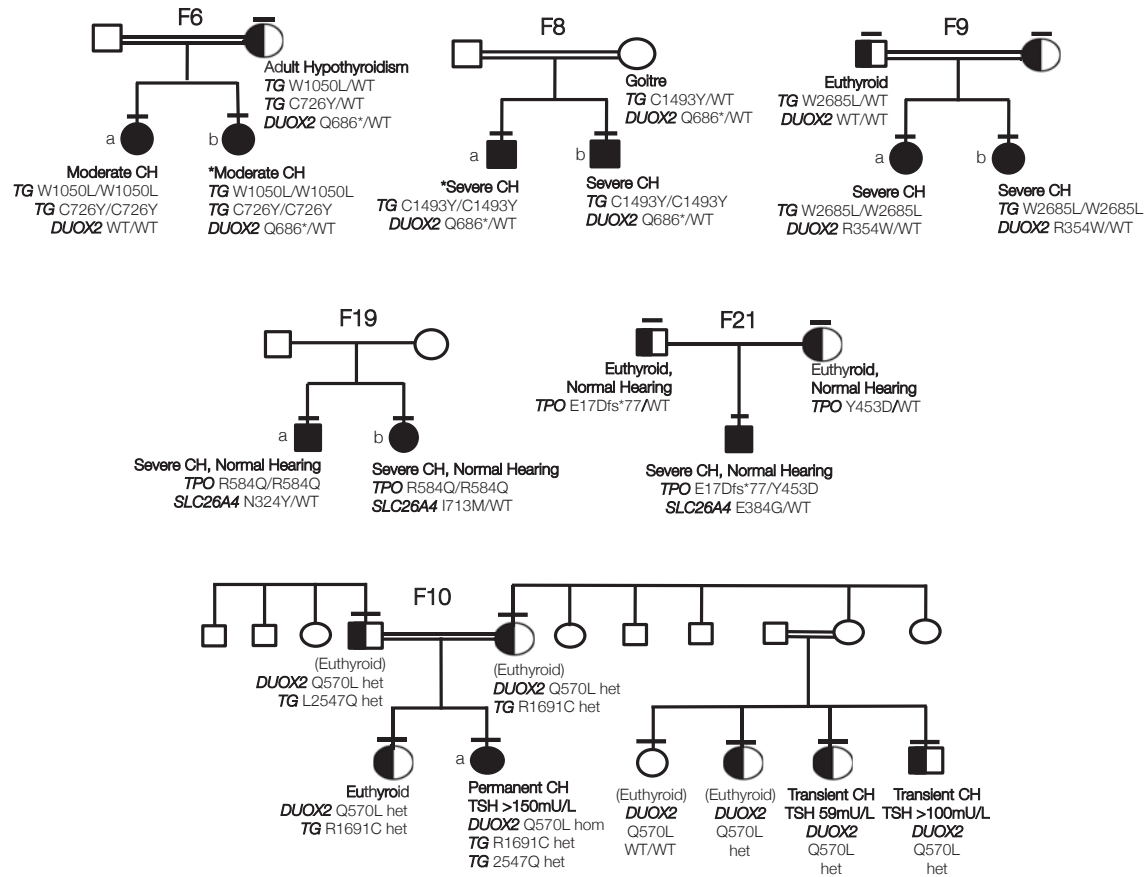


Figure 2.10 Genotype-phenotype segregation in six families with oligogenic variants.

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 ($11 \times 0.06 = 0.66$), 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 $\sim 81,000$ 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 H₂O₂ 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-T₄ levels [474]. Such hypotheses will be explored, via exome-sequencing analyses of these and additional CH samples, in the following chapter.

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].

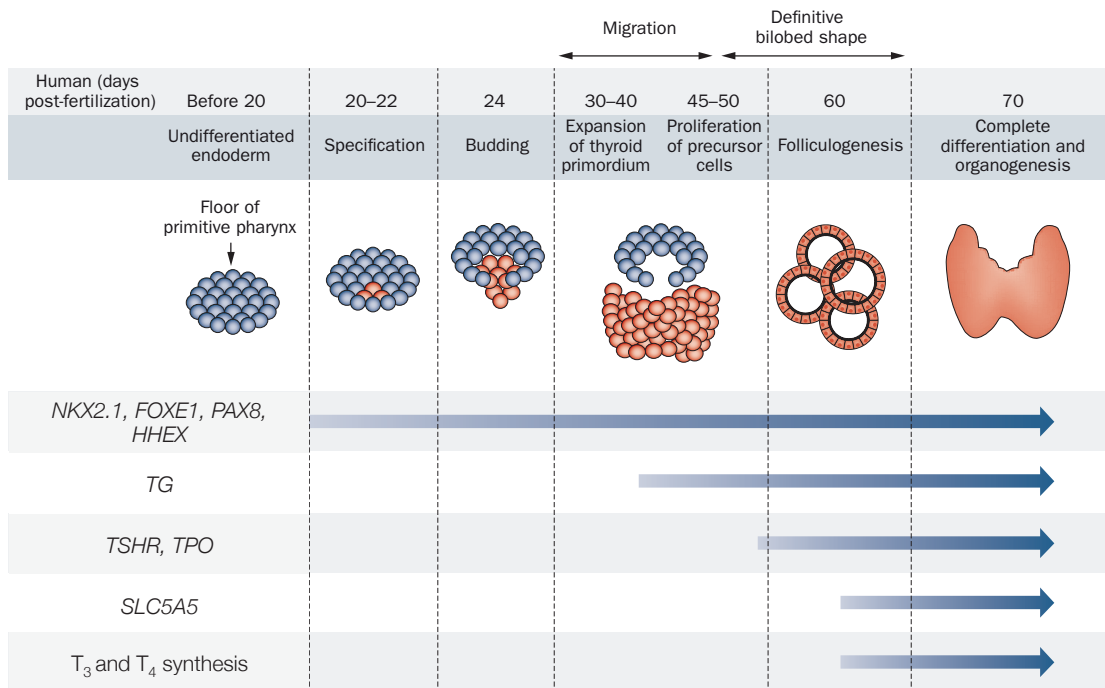


Figure 3.1 Gene expression during the stages of thyroid gland development in humans.

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
<i>NKX2.1</i>	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 <i>de novo</i> 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.
<i>FOXE1</i>	9q22.33	Tongue, palate, oesophagus and in ectoderm-derived organs: anterior pituitary, 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.
<i>PAX8</i>	2q14.1	Kidney, excretory system, endometrium, ovary, fallopian tube, pancreatic islet cells and lymphoid cells.	Thyroid agenesis, hypoplasia, ectopia (sublingual 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 (<i>Mus musculus</i>)	Gene in human (<i>Homo sapiens</i>)	Description	Thyroid phenotype	Extrathyroidal features
<i>Shh</i>	<i>SHH</i>	Sonic Hedgehog	Bilobation defect	Holoprosencephaly, cardiac outflow tract defects
<i>Foxe1</i>	<i>FOXE1</i>	Forkhead Box E1	Ectopia or agenesis	Cleft palate
<i>Chrd</i>	<i>CHRD</i>	Chordin	Hypoplasia	Cardiac outflow tract defects, hypoplasia of thymus, parathyroid
<i>Edn1</i>	<i>EDN1</i>	Endothelin 1		Craniofacial, cardiac and thymus defects
<i>Eya1</i>	<i>EYA1</i>	EYA Transcriptional Coactivator And Phosphatase 1		Aplasia of kidneys, thymus, parathyroid
<i>Fbln1</i>	<i>FBLN1</i>	Fibulin 1		Craniofacial, cardiac and thymus defects
<i>Hes1</i>	<i>HES1</i>	Hes Family BHLH Transcription Factor 1		Craniofacial, cardiac and thymus defects
<i>Hoxa5</i>	<i>HOXA5</i>	Homeobox A5		Cardiovascular and skeletal defects
<i>Isl1</i>	<i>ISL1</i>	ISL LIM Homeobox 1		Heart, pancreas and neural defects
<i>Nkx2-5</i>	<i>NKX2-5</i>	NK2 Homeobox 5		Cardiac defects
<i>Frs2</i>	<i>FRS2</i>	Fibroblast Growth Factor Receptor Substrate 2	Hypoplasia plus bilobation defects	Thymus and parathyroid defects
<i>Hoxa3</i>	<i>HOXA3</i>	Homeobox A3		Cardiovascular and skeletal defects
<i>Hoxb3</i>	<i>HOXB3</i>	Homeobox B3		Cardiovascular and skeletal defects
<i>Hoxd3</i>	<i>HOXD3</i>	Homeobox D3		Thymus and parathyroid defects
<i>Pax3</i>	<i>PAX3</i>	Paired Box 3		Cardiac outflow tract defects, hypoplasia of thymus, parathyroid
<i>Tbx1</i>	<i>TBX1</i>	T-Box 1		Cardiac outflow tract defects, hypoplasia of thymus, parathyroid
<i>Fgfr2</i>	<i>FGFR2</i>	Fibroblast Growth Factor Receptor 2	Agenesis	Atresia of the lungs
<i>Fgf10</i>	<i>FGF10</i>	Fibroblast Growth Factor 10		Aplasia of limbs, lungs, pituitary, salivary glands
<i>Hhex</i>	<i>HHEX</i>	Hematopoietically Expressed Homeobox		Forebrain truncations, liver aplasia, cardiac defects
<i>Nkx2-1</i>	<i>NKX2-1</i>	NK2 Homeobox 1		Pulmonary atresia, neural defects
<i>Pax8</i>	<i>PAX8</i>	Paired Box 8		Reproductive tract defects
<i>Twsg1</i>	<i>TWSG1</i>	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*, encoding 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-T₄ 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.

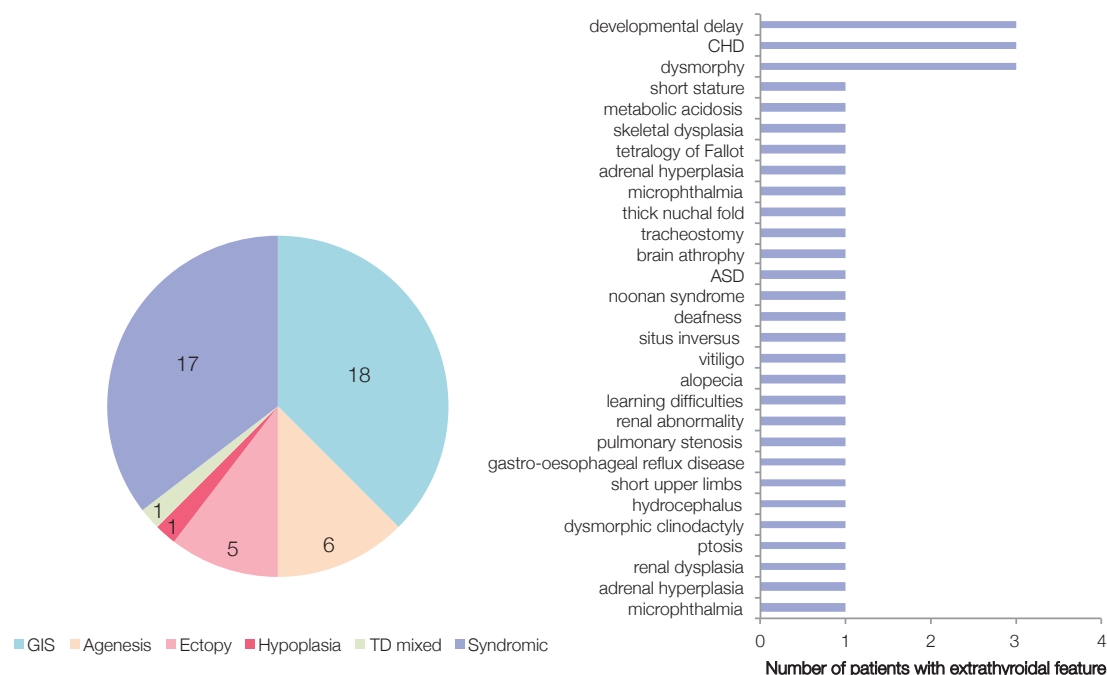


Figure 3.2 Phenotype categories within the CH cohort.

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	
<i>DUOX1</i>	Thyroid synthesis biological pathway	
<i>DUOXA1</i>		
<i>KCNQ1</i>		
<i>KCNE2</i>		
<i>LHX3</i>	GWAS hits for TSH and free T ₄ levels	
<i>PDE8B</i>		
<i>SLC26A7</i>	Mouse knockout exhibits CH	
<i>TBX1</i>	Established mouse models	
<i>ISL1</i>		
<i>NKX2.5</i>		
<i>EDN1</i>		
<i>FBLN1</i>		
<i>HOXA3</i>		
<i>FGF8</i>		
<i>HAND2</i>		
<i>DICER1</i>		Involved in thyroid carcinoma
<i>TPST2</i>		Mouse knockouts exhibits CH
<i>MCHR1</i>		
<i>GLIS3</i>	Neonatal diabetes with CH, OMIM 610199	
<i>WWTR1</i>	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 transitions 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.

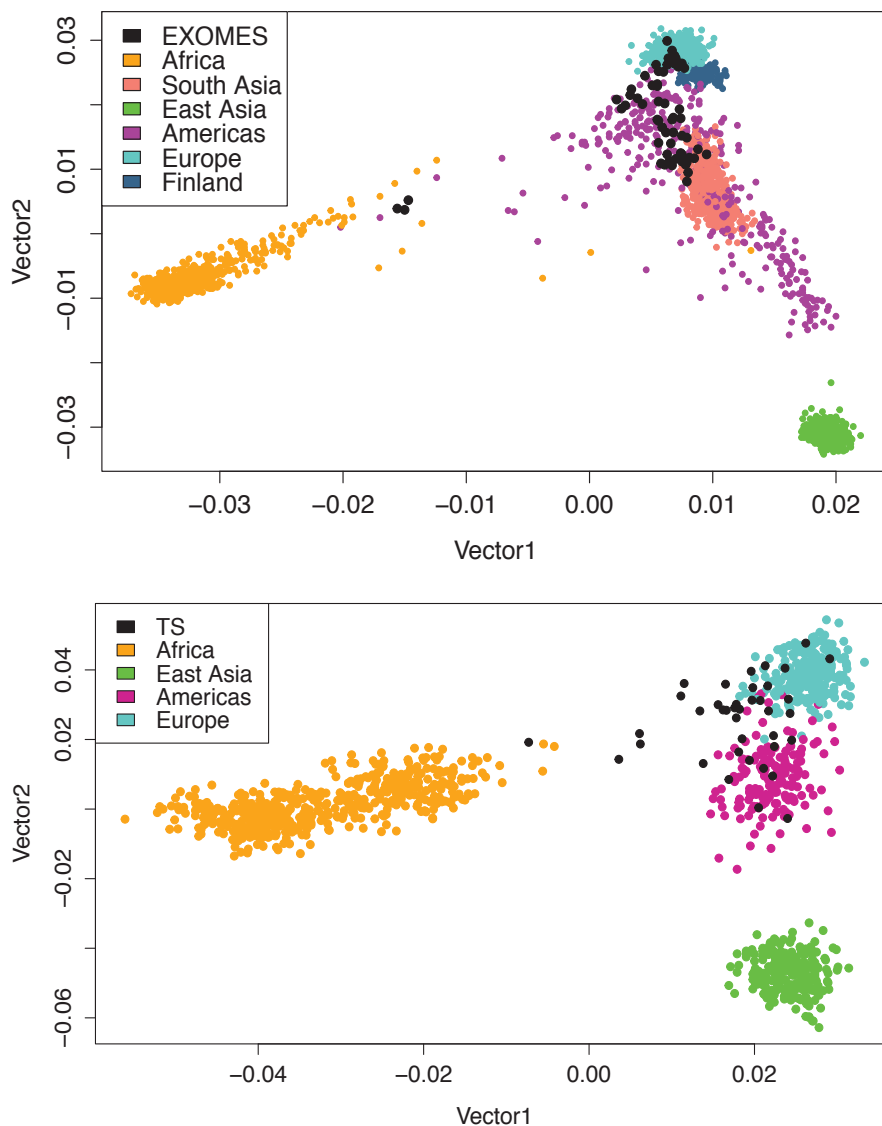


Figure 3.3 Principal component analysis of WES and targeted-sequencing CH samples. **A)** 75 WES samples plotted against 2,504 1KG Phase 3 samples. **B)** 33 targeted-sequencing (TS) samples plotted against 1,192 HapMap3 samples.

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

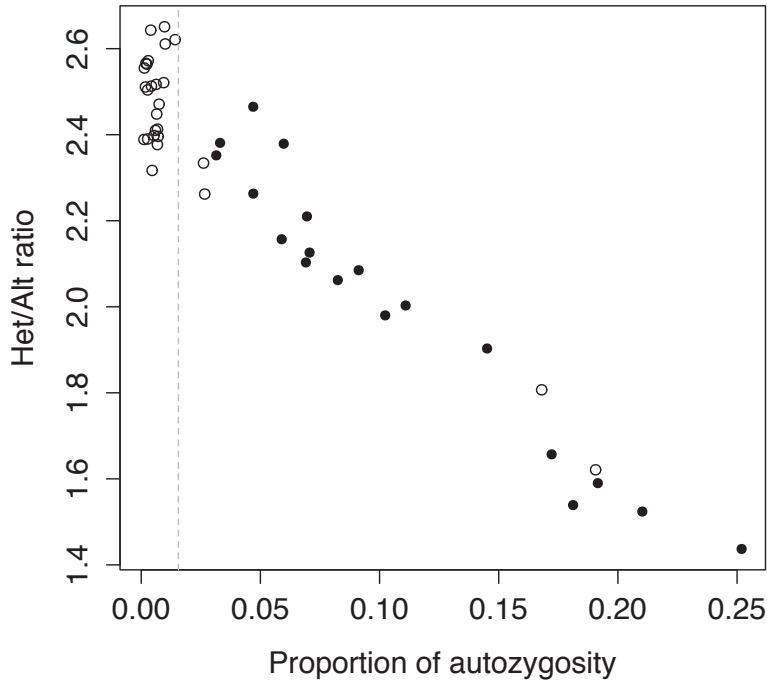


Figure 3.4 Estimation of consanguinity status for CH cases. Het/Alt ratio: the ratio of heterozygous to homozygous SNVs. Filled circles denote the individuals which were reported to be consanguineous upon sample recruitment whereas empty circles represent samples that were reported to be non-consanguineous. 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, $\sim 94\%$ 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 $\sim 16\%$ considered to be damaging by both tools and a further $\sim 5\%$ 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).

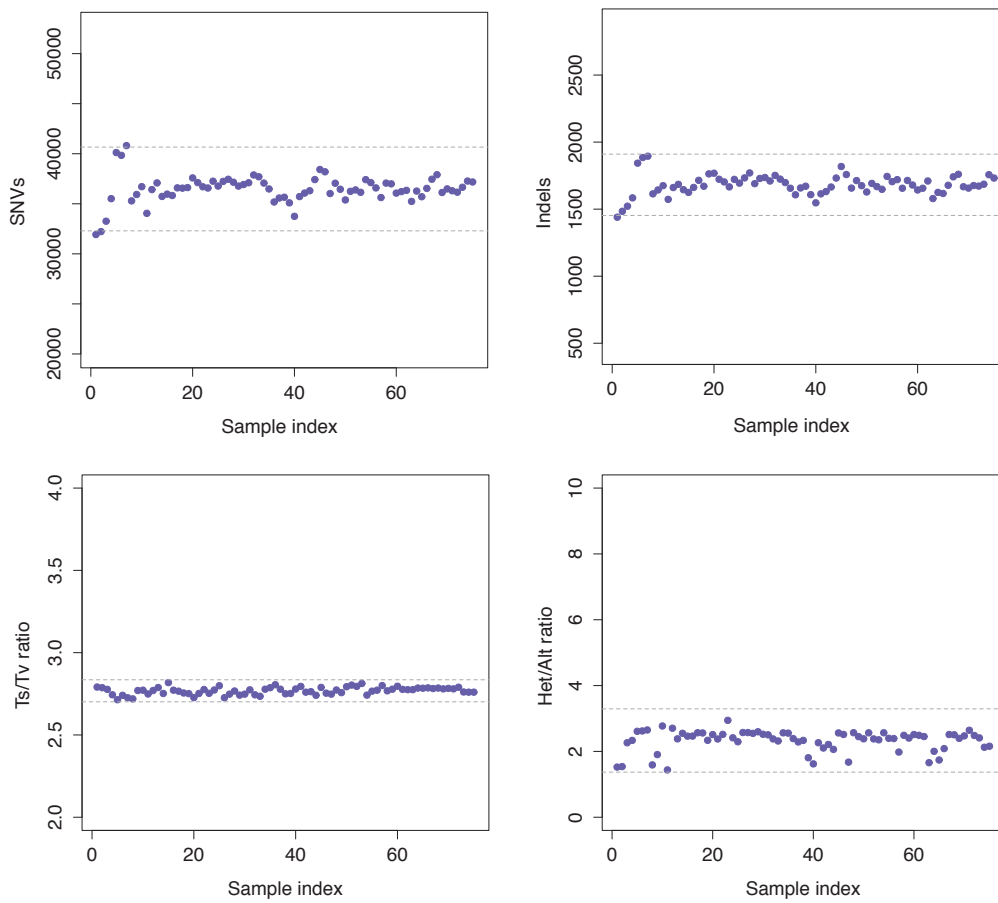


Figure 3.5 Quality control metrics for the WES experiment. **A)** Number of SNVs called and passing the variant QC per sample (see previous chapter for details). The cluster of three samples with higher numbers correspond to individuals of African ethnicity; **B)** Number of indels called and passing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

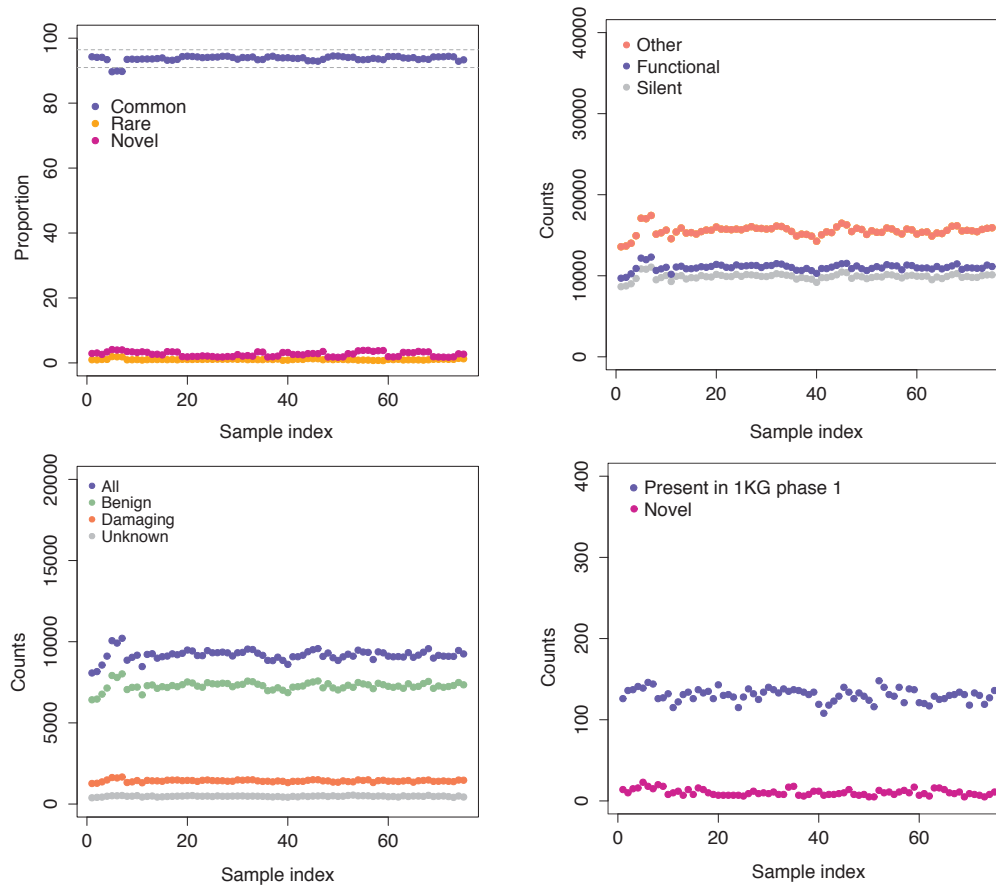


Figure 3.6 Population genetics metrics for the WES experiment. **A)** Proportion of SNVs that are common ($\geq 5\%$ in 1KG Phase 1), rare ($\leq 1\%$ in 1KG Phase 1) and novel (absent from both in 1KG Phase 1 and dbSNP137) per sample. The cluster of three samples with lower common variants correspond 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 variants 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.

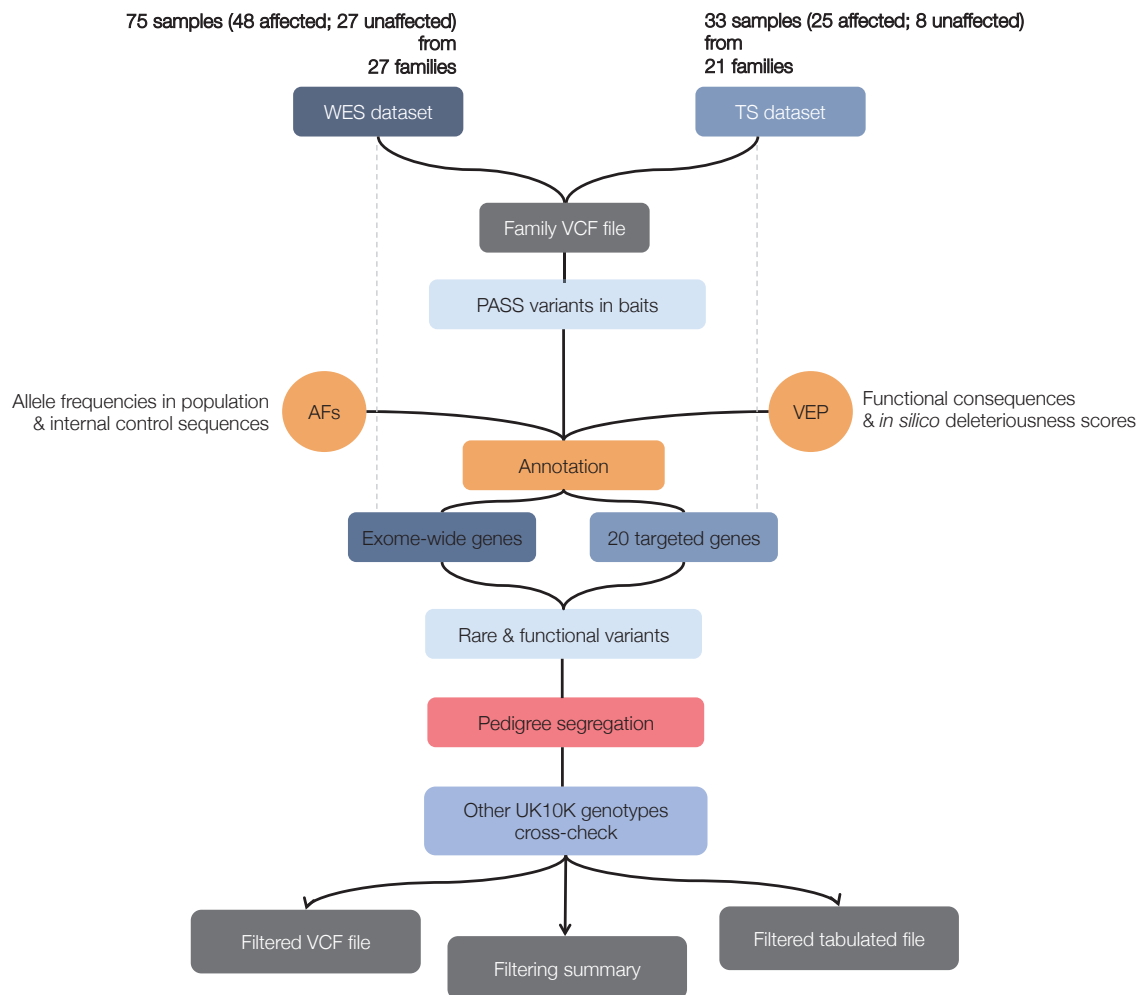


Figure 3.7 Variant filtering pipeline.

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	<i>De novo</i> 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 sequenced 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 sequenced 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 overlapped 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	<i>GLIS3</i>	0/1	MI	H837R	10/11	.	5.93
<i>Gland-in-situ</i> CH	R13	<i>DUOX1</i>	0/1	MI	K653N	18/35	.	0.74	.	.	5.00E-05	.	.
		<i>TBX1</i>	0/1	SD	.	.	4/8	5.08
Syndromic (CH and congenital heart disease)	R22	<i>NKX2-5</i>	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 T₃ and T₃, 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 dysmorphogenesis 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.

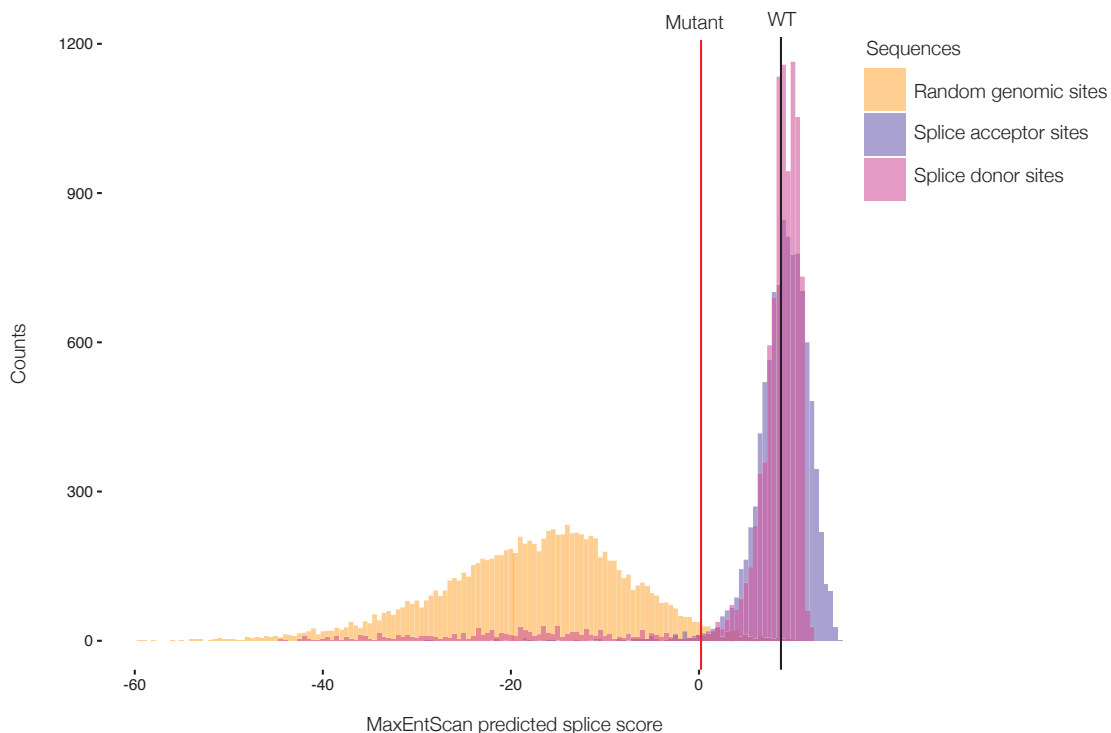


Figure 3.8 Predicted MaxEntScan scores for wild-type and mutant splice donor sequences of *TBX1* intron 4. To put the WT and mutant *TBX1* splice donor scores into context, MaxEntScan scores were also generated for a random set of 10,000 genomic regions, splice acceptor and splice donor regions occurring 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 co-occurrence 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 ($\sim 85\%$) were missense and very few ($\sim 4\%$) represented LoF variants (**Figure 3.9**). In total, the 800 variants identified were distributed across 678 unique genes, the majority of which ($\sim 51\%$) fit the dominant model of inheritance.

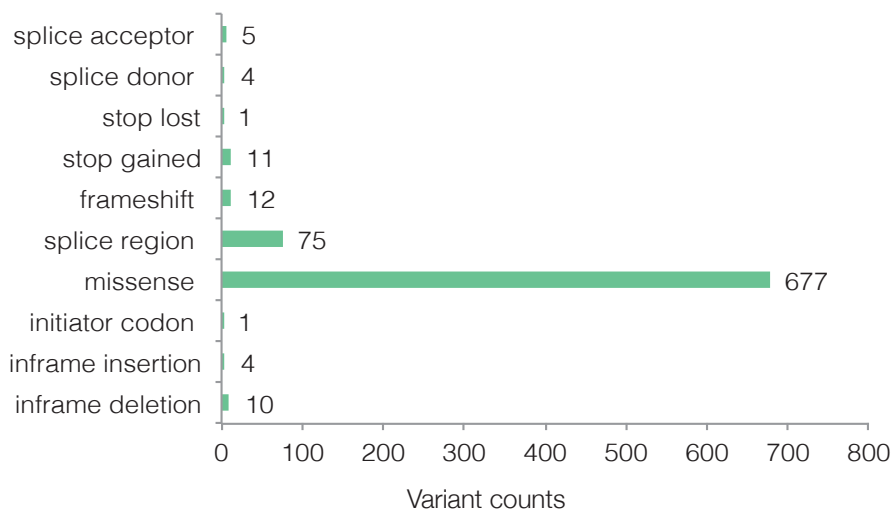


Figure 3.9 Functional consequences of the 800 inherited variants identified in 27 WES CH families.

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.

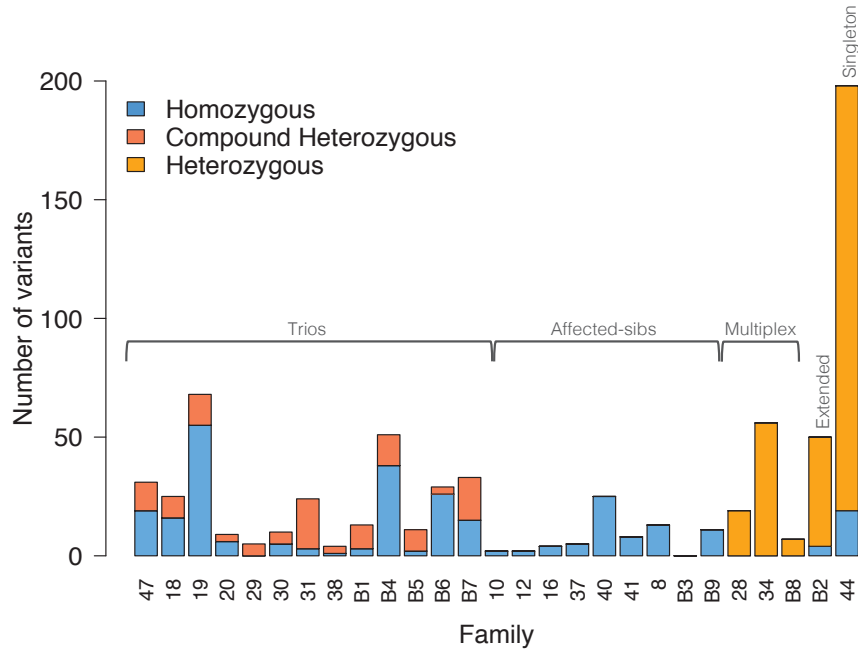


Figure 3.10 Distribution of inherited rare functional variants across families. Homozygous: variants that are homozygous for the non-reference allele; Compound heterozygous variants: in genes that have at least two variants, each of which inherited from exactly one parent. Trios: unaffected parents and 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.

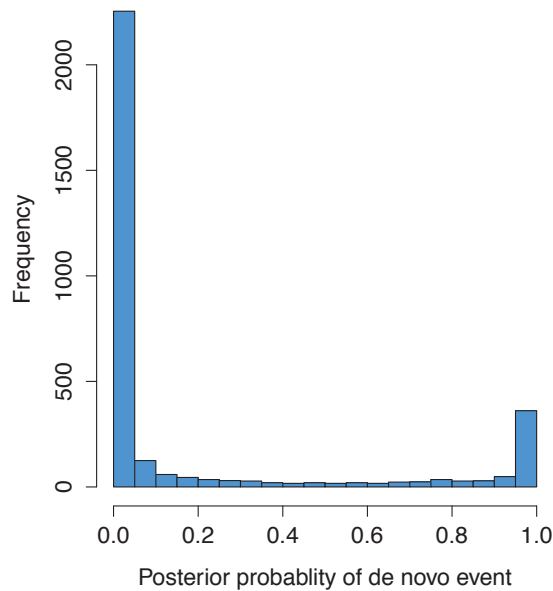


Figure 3.11 Histogram of posterior probabilities of *de novo* events called by DeNovoGear.

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 (<i>gland-in-situ</i> CH, skeletal dysplasia)	B4	239	34	5	4	2	0	0
Syndromic (<i>gland-in-situ</i> CH, developmental delay, dysmorphic, metabolic acidosis)	B5	268	33	2	2	1	0	0
Syndromic (<i>gland-in-situ</i> CH, <i>situs inversus</i> , CHD)	47	203	22	5	3	2	0	0
<i>Gland-in-situ</i> 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
Agnesis	19	ATXN2L	T>C	1	MI	V130A	3/24	.	4.88
Agnesis	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	HNRNPB	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	MYNR1	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	ZFRANB2	C>A	1	MI	C71F	3/10	.	5.63

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-uridylylate-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.

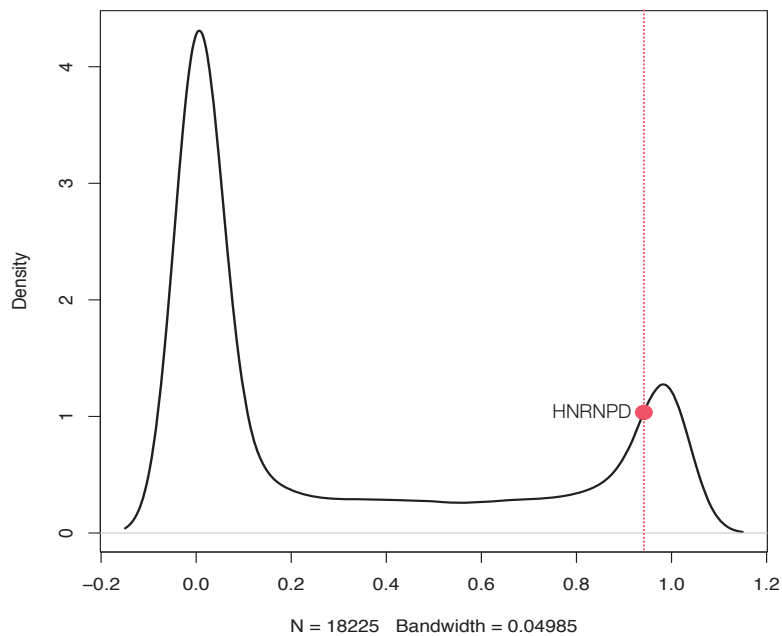


Figure 3.12 Distribution of intolerance to loss-of-function (LoF) mutation (ExAC pLI score) for all genes in the genome (N=18,225). The red dot represents *HNRNPD*, with a score of 0.96. This graph made use of data taken from Samocha *et al* [431].

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**).

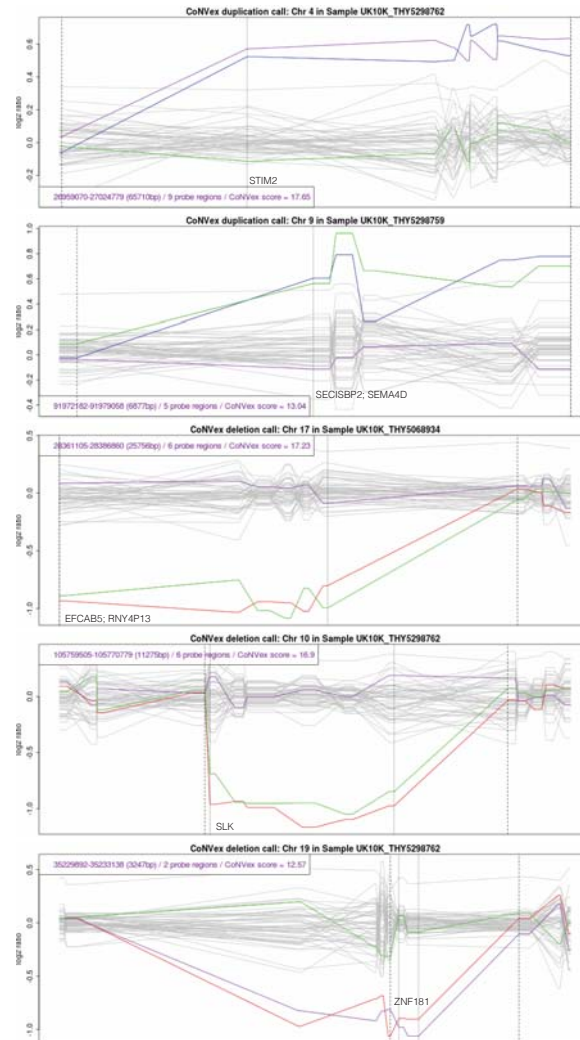


Figure 3.13 Log₂ ratios of rare (<1% AF) CNVs identified by CoNVex. (A) F31, (B) F30, (C) F12, (D) F31 and (E) F31.

The x-axis indicates the genomic coordinates. The y-axis indicates the normalised log₂ ratio of the exome read depth, compared to a group of controls. The red line shows the log₂ 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 log₂ ratio of the mother, and the green line shows the log₂ ratio of the father. The grey lines show the log₂ 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 ($\sim 96\%$) 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 ~ 2.7), genotype qualities (median ~ 94) and Het/Alt ratios (median ~ 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.

Gene	Cases		Controls		P-value	OR
	Carriers	Non-carriers	Carriers	Non-carriers		
<i>TRBV6-9</i>	2	25	1	2119	4.54E-04	164.92
<i>GGT1</i>	2	25	4	2116	2.22E-03	41.80
<i>ZNF623</i>	2	25	4	2116	2.22E-03	41.80
<i>CBFA2T2</i>	2	25	8	2112	6.44E-03	20.97
<i>MUC4</i>	3	24	41	2079	0.0167	6.32
<i>FILIP1</i>	2	25	16	2104	0.0206	10.48
<i>KLHDC4</i>	2	25	17	2103	0.0228	9.86
<i>ATP2B3</i>	2	25	18	2102	0.0252	9.31
<i>SCNN1A</i>	2	25	18	2102	0.0252	9.31
<i>ADAMTS15</i>	2	25	19	2101	0.0276	8.82
<i>DMBT1</i>	2	25	20	2100	0.0301	8.37
<i>AHNAK2</i>	3	24	79	2041	0.0813	3.23
<i>PTPRB</i>	2	25	37	2083	0.0849	4.50
<i>TLN2</i>	2	25	40	2080	0.0964	4.15
<i>CACNA1A</i>	2	25	52	2068	0.1464	3.18
<i>MUC17</i>	2	25	59	2061	0.1776	2.79
<i>DUOX1</i>	2	46	35	2085	0.1967	2.59
<i>DCHS2</i>	2	25	67	2053	0.2145	2.45
<i>LRP1B</i>	2	25	69	2051	0.2239	2.38
<i>SPTBN5</i>	2	25	72	2048	0.2380	2.27
<i>LRP2</i>	2	25	85	2035	0.2996	1.91
<i>HSPG2</i>	2	25	102	2018	0.3792	1.58
<i>FSIP2</i>	2	25	117	2003	0.4467	1.37
<i>MUC16</i>	3	24	237	1883	0.5957	0.99
<i>OBSCN</i>	2	25	228	1892	0.8029	0.66
<i>TTN</i>	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 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 <i>FOXE1</i>	52
Targets of <i>PAX8</i>	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-T₄ 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 dysplasia. Both parents were obligate carriers. *HSPG2* is a DDG2P gene that is responsible for skeletal dysplasia 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 54th 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 T₄ 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.

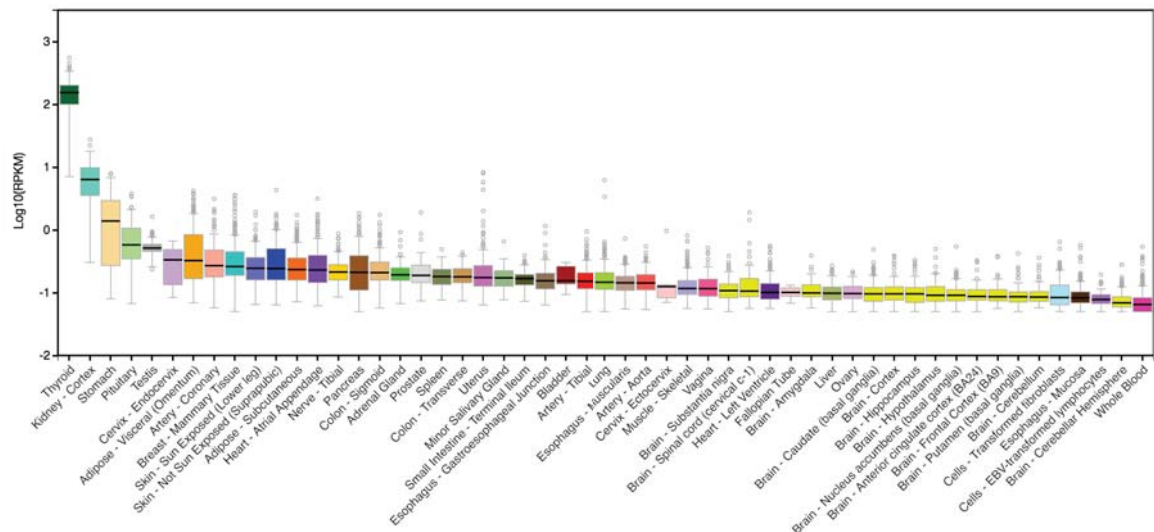


Figure 3.14 *SLC26A7* expression in GTEx tissues (www.gtexportal.org)

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 $\sim 21,000$ protein-coding genes in the exome, the search space for gene-pairs would be huge (2.1×10^8 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.

Chapter 4

The genetic architecture of *very-early-onset* inflammatory bowel disease

4.1 Introduction

4.1.1 What is inflammatory bowel disease?

Inflammatory bowel disease (IBD), comprising Crohn's disease (CD) and Ulcerative colitis (UC), is a chronic inflammatory condition that affects the gastrointestinal tract leading to epithelial injury. It is currently estimated to affect 2.2 million people in Europe [17] and millions more worldwide (**Figure 4.1**) [331, 353].

CD and UC constitute debilitating conditions that can ultimately be fatal. They both develop in the second or third decades of life and present with similar remission-relapse cycles. Patients experience an array of symptoms including abdominal pain, cramping, fever, vomiting, diarrhoea, rectal bleeding, anaemia, weight loss and fatigue [348]. No cure is currently available and symptoms are usually managed via anti-inflammatory steroids or immunosuppressants to reduce inflammation, dietary changes to minimize environmental triggers and, in severe cases, surgery to remove damaged portions of the bowel [73].

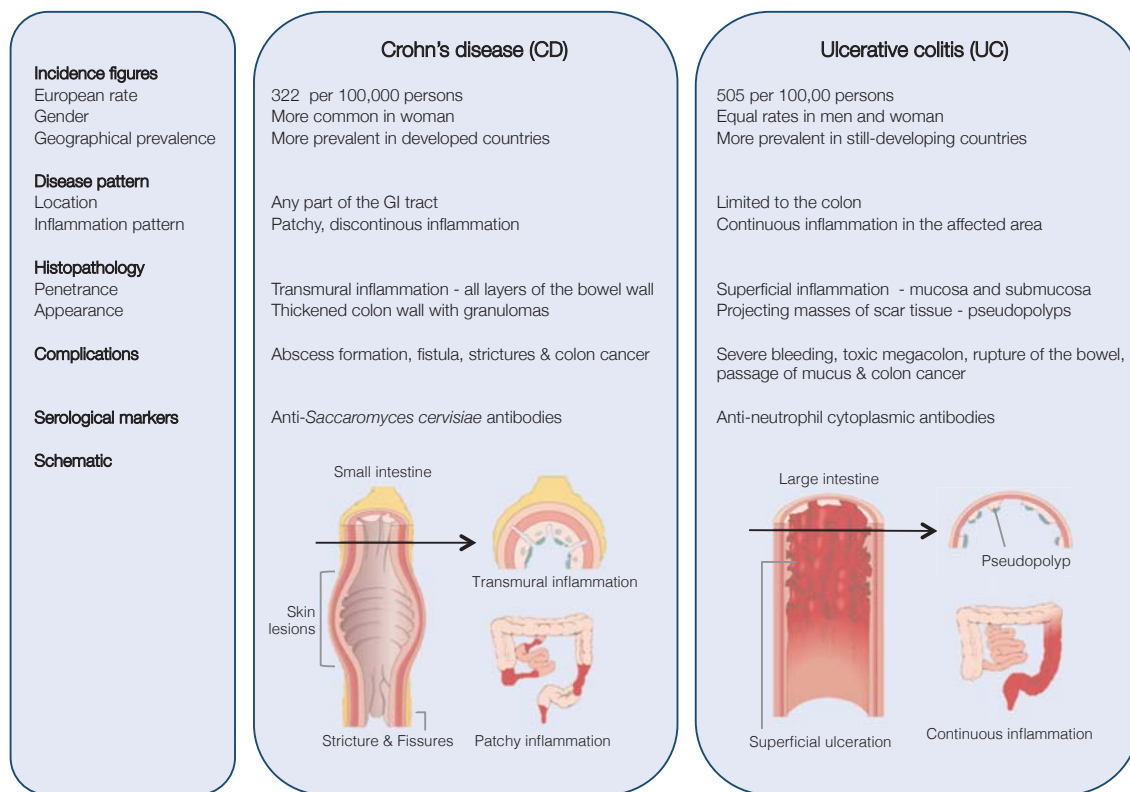


Figure 4.1 Epidemiological and clinical features of the two inflammatory bowel disease subtypes: Crohn's disease and ulcerative colitis [109] [35]. GI: gastrointestinal tract.

4.1.2 The genetics of IBD

IBD is a complex disease thought to arise from inappropriate activation of the intestinal mucosal immune system in response to commensal bacteria in a genetically susceptible host [232]. Large GWAS meta-analyses conducted by Jostins *et al* and Liu *et al* have uncovered 231 genomic signals associated with IBD [290]; together, they explain 13.1% and 8.2% of the variance in disease liability for CD and UC, respectively [290]. Such studies have also demonstrated that the genetic risk for CD and UC substantially overlap, with ~70% of the loci associated with both phenotypes [232, 290]. Similar to other complex diseases, the majority of the associated variants are common frequency alleles of modest effects (**Figure 4.2**), with an average increase in odds of developing the disease (OR) of 1.12 [290]. Despite the diversity in their roles in the immune system, many of the genes overlapping with the associated regions can be broadly split into 11 categories under the umbrella of the innate or adaptive immune systems (**Table 4.1**). **Figure 4.3** illustrates the role of some of these categories and constituent proteins within the intestinal immune system in health and disease.

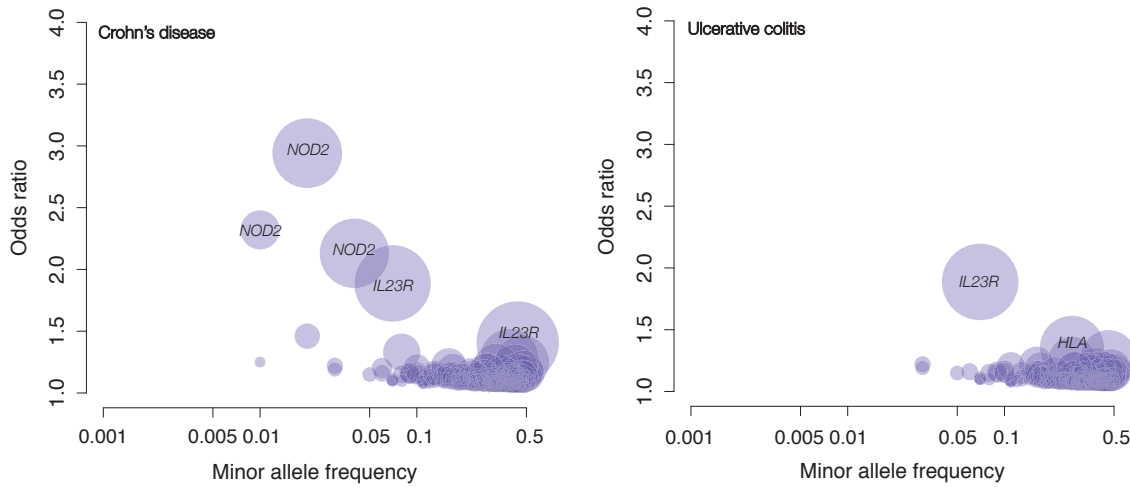


Figure 4.2 The genetic architecture of Crohn's disease and ulcerative colitis. Known CD/UC-associated variants are plotted according to their minor allele frequency and odds ratio (OR) [290]. The OR of protective alleles were inverted for illustrative purposes. The size of the circles represents the amount of variance in disease liability explained by that variant.

The loci with the largest effects on CD, IBD and UC are *NOD2*, *IL23R* and the *HLA*, respectively. The *NOD2* signal was initially identified through linkage studies [218, 219, 359], and is driven by three low-frequency coding variants (R702W, G908R, L1007fs) with allele frequencies of ~ 0.03 , ~ 0.01 and ~ 0.02 in European individuals, respectively [218, 359]. Homozygosity at any of these three alleles confers a 20-40-fold increase in CD risk while heterozygosity is associated with a more modest rise in risk (2-4 OR) [313], although still strikingly high for a complex disease association. *NOD2* encodes a key intracellular pattern recognition receptor that ensures antimicrobial activity at the surface of intestinal epithelial cells [3] (**Figure 4.3**). The *IL23R* locus is protective for both CD and UC. It encodes a cytokine receptor embedded in the cell membrane of activated T-cells, particularly Th17 cells (i.e. those that produce interleukin-17), and is important for their proliferation and survival [392], both of which are paramount for host defence at the mucosal surface (**Figure 4.3**). Finally, the classical human leukocyte antigen (*HLA*), the strongest UC-specific signal, contains genes that encode antigen-presenting proteins on the surface of the cell, and plays a crucial role in the regulation of the adaptive immune system.

Biological process	Context	Genes
Innate immunity	Provides an initial and quick response to microbes through pattern-recognition receptors	
Epithelial barrier function and repair	Maintains a physical and chemical barrier to commensal and pathogenic microorganisms	<i>HNF4A, CDH1, LAMB1, OSMR, ESR1, GNA12, IRL1, MUC19, PLA2G2E, PTGER4, REL, STAT3, NKX2-3</i>
Innate mucosal defence	Cell-surface receptors or adaptor proteins involved in mediating innate immune response signalling	<i>NKX2-3, CARD9, FCGR2A, IL18RAP, IRL1, NOD2, REL, SLC11A1</i>
Autophagy pathway	Intracellular degradation and recycling system for clearing intracellular organelles, proteins and macromolecular complexes. Crucial for cellular activity and protein recycling	<i>ATG16L1, CUL2, DAP, IRGM, LRRK2, NOD2, ATG4B, PARK7</i>
Apoptosis	Important mechanism of peripheral immune tolerance that controls programmed cell death of mucosa cells	<i>DAP, FASLG, MST1, PUS10, TH1A, TNFSF15</i>
Adaptive immunity	Highly specialised cells and processes tailored to eliminate or prevent specific pathogens, while also developing immunological memory (through memory B and memory T cells)	
Activation		
IL23R pathway	Important for Th17 proliferation and survival. Crucial for host defence against foreign pathogens at the mucosal surface	<i>CCR6, IL12B, IL21, IL23R, JAK2, STAT3, STAT4, TYK2, PTPN2</i>
NF- κ B pathway	Involved in cellular inflammatory responses in the mucosal environment, which leads to subsequent production of cytokines (TNF and IL-1 β) and antimicrobial peptides (defensins) by T helper cells, subsequently activating the adaptive immune system	<i>NFKB1, REL, TNFAIP3, TNIP1, NFKBIZ, TNFSF15, TNFRSF9, RIPK2</i>
Aminopeptidases	Involved in the generation of HLA class II-binding peptides. HLA-II proteins display short peptides derived from pathogens in their cell surface for recognition by the appropriate T-cells	<i>ERAP1, ERAP2</i>
IL2 and IL21 dependent T-cell activation	Cytokine growth factors that optimise T-cell responses. Produced by CD4+ T-helper cells and CD8+ cytotoxic T-cells upon antigen-induced activation	<i>IL2, IL21, IL2RA</i>
Regulation		
Th17 cell differentiation	These cells are abundant in the intestine, especially in the terminal ileum and control epithelial cell proliferation, wound healing and the production of defensins	<i>AHR, CCR6, IL2, IL22, IL23R, IRF4, JAK2, RORC, STAT3, TNFSF15, TYK2</i>
T-cell regulation		<i>ICOSLG, IFNG, IL12B, IL2, IL21, IL23R, IL2RA, IL7R, NDFIP1, PIM3, PRDM1, TAGAP, TNFRSF9, TNFSF8, LY75, CD28, NFATC1, CCL20, IL10</i>
B-cell regulation		<i>BACH2, IKZF1, IL5, IL7R, IRF5, NFATC1, IL10</i>

Table 4.1 Biological processes involved in the pathology of IBD. Example processes and pathways implicated in inflammatory bowel disease pathogenesis via genome-wide association studies. Genes belonging to these categories and falling within IBD-associated loci are listed. Note however that, in some cases, the specific genes have not yet been identified as causal, and as many loci contain multiple signals spanning multiple genes, these should not be considered as confirmed. Table adapted from De Lange *et al* [109].

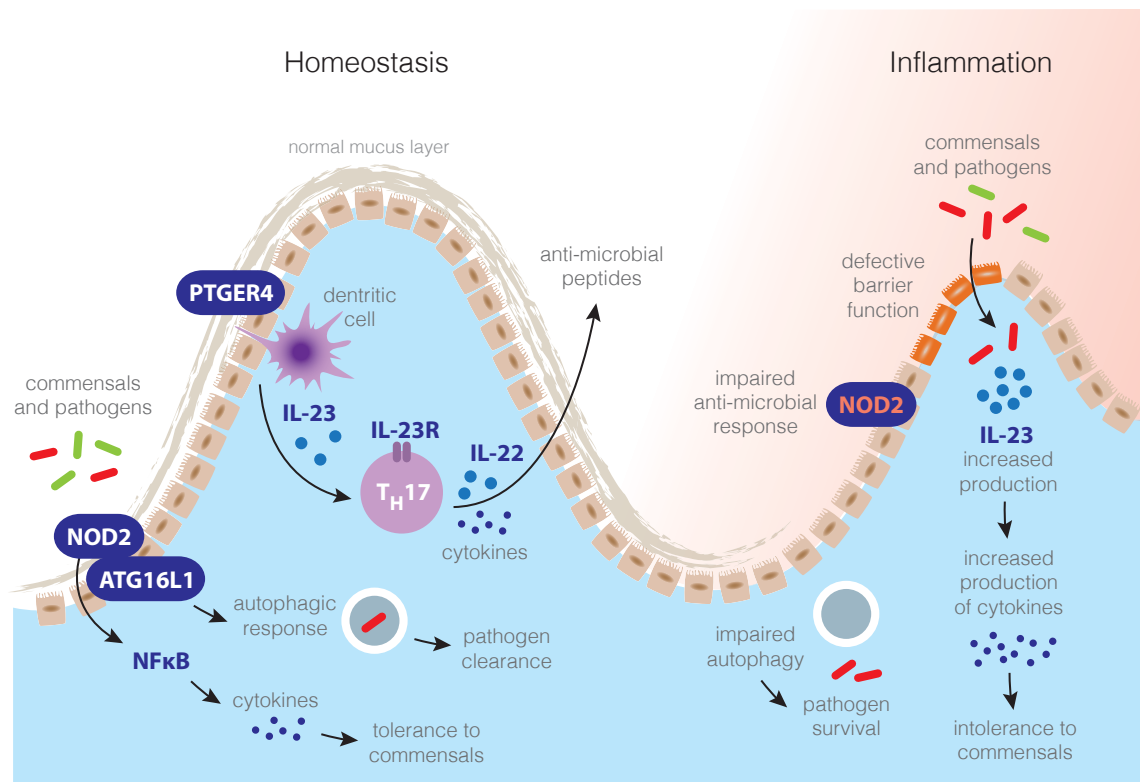


Figure 4.3 A simplified overview of the intestinal immunity in health and disease.

The gastrointestinal tract has a large mucosal surface (300m²) where intestinal epithelial cells, innate and adaptive immune cells interact to scrutinise foreign bodies transiting along the tract. Barrier permeability permits microbial invasion, which is immediately detected by the innate immune system. The result is either an active tolerogenic response, for example, towards dietary and commensal antigens, or an immunoinflammatory response against pathogens. Extracellular mediators such as cytokines and antimicrobial peptides mediate these two responses and the appropriate balance between the anti-inflammatory and pro-inflammatory signals maintains intestinal immune homeostasis. By sensing bacterial peptidoglycans, *NOD2* activates the NF- κ B pathway, which in turn leads to the production of cytokines and antimicrobial peptides that provide a barrier between microorganisms and the epithelial layer. Interaction between *ATG16L1* and *NOD2* activates the autophagy pathway in epithelial cells, which results in immediate pathogen clearance upon invasion. Dendritic cells are active participants in maintaining immunologic tolerance within the intestine, continuously sampling external and internal contents via podocytes extending through the epithelium. After activation of *NOD2* by bacterial products, *PTGER4* promotes the release of IL-23 cytokine from dendritic cells. This favours the development of Th17 cells (i.e. a subset of T-helper cells that produce IL17), which in turn secrete a series of pro-inflammatory cytokines, including IL17 and IL-22, for which receptors are expressed in the epithelium. IL-22 acts as an epithelial barrier protection factor, as it increases the production of anti-microbial peptides by certain epithelial cell types. IBD-associated variants perturb many aspects of intestinal homeostasis, shifting it to an inappropriate state of chronic inflammation characterised by intolerance to microbiota. Several events contribute to this IBD state including: disruption of the mucus layer, dysregulation of epithelial tight junctions, epithelial cell apoptosis, increased epithelial permeability, bacteria translocation, impaired sensing of pathogens by *NOD2*, and increased and sustained production of chemokines and cytokines.

4.1.3 Paediatric-IBD

Although precise epidemiological data are still lacking, approximately 10-15% of patients with IBD develop intestinal inflammation before 18 years of age [356], with this proportion growing worldwide [420, 496]. Even though this earlier IBD phenotype may apparently look similar in terms of symptoms and general treatment, it is becoming clear that the pathology of IBD in certain age groups in children presents unique challenges not encountered in adults. Children with a diagnosis of IBD before the age of six exhibit a more severe phenotype and disease course when compared to adolescents and adults [173, 181, 433]. On the other hand, most children diagnosed from seven years onwards, present with a more ordinary disease course and pathology, similar to that seen in adult-onset cases [202]. This increasing understanding of age-specific characteristics has led to changes in the traditional classification of paediatric IBD (the Montreal system), with a new classification system comprised of five major age groups [496] (**Table 4.2**).

Group	Age range
Paediatric-onset IBD	< 17 yrs
EO-IBD	< 10 yrs
VEO-IBD	< 6 yrs
Infantile-onset IBD	< 2 yrs
Neonatal-onset IBD	< 38 days of age

Table 4.2 Subgroups of paediatric-IBD according to age. Table adapted from Uhlig *et al* [496].

4.1.4 Very-early-onset IBD

Very-early onset IBD (VEO-IBD) represents a distinct group of children with a diagnosis before the age of six years (**Table 4.2**). It has an estimated incidence of 4.37 per 100,000 children and a prevalence of 14 per 100,000 children [496]. In comparison to later forms of intestinal inflammation, VEO-IBD presents with higher rates of affected first-degree relatives [202], higher concordance in disease location [91], and a higher male-to-female ratio [45, 181].

Three main features are thought to characterise VEO-IBD: first, approximately 1/5 of children with IBD younger than six years of age and 1/3 of children with IBD younger than three years of age have an undetermined type of colitis (U-IBD) [399].

In comparison, the rate of U-IBD is less than 5% in the adult IBD population [495]. This disparity reflects the difficulty in classifying VEO-IBD patients into discrete CD or UC categories, and the lack of a refined phenotype to further subgroup individuals within the VEO-IBD group.

Second, VEO-IBD phenotypes display a different anatomic distribution when compared to adult IBD, exhibiting more extensive intestinal involvement [356]. The extent of disease in VEO-CD is manifested by penetrating histologic abnormalities throughout the GI tract, whereas in VEO-UC, it is reflected by a pancolitis (i.e. inflammation of the entire colon) rate of 80-90%, compared to the 24% rate documented in adults [45]. This extreme manifestation in VEO-IBD is perhaps not surprising: the first years of life are a critical and vulnerable period in the initiation of a normal host immune response toward the external environment, with the mucosal immune system and the intestinal flora still under development [356].

Lastly, VEO-CD patients have an unpredictable and a more complicated disease path, quickly progressing to a severely structuring phenotype over time [181]. More importantly, there is a high rate of resistance to conventional therapies including second-line immunosuppressive drugs [500]. Because of this, therapeutic approaches often need to be aggressive, encompassing multiple drugs, injection with monoclonal antibodies against TNF- α (Infliximab) [68] and, in very extreme cases, allogeneic bone marrow transplantation [495]. Ultimately, VEO-IBD results in an increased probability of colectomy [495], severe growth impairment as a complication of the chronic colitis and/or its treatment and, sometimes, death.

4.1.5 The genetics of VEO-IBD: the rare-variant hypothesis

The aetiology of paediatric forms of IBD, including VEO-IBD, has been less well studied than its adult counterpart and the exact genetic determinants of VEO-IBD remain largely unexplored. The relatively short exposure time to environmental triggers in VEO-IBD and the higher familial clustering observed [202], suggests VEO-IBD might represent a more genetically influenced group of affected individuals, with a phenotype driven by rare penetrating variants of large effect – this has been the most popular hypothesis in the field, with researchers often viewing it as a Mendelian form of IBD. Indeed, children with premature onset of intestinal inflammation may not only represent a distinct phenotype with an atypical presentation, but also a genetic

architecture distinct from the general and polygenic IBD forms, and thus not amenable to GWAS.

The suspicion of a monogenic cause underlying VEO-IBD was confirmed in 2009, via linkage, with the discovery of fully penetrant mutations in the interleukin-10 (*IL10*) receptors alpha (*IL10RA*) and beta (*IL10RB*) [174, 255] in patients presenting with VEO-IBD at an average age of 7.5 months. Subsequent candidate-gene studies identified additional LoF mutations in *IL10* receptors and in *IL10* itself [39, 121, 255, 341], a locus in which common variants have already demonstrated association with adult-IBD (**Table 4.1**).

IL10 encodes a potent anti-inflammatory cytokine that counteracts hyperactive immune responses in the human body [333]. Its anti-inflammatory effects are mediated via binding to IL10 receptors, which then fuels a downstream signalling cascade to block pro-inflammatory loci and cytokine production [340]. A clear disruption of this pathway is evident in IL10 and IL10RA/B-mutant patients [174], and is consistent with the well-known phenotype of *Il10*-deficient mice, which is marked by spontaneous colitis with systemic outburst of cytokines [447]. Together, these studies confirmed that the loss of *IL10* and its negative-feedback signalling drive excessive inflammatory responses forward, leading to gut mucosal injury [262].

Exome sequencing in VEO-IBD: *XIAP*, *TTC7A*, *FOXP3* and other stories

There are a few success stories resulting from exome-sequencing of individual VEO-IBD cases or a few affected families. However, next-generation sequencing has not yet been extensively employed in the diagnosis of VEO-IBD, nor in research studies aiming at identifying novel causative disease-genes.

Table 4.3 summarises the main findings of the first WES studies conducted in VEO-IBD patients. Collectively, the identified mutations disrupted key genes (*XIAP*, *TTC7A* and *FOXP3*) previously known to be associated with rare and severe monogenic disorders of the immune system [24, 360, 525]. Functional studies of the mutant proteins and assessment of the immunological profiles of the studied patients, suggested their early gastrointestinal pathologies represented new manifestations of these immune-related syndromes or milder forms of disease, marked by atypical-IBD phenotypes. This finding is reminiscent of many other disease phenotypes that seemed novel at first, but that were subsequently reassigned as atypical or unusually complex presentations of well established Mendelian disorders [55, 77, 323].

Gene	Mutations identified	Patients studied	Gene function	Known condition
<i>XIAP</i>	Hemizygous missense mutation: C203Y	Male child presenting with intractable IBD at 15 months.	Activator of <i>NOD2</i> and the NF- κ B pathway, with a critical role in the apoptosis of defective intestinal epithelial cell. Important for commensal tolerance.	X-linked lymphoproliferative syndrome 2 (XLP2), a disorder of the immune system characterised by dysgammaglobulinemia and hemophagocytic lymphohistiocytosis, usually associated with an exaggerated response to the Epstein-Barr virus.
<i>TTC7A</i>	Compound heterozygous or homozygous mutations: E71K + Q526X c.844-1 G>T + c.1204-2 A>G; A832T	Five patients from three families presenting with severe apoptotic enterocolitis before 1 yr of age.	Maintains lymphocyte homeostasis by regulating cell adhesion, migration and proliferation. Important for the intestinal epithelial barrier.	Multiple intestinal atresia with severe combined immunodeficiency (SCID), characterised by increased susceptibility to bacterial infections.
<i>FOXP3</i>	Hemizygous missense mutation: C232G	Multiplex family composed of an affected mother and three affected sons presenting with atypical chronic gastroenteritis before 2 yrs of age.	Master transcription factor of CD4+ T cells, which promote tolerance to the flora and dietary products at the intestinal mucosa.	X-linked immune dysregulation, polyendocrinopathy, enteropathy syndrome (IPEX), characterised by systemic autoimmunity typically beginning in the first year of life.

Table 4.3 Summary of mutations and disease-causative genes discovered in the first three WES studies of VEO-IBD patients [24, 360, 525].

Exome sequencing in larger VEO-IBD cohorts was conducted in three recent studies, all of which ended up focusing on a smaller number of genes, including known IBD (N=169) [78], autoimmune (N=33) [20] or primary immunodeficiency (N=400) loci [244]. The first two studies analysed eight and 18 paediatric-IBD patients, respectively, with ages at diagnosis ranging from 2-16 years. These two reports did not convincingly identify any gene enriched for mutations in patients in the analysed set of genes. The third and largest study to date, performed by Kelsen *et al* [244], analysed exome data from 125 VEO-IBD children diagnosed before four years of age. The authors limited their analysis to rare variation (<0.1% AF) present in genes associated to primary immunodeficiencies (PIDs) and related pathways (n=400) and their findings suggested an over-representation of damaging variants in such genes in their cohort.

Monogenic disorders with IBD-like inflammation

Inspired by stories such as *XIAP*, *TTC7A* and *FOXP3*, there is now increasing awareness that several monogenic disorders present with overlapping pathology with CD or UC and, most frequently, their histological and endoscopic information does not allow a

clear distinction to IBD [5]. These conditions have been termed to exhibit an ‘IBD-like inflammation’ with varying levels of penetrance of the IBD phenotype, which has been estimated to range from 2 to 30% [495]. A total of 59 IBD-like conditions have been identified and associated to IBD-like inflammation [495, 496], the majority of which represent PIDs caused by familial defects in key components of the immune system. Most of these abnormalities are recessively inherited (~72%) and can be divided into distinct subtypes depending on the biological mechanisms by which they affect intestinal immune homeostasis (**Figure 4.4**). Collectively, they disturb multiple layers of immune competence that severely compromise intestinal immunity.

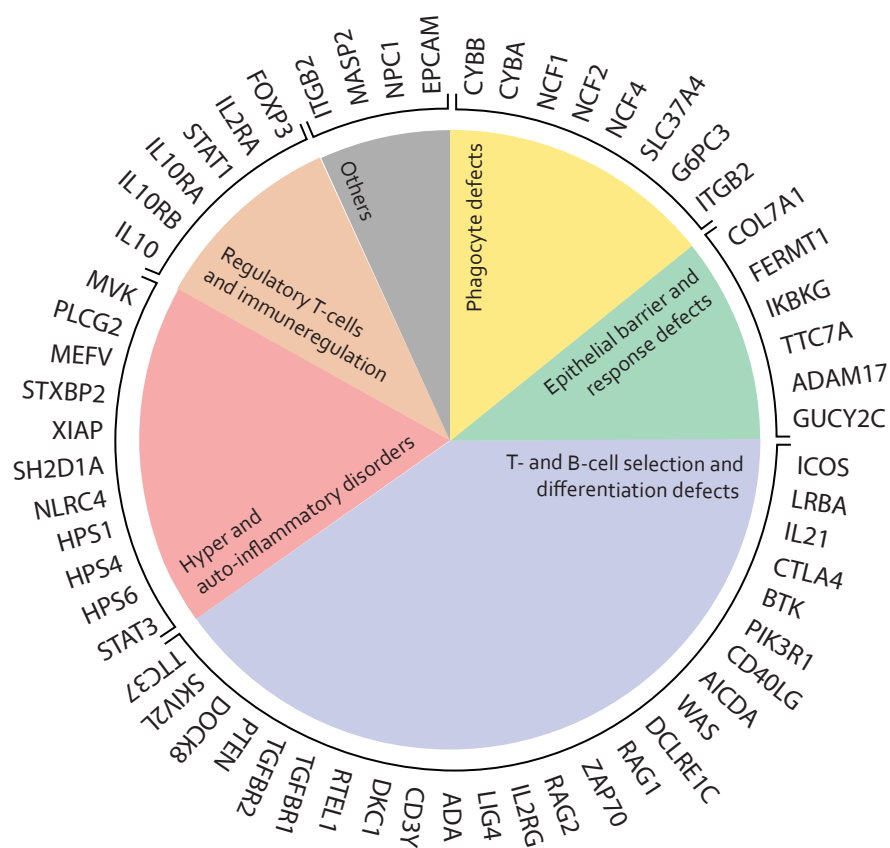


Figure 4.4 Monogenic defects (n=59) associated with VEO-IBD and IBD-like immunopathology stratified by biological category [495, 496].

It has been suggested that the intestinal immune dysfunction seen in these 59 conditions has similarities to those seen in VEO-IBD and that rare, penetrating defects within any of these loci may underlie the many VEO-IBD cases that still await a genetic diagnosis [495]. Following this hypothesis, recent guidelines on the diagnostic approach

to VEO-IBD recommended excluding other possible causes of early-onset inflammation, such as these immune function defects [496], which would also allow for more targeted treatment strategies. Given that adult-IBD associated loci are enriched 4.9-fold for PID genes [232], the importance of these 59 loci may not be limited to the VEO-IBD phenotype, but may extend to the biology of IBD in general, albeit with a different magnitude of effect. Moreover, a substantial proportion of proteins encoded by the genes mutated in these disorders directly, or indirectly, interact with loci that confer susceptibility to IBD, suggesting common signalling pathways predisposing to colitis [495]. Examples include *XIAP* and *IKBKG*, which interact with IBD-loci such as *RIPK2* and *NOD2* [256], as well as *CYBA* and *CYBB* that interact with *RELA* and *NFBK1* [196].

Despite this growing appreciation of the possible defects underlying VEO-IBD, known mutations still account for only a small fraction of VEO-IBD cases [496], and the true fraction of VEO-IBD incidences caused by this type of variation is unknown. Many studies that linked molecular defects in IBD-like genes to VEO-IBD phenotypes may have had a strong selection bias towards an expected clinical and molecular subphenotype, and may have therefore overestimated the frequency of specific defects as a cause of VEO-IBD. Large-scale analysis of multi-centre, population-based cohorts is thus warranted to determine the true proportion of VEO-IBD caused by defects in IBD-like genes or other loci, and to estimate their penetrance.

4.1.6 Another hypothesis for the aetiology of VEO-IBD

In addition to rare-variant studies, several analyses have been conducted to understand whether common genetic variation plays a role in paediatric manifestations of IBD. This started with studies [129, 164, 387] focusing on individual genes that had already been implicated in adult-IBD via GWAS (*NOD2*, *IL23R*, *ATG16L1*, *IRGM*, *NKX2-3*, *PTPN2*) [192, 374, 528] and exploring their specific effects in paediatric-IBD (at the time defined as <19 yrs of age). These early studies suggested similar genetic determinants acting in a polygenic fashion, with similar effect sizes and direction of effects, in both adult and juvenile phenotypes.

Two subsequent GWAS, focusing solely on children with a mean age of 12 years but less than 18 years, confirmed the association of *NOD2* and *IL23R* and identified children-specific loci overlapping with *TNFRSF6B*, *PSMG1* and *IL27* [222, 261]. However, all

of these associations have been subsequently replicated in a larger meta-analysis of adult IBD cohorts [153], leaving no paediatric-specific loci behind.

Rare penetrating variants of large effects have been hypothesised to be the most likely genetic contributors to VEO-IBD phenotypes and, as I outlined above, convincing causative defects have indeed been identified in some cases. VEO-IBD patients, however, have been relatively uncommon in paediatric GWAS studies and no single well-powered GWAS has been conducted for this specific age group. As such, common polymorphisms influencing susceptibility to VEO-IBD have not yet been comprehensively investigated, meaning the existence and extent of a contribution of common variants to VEO-IBD aetiology is unknown. Results from paediatric GWAS studies suggest polygenic variation may play a role in the genetic architecture of disease, as alluded to above. Perhaps more likely, VEO-IBD children might harbour a higher load of common alleles predisposing to adult-IBD, yet this possibility has not yet been explored in IBD patients in such a young age group. A weak but statistically significant (e.g. $P < 0.03$, $R^2 = 0.00741$) relationship between a polygenic risk score (derived from either 30 of the 32 CD loci [33] or 158 of the 163 IBD loci at the time [232]) and age of onset in CD has been documented in two paediatric-IBD cohorts (mean age: 12 yrs, maximum age: 17 or 19 yrs) [102, 129], which makes the polygenic burden hypothesis for VEO-IBD a timely investigation.

4.2 Aims

The research presented here describes a set of exome and genotyping-based analyses conducted in a multi-centre cohort of 146 VEO-IBD children. The overall aims of this project were fourfold. The first aim was to investigate whether pathogenic variants in known IBD-like inflammatory genes account for disease in this cohort. The second aim was to identify novel VEO-IBD causing genes. The third aim was to determine whether there is a significant enrichment of rare variants in biologically relevant genesets or pathways in VEO-IBD patients compared to controls. Finally, the last aim was to evaluate the role of common CD and UC-susceptibility alleles in the pathogenesis of VEO-IBD. Specifically, by generating polygenic risk scores based on the effects estimated from adult-IBD GWAS, I wanted to investigate whether VEO-IBD children harbor a higher load of such alleles when compared to a large collection of adult-IBD and healthy individuals.

4.3 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). This work was done in close collaboration with other colleagues at the University of Oxford, namely Professor Holm Uhlig and Dr Tobias Schwerd.

4.4 Methods

4.4.1 Patients

All investigations conducted in this work were part of an ethically approved protocol and were undertaken with the consent from patients and/or next of kin. A total of 146 patients were enrolled in this study. These patients were recruited by Professor Holm Uhlig as part of the COLORS study (COLitis of early Onset - Rare diseases withIN IBD). Samples were referred from participating centres in the UK (Cambridge, Liverpool, Great Ormond Street Hospital, Oxford and Edinburgh), Switzerland, Poland and Germany. The average age of onset of the affected children was 3.5 years (± 1.8) and ranged from 4 weeks to 7 years. Detailed demographics and immunophenotype characteristics of the VEO-IBD cohort are provided in Appendix **Table A.7**.

Briefly, 46% of patients were characterised as CD-like, 35% as UC-like and the remaining as U-IBD. 64% of CD-like patients had ileocolonic disease (i.e. involvement of both the terminal ileum and colon) and 84% of UC-like and U-IBD patients had pancolitis. 35% of all patients have been treated with anti-TNF- α therapy, and at least $\sim 71\%$ with immunomodulators. 16% of patients had undergone colectomy. There was a positive family history for IBD in at least one first-degree relative in $\sim 21\%$ (29/137) of the children. Finally, there were no identified genetic defects in any individual prior to enrollment in this study.

4.4.2 Controls

A total of 4,436 healthy individuals sequenced as part of the INTERVAL study (www.intervalstudy.org.uk/) were used here as controls, as they were sequenced in parallel to the VEO-IBD patients at the WTSI, using the same sequencing machines, chemistry and pull-down assays.

4.4.3 Exome sequencing and variant calling

Whole-exome sequencing of both cases and controls was performed and processed at the WTSI by the Sanger Institute Core Sequencing pipeline. Genomic DNA (1-3µg) extracted from blood 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 targeted sequencing (Agilent Technologies, Santa Clara, CA, USA; Human All Exon 50 Mb – ELID S04380110) according to the manufacturer’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 manufacture’s protocol. The Burrows-Wheeler Aligner [280] was used for alignment to the human reference genome build UCSC hg19/Grch37. Variants were first called at the single sample level using GATK Haplotype Caller (version 3.4) [116] and then joint-called across all cases and controls using GATK CombineVCFs and GenotypeVCFs at default settings. These steps were performed by the Human Genetics Informatics team at the WTSI.

4.4.4 Data quality control

Before embarking on downstream genetic analyses, I conducted a series of QC assessments on BAM and VCF files to ensure the sequencing data were of high quality at both the individual and variant levels.

Individual-level QC

1. **Detection of cross-sample contamination:** Cross-sample contamination due to technical issues during sample management, library-preparation and/or se-

quencing can reduce the accuracy of variant calls [233]. This can result in a higher number of variants being called, poor genotype estimates and inflated heterozygosity levels, leading to unexpected levels of relatedness between samples and, more importantly, downstream false positive signals. To investigate whether sequencing data showed evidence of contamination with another sample(s), I calculated the FREEMIX value using VerifyBAMID (version 1.1.0) [233]. This value is an estimation of the proportion of non-reference bases at reference sites, and thus gives an indication of the level of contamination of a given sample. To gain further evidence of contamination, I made use of additional metrics which I calculated from the data myself. One of such metrics quantified the fraction of heterozygous sites for which the ratio of reference to alternative reads was shifted away from the expected 50%, with the thought being that an unbalanced proportion of reads would likely indicate sites affected by contamination. Similar to Walter *et al* [507], an heterozygous site was termed to have an extreme frequency of alternative reads if their frequency were greater than 0.8 or lower than 0.15. I also calculated the global ratio of heterozygous to alternative-homozygous alleles (Het/Alt ratio), as well as the estimated number of relationships greater than third-degree relatives between samples. The latter is useful because contaminated samples will display a pattern of low-level relatedness to many people. To calculate the relationship between samples, I used PLINK2 [406], which estimates, in a pair-wise manner, the genome-wide proportion of alleles identical-by-descent (IBD) between samples, i.e. the IBD-sharing coefficient or IBD Pihat. Because parents and children obligatory share 0.5 of their genome in IBD [18], and because for each degree of pedigree relationship the expected IBD sharing decreases by a factor of 0.5, third degree-cousins were defined as samples with a IBD Pihat greater than 0.125 [227].

This investigation flagged 113 control samples that appeared to be contaminated (**Figure 4.5**) due to a higher FREEMIX fraction than the recommended value of 0.03 [233]. Most of these samples also represented outliers for the empirically derived fraction of skewed heterozygous sites (>0.035), exhibited Het/Alt ratios greater than 3 standard deviations (SD) from the mean, and appeared to have a much higher number of estimated relationships at IBD >0.125 for supposed unrelated individuals (**Figure 4.5**), all of which combined, supported the exclusion of these samples.

2. **Inferring ethnicity:** I evaluated the ethnicity of case and control exomes via a principal component analysis (PCA) using 1KG phase 3 individuals and following

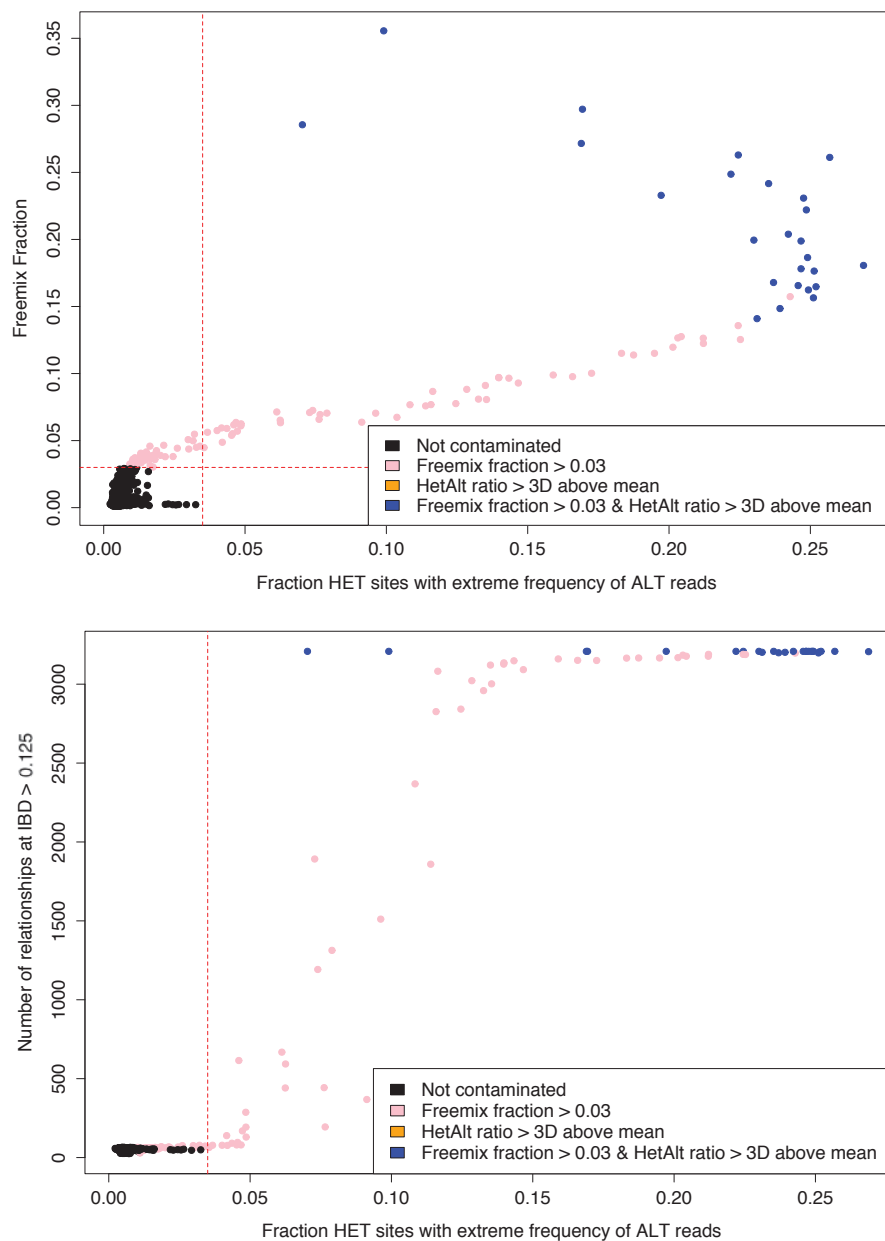


Figure 4.5 Contamination metrics. **A)** Fraction of heterozygous sites with extreme frequency of alternative reads vs. the freemix fraction. A heterozygous (HET) site was termed to have an extreme frequency of alternative (ALT) reads if the frequency of ALT reads were greater than 0.8 or lower than 0.15. Vertical dashed red line marks the empirically-derived threshold of 0.035 for the fraction of heterozygous sites with extreme frequency of alternative reads. The horizontal dashed red line marks the recommended threshold of 0.03 for freemix [233]. **B)** Fraction of heterozygous sites with extreme frequency of alternative reads vs. the estimated number of relationships at IBD > 0.1 (equivalent to third-degree relatives). Vertical dashed red line marks the empirically-derived threshold of 0.035 for the fraction of heterozygous sites with extreme frequency of alternative reads. The freemix value is an estimation of the proportion of non-reference bases at reference sites.

the same methodology outlined in the previous chapter. A total of 12,954 SNVs were used to construct the PCA.

This analysis identified a well defined cluster of samples (104 cases and 4,073 controls) that overlapped with the 1KG European populations and another smaller cluster of individuals of South Asian ancestry (21 cases and 68 controls). The remaining samples, most of which were cases, were of African, East Asian or mixed ancestries (**Figure 4.6**). This PCA analysis was intended to identify case and control groups, matched on ethnicity, that could be used in downstream case-control enrichment analyses (explained below). Thus, apart from the screening of IBD-like inflammatory genes outlined in section 4.4.6, which was performed for all VEO-IBD individuals regardless of ethnicity, all case-control analyses were restricted to only the European case-control group, or were performed within each of the two case-control groups (European and South Asian) and then meta-analysed.

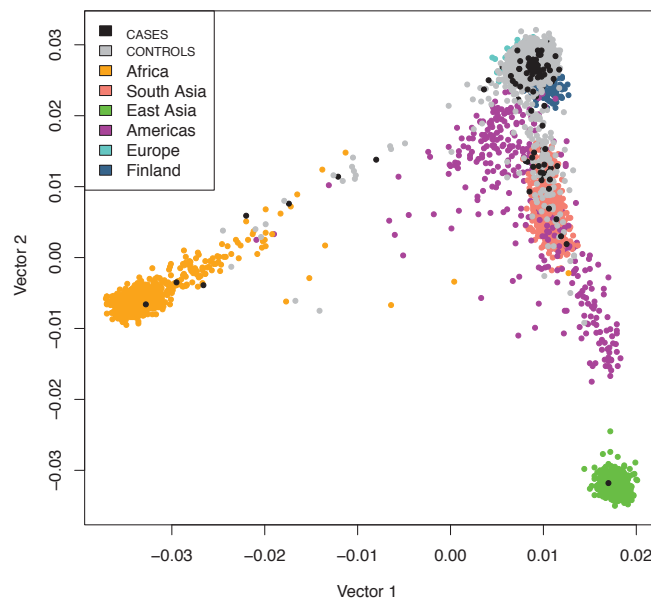


Figure 4.6 Principal component analysis (PCA) of VEO-IBD cases and INTERVAL controls with 1KG Phase 3 reference populations.

- 3. Identification of outlying samples:** This analysis aimed at identifying poorly performing samples for mean genotype quality (GQ), mean depth (DP) and genotype missingness rate, i.e. the proportion of non-called genotypes per sample.

As thresholds, I required a minimum mean GQ of 85.4, representing 3SD from the mean, a minimum DP of 40x (**Figure 4.7**), and a maximum genotype missingness rate of 0.002. The two latter thresholds were empirically derived by looking at the distribution of the data.

A total of 291 controls were outliers for at least one of these metrics, and one case had a considerably high rate (~ 0.07) of non-called genotypes (data not shown). All of these samples were removed from the dataset. Importantly, this analysis also revealed cases and controls were sequenced at mean depths of 69x and 53x, respectively, which represented a significant difference ($P\text{-value} < 2.2 \times 10^{-16}$) that needed to be corrected, if possible, with further variant-QC (explained below), or accounted for in downstream analyses.

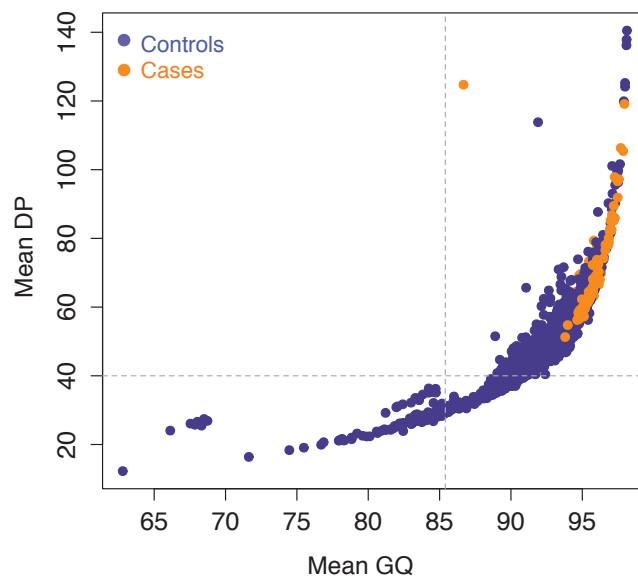


Figure 4.7 Mean genotype quality (GQ) and mean depth (DP) per sample. Grey dashed lines represent the applied thresholds of $GQ=85.4$ and $DP=40$. The GQ threshold represented 3SD from the mean and the DP threshold was empirically derived.

Finally, I also ensured samples were consistent at various population genetics metrics such as the Ts/Tv and Het/Alt ratios, which were within the expected values for exome datasets [71, 116, 187] (data not shown). As expected, non-EU samples revealed a higher number of variants called throughout the frequency spectrum. However I noted that a small number ($\sim 4\%$) of EU individuals, comprised of both cases and controls, harbored a greater than average rate of singletons variants (**Figure 4.8**), which could potentially represent sequencing

Chapter 5

A meta-analysis to map loci associated with age at IBD diagnosis

5.1 Introduction

As described in the previous chapter, large GWAS meta-analyses have uncovered a total of 231 genomic signals associated with the risk of IBD, and this has substantially advanced our understanding of the processes implicated in disease development. However, disease risk is only one aspect of disease biology, and the extent to which these (or novel) association signals also influence other aspects of disease, such as disease severity, disease location, response to treatments, or age at disease onset, is still poorly understood. The identification of genes modulating these aspects of disease can be of great importance from a clinical standpoint [275], as it may ultimately have important implications for drug development, diagnosis testing and risk stratification.

5.1.1 The role of genetic variation in the age at IBD diagnosis

Contrary to other examples of complex diseases, such as Alzheimer disease and Parkinson disease [235, 284], the estimated heritability of age of onset of IBD has not yet been quantified. However, family studies have shown that age at disease diagnosis (ADD), an imperfect proxy for age at onset, is highly concordant ($r = 0.69$, $P = 0.0001$) within families [190, 380], suggesting genetic modifiers for IBD age at onset may indeed exist.

The first study aimed at identifying polymorphisms associated with age at diagnosis of CD and UC outside individual genes such as *NOD2*, focused on 332 known IBD-associated SNPs and 329 CD and 294 UC patients, respectively [93]. Using the age at diagnosis as a continuous trait, and by comparing the mean age between genotypes, the authors identified rs2076756 in *NOD2* to be associated with a younger age of onset for CD ($P = 0.0002$): patients with the AA wild-type genotype were diagnosed at 31.9 ± 1.23 years, AG heterozygotes at 25.6 ± 0.99 years and GG homozygotes at 22.6 ± 1.32 years. In addition, depending on the age subgroups further compared, SNPs in *POU5F1*, *TNFSF15* and *HLA-DRB1*501* were found to be associated with age of Crohn's disease diagnosis, and a variant in *LAMB1* with the age of UC diagnosis.

A much larger study conducted by Cleynen *et al* last year [87], made use of 16,902 CD and 12,597 UC patients genotyped on the Immunochip, a dense custom-design array of 195,806 polymorphisms located in 186 regions with known association with one or more of 12 immune-related diseases, including CD and UC [288, 375]. Apart from *NOD2*, none of the signals identified in the previous study replicated in this analysis, despite all being typed on the Immunochip. As new findings however, two loci (rs3197999 in *MST1* and rs2066847 in *NOD2*) achieved genome-wide significance for the association with age at CD diagnosis, and one SNP (rs3129891) in the major histocompatibility complex (MHC) was found to be associated at genome-wide significance with age at UC diagnosis. Together, these findings confirmed that the general timing of CD and UC onset itself is influenced by genetic variation.

5.2 Aims

The aim of the research presented in this chapter was to build on previous findings of other colleagues, who identified variation in known or immune-related regions to be associated with age at IBD diagnosis, and conduct the first association analysis to date that interrogates the entire genome of $\sim 5,400$ CD and $\sim 4,400$ UC individuals to identify genetic modifiers of age at disease diagnosis.

5.3 Methods

5.3.1 Association analyses

The association analysis for age at disease diagnosis was conducted using the UKIBDGC CD and UC cases for which information on age at disease diagnosis was available (5,403 CD and 4,490 UC individuals). As mentioned in the previous chapter, these samples originally came from three independent GWAS studies (GWAS1, GWAS2 or GWAS3) genotyped on different platforms or from a low-coverage whole-genome sequencing study (IBDSeq, **Table 5.1**). To leverage the whole-genome sequencing data, and thus survey lower frequency variants ($1\% < \text{MAF} < 5\%$) not well represented in the GWAS arrays, the reference panel containing haplotypes drawn from the low-coverage whole-genome IBD samples ($N=4,445$), as well as the UK10K ($N=3,652$) and 1000 Genomes (1KG) Phase 3 control sequences ($N=2,505$) were imputed into the GWAS cohorts [110, 295].

Studies	CD samples	UC samples
GWAS1	1,116	.
GWAS2	.	1,060
GWA3_CD	2,683	.
GWAS3_UC	.	2,165
IBDSeq_CD	1,604	.
IBDSeq_UC	.	1,265

Table 5.1 UKIBDGC sample breakdown per contributing study. The studies that contributed samples to the UKIBDGC dataset are given. Total of 5,403 CD and 4,490 UC samples.

To test for association between age at disease diagnosis and genetic variation, I carried out separate linear regression analyses within each of the three studies of each trait (CD and UC, **Table 5.1**). I tested all the variants that passed all UKIBDGC quality control procedures [110, 295] after excluding sites with $\text{MAF} < 1\%$ (in UKIBDGC control samples only) and $\text{INFO} < 0.4$, as recommended by Marchini *et al* [309]. The MAF threshold of 1% was chosen because the power to detect single-variant associations below this frequency is very low at current sample sizes [295] and because false-positives will be increased below this frequency threshold as imputation does not work as effectively at rare variant sites [309]. The INFO threshold of 0.4, as routinely used in GWAS [286, 309, 507, 539, 540], was chosen to minimise false positive associations arising from high genotype uncertainty post-imputation. **Table 5.2** lists the total number of variants tested in each study dataset.

Studies	CD SNPs	UC SNPs
GWAS1	8,123,580	.
GWAS2	.	8,113,309
GWA3_CD	8,141,056	.
GWAS3_UC	.	8,140,904
IBDSeq_CD	7,991,854	.
IBDSeq_UC	.	7,955,914

Table 5.2 Number of high-quality SNPs tested in each UKIBDGC study. The studies that contributed samples to the UKIBDGC dataset are listed, along with the number of SNPs tested in each association analysis for age at CD or UC diagnosis. High-quality SNPs were defined as those that passed all UKIBDGC QC procedures, and had MAF >1% and INFO >0.4. For details of QC procedures see De Lange *et al* [110] and Luo *et al* [295].

Because all the UKIBDGC samples were imputed, the probabilistic nature of the genotypes meant the association testing needed to take the uncertainty of the imputed genotypes into account. To do so, I used the regression framework implemented in SNPTEST v2 [309]. This model uses well-established statistical theory for missing data problems, in which an observed data likelihood is used where the contribution of each possible genotype is weighted by its imputation probability. The test was run assuming an additive genetic model [85], where the effect is increased by β -fold for genotype Aa (or 1) and by 2β -fold for genotype AA (or 2), and contained the first 10 principal components for ancestry to adjust for potential population structure (PCs were calculated and provided by Katie De Lange):

$$E(Y_i) = \mu + \beta_G * G_i + \eta z_i + \varepsilon \quad (5.1)$$

where $E(Y_i)$ denotes the phenotypic value for each individual, μ denotes the baseline effect for the non-effect genotype, β_G denotes the estimated effect due to each copy of the effect allele, G_i denotes the observed genotype for each individual (coded as 0, 1 or 2, according to the number of copies of the effect allele), z is a matrix of covariates and ε is a residual error.

When performing a regression on a continuous rather than in a binary phenotype (i.e. case-control), the quantitative phenotype is generally either standardized or quantile normalized to fit a normal distribution [37, 507]. I decided to use the quantile normalization available in SNPTEST v2 in this case, because it was the transformation previously used in the Immunochip study reported by Cleyne *et al* [87], and because I wanted to compare the effect size estimates between the two studies.

5.3.2 Meta-analysis within CD and UC studies

After performing association analysis for each SNP in each study individually, I conducted a meta-analysis to obtain pooled estimates of the effect of each SNP on the age at disease diagnosis across all studies of each trait (CD and UC). I used the fixed-effects methodology implemented in METAL [520], in which the study-specific effect estimates and standard errors derived from the regression analysis of each cohort are combined in an inverse variance-weighted fixed effects meta-analysis, the most powerful and commonly used method for discovering phenotype-associated SNPs [130, 390]. METAL assumes a given allele exerts similar effects across datasets, and calculates the combined allelic effect (B) across all studies at each marker as:

$$B = \frac{\sum_{i=1}^k \omega_i \beta_i}{\sum_{i=1}^k \omega_i} \quad (5.2)$$

where k is the number of studies, β_i is the effect size from study i and ω_i represents the inverse of the variance of the estimated allelic effect, which is given by $SE(\beta_i)^2$.

A fundamental principle of meta-analysis is that all studies tested the same hypothesis using near-identical procedures for QC, covariate adjustment and statistical test, for example, all of which were the case here.

5.3.3 Meta-analysis for IBD

The analysis for age-at-disease diagnosis for IBD was conducted by meta-analysing summary statistics from the CD and UC meta-analysis, similar to Cleynen *et al* [87]. This approach is also generally followed because CD and UC have slightly different age distributions (mean CD age: 27 yrs; mean UC age: 36 yrs, **Figure 5.1**), which would look bimodal if the samples were combined.

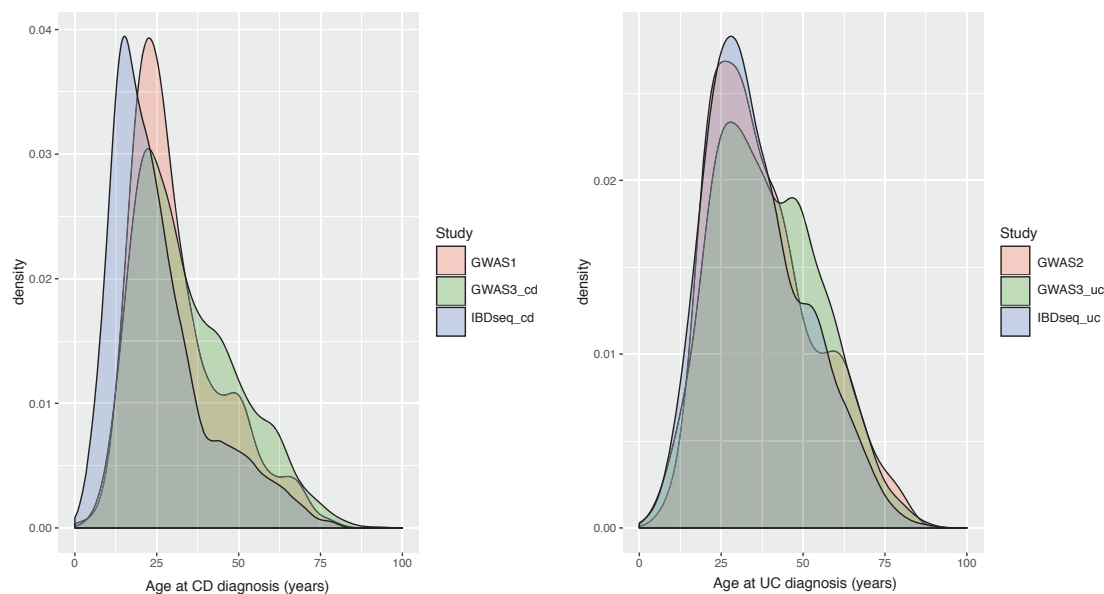


Figure 5.1 Distribution of age at disease diagnosis across the different studies. Prior to association testing, the quantile normalization was performed so that the age distributions within the CD and UC studies were forced to have the same statistical properties (mean and standard deviation), a procedure that is normally conducted when different studies are to be meta-analysed.

5.3.4 Post meta-analysis quality control

To control for between-study/traits heterogeneity in effect sizes, I excluded SNPs for which the I^2 metric was greater than 90%, similarly to what others have done [290]. Briefly, the I^2 measures the degree of inconsistency in the studies' results, and describes the percentage of total variation across the studies that is due to heterogeneity rather than chance [130, 205]. As additional filtering post-meta-analysis, I excluded SNPs that: 1) were present solely in one study/trait out of all that were meta-analysed, 2) the meta-analysis P -value (P_{META}) was greater than the individual studies' P -values and 3) the INFOs of the studies driving the signal (at $\alpha = 0.05$) were < 0.6 .

5.3.5 Power to detect previous ImmunoChip signals

Finally, I conducted an analysis to determine the statistical power of my study to detect, at genome-wide significance, the previous genome-wide signals associated with either the age at CD or UC diagnosis reported in the ImmunoChip study of Cleynen *et al* [87]. Because power is determined by both the frequency and the effect size of the risk allele [22], I calculated the power for each variant separately using the method

derived by Sham and Purcell [443], which assumes the non-centrality parameter (NCP) of the chi-squared distribution for a single SNP under the additive genetic model is:

$$NCP = N * h^2 \quad (5.3)$$

where N is the total number of studied individuals and h^2 is the fraction of phenotypic variance explained by the marker, which I calculated as follows:

$$h^2 = 2p(1 - p)\beta^2 \quad (5.4)$$

where p is the frequency of the effect allele assuming Hardy-Weinberg equilibrium and β is its additive effect, defined as the regression coefficient of the linear model [545].

5.4 Results

Association of variants with age at disease diagnosis of CD and UC was tested using linear regression of the quantitative phenotype (**Figure 5.1**), in a total of 5,403 and 4,490 UKIBDGC cases, respectively, each split across three different studies.

Figure 5.2 shows the QQ plot for comparison of the observed and expected P-values distributions for the average of 9 million variants with $MAF > 0.5\%$ that were tested in each of the six studies. All QQ plots demonstrate evidence of genetic associations at the tail of the distribution. Importantly, the QQ plots demonstrate no evidence of population stratification, as none exhibit a global excess of higher observed p-values than expected throughout the distribution, and as measured by the inflation factor ($\lambda \sim 1$ in all studies). The λ value represents the degree of deviation from the expected distribution and was calculated as the ratio of the median association test statistic over the theoretical median test statistic of the χ^2 distribution (0.675²).

The QQ plots resulting from the meta-analyses combining the effect sizes across the studies within each disease entity (CD, UC and IBD) are illustrated in **Figure 5.3**. No genome-wide significant signals remained after the QC procedure applied post meta-analysis (**Figure 5.3**), however, a total of four signals showed suggestive levels of association ($P_{META}\text{-value} \leq 5 \times 10^{-7}$) with either CD, UC or IBD (**Table 5.3**). All of

these signals were driven by common variants with MAF >1% and were present in all meta-analysed studies, therefore being supported by different genotyping platforms. Moreover, all signals also showed consistency in direction and magnitude of effects across all studies within each trait. I will describe these four associations in greater detail below, however these findings should not be taken as definitive, as additional validation in independent and larger studies will be necessary. Approximately 73% of the associations with borderline significance are successfully replicated when additional data are acquired [370], therefore some of the signals I report here likely contain true associations that may be replicated in future analyses.

Disease	Signal	Locus	REF/EA	META			I^2	INFOs	EAF*	rsID	Genes in region
				Direction	β (SE)	P-value					
CD	1	2:28606778	C/T	---	-0.10 (0.01)	1.89×10^{-7}	47.8	0.96; 0.96; 0.99	0.450	rs2879179	<i>FOSL2</i> (intron), <i>BRE</i> , <i>PLB1</i> , <i>PPP1CB</i> (+8)
		2:28608504	C/T	---	-0.10 (0.01)	1.95×10^{-7}	55.2	0.96; 0.96; 0.99	0.448	rs4666067	
		2:28612213	G/C	---	-0.09 (0.01)	4.16×10^{-7}	5.4	0.96; 0.97; 0.99	0.493	rs1509396	
		2:28623047	T/C	---	-0.09 (0.01)	3.21×10^{-7}	30.4	0.96; 0.99; 0.99	0.476	rs4617998	
UC	2	1:245581534	C/T	+++	0.10 (0.02)	3.60×10^{-7}	47.4	0.97; 0.97; 0.99	0.501	rs1148919	<i>KIF26B</i> (intronic), <i>SMYD3</i> , <i>EFCAB2</i> (+1)
		22:40382249	T/C	+++	0.12 (0.02)	2.23×10^{-7}	0	0.90; 0.91; 0.98	0.295	rs2958654	<i>FAM83F</i> , <i>GRAP2</i> , <i>ENTHD1</i> , <i>TNRC6B</i> (+9)
		22:40389007	T/G	+++	0.12 (0.02)	2.24×10^{-7}	0	0.89; 0.92; 0.99	0.285	rs2958658	
UC	3	22:40390238	G/A	+++	0.12 (0.02)	2.82×10^{-7}	0	0.89; 0.92; 0.99	0.295	rs28607928	
		20:29904377	G/A	++	0.11 (0.02)	1.02×10^{-7}	0	0.77; 0.79; 0.91 0.84; 0.82; 0.85	0.165	rs6141273	<i>DEFB115</i> , <i>DEFB119</i> , <i>DEFB116</i> (+17)
IBD	4	20:29904377	G/A	++	0.11 (0.02)	1.02×10^{-7}	0	0.77; 0.79; 0.91 0.84; 0.82; 0.85	0.165	rs6141273	<i>DEFB115</i> , <i>DEFB119</i> , <i>DEFB116</i> (+17)

Table 5.3 Genetic loci associated at suggestive significance ($P_{\text{META-value}} \leq 5 \times 10^{-7}$) with age at CD, UC or IBD diagnosis. REF: reference allele; EA: effect allele; Direction denotes either the positive (+) or negative (-) effect of the effect allele on the phenotype and it includes the direction of the effect in the three independent studies (ordered by GWAS1, GWAS3, IBDSeg for CD and GWAS2, GWAS3 and IBDSeg for UC) or, in the case of IBD, in the two traits (CD and UC); SE: standard error around the beta estimate; EAF: effect allele frequency calculated from the largest control group (GWAS3, N=9,454 individuals). I^2 measures the degree of inconsistency in the studies' results. INFOs correspond to the INFO of the individual studies that were meta-analysed (studies ordered similarly as above for CD and UC; for IBD I give the INFOs of all CD and UC studies). Location given by Ensembl VEP v75. Table is sorted by genomic location within each disease entity. All variants represent common variants and all show consistent direction of effects across studies/traits.

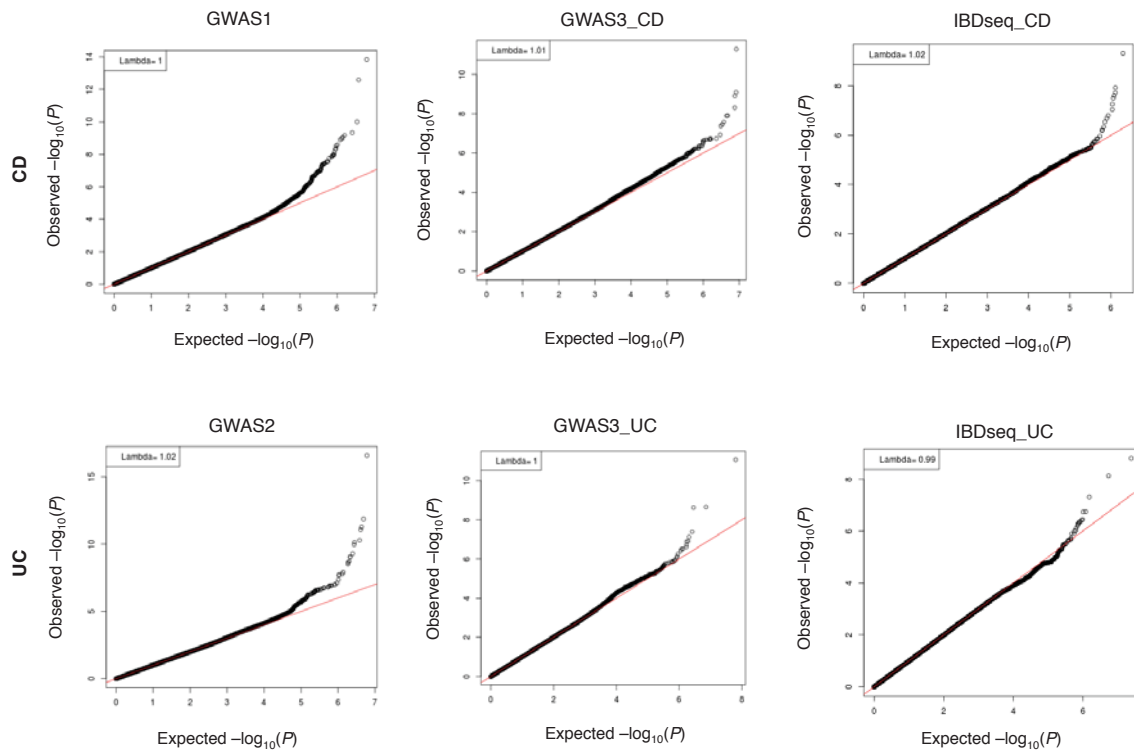


Figure 5.2 Quantile-quantile plots of the individual CD and UC association studies. The red line shows the distribution under the null hypothesis, where the observed p-values correspond exactly to the expected p-values. The inflation at the end of the tail reveals there is evidence of genetic associations. There is no evidence of inflation caused by population stratification, as all lambda values (λ) are close to 1 in all studies. Variants included in the association tests and the QQ plots are those that passed all UKIBDGC QC procedures (see [110, 295]), and had MAF $>0.5\%$ (derived from controls of each study) and INFO >0.4 .

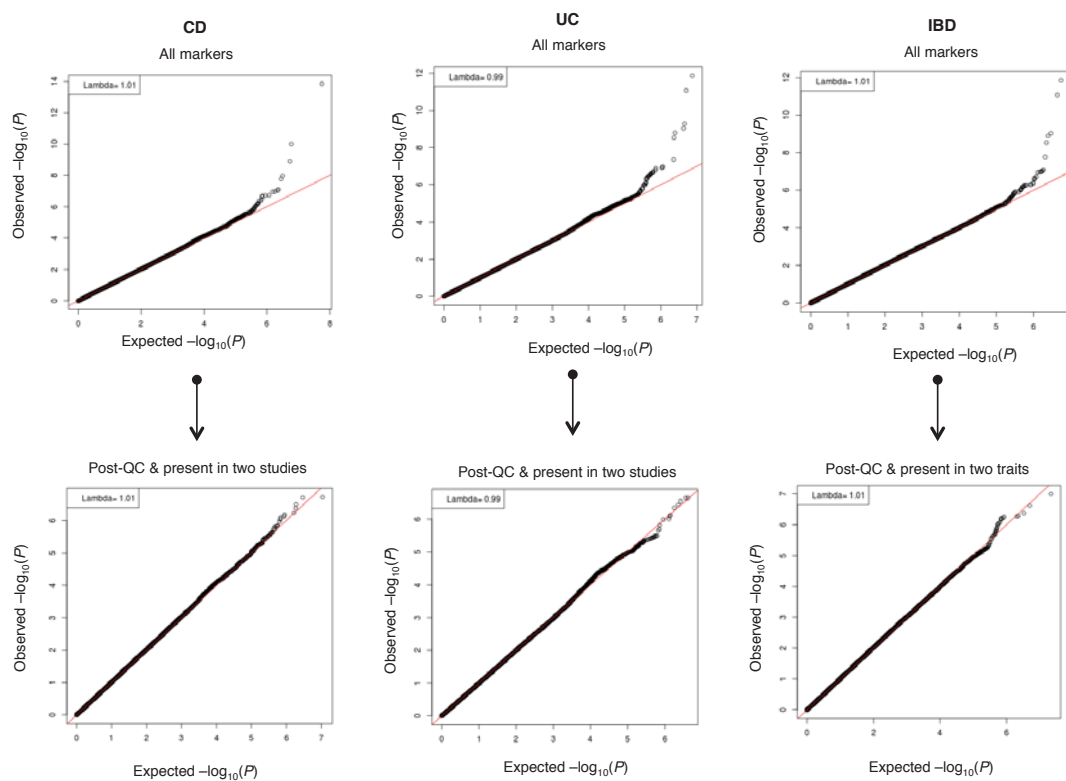


Figure 5.3 Quantile-quantile plots of the meta-analysis results for CD, UC and IBD. The red line shows the distribution under the null hypothesis, where the observed p-values correspond exactly to the expected p-values. The inflation at the end of the tail reveals there is evidence of genetic associations. There is no evidence of inflation caused by population stratification, as all lambda values (λ) are close to 1 in all studies. Variants included in QQ plots are those that passed all UKIBDGC QC procedures (see [110, 295]), had MAF $>0.5\%$ (derived from controls of each study), imputation INFO >0.4 and displayed between-study heterogeneity in effect sizes (I^2) below 90%. Note in the top panel, the significance of each marker is not necessarily supported by all the meta-analysed datasets, i.e. the p-value might be driven by a single dataset. Bottom panel illustrates the markers that are present, post meta-analysis QC, in two of the three meta-analysed studies (in the case of CD and UC) and by both traits (CD and UC) in IBD. There is no evidence of genome-wide significant signals.

5.4.1 Suggestive association for the age at CD diagnosis

Three common frequency (EAF=45%) intronic variants in *FOSL2* were associated at suggestive levels of significance with age at CD diagnosis. The lead SNP driving this signal (rs2879179, $P_{\text{META_CD}} = 1.89 \times 10^{-7}$, **Table 5.3**) was associated with a decrease in the age at CD diagnosis with a per-allele effect beta of -0.10 (SE=0.01). The regional plot for this signal (**Figure 5.4**), including all the SNPs within 500kb on either side of this variant, revealed multiple SNPs with varying degrees of association due to local LD patterns, which decrease the chance that genotyping artifacts are driving this suggestive association. In addition, the genotyping clusters for this SNP in all UKIBDGC individuals were well defined (**Figure 5.5**), which again argues against poor genotyping at this SNP.

Interestingly, the *FOSL2* locus has been previously reported by Jostins *et al* [232] and Liu *et al* [290] to be associated with the risk of IBD via rs925255, a SNP in high LD ($r^2 = 0.71$) with rs2879179. Both studies reported P-values of 2.67×10^{-15} , and 1.07×10^{-16} for rs925255, respectively, and ORs of 1.09 (CI: 1.09 - 1.16) and 1.11 (CI: 1.09 - 1.12). The current UKIBDGC-GWAS analysis [110] replicated that signal and identified another lead SNP (rs11677002) in perfect LD ($r^2 = 1$) with that of Jostins and Liu for that IBD-risk association (**Figure 5.4**), which is actually stronger in CD ($\beta = -0.14$, SE=0.02) than in UC ($\beta = -0.07$, SE=0.02). More interestingly, this latter study also showed that rs2879179, here associated with the age at CD diagnosis, is also associated with the risk of developing IBD ($P = 2.8 \times 10^{-9}$), again with a stronger effect on CD ($P = 2.2 \times 10^{-12}$, **Figure 5.4**). This cross-phenotype association at the same locus is intriguing and is reminiscent of what is known for *NOD2*, which has the largest effect in susceptibility for CD while also being associated with an earlier age of CD onset (rs2066847, p.L1007fsX, $\beta = -0.17$, $P = 2.04 \times 10^{-16}$) [87].

To contrast my *FOSL2* finding with the previous Immunochip ADD analysis, I inspected whether this locus showed nominal significance (at $\alpha = 0.05$) in the summary results kindly provided by Dr Isabel Cleynen. rs2879179 was not directly typed in the Immunochip. In fact, this whole locus was not densely represented on the chip because its association with IBD-risk was unknown at the time of design, meaning it was not included in the fine-mapping regions that were typed on the chip. Still, subsequent inspection for possible proxies of rs2879179, revealed one marker with an $r^2 > 0.7$, and the SNP showed nominal significance ($P = 0.03$). The evidence of replication is perhaps not as strong as we may expect given the LD between the two variants, however the

poor representation of SNPs in higher LD with my SNP in the Immunochip prevents me to make further comparisons.

FOSL2 is part of a family of transcription factors composed of three other members (*FOSL1*, *FOSB*, *FOS*) that together form the AP-1 (activator protein-1) transcription factor complex. Amongst a plethora of functions, such as regulation of cell proliferation, death, survival and differentiation [444], AP-1 has been shown to be a positive regulator of inflammation, containing transcriptional regulator binding sites for numerous inflammatory mediators (IL6, IL8, TNF-*a*), and capable of binding to promoters independently of NF- κ B [510].

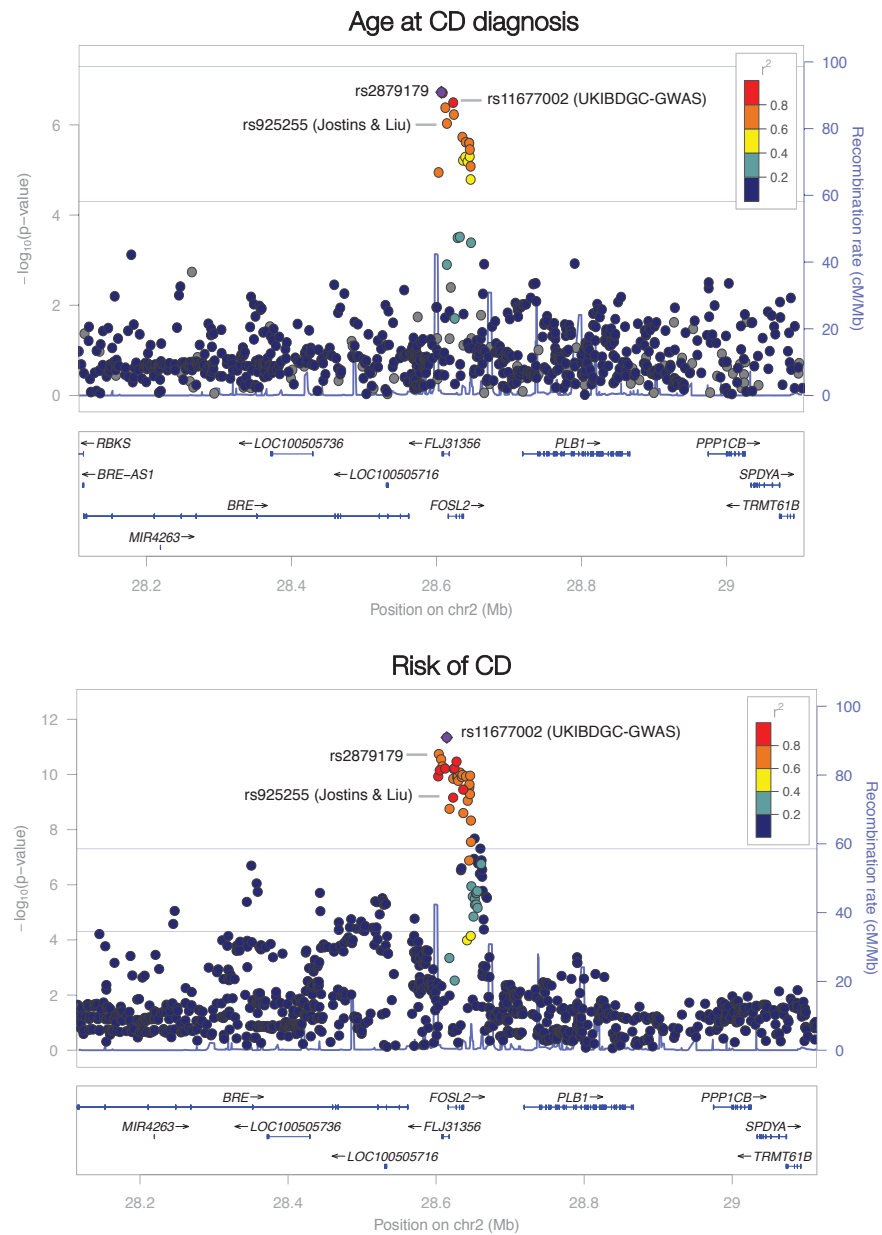


Figure 5.4 **A)** Regional association plot for 2p28, including the best SNP (rs2879179) for age at CD diagnosis (GWAS3_CD dataset). Plot also illustrates the SNPs of this locus that were previously reported to be associated with the risk of IBD in the analyses of Jostins Liu (rs925255) and UKIBDGC-GWAS (rs1167702). **B)** Regional association plot for 2p28, but for the SNPs associated with risk of CD in the UKIBDGC-GWAS analysis (GWAS3_CD dataset, data provided by Dr Loukas Moutsianas). Plot also shows where my SNP (rs2879179) and Jostins Liu (rs925255) lie in this associated signal. The lead SNP for age at CD diagnosis is also associated ($P = 2.2 \times 10^{-12}$) with susceptibility to CD. The $-\log_{10}$ P-values for the associated SNPs are shown on the upper part of each plot. SNPs are colored based on their r^2 with the labeled lead SNP (purple symbol), which has the smallest P-value in the region. r^2 was calculated from the 1KG Phase 3 European panel. The bottom section of each plot shows the fine scale recombination rates estimated from individuals in the HapMap population, and genes are marked by horizontal blue lines. Genes within the recombination region of the hit SNPs are labeled. Figures were generated using LocusZoom [404].

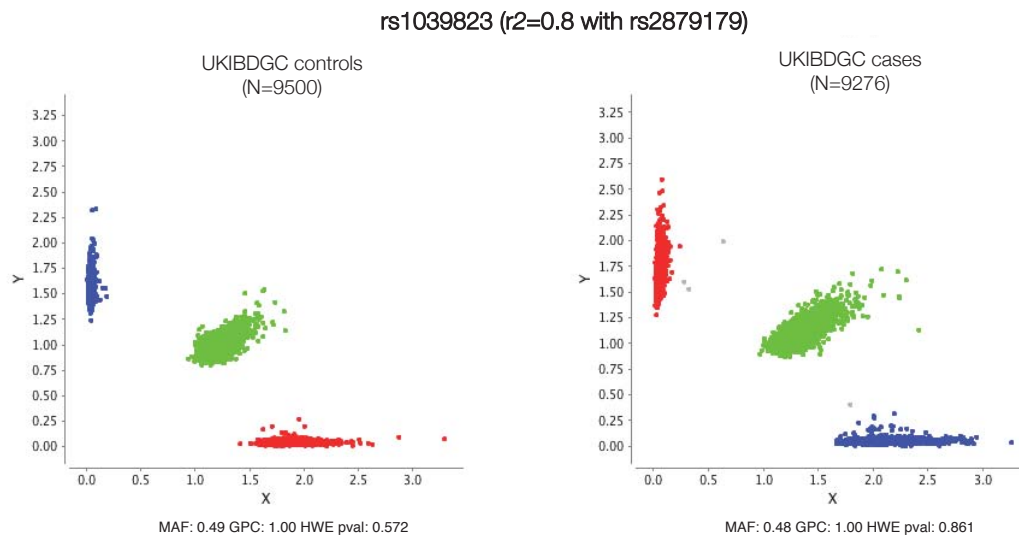


Figure 5.5 Genotype cluster plot for a directly genotyped proxy of rs2879179. The plots represent the raw intensity data from the probes used during genotyping for each UKIBDGC individual. Because rs2879179 was imputed, a proxy (rs1039823) in high LD ($r^2=0.8$) was chosen for plotting. The plot demonstrates genotypes are of high quality, with genotypes of the same class clustering together and with clusters consistent across UKIBDGC case and control groups. Plot generated by Daniel Rice using Evoker [337].

5.4.2 Suggestive associations for the age at UC diagnosis

Two signals driven by rs2958654 and rs1148919, respectively, were associated at suggestive significance with an increase in the age at which UC presents (**Table 5.3**). Both SNPs had high INFO scores (>0.89) in all meta-analysed datasets, and rs2958654 showed no evidence of heterogeneity between studies ($I^2=0$, **Table 5.3**). The genes spanning the two associated regions are illustrated in **Figure 5.6**. Again, the regional association plot demonstrates multiple correlated markers with comparable evidence of association, suggesting the signals are less likely to represent type-I errors. rs2958654 and all its proxies were imputed SNPs hence cluster plots could not be generated. The cluster plot for a proxy of rs1148919, with $r^2=0.86$, showed well defined genotypes (data not shown).

While the *FOSL2* signal described above overlaps with a gene with strong biological candidacy, the relevance of the loci located within these two associated regions is unclear. The closest gene to rs2958654 encodes a protein of unknown function (FAM83F) whereas rs1148919 is located in an intronic sequence of *KIF26B*, an intracellular motor protein involved in microtubule-based processes [197].

The two SNPs identified herein were not directly typed in the ImmunoChip study of Cleyne *et al* [87], nor were proxies with sufficient and informative LD ($r^2 > 0.1$), which precludes comparisons between the two studies.

5.4.3 Suggestive association for the age at IBD diagnosis

The search for genetic determinants for age at IBD diagnosis was conducted by meta-analysing the results from the CD and the UC meta-analyses, similar to Cleynen *et al* [87]. This approach yielded one common, imputed variant (rs6141273) with suggestive association for an increase in the age at IBD diagnosis (**Table 5.3**, $\beta=0.11$, $SE=0.02$, $P_{META_CD} = 6.7 \times 10^{-6}$ and $P_{META_UC} = 3.0 \times 10^{-3}$).

rs6141273 is located in a region near the centromere of chromosome 20, where a cluster of evolutionarily conserved β -defensins lie (**Figure 5.7**). These proteins are produced at a variety of epithelial surfaces, including the intestinal mucosa, and are predominantly considered to act as antimicrobial peptides that activate the NF- κ B pro-inflammatory pathway [411].

Similarly as above, the associated SNP identified herein was not directly typed in the ImmunoChip study of Cleynen *et al* [87], nor were proxies with sufficient and informative LD ($r^2 > 0.1$), which precludes comparisons between the two studies.

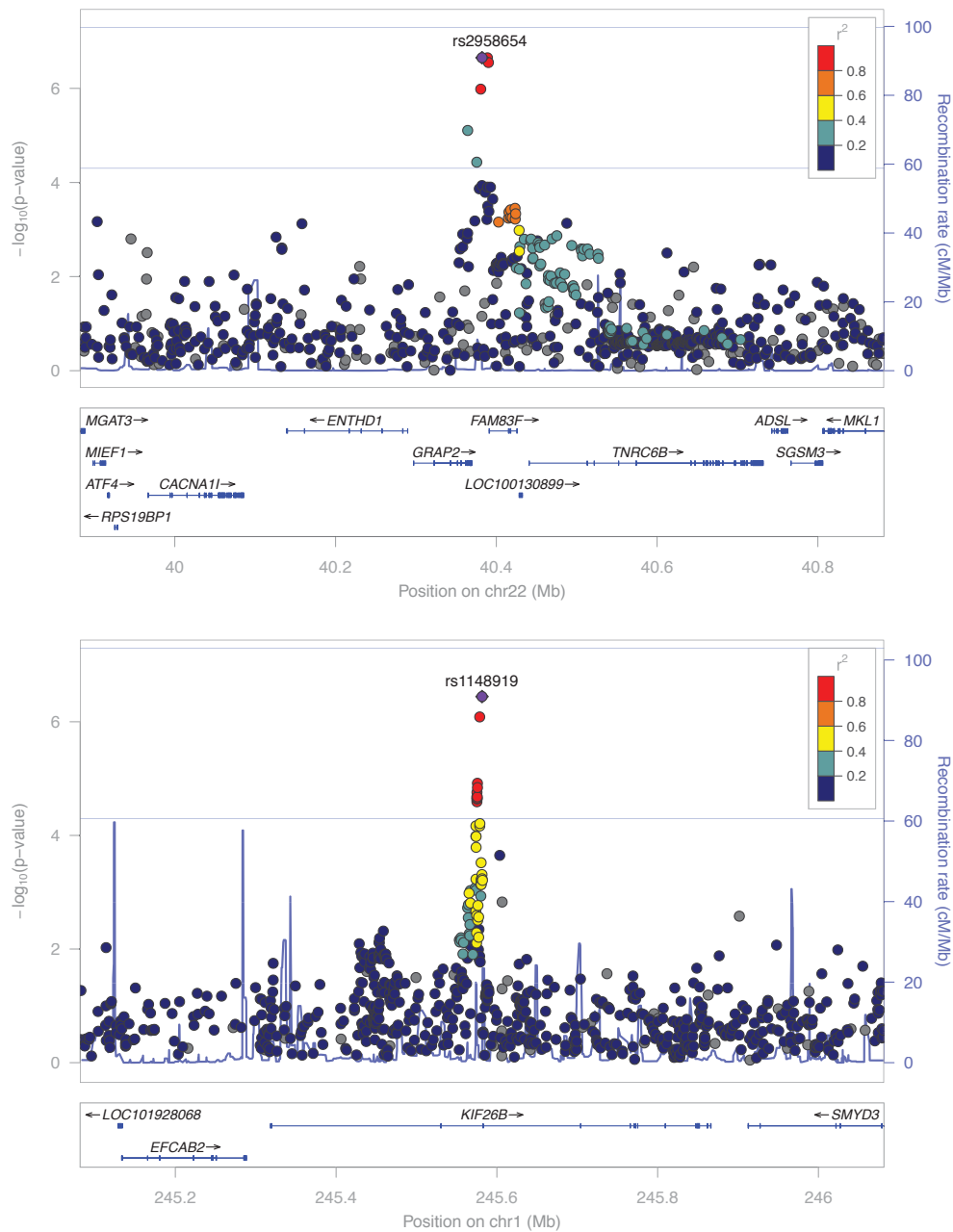


Figure 5.6 Regional association plots for the common frequency signals with suggestive association with age at UC diagnosis (GWAS3_CD dataset). **A)** rs2958654. **B)** rs1148919. The $-\log_{10}$ P-values for the associated SNPs are shown on the upper part of each plot. SNPs are colored based on their r^2 with the labeled lead SNP (purple symbol), which has the smallest P-value in the region. r^2 was calculated from the 1KG Phase 3 European panel. The bottom section of each plot shows the fine scale recombination rates estimated from individuals in the HapMap population, and genes are marked by horizontal blue lines. Genes within the recombination region of the hit SNPs are labeled. Figures were generated using LocusZoom [404].

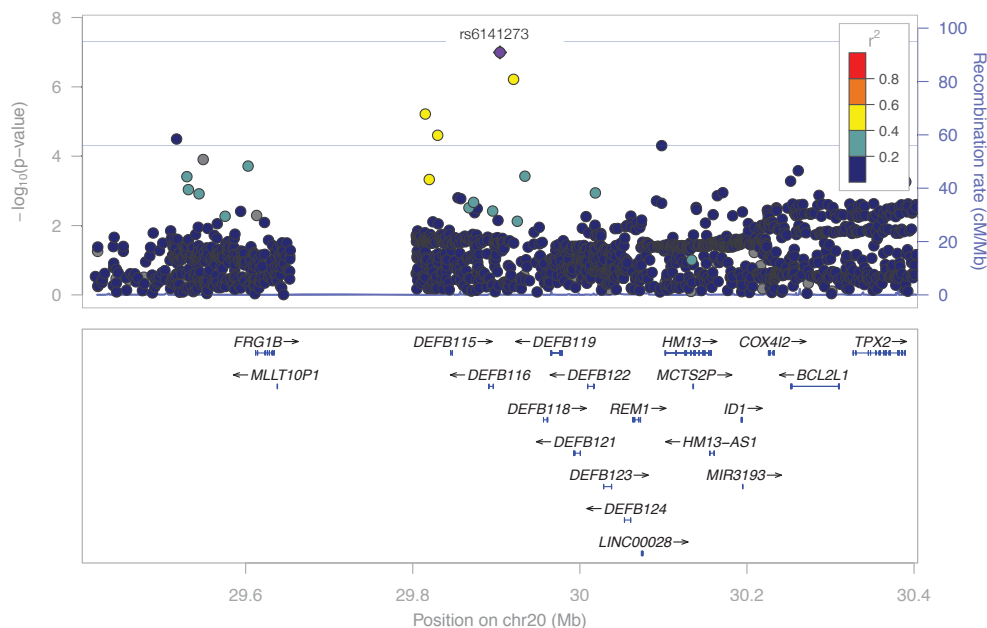


Figure 5.7 Regional association plots for the common frequency signal (rs6141273) with suggestive association with age at IBD diagnosis (GWAS3_CD dataset). The $-\log_{10}$ P-values for the associated SNPs are shown on the upper part of each plot. SNPs are colored based on their r^2 with the labeled lead SNP (purple symbol), which has the smallest P-value in the region. r^2 was calculated from the 1KG Phase 3 European panel. The bottom section of each plot shows the fine scale recombination rates estimated from individuals in the HapMap population, and genes are marked by horizontal blue lines. Genes within the recombination region of the hit SNPs are labeled. Figures were generated using LocusZoom [404].

5.4.4 Comparison with the previous ADD Immunochip study

As mentioned in the introduction to this chapter, three loci have been reported to be associated, at genome-wide significance, with either the age of CD (*NOD2* and *MST1*) or UC (*MHC*) diagnosis. These associations were uncovered in a well-powered Immunochip-based GWAS study comprised of 16,902 CD and 13,597 UC patients [87]. As none of these regions featured in my list of suggestive signals, I hypothesised this could potentially be attributable to the much smaller sample size available here (CD: 5,403; UC 4,490), something that would necessarily hinder the statistical power of this study, i.e. the probability of rejecting the null hypothesis when the alternative hypothesis is true [443].

According to my power calculations, the UC meta-analysis conducted here (N=4,490) was underpowered to detect the *MHC* association with age at UC diagnosis at genome-

wide significance (power = 2.3×10^{-7}). The same was true for my CD meta-analysis (N=5,403), which had only 0.3% and 1.5% power to detect an association, at an $\alpha = 5 \times 10^{-8}$, with *NOD2* and *MST1*, respectively (**Table 5.4**). Out of these two loci, only *NOD2* achieved nominal significance ($P_{\text{META}} = 2.08 \times 10^{-4}$), whereas *MST1* did not. A closer look at the *NOD2* signal in my data, which showed no significant evidence of heterogeneity of effect across the studies ($I^2 = 0$), revealed my point estimate of the effect size was consistent with the previous finding, as it fell within the 95% confidence intervals reported in the more highly powered study (**Figure 5.8**). The reason why *MST1* did not show nominal significance is unclear, however the associated alleles for this region, as well as for *MHC*, showed the same direction of effect as previously reported.

Disease	rsID	Locus	ImmunoChip data				Current study			
			Effect allele	EAF	P-value	β (SE)	h^2	P-value	β (SE)	POWER
CD	rs3197999	3:49721532				-0.07			-0.03	
		<i>MST1</i>	A	0.281	2.37×10^{-8}	(0.01)	0.20%	0.097	(0.02)	1.5%
CD	rs5743293	16:50763778				-0.17			-0.16	
		<i>NOD2</i>	GC	0.024	2.04×10^{-16}	(0.02)	0.14%	2.08×10^{-4}	(0.04)	0.3%
UC	rs3129891	6:32415080				-0.01			-0.02	
		<i>MHC</i>	A	0.209	1.43×10^{-8}	(0.02)	0.003%	0.323	(0.03)	2.3×10^{-7}

Table 5.4 Power to detect previous loci associated with age at CD and UC diagnosis. Table lists the three loci previously detected at genome-wide significance in Cleyne *et al* [87] to be associated with either CD or UC age at diagnosis. EAF: effect allele frequency in control samples of my study (GWAS3, N=9,459); SE: standard error of the effect size (β); h^2 : phenotypic trait variance explained by the SNP. Power calculated for an $\alpha = 5 \times 10^{-8}$.

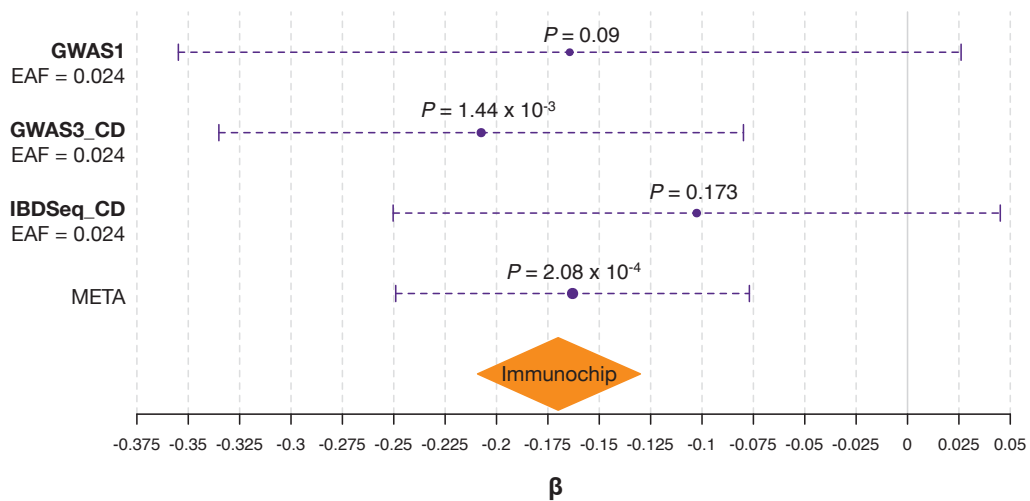


Figure 5.8 Effect size estimations for *NOD2* rs5743293 across the studies.

5.5 Discussion

To identify genetic modifiers for the age of onset of CD, UC and IBD, I conducted three GWAS studies, followed by meta-analyses, using the age at diagnosis reported across a total of 5,403 CD and 4,490 UC UKIBDGC patients. This study is the first to conduct such an analysis in a genome-wide manner, with two previous reports focusing either on 332 known IBD-risk loci [93], or on 186 known immune-associated regions that were included in the ImmunoChip platform at the time of design [87].

5.5.1 The advantage of imputation

While this study is one order of magnitude smaller than the previous ImmunoChip analysis [87], which used 16,902 CD and 12,597 UC patients, a much larger set of SNPs were available for testing after the imputation effort conducted by the UKIBDGC (9 million vs. 156,154). This imputation procedure, leveraging $\sim 10,600$ whole-genome sequences drawn from IBD as well as from healthy individuals included in the UK10K and the 1KG projects, also meant that I could examine a much larger frequency space than previous studies, with about $\sim 40\%$ of the total sites representing low-frequency variants with MAFs between 1% to 5%. In contrast, the ImmunoChip was designed using the early 1KG Pilot data, which has incomplete coverage particularly of lower frequency variation [1, 375]. The imputation step conducted here therefore clearly demonstrates the value of incorporating genomes of IBD patients and UK population controls and using that information to build a specific reference panel to which independent GWAS samples can be imputed in. The incorporation of UK10K haplotypes in the imputation panel was particularly beneficial, as this resource has been demonstrated to greatly increase the accuracy and coverage of low-frequency and rarer variants compared to existing panels such as the 1KG, because it contains 10-fold more European samples [507].

5.5.2 The pitfall and advantage of my genome-wide analysis

This study was underpowered, at current sample sizes, to identify associations statistically significant at the genome-wide level. This reflects a disadvantage of employing GWAS arrays instead of custom-designed platforms such as the ImmunoChip, which allow far more individuals to be genotyped since the cost is approximately 20% of that of contemporary GWAS chips [375]. As the number of loci identified strongly

correlates with sample size [1], using genome-wide genotyping platforms in smaller sample cohorts due to cost constraints can ultimately compromise the power of association discovery, as observed here. Despite this, however, my genome-wide analysis did yield three loci with suggestive evidence of association ($P_{\text{META-value}} \leq 5 \times 10^{-7}$) that are worth of follow-up in additional replication studies. Importantly, none of the newly associated regions were represented in the Immunochip, which highlights the usefulness of conducting a genome-wide analysis for ADD. The three newly identified signals were of high quality and showed consistent effects across all meta-analysed studies, providing technical validation in multiple independent platforms. As expected, these three associations were driven by common-frequency variants with modest effect sizes (mean=0.10). Unsurprisingly, none of the lead SNPs represented functional variants such as missense or splice disrupting alleles, nor were they in LD with such variants, which is also reminiscent of most GWAS associations [317, 379].

5.5.3 The possible pleiotropy of *FOSL2*

A particularly intriguing result yielded by this analysis is the suggestive association observed for a variant in the *FOSL2* gene and age at CD diagnosis ($P_{\text{META}}=1.89 \times 10^{-7}$, $\beta=-0.1$). As previously mentioned, *FOSL2* is part of a protein complex (AP-1) [510] which has been shown to upregulate genes involved in immune and pro-inflammatory responses during the pathogenesis of IBD [19, 224, 336]. More specifically, *FOSL2* is a core regulator of plasticity and a repressor of Th17 cells [81], which have emerged as major players in the tissue-specific immune pathology of IBD [162, 193, 232, 332]. Because of this, *FOSL2* has been suggested as an ideal candidate for the development of new therapeutic options aiming to target this Th17 cell population [368].

In addition to its obvious biological candidacy, this locus also showed association with IBD case/control status in three large IBD meta-analyses [110, 232, 290]. This finding is intriguing and opens up the possibility, if successfully replicated in future studies, for a locus to modulate both the risk and the age at which CD presents. A similar mechanism is already known for *NOD2*, which is associated with both the risk [217] and the age of onset of CD symptoms [87]. Another example is for Alzheimer's disease, where the major risk factor, the apolipoprotein E (*APOE*) gene, in addition to affecting the risk of Alzheimer's [100], also has a significant impact on the age at onset, explaining about 10% of its variation [235]. More generally, cross-phenotype associations, sometimes even in seemingly distinct traits [455], have been widely observed, particularly across immune diseases and psychiatric traits. Notable examples include: *IL23R* for IBD [126],

ankylosing spondylitis [132] and psoriasis [165]; *PTPN22* for rheumatoid arthritis [395], CD [33], systemic lupus erythematosus [265] and type 1 diabetes [488]; and *CACNA1C* for bipolar disorder and schizophrenia [453].

Genotyping of the *FOSL2* locus in additional IBD cases with available information on age at diagnosis is currently ongoing. If the observed association with CD-onset is successfully replicated and reaches genome-wide significance, it will be interesting to conduct further regional analysis to try and disentangle the cross-phenotype association seen at this locus. There are several possible scenarios that can underlie the (apparent) pleiotropic genetic effect observed here. One possibility is that *FOSL2* affects both the risk and age-at-onset via the same causal SNP (i.e. allelic pleiotropy). Another hypothesis is that *FOSL2* affects both phenotypes via different and independent causal variants (i.e. genetic pleiotropy). These two possibilities can, in theory, be evaluated through fine-mapping strategies conducted within each phenotype, which would help to refine the associated signals and locate the most likely causal variant (or variants) driving each association [456]. For the case of *FOSL2* however, this is likely to be challenging, because the identified SNPs are in high LD with many others, which will make their effects indistinguishable when conducting conditional analysis, preventing confident fine-mapping. An alternative approach would be to use colocalisation methods such as the one applied by Fortune *et al* [152], which is a Bayesian framework that derives the posterior support for each of five hypotheses describing the possible associations of a given region with two phenotypes. Here, the two hypothesis of greatest interest are: both traits are associated with the region via different causal variants or both traits are associated with the region and share a single causal variant.

An alternative hypothesis for the cross-phenotype association observed here could be mediated pleiotropy. Under that scenario, *FOSL2* could be indirectly associated with the risk of CD via a primary association with age at diagnosis or vice versa, which means the locus would be necessarily associated with both phenotypes if tested separately [455]. To explore this hypothesis, it will be interesting to re-test for an IBD case/control status in the UKIBDGC-GWAS samples while adjusting or stratifying the cases by the age at diagnosis of CD, for example. If the association with IBD-risk persists, then the cross-phenotype association is probably not fully mediated. Alternatively, one can also use another approach which is able to infer whether a given SNP directly influences a given phenotype through a path that does not involve a second correlated trait [501]. When conducting adjustment analyses, it will also be important to evaluate the effect of other sub-phenotypes that may equally affect the observed associations [393]. For

the case of IBD, one could account for information such as disease location at onset (i.e. ileal/colonic), disease behaviour (i.e. penetrating/stricturing/inflammatory) [87] and smoking status, a known environmental modulator [17]. However, such covariates should not be included in discovery efforts, as they can substantially reduce power for the identification of associated variants [393]. Instead, they can be accounted for afterwards, to deconvolve the associated signals. Several examples of sub-phenotype associations driving primary signals exist. For instance, an association of *NOD2* with disease behaviour has been shown to be driven almost entirely by its phenotypic correlation with location and age at diagnosis [87]. Another notable example is *FTO*. This gene was initially discovered to be associated with type 2 diabetes but subsequent correction for body-mass-index (BMI) abolished the signal, suggesting *FTO*-mediated susceptibility to type 2 diabetes was in fact driven through a relationship between *FTO* and obesity [142].

Chapter 6

Conclusions and future prospects

6.1 Summary of my research

This dissertation described four distinct projects in which NGS technologies were employed to identify genetic determinants of human diseases, or aspects of disease biology, that have been poorly studied thus far. Each research chapter contained a unique dataset with a unique study design. Overall, these projects focused on SNVs and small indels solely, as large structural variants (i.e. >50bp) [473], triplet repeat expansions [294, 450], mosaic [161] and uniparental disomy (UPD) [249] events remain challenging to assay using conventional short-read sequencing.

In Chapter 2, I conducted a comprehensive NGS-based screening of known causative genes in 49 cases with congenital hypothyroidism and *gland-in-situ*. Genetic screening of such patients has been traditionally limited by the cost and labour implications of Sanger-sequencing multiple exons, meaning many have remained genetically undiagnosed. By combining a stringent variant filtering pipeline with pedigree segregation analyses and *in silico* predictions of pathogenicity for candidate variants, we successfully attained a solid genetic diagnosis for 59% of the patients. This project explored, for the first time, the utility of NGS methods for genetic diagnosis of CH with GIS, and paved the way for the development of a gene panel which will hopefully move across into the NHS domain at Addenbrooke's Hospital in Cambridge (UK) in the near future.

In Chapter 3, I described a family-based study in which exome and targeted-sequencing were used, also for the first time, to identify novel disease genes in a phenotypically heterogeneous CH cohort comprised of 48 families. This condition has been refractory to traditional gene-mapping techniques, meaning it is poorly understood. By implementing

distinct variant filtering pipelines, I identified rare inherited variation segregating with CH within families, as well as *de novo* and CNVs events. Due to scant sharing of genetic causes across CH families, this study was unable to robustly implicate a novel gene for this condition. However, by adopting a candidate-focused approach, screening for likely pathogenic variants in long-standing CH candidate genes, I identified a homozygous loss-of-function variant in *SLC26A7* which was subsequently observed in two different haplotypes of two additional CH families. *SLC26A7* therefore emerges as a putative causative gene for CH with *gland-in-situ*. Experimental studies are ongoing to confirm the pathogenic status of the variant identified herein and to elucidate its role in the pathogenesis of disease.

In Chapter 4, I described an analysis leveraging exome and genotyping data from 145 children with very-early-onset IBD, the largest cohort recruited to date. This condition is still incompletely understood and is thought to be caused by highly penetrant variants. Using a conservative variant screening procedure, we identified likely causative mutations in *XIAP*, *CYBA* or *SH2D1A* in four patients. This finding added further strength to a growing body of recent evidence suggesting defects in loci associated with primary immunodeficiencies can underlie VEO-IBD phenotypes, and suggested targeted-sequencing of such genes is likely to be a fruitful prospective tool for the molecular diagnosis of VEO-IBD children. Moving beyond rare variants, I calculated polygenic risk scores for each proband using the estimated effect sizes of established adult IBD-risk alleles, and showed that the majority of VEO-IBD cases do have, in fact, a polygenic load similar to that seen in adult-onset IBD cases. This study therefore provided important insights into the genetic architecture of this condition that suggested, for the first time, that if highly penetrating variants contribute to VEO-IBD, they likely do so on an already IBD-susceptible genetic background (at least in a large proportion of the cases).

Lastly in Chapter 5, I meta-analysed three distinct GWAS datasets and low-coverage whole-genome sequences to identify genetic modifiers of the age of onset of CD, UC and IBD. While this study did not detect loci associated at genome-wide significance, I identified three suggestive associations worthy of follow-up in replication studies. Importantly, the signal associated with a decrease in the age at diagnosis of CD overlapped with an established CD-susceptibility locus (*FOSL2*) known to modulate immune and pro-inflammatory responses involved in the pathogenesis of IBD. If associated at genome-wide significance in future analyses, *FOSL2* will represent yet another example of biological or mediated genetic pleiotropy occurring across human traits.

6.2 NGS: from bench to bedside

In addition to revealing insights into the pathology of disease, much of the research presented in this dissertation had one important outcome that cannot be overlooked – a direct impact on patients lives. Importantly, this underscores the key role of all clinicians who were actively involved in these projects, without whom the clinical interpretation and translation of my findings would not have been possible.

The conclusive molecular diagnosis reached in 59% of the cases included in Chapter 2 allowed for genetic counselling, discussion of recurrence risk with families and the identification of asymptomatic mutation carriers at risk of developing CH. The diagnosis of these patients will also now enable early identification of subsequent cases in the same family, and help to avoid the negative consequences on mental development associated with delayed diagnosis and treatment of hypothyroidism. This is especially informative for our patients residing in countries where no national screening programme for CH is available, such as Saudi Arabia and Turkey. Even though in the majority of cases the genetic ascertainment of CH with GIS does not directly affect clinical management, the confirmation of *DUOX2* mutations in some of our patients alerted their clinicians to the fact their phenotype may be transient. Consequently, this will now enable them to look out for children whose treatment dose requirements for levothyroxine are modest and to do a carefully monitored trial off treatment at this age.

Two syndromic-CH cases included in Chapter 3 harbored likely causative variants in genes associated with congenital heart defects (*NKX2.5*) or with skeletal dysplasias (*HSPG2*), the exact two extrathyroidal phenotypes documented in each of these cases, respectively. These results were informative to patients and their families because it suggested their thyroid phenotype is independent from their other congenital abnormalities, and this ended up being especially relevant for their corresponding siblings who presented with the extrathyroidal malformations in the absence of CH.

The identification of *CYBA* and *XIAP* defects in three patients studied in Chapter 4, directly informed their treatment options and opened up new avenues for disease-specific treatment. The two *CYBA*-deficient siblings were referred to an immunological clinic and treatment will be decided based on a multidisciplinary team consensus. By default, anti-TNF α therapy (infliximab) is the usual course of treatment for chronic granulomatous disease (CGD) patients, however it is sometimes contraindicated because it is often accompanied by life-threatening infections and complications [497]. Recently, treatments targeting IL1B using a IL1-receptor antagonist (anakinra) have shown

promise in the management of CGD [112] and our patients may eventually benefit from such options in addition to allogenic haematopoietic stem cell transplantation (HSCT), the new therapeutic strategy for refractory VEO-IBD.

HSCT was already initiated in the patient with *XIAP* deficiency, and the hope is that this treatment will now finally enable clinical remission. The patient was diagnosed shortly after turning six, but it took 14 years (and three major GI operations) for him to get a conclusive genetic diagnosis. The same *XIAP* mutation was reported recently by Wada *et al* [506] in a five month old child who had the chance to undergo cord blood transplantation much earlier in life and has since been in remission. These two different stories demonstrate the importance of establishing a timely molecular diagnosis early and accurately in VEO-IBD children to avoid unnecessary surgery and instead proceed with appropriate curative approaches such as HSCT. Also, the confirmation of a *XIAP*-defect in our patient means he will now undergo regular infection screening to prevent the potentially fatal EBV-triggered haemophagocytic lymphohistiocytosis (HLH) that is commonly developed in such patients.

6.3 Common themes emerging from my research

Several common themes have emerged as relevant to most (if not all) projects presented in this dissertation, and these can be extended to the field of rare and/or complex diseases more broadly. In the following pages, I will discuss how these topics impacted my research and will present some solutions that are increasingly being adopted to address them and to improve the analysis of NGS data in gene-mapping experiments. Finally, I will then look to the types of studies (many of which are already underway) that will shape human disease research over the coming years and discuss how they will provide important clues in the road towards personalised medicine.

6.3.1 Sample size

Sample size is crucial for all genetic studies of human disease. In all analyses conducted in this dissertation, a larger sample size would have solved a great part of the limitations that were already mentioned in the corresponding chapters.

For the novel-gene discovery aims of Chapter 3 and 4, larger patient cohorts would have permitted statistically significant recurrence of mutations in individual genes across

independent families or patients. Even though gene discovery for disorders with low locus heterogeneity and fully penetrant mutations is occasionally possible by sequencing a single family [230], most gene-discovery applications do require substantially larger sample sizes, and this is especially paramount if genetic heterogeneity is suspected (as in congenital hypothyroidism and very-early-onset IBD, as discussed previously). My studies represent the largest that have been conducted for such conditions, however larger sample sizes are needed to robustly implicate novel disease-associated genes in these conditions. In most rare disease studies, the sample size needed is seldom known in advance and it depends on the (presumed) genetic architecture of disease, which is also poorly understood in many instances. For some disorders, the sample size needed may possibly approach or even exceed those needed for GWAS, as illustrated by Singh *et al* [449]. The authors started with data from 1,745 patients with schizophrenia and 6,789 controls, and then added 2,591 extra published cases [407] and 2,554 controls. Yet, even with more than 4,200 cases, no gene attained exome-wide significance. They then combined the rare LoF variants (MAF <0.1%) seen in their cases with *de novo* mutations of 1,077 schizophrenia probands from seven published studies [155, 172, 185, 186, 320, 469, 530]. Altogether, this yielded three *de novo* events and seven LoF variants in *SETD1A*, while none were found in 20,000 control exomes, providing a $P = 3.3 \times 10^{-9}$ and an estimated OR of 32 (CI: 4.5 – 4.528). This study highlights the enormous importance of data sharing. Future NGS studies of VEO-IBD and CH aiming to discover novel disease-associated genes should therefore embrace the value of collaborative research, as this will permit more rapid accumulation of evidence for novel disease-associated genes. In VEO-IBD in particular, given that these patient populations are studied worldwide and usually in very small numbers [243], an international registry containing sequence data, immunological and environmental data of such patients could prove beneficial to make reliable inroads into better understanding the mechanisms underlying disease and resolve the monogenic-polygenic interface of the phenotype. When sharing data however, researchers should be mindful of systematic differences among patient cohorts stemming from population stratification and technical biases. Such disparities may require careful and extensive quality control investigations, as well as study design considerations, before pooling individual data or meta-analysing patient cohorts.

Examples of successful data sharing initiatives in rare disease already exist in the field of copy number variation with the DECIPHER database [149] and the International Standards for Cytogenomic Arrays Consortium (<https://www.iscaconsortium.org/>), and several ambitious efforts to establish global standards for genomic data sharing have been initiated (e.g. Global Alliance) [57]. The accumulation of evidence for novel

disease-associated genes and therefore the end of the "N-of-1 problem" can also be greatly facilitated by the use of recently developed tools such as GeneMatcher [454]. This resource is freely accessible and is designed to enable connections between clinicians who share patients with variants in the same candidate gene. Using GeneMatcher, researchers can also connect with other scientists with special expertise and/or model organisms with defects in the orthologous gene(s), which may ultimately expedite the development of follow-up functional studies to elucidate the pathological mechanisms of disease.

Apart from facilitating novel-gene discovery, a larger sample size in Chapter 4 would have also increased the power to conduct more specific (and therefore more useful) comparisons between the polygenic component of VEO-IBD and UKIBDGC cases. Importantly, VEO-IBD children could have been stratified by their IBD status (i.e. CD-like/UC-like/U-IBD), for example, and UKIBDGC cases could have been stratified by their age at diagnosis as well. Collectively, these analyses could have potentially revealed important similarities (or differences) between these multiple clinical entities and different ages of onset of disease. Finally, a larger sample size in Chapter 5 would have provided greater power to detect associations with age at IBD diagnosis and to fine-map causal variants in *FOSL2*, which would have helped us to better understand the apparent pleiotropic mechanism observed at that locus.

6.3.2 Phenotypic heterogeneity

The issue of phenotypic heterogeneity goes hand-with-hand with sample size. Both CH and VEO-IBD have a broad phenotypic spectrum which presents a challenge for gene-mapping applications because it suggests genetic heterogeneity, which is hard to deal with in genetic analyses when total sample sizes are small. As mentioned previously, rare variant association methods testing biological units other than single genes can leverage genetic heterogeneity and provide important insights into disease pathology without implicating individual loci. However, this type of approach was still underpowered to reveal a significant enrichment (if indeed one exists) of rare disruptive variants in biologically relevant genesets or pathways in VEO-IBD children (N=124) in Chapter 4, and was not attempted in Chapter 3 due to the more complex nature of the study design (i.e. family-based rather than solely unrelated cases) and the multiplicity of phenotypic categories (i.e. agenesis, ectopia, hypoplasia, syndromic-CH, *gland-in-situ* CH).

The use of Human-Phenotype-Ontology (HPO) terms is one strategy increasingly being adopted to leverage the heterogeneity of large cohorts of rare heterogeneous disorders [150, 391, 517] and to use phenotypic information effectively. The HPO represents a standardised vocabulary to describe rare disease phenotypes, where terms are connected to each other through semantic relations and organised hierarchically [253]. Rather than describing individual disease entities, the HPO describes the phenotypic abnormalities associated with them. Combined with deep phenotyping of the individuals being sequenced, the use of HPO annotations enables researchers to apply statistical clustering approaches to guide and aid gene-discovery [7, 180]. For instance, sub-groups of patients who cluster strongly on the basis of their HPO-encoded phenotypes can be identified, and these are then more likely to share mutations in the same or in functionally related genes. Inversely, the degree of phenotypic similarity in groups of cases sharing rare protein-altering mutations in the same gene can be calculated, which in turn can help reveal important genotype-phenotype relationships. Both of these strategies were first applied to heritable bleeding and platelet disorders [517], and thus proved their usefulness for rare diseases with heterogeneous clinical characteristics that very often encompass multi-organ abnormalities, precisely as CH. The non-specific and poorly defined nature of the VEO-IBD phenotype, on the other hand, may prove more challenging to study via these strategies but should nevertheless be attempted when large cohorts become available.

Apart from rare diseases, the extension of HPO to complex disease was also already initiated, with some researchers suggesting it will equally be an invaluable resource to more efficiently leverage the available phenotypic information [182]. Specifically, the HPO will enable phenotypic networks of common diseases to be created and similarities between etiologically related disease groups that show overlapping phenotypes to be identified. Collectively, these strategies will boost our understanding of complex diseases in general and help to map additional genetic risk factors [182].

6.3.3 Diverse ethnic origin

With the exception of Chapter 5, which contained a large cohort of European individuals, the diverse ethnic background of the patients included in my research projects posed challenges and had implications on my downstream genetic analyses. In Chapters 2 and 3, the multiplicity of ethnicities meant that the null-models used to derive the statistical significance of *TG-DUOX2* digenicity and the enrichment of rare variants per-gene, respectively, were not derived from appropriately ancestry-matched controls,

as none were internally or publicly available. In Chapter 4, 14% of VEO-IBD cases were ignored in rare-variant enrichment analyses because they could not be placed in one of the two ancestry-based case-control clusters (European or South Asian) identified via PCA. Finally, all non-European VEO-IBD cases (30% of the whole cohort) were also not taken into account in my polygenic risk score analyses because our understanding of the IBD-susceptibility factors in non-European populations is very poor.

There are thus two important take-home messages. First, sequencing and genotyping data from populations of various demographic backgrounds need to become available to the research community at an increasing pace. Rare disease studies usually recruit patients from around the world, however, the combination of natural cost-constraints with the fact most studies are still case-only, means sequencing of appropriately large numbers ($>2,500$ [300]) of control individuals from the same population as cases is rarely conducted. This has contributed to the general lack of ethnically-diverse control sequences that can be used by researchers; significant improvements are expected in the near future as NGS becomes more affordable and with researchers more actively participating in responsible data-sharing initiatives. An important incentive to accelerate the accumulation of data from diverse ethnic populations is that it will ultimately improve our ability to do accurate variant interpretation. Ideally, variants identified by NGS should be evaluated both at the level of the patient's ancestry background but also in a large number of ethnically diverse populations [300, 308]. The tremendous effort of the ExAC project [135] has made this possible, however many populations remain underrepresented or even absent in this database. With the accrual of more diverse sequences in the future, we will be better able to conduct family-based as well as case-control analyses in appropriately matched groups, identify benign and common variants and thus reduce false-positive findings, reassess previous disease-variant associations, corroborate truly pathogenic mutations and find population-specific pathogenic variants.

Second, complex disease studies need to be more rapidly extended to non-European populations for us to better understand population-specific effects of genetic variation. This has already been initiated in some quantitative traits and common disorders, including IBD [242, 287, 290, 303], but non-European cohorts are still disproportionately smaller than European ones, and many populations have not yet been included in trans-ethnic meta-analyses. Apart from enabling heterogeneity of effect to be detected across population cohorts, an important benefit of cross-population analyses is that they actually boost the overall power to detect associations because many common-frequency risk-alleles will still be shared across many populations [290, 303]. Moreover,

trans-ethnic analysis empower fine-mapping of causal variants by taking advantage of the different LD structures that often exist between populations. The African Genome Variation Project [188] is one notable example that has taken the lead to address the lack of GWAS studies in African populations and it is anticipated that many similar initiatives will now follow.

6.4 Future studies of rare and complex diseases

Several ambitious and large-scale genetic studies of human diseases are already underway. Many of these projects were made possible with the development of ultra-high-throughput instruments (e.g. Illumina X Ten) dedicated to population-level sequencing, now offering \$1,000 per 30X human genome [177]. Also importantly, many of these studies were driven by the growing recognition among governments and health policymakers of the benefits of using genomic information to diagnose, understand and cure human disease.

In the UK, the 100,000 Genomes Project delivered by Genomics England (GEL), is a government-funded initiative that aims to sequence 100,000 whole-genomes from National Health Service (NHS) patients with rare disorders, cancers and infectious diseases. Over 190 distinct rare diseases were included in the project as of June 2016. For these conditions, DNA from the two closest blood relatives of patients will also be sequenced and for each cancer patient, two genomes will be sequenced – one from the tumour and one from the healthy tissue [451]. Overall, the project aims to find clinically significant findings by linking extensive and long-term clinical information with precise molecular signatures; to generate improved diagnostic tests; to identify treatments that can be tailored to individual patients; and, ultimately, to set up a genomics service for the NHS where genome sequencing becomes an integral and routine part of medical care. To achieve these goals, the project will require complex statistical analyses of large amounts of clinical and molecular data, fast and efficient variant annotation and interpretation pipelines and the development of high standards of ethical practice based on informed consent. In addition, tight collaboration with research scientists will be necessary to elucidate project results where a connection to disease has not yet been well established i.e. putative disease-associated genes. Finally, a close collaboration with the pharma and biotech industry will also be required to develop new diagnostics and treatments that can then be deployed to participating patients [318]. The GEL project is the first genomics initiative to be tightly integrated with a health-care system [427],

and its hoped that benefits are propagated to clinical practice much faster than usual. In the future, the study will also collect other biological material including serum and plasma for proteomics and metabolomics, RNA for transcriptomics, lymphocyte DNA for epigenetics and tumour samples for RNA expression profiles, tumour epigenetics and proteomics [167]. Collectively, these data are expected to fuel a series of downstream studies that will apply multi-omics integrated approaches to study cancer and the rare disorders included in the GEL initiative.

Gene-mapping of complex diseases and quantitative traits will benefit greatly from recent efforts conducting deep-phenotyping and follow-up of large collections of individuals in longitudinal settings. Examples of such initiatives include the 100K Wellness project and the UK Biobank. The 100K Wellness project is whole-genome sequencing 100,000 healthy volunteers as well as periodically collecting biological (e.g. blood, saliva, stool), environmental (e.g. diet) and physiological (e.g. heart rate, blood pressure, weight and sleep quality) parameters [212]. The project aims to identify metrics that correlate with well being; to identify the early origins and mechanisms of common diseases developing in study participants; and to follow disease progression long-term. To do so, blood metabolites, organ-specific blood proteins, white blood cells and the gut microbiome will be intensely monitored to identify changes in health occurring in participants over 25 years [168]. In the end, the project is expected to contribute to the development of powerful biomarkers of disease and well being, and to provide important information on how to predict, as well as prevent, some common disorders.

The UK Biobank, backed by the Medical Research Council and the Wellcome Trust, is an open-access population-based prospective cohort containing biological, demographic and physiological data from half a million people in the UK aged 40–69 years. The resource consists of diverse biological material (e.g. blood, saliva, urine, faeces) that will allow a variety of assays to be conducted including genetic, proteomic, metabolomic and biochemical. All participants have already been genotyped using a bespoke genome-wide microarray containing $\sim 820,000$ variants, 18% of which are rare ($>0.02\%$) non-synonymous variants. The participants also gave consent to have their future health records linked to their data, a process which is ongoing. Together, these data will contribute towards the identification of novel genetic factors influencing anthropometric and cardiometabolic traits, as well as novel predisposing factors for major diseases of middle and old age (e.g. age-related macular degeneration, dementia, irritable bowel syndrome, hypertension, multiple sclerosis and cardiovascular and autoimmune diseases) [465]. For many quantitative traits (e.g. height, BMI, lipid levels) and common diseases (e.g. osteoarthritis), the UK Biobank will grant significant increase

in sample size compared with the largest GWAS analyses previously conducted on such phenotypes [293, 523, 538], ultimately increasing the power to detect novel associations. This resource therefore exemplifies how array-based studies with ever larger sample sizes will continue to play an active role in locus discovery of many traits and complex diseases, in parallel to NGS technologies. In addition to providing opportunities for better powered GWASs, this resource will also support a variety of other studies such as epidemiological and exposure-outcome analyses, cross-sectional studies of genotype-phenotype correlations, mendelian randomisation studies, and prospective analyses combining the joint effects of genetics, lifestyle and environmental variables [250, 452]. With sequencing costs continuing to plummet it is likely the UK Biobank will eventually adopt whole-genome sequencing.

Array-based and low-coverage WGS studies of complex diseases will also greatly benefit from efforts such as the Haplotype Reference Consortium (HRC). By collating whole-genome sequence data from 20 studies of predominantly European ancestry, the HRC generated the largest imputation reference panel to date, containing $\sim 65,000$ haplotypes and ~ 40 million SNPs [319]. Studies using this resource will be able to get high quality and accurate genotype imputation at minor allele frequencies as low as 0.1%, greatly increasing the number of SNPs tested in association studies and the power to fine-map causal variants. In the near future, the HRC plans to incorporate the high coverage genomes generated by Genomics England and to increase the ethnic diversity of the panel by incorporating data from sequencing studies in world-wide sample sets such as the African Genome Variation project [188], the CONVERGE study of Han Chinese individuals [95] and the Human Genome Diversity project [429], which studied 51 different populations from Africa, Europe, the Middle East, South and Central Asia, East Asia, Oceania and the Americas.

6.5 From variant discovery to disease mechanisms

The studies just mentioned, encompassing both rare and complex diseases, will soon yield numerous candidate disease-causing mutations and disease-associations that will eventually lead to personalised medicine opportunities, as alluded to above. However, although variant discovery is an important breakthrough towards this vision, it is only the first step; understanding the biological mechanisms by which mutations and disease-susceptibility alleles contribute to disease pathogenesis is certainly a greater challenge. Indeed, understanding the functional effects of genetic variation is a challenging area in

need of intense research. Firstly, we must improve our ability to interpret the biological consequences of variants of unknown significance [419]. These are incredibly common in genome sequences [240, 418] and include newly identified alleles that affect the coding sequences of genes previously unlinked to disease or genes of unknown function. Secondly, we must also improve our ability to map and assign regulatory activity to non-coding regions of the genome, and enhance our understanding of their mechanistic effects in disease development. This is already an issue for the majority of common variants found on GWAS but will be of particular importance with the widespread use of WGS, as non-coding variants will be more frequently seen in gene-mapping studies of both rare and complex diseases. Interrogating expression quantitative trait locus (eQTL) studies conducted in relevant cells types can be useful to determine whether non-coding variants influence expression levels of nearby genes [521], and thereby generate initial hypotheses about how these variants lead to disease susceptibility. Generating more eQTL studies in relevant cell types and cellular pathophysiological states should be a priority in the years ahead. Potential regulatory function can also be predicted on the basis of overlap with particular genomic features such as protein-binding motifs, transcription factor-binding sites, histone modification marks and open chromatin regions. These genomic features are increasingly being measured with targeted biochemical assays and NGS technologies in large-scale genomic studies such as the ENCODE [478], FANTOM5 [151] and the NIH Roadmap Epigenomics projects [425] and many more studies, focusing in diverse and more specific cellular contexts are currently being conducted. Ultimately, linking non-coding variants to genes and genomic regions of interest can facilitate the design of downstream functional studies, which are still required to inform, and convincingly demonstrate, the causal mechanisms of disease.

Emerging genome-editing tools such as the CRISPR-Cas9 system, enabling genome modifications at single-nucleotide resolution, will play an important role in future functional studies that aim to understand variant effects, gene function and disease mechanisms. This system involves guiding a Cas-cleavage enzyme to a specific genomic locus that is then cleaved and imprecisely repaired to allow the introduction of a specific mutation [489]. The approach is highly specific and efficient, and can be multiplexed to enable simultaneous editing at multiple genomic sites. In addition, it can be used in a variety of *in vitro* and *in vivo* biological models (e.g. human cell lines, primary cells, induced pluripotent stem cells (iPSC) and animals), therefore it can be applied in many different kinds of experimental studies.

In rare disease studies, the CRISPR-Cas9 system will enable researchers to quickly create site-specific mice knockout models, which sometimes can be more appropriate to study than available full-gene knockouts [90], i.e. if one wants to discriminate between the effect of a specific variant and the biological function of the gene. This will accelerate the interpretation of variants of uncertain significance. Researchers will also be able to investigate digenic mutations and tissue-specific mutations in genes, which will be incredibly useful when studying oligogenic inheritance of disease and the effects of somatic mutations, respectively. Allied with this technology, knockout mice will remain a crucial biological model to investigate the functional effects of mutations, to place putative disease-associated genes into a biological context, and to elucidate the mechanistic basis of rare disorders [204].

In complex disease studies, mouse experiments that either knockdown or over-express genes located in associated GWAS regions, using Cas9 fused with repression or activation transcription domains respectively [383], can also help in identifying the biological function of many candidate loci. The combination of CRISPR-Cas9 with human iPSC cells derived from patients with a specific risk allele, will offer new opportunities to model complex diseases that have been extremely challenging to study via conventional models such as cell lines and/or transgenic animals [463]. In addition, the simultaneous study of patient-specific iPSC lines with different risk-variants can aid in our understanding of how various disease-associated loci interact to produce a phenotype [440].

Finally, investigating the functional relevance of non-coding variants that overlap with known (or putative) regulatory elements and epigenetic marks is also becoming feasible with the availability of novel Cas9-based techniques. One recent example, developed by Hilton *et al*, consists of a Cas9-based protein fused to the catalytic core of the human acetyltransferase p300 [206]. This fusion protein catalyses acetylation of histone H3 lysine 27 at its target sites, leading to transcriptional activation of target genes from promoters and proximal and distal enhancers with high spacial and temporal specificity. Modifications of DNA molecules by cytosine methylation using a Cas9-DNA methyltransferase 3A (DNMT3A) fusion protein is another example [505]. These epigenome-editing proteins can therefore be targeted to candidate regulatory elements in order to modify local chromatin structure and determine the role of these elements in influencing gene expression and pathological mechanisms of disease.

6.6 Translation

The ultimate goal of gene-mapping studies is to translate scientific gains into the clinic, by providing drug targets, allowing personalised treatment based on genetic information, and perhaps predicting disease risk to leverage preventive medicine strategies.

6.6.1 Novel drug targets

Since the publication of the human genome in 2003, genetic studies of Mendelian and complex diseases have been providing an invaluable source of knowledge into potential drug targets. Genetic-based evidence is now proving to be invaluable in evaluating and developing novel drug targets. More than 50% of the drugs that undergo clinical trials fail, commonly due to a lack of efficacy. In contrast, it has recently been reported that targets entering clinical development with genetic-based support (based on GWAS or OMIM catalogues) were twice as likely to achieve drug approval [487]. In addition, FDA-approved drugs were fourfold more likely to have genetic support than drugs in phase I trials, suggesting drugs with genetic support are indeed more likely to advance into later stages of the development pipeline. Strikingly, the lack of genetic support in clinical development had the greatest impact earlier in the drug development process, suggesting the use of genetic evidence in the initial selection process is vital. This is the vision behind the recently established Open Targets initiative (www.opentargets.org/), a public-private partnership between academic scientists and pharmaceutical companies, that was established to catalyse the validation of novel drug targets based on genome-scale experiments and analysis. Thus, by increasing the proportion of discovery and development activities focused on targets with genetic support, and by allowing genetic data to guide the selection of the most promising molecules, initiatives such as this are expected to drive better and more effective treatments in the future.

6.6.2 Personalised treatments

In addition to providing greater efficacy in drug development, genetics can also inform personalised treatments. The ten best-selling drugs in the US are only able to improve the condition of 4-25% people who take them, and some drugs can even be harmful to certain ethnic groups [437]. Indeed, there is a great heterogeneity in the way individual people respond to medication, in terms of both treatment efficacy and toxicity. Although there are several clinical variables that can influence these varied

responses (e.g. disease severity, individual's age, nutritional status), it has long been recognised that polymorphisms in genes (e.g. *CYP2D6*, *ADH*, *CYP3A4*) that impact drug metabolism, disposition and response can have an even greater influence [134]. To accelerate the development of treatment strategies that take individual variability into account, the US launched a national Precision Medicine Initiative that includes the establishment of a national database of the genetic and clinical data of one million volunteers. Ultimately, the hope is that personalised pharmacogenetic information of drug response phenotypes can be incorporated into treatment selection in the future, to identify the correct medication and optimal dosage on an individual basis.

6.6.3 Genetic risk prediction

Whole-genome sequencing will perhaps soon be the universal diagnostic and public health tool, allowing us to more rapidly diagnose disease and predict its onset. WGS is already having a tremendous impact in the diagnostic testing of patients with genetic disorders caused by disruption of single genes or chromosomal regions and, in the near future, it will likely allow the much earlier detection of such disorders. The increasing deployment of WGS in a variety of clinical settings will facilitate the assessment of its overall benefits and limitations, and the diagnostic yield in different genetic disorders, which will inform and guide the widespread use of WGS in the clinic. In addition, in the subset of diseases for which preventive measures are available, detection of deleterious and medically actionable variants before the onset of disease in asymptomatic patients could also be highly beneficial. In total, it has been estimated that ~1% of Americans likely harbor deleterious and medically actionable variants [42], therefore the public health impact of WGS could be considerable. Another anticipated application of WGS in asymptomatic individuals is the identification of carrier status for many autosomal recessive disorders [40], which could be of great importance to couples for family planning.

The use of WGS for complex disease risk prediction in a clinical setting remains a hot topic of debate. Hundreds of GWAS meta-analyses have discovered a lengthening list of alleles associated with complex disease, which can be used to construct disease predictors in the form of polygenic risk scores (PRS). Studies have shown that the predictive ability of PRSs, measured as the area under the ROC curve (AUC), varies greatly across 18 common diseases, based on the current genetic knowledge of those conditions [133, 231]. Psychiatric diseases and cancers cannot be well predicted, but others disorders including type 1 diabetes, Crohn's disease and age-related macular

degeneration have relatively good predictive power. This variation in AUCs depends on several factors such as the heritability of disease, the effect sizes of the risk loci, the clinical heterogeneity of the phenotype and the statistical power of the GWAS that identified the risk alleles in the first place. Perhaps not surprisingly, given the generally weak effects of disease-associated loci, the range of AUCs for those 18 disorders is very similar to that of classic, non-genetic risk predictors of disease (e.g. body mass index, cholesterol levels, smoking status, family history) [62, 291], and for the majority of disorders, the combination of genetic with traditional predictions offers poor improvements in terms of clinical utility [231]. It remains to be seen whether with a more complete understanding of the genetic architecture of complex disease, propelled by future WGS studies, genetic risk prediction will be a realistic utility. In principle, an advantage of using genetic predictors over classical alternatives is that risk prediction would be stable over time, as a person's genetic code is essentially constant throughout their lifetime. This could allow for risk prediction to be performed much further in advance than is currently possible, which could be especially important for cases where prevention strategies are more effective if started long before disease onset, or if carried out over a long time period. Assessing the potential clinical utility of PRSs, and deciding on the optimal way to use it in disease risk prediction, will be a serious challenge for medical practice in the future.

6.7 Concluding remarks

It is remarkable how in a relatively short amount of time, the 13 years since the publication of the human genome, the field of human genetics experienced an extraordinary leap in our knowledge of the genetic determinants and pathological mechanisms of disease. Next-generation sequencing is now dramatically altering our ability to conduct gene-mapping studies and is yielding unprecedented biological insights that are truly driving a revolution in health care. As human geneticists we sit in an enviable position, one with a history of considerable success behind us, and a future that holds promise for significant gains.

References

- [1] Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. a., Hurles, M. E., and McVean, G. a. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–73.
- [2] Abou Hassan, O. K., Fahed, A. C., Batrawi, M., Arabi, M., Refaat, M. M., DePalma, S. R., Seidman, J. G., Seidman, C. E., Bitar, F. F., and Nemer, G. M. (2015). NKX2-5 Mutations in an Inbred Consanguineous Population: Genetic and Phenotypic Diversity. *Scientific Reports*, 5:8848.
- [3] Abraham, C. and Cho, J. H. (2006). Functional consequences of NOD2 (CARD15) mutations. *Inflammatory Bowel Diseases*, 12(7):641–650.
- [4] Adzhubei, I., Schmidt, S., Peshkin, L., Ramensky, V., and Sunyaev, S. (2010). A method and server for predicting damaging missense mutations. *Nature Methodshods*, 7(4):248–249.
- [5] Agarwal, S. and Cunningham-Rundles, C. (2009). Autoimmunity in Common Variable Immunodeficiency. *Current Allergy and Asthma Reports*, 9(5):347–352.
- [6] Aguilar-Bryan, L. and Bryan, J. (2008). Neonatal diabetes mellitus. *Endocrine Reviews*, 29(3):265–91.
- [7] Akawi, N., McRae, J., Ansari, M., Balasubramanian, M., Blyth, M., Brady, A. F., Clayton, S., Cole, T., Deshpande, C., Fitzgerald, T. W., Foulds, N., Francis, R., Gabriel, G., Gerety, S. S., Goodship, J., Hobson, E., Jones, W. D., Joss, S., King, D., Klena, N., Kumar, A., Lees, M., Lelliott, C., Lord, J., McMullan, D., Osio, D., Piombo, V., Prigmore, E., Rajan, D., Rosser, E., Sifrim, A., Smith, A., Swaminathan, G. J., Turnpenny, P., Whitworth, J., Wright, C. F., Firth, H. V., Barrett, J. C., Lo, C. W., FitzPatrick, D. R., and Hurles, M. E. (2015). Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families. *Nature Genetics*, 47(11):1363–1369.
- [8] Al-Bousafy, A. (2006). Libyan Boy with Autosomal Recessive Trait (P22-phox Defect) of Chronic Granulomatous Disease. *Libyan J Med*, 1(2).
- [9] Al Taji, E., Biebermann, H., Límanová, Z., Hníková, O., Zikmund, J., Dame, C., Grüters, A., Lebl, J., and Krude, H. (2007). Screening for mutations in transcription factors in a Czech cohort of 170 patients with congenital and early-onset hypothyroidism: identification of a novel PAX8 mutation in dominantly inherited early-onset non-autoimmune hypothyroidism. *European Journal of Endocrinology*, 156(5):521–9.

- [10] Al Turki, S. (2013). Genetic investigations of Tetralogy of Fallot in trios. *University of Cambridge*, PhD Thesis.
- [11] Al Turki, S., Manickaraj, A. K., Mercer, C. L., Gerety, S. S., Hitz, M.-P., Lindsay, S., D'Alessandro, L. C. a., Swaminathan, G. J., Bentham, J., Arndt, A.-K., Low, J., Breckpot, J., Gewillig, M., Thienpont, B., Abdul-Khaliq, H., Harnack, C., Hoff, K., Kramer, H.-H., Schubert, S., Siebert, R., Toka, O., Cosgrove, C., Watkins, H., Lucassen, A. M., O'Kelly, I. M., Salmon, A. P., Bu'lock, F. a., Granados-Riveron, J., Setchfield, K., Thornborough, C., Brook, J. D., Mulder, B., Klaassen, S., Bhattacharya, S., Devriendt, K., Fitzpatrick, D. F., Wilson, D. I., Mital, S., and Hurler, M. E. (2014). Rare variants in NR2F2 cause congenital heart defects in humans. *American Journal of Human Genetics*, 94(4):574–85.
- [12] Albert, T. J., Molla, M. N., Muzny, D. M., Nazareth, L., Wheeler, D., Song, X., Richmond, T. a., Middle, C. M., Rodesch, M. J., Packard, C. J., Weinstock, G. M., and Gibbs, R. a. (2007). Direct selection of human genomic loci by microarray hybridization. *Nature Methods*, 4(11):903–905.
- [13] Altshuler, D., Daly, M. J., and Lander, E. (2009). Genetic Mapping in Human Disease. *Science*, 322(5903):881–888.
- [14] Altshuler, D. M., Gibbs, R. A., Peltonen, L., Dermitzakis, E. T., Schaffner, S. F., Yu, F., Bonnen, P. E., de Bakker, P. I. W., Deloukas, P., Gabriel, S. B., Gwilliam, R., Hunt, S., Inouye, M., Jia, X., Palotie, A., Parkin, M., Whittaker, P., Chang, K., Hawes, A., Lewis, L. R., Ren, Y., Wheeler, D., Muzny, D. M., Barnes, C., Darvishi, K., Hurler, M., Korn, J. M., Kristiansson, K., Lee, C., McCarroll, S. A., Nemesh, J., Keinan, A., Montgomery, S. B., Pollack, S., Price, A. L., Soranzo, N., Gonzaga-Jauregui, C., Anttila, V., Brodeur, W., Daly, M. J., Leslie, S., McVean, G., Moutsianas, L., Nguyen, H., Zhang, Q., Ghorri, M. J. R., McGinnis, R., McLaren, W., Takeuchi, F., Grossman, S. R., Shlyakhter, I., Hostetter, E. B., Sabeti, P. C., Adebamowo, C. A., Foster, M. W., Gordon, D. R., Licinio, J., Manca, M. C., Marshall, P. A., Matsuda, I., Ngare, D., Wang, V. O., Reddy, D., Rotimi, C. N., Royal, C. D., Sharp, R. R., Zeng, C., Brooks, L. D., and McEwen, J. E. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–8.
- [15] Alzahrani, A. S., Baitei, E. Y., Zou, M., and Shi, Y. (2006). Metastatic Follicular Thyroid Carcinoma Arising from Congenital Goiter as a Result of a Novel Splice Donor Site Mutation in the Thyroglobulin Gene. *The Journal of Clinical Endocrinology and Metabolism*, 91(August):740–746.
- [16] Amendola, E., Luca, P. D., Macchia, P. E., Terracciano, D., Rosica, A., Chiappetta, G., Kimura, S., Mansouri, A., Affuso, A., Arra, C., Macchia, V., Lauro, R. D., and Felice, M. D. (2005). A Mouse Model Demonstrates a Multigenic Origin of Congenital Hypothyroidism. *The Journal of Clinical Endocrinology and Metabolism*, 146(August):5038–5047.
- [17] Ananthkrishnan, A. N. (2015). Epidemiology and risk factors for IBD. *Nature Reviews. Gastroenterology & hepatology*, 12(April):205–217.

- [18] Anderson, C. a., Pettersson, F. F. H., Clarke, G. G. M., Cardon, L. R., Morris, A. P., and Zondervan, K. T. (2010). Data quality control in genetic case-control association studies. *Nature Protocols*, 5(9):1564–73.
- [19] Andoh, A., Zhang, Z., Inatomi, O., Fujino, S., Deguchi, Y., Araki, Y., Tsujikawa, T., Kitoh, K., Kim-Mitsuyama, S., Takayanagi, A., Shimizu, N., and Fujiyama, Y. (2005). Interleukin-22, a member of the IL-10 subfamily, induces inflammatory responses in colonic subepithelial myofibroblasts. *Gastroenterology*, 129(3):969–984.
- [20] Andreoletti, G., Ashton, J. J., Coelho, T., Willis, C., Haggarty, R., Gibson, J., Holloway, J., Batra, A., Afzal, N. a., Beattie, R. M., and Ennis, S. (2015). Exome analysis of patients with concurrent pediatric inflammatory bowel disease and autoimmune disease. *Inflammatory Bowel Diseases*, 21(6):1.
- [21] Arico, M., Imashuku, S., Clementi, R., Hibi, S., Teramura, T., Danesino, C., Haber, D. a., and Nichols, K. E. (2001). Brief report Hemophagocytic lymphohistiocytosis due to germline mutations in SH2D1A , the X-linked lymphoproliferative disease gene. *Blood*, 97(4):1131–1133.
- [22] Asimit, J. and Zeggini, E. (2010). Rare variant association analysis methods for complex traits. *Annual Review of Genetics*, 44:293–308.
- [23] Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flück, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurler, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., Lehrach, H., Mardis, E. R., Marth, G. T., McVean, G. A., Nickerson, D. A., Schmidt, J. P., Sherry, S. T., Wang, J., Wilson, R. K., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., Kovar, C., Lee, S., Muzny, D., Reid, J. G., Zhu, Y., Chang, Y., Feng, Q., Fang, X., Guo, X., Jian, M., Jiang, H., Jin, X., Lan, T., Li, G., Li, J., Li, Y. Y., Liu, S., Liu, X., Lu, Y., Ma, X., Tang, M., Wang, B., Wang, G., Wu, H., Wu, R., Xu, X., Yin, Y., Zhang, D., Zhang, W., Zhao, J., Zhao, M., Zheng, X., Gupta, N., Gharani, N., Toji, L. H., Gerry, N. P., Resch, A. M., Barker, J., Clarke, L., Gil, L., Hunt, S. E., Kelman, G., Kulesha, E., Leinonen, R., McLaren, W. M., Radhakrishnan, R., Roa, A., Smirnov, D., Smith, R. E., Streeter, I., Thormann, A., Toneva, I., Vaughan, B., Zheng-Bradley, X., Grocock, R., Humphray, S., James, T., Kingsbury, Z., Sudbrak, R., Albrecht, M. W., Amstislavskiy, V. S., Borodina, T. A., Lienhard, M., Mertes, F., Sultan, M., Timmermann, B., Yaspo, M.-L., Fulton, L., Fulton, R., Ananiev, V., Belaia, Z., Beloslyudtsev, D., Bouk, N., Chen, C., Church, D., Cohen, R., Cook, C., Garner, J., Hefferon, T., Kimelman, M., Liu, C., Lopez, J., Meric, P., O’Sullivan, C., Ostapchuk, Y., Phan, L., Ponomarov, S., Schneider, V., Shekhtman, E., Sirotkin, K., Slotka, D., Zhang, H., Balasubramaniam, S., Burton, J., Danecek, P., Keane, T. M., Kolb-Kokocinski, A., McCarthy, S., Stalker, J., Quail, M., Davies, C. J., Gollub, J., Webster, T., Wong, B., Zhan, Y., Campbell, C. L., Kong, Y., Marnett, A., Yu, F., Antunes, L., Bainbridge, M., Sabo, A., Huang, Z., Coin, L. J. M., Fang, L., Li, Q., Li, Z., Lin, H., Liu, B., Luo, R., Shao, H., Xie, Y., Ye, C., Yu, C., Zhang, F., Zheng, H., Zhu, H., Alkan, C., Dal, E., Kahveci, F., Garrison, E. P., Kural, D., Lee, W.-P., Fung Leong, W., Stromberg, M., Ward, A. N., Wu, J., Zhang, M., Daly, M. J., DePristo, M. A., Handsaker, R. E., Banks, E., Bhatia, G., del Angel, G., Genovese, G., Li, H., Kashin, S., McCarroll, S. A., Nemesh, J. C., Poplin, R. E., Yoon, S. C., Lihm, J., Makarov, V., Gottipati,

- S., Keinan, A., Rodriguez-Flores, J. L., Rausch, T., Fritz, M. H., Stütz, A. M., Beal, K., Datta, A., Herrero, J., Ritchie, G. R. S., Zerbino, D., Sabeti, P. C., Shlyakhter, I., Schaffner, S. F., Vitti, J., Cooper, D. N., Ball, E. V., Stenson, P. D., Barnes, B., Bauer, M., Keira Cheetham, R., Cox, A., Eberle, M., Kahn, S., Murray, L., Peden, J., Shaw, R., Kenny, E. E., Batzer, M. A., Konkel, M. K., Walker, J. A., MacArthur, D. G., Lek, M., Herwig, R., Ding, L., Koboldt, D. C., Larson, D., Ye, K. K., Gravel, S., Swaroop, A., Chew, E., Lappalainen, T., Erlich, Y., Gymrek, M., Frederick Willems, T., Simpson, J. T., Shriver, M. D., Rosenfeld, J. A., Bustamante, C. D., Montgomery, S. B., De La Vega, F. M., Byrnes, J. K., Carroll, A. W., DeGorter, M. K., Lacroute, P., Maples, B. K., Martin, A. R., Moreno-Estrada, A., Shringarpure, S. S., Zakharia, F., Halperin, E., Baran, Y., Cerveira, E., Hwang, J., Malhotra, A., Plewczynski, D., Radew, K., Romanovitch, M., Zhang, C., Hyland, F. C. L., Craig, D. W., Christoforides, A., Homer, N., Izatt, T., Kurdoglu, A. A., Sinari, S. A., Squire, K., Xiao, C., Sebat, J., Antaki, D., Gujral, M., Noor, A., Ye, K. K., Burchard, E. G., Hernandez, R. D., Gignoux, C. R., Haussler, D., Katzman, S. J., James Kent, W., Howie, B., Ruiz-Linares, A., Dermitzakis, E. T., Devine, S. E., Min Kang, H., Kidd, J. M., Blackwell, T., Caron, S., Chen, W., Emery, S., Fritsche, L., Fuchsberger, C., Jun, G., Li, B., Lyons, R., Scheller, C., Sidore, C., Song, S., Sliwerska, E., Taliun, D., Tan, A., Welch, R., Kate Wing, M., Zhan, X., Awadalla, P., Hodgkinson, A., Li, Y. Y., Shi, X., Quitadamo, A., Lunter, G., Marchini, J. L., Myers, S., Churchhouse, C., Delaneau, O., Gupta-Hinch, A., Kretzschmar, W., Iqbal, Z., Mathieson, I., Menelaou, A., Rimmer, A., Xifara, D. K., Oleksyk, T. K., Fu, Y. Y., Liu, X., Xiong, M., Jorde, L., Witherspoon, D., Xing, J., Browning, B. L., Browning, S. R., Hormozdiari, F., Sudmant, P. H., Khurana, E., Tyler-Smith, C., Albers, C. A., Ayub, Q., Chen, Y., Colonna, V., Jostins, L., Walter, K., Xue, Y., Gerstein, M. B., Abyzov, A., Balasubramanian, S., Chen, J., Clarke, D., Fu, Y. Y., Harmanci, A. O., Jin, M., Lee, D., Liu, J., Jasmine Mu, X., Zhang, J., Zhang, Y. Y., Hartl, C., Shakir, K., Degenhardt, J., Meiers, S., Raeder, B., Paolo Casale, F., Stegle, O., Lameijer, E.-W., Hall, I., Bafna, V., Michaelson, J., Gardner, E. J., Mills, R. E., Dayama, G., Chen, K., Fan, X., Chong, Z., Chen, T., Chaisson, M. J., Huddleston, J., Malig, M., Nelson, B. J., Parrish, N. F., Blackburne, B., Lindsay, S. J., Ning, Z., Zhang, Y. Y., Lam, H., Sisu, C., Challis, D., Evani, U. S., Lu, J., Nagaswamy, U., Yu, J., Li, W., Habegger, L., Yu, H., Cunningham, F., Dunham, I., Lage, K., Berg Jaspersen, J., Horn, H., Kim, D., Desalle, R., Narechania, A., Wilson Sayres, M. A., Mendez, F. L., David Poznik, G., Underhill, P. A., Coin, L. J. M., Mittelman, D., Banerjee, R., Cerezo, M., Fitzgerald, T. W., Louzada, S., Massaia, A., Ritchie, G. R. S., Yang, F., Kalra, D., Hale, W., Dan, X., Barnes, K. C., Beiswanger, C., Cai, H., Cao, H., Henn, B., Jones, D., Kaye, J. S., Kent, A., Kerasidou, A., Mathias, R. A., Ossorio, P. N., Parker, M., Rotimi, C. N., Royal, C. D., Sandoval, K., Su, Y., Tian, Z., Tishkoff, S., Via, M., Wang, Y., Yang, H., Yang, L., Zhu, J., Bodmer, W., Bedoya, G., Cai, Z., Gao, Y., Chu, J., Peltonen, L., Garcia-Montero, A., Orfao, A., Dutil, J., Martinez-Cruzado, J. C., Mathias, R. A., Hennis, A., Watson, H., McKenzie, C., Qadri, F., LaRocque, R., Deng, X., Asogun, D., Folarin, O., Happi, C., Omoniwa, O., Stremlau, M., Tariyal, R., Jallow, M., Sisay Joof, F., Corrah, T., Rickett, K., Kwiakowski, D., Kooner, J., Tnh Hiê'n, T., Dunstan, S. J., Thuy Hang, N., Fonnies, R., Garry, R., Kanneh, L., Moses, L., Schieffelin, J., Grant, D. S., Gallo, C., Poletti, G., Saleheen, D., Rasheed, A., Brooks, L. D., Felsenfeld, A. L., McEwen, J. E., Vaydylevich, Y., Duncanson, A., Dunn, M., and Schloss, J. A. (2015). A

- global reference for human genetic variation. *Nature*, 526(7571):68–74.
- [24] Avitzur, Y., Guo, C., Mastropaolo, L. a., Bahrami, E., Chen, H., Zhao, Z., Elkadri, A., Dhillon, S., Murchie, R., Fattouh, R., Huynh, H., Walker, J. L., Wales, P. W., Cutz, E., Kakuta, Y., Dudley, J., Kammermeier, J., Powrie, F., Shah, N., Walz, C., Nathrath, M., Kotlarz, D., Puchaka, J., Krieger, J. R., Racek, T., Kirchner, T., Walters, T. D., Brumell, J. H., Griffiths, A. M., Rezaei, N., Rashtian, P., Najafi, M., Monajemzadeh, M., Pelsue, S., McGovern, D. P. B., Uhlig, H. H., Schadt, E., Klein, C., Snapper, S. B., and Muise, A. M. (2014). Mutations in tetratricopeptide repeat domain 7A result in a severe form of very early onset inflammatory bowel disease. *Gastroenterology*, 146(4):1028–39.
- [25] Awadalla, P., Gauthier, J., Myers, R. a., Casals, F., Hamdan, F. F., Griffing, A. R., Côté, M., Henrion, E., Spiegelman, D., Tarabeux, J., Piton, A., Yang, Y., Boyko, A., Bustamante, C., Xiong, L., Rapoport, J. L., Addington, A. M., DeLisi, J. L. E., Krebs, M.-O., Joober, R., Millet, B., Fombonne, E., Mottron, L., Zilvermit, M., Keebler, J., Daoud, H., Marineau, C., Roy-Gagnon, M.-H., Dubé, M.-P., Eyre-Walker, A., Drapeau, P., Stone, E. a., Lafrenière, R. G., and Rouleau, G. a. (2010). Direct measure of the de novo mutation rate in autism and schizophrenia cohorts. *American Journal of Human Genetics*, 87(3):316–24.
- [26] Badano, J. L. and Katsanis, N. (2002). Beyond Mendel: an evolving view of human genetic disease transmission. *Nature Review Genetics*, 3(10):779–789.
- [27] Bain, S. C., Prins, J. B., Hearne, C. M., Rodrigues, N. R., Rowe, B. R., Pritchard, L. E., Ritchie, R. J., Hall, J. R., Undlien, D. E., Ronningen, K. S., and al., E. (1992). Insulin gene region-encoded susceptibility to type 1 diabetes is not restricted to HLA-DR4-positive individuals. *Nature Genetics*, 2(3):212–215.
- [28] Bakker, B., Bikker, H., Vulsma, T., Randamie, J. S. E. D. E., Wiedijk, B. M., and Vijlder, J. A. N. J. M. D. E. (2000). Two Decades of Screening for Congenital Hypothyroidism in the Netherlands: TPO Gene Mutations in Total Iodide Organification Defects (an Update). *The Journal of Clinical Endocrinology and Metabolism*, 85(10):3708–3712.
- [29] Balaesque, P. L., Ballereau, S. J., and Jobling, M. A. (2007). Challenges in human genetic diversity: Demographic history and adaptation. *Human Molecular Genetics*, 16(R2):134–139.
- [30] Baldini, A. (2005). Dissecting contiguous gene defects: TBX1. *Current Opinion in Genetics and Development*, 15(3 SPEC. ISS.):279–284.
- [31] Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. a., and Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics*, 12(11):745–55.
- [32] Barrett, J. C. and Cardon, L. R. (2006). Evaluating coverage of genome-wide association studies. *Nature Genetics*, 38(6):659–662.

- [33] Barrett, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H., Duerr, R. H., Rioux, J. D., Brant, S. R., Silverberg, M. S., Taylor, K. D., Michael, M., Bitton, A., Dassopoulos, T., Datta, L. W., Green, T., Griffiths, M., Kistner, E. O., Murtha, M. T., Regueiro, M. D., Jerome, I., Schumm, L. P., Steinhart, a. H., Targan, S. R., Ramnik, J., Prescott, J., Onnie, C. M., Fisher, S. a., Marchini, J., and Ghori, J. (2009). Genome-wide association defines more than thirty distinct susceptibility loci for Crohn's disease. *Nat Genet.*, 40(8):955–962.
- [34] Bartlett, H. and Weeks, D. L. (2011). Examining the Cardiac NK-2 Genes in Early Heart Development. *Development*, 31(3):335–341.
- [35] Baumgart, D. C. and Sandborn, W. J. (2007). Inflammatory bowel disease: clinical aspects and established and evolving therapies. *Lancet*, 369(9573):1641–1657.
- [36] Beardsall, K. and Ogilvy-Stuart, A. L. (2004). Congenital hypothyroidism. *Current Paediatrics*, 14(5):422–429.
- [37] Beasley, T. M., Erickson, S., and Allison, D. B. (2009). Rank-based inverse normal transformations are increasingly used, but are they merited? *Behavior Genetics*, 39(5):580–595.
- [38] Beaulieu, C. L., Majewski, J., Schwartzentruber, J., Samuels, M. E., Fernandez, B. A., Bernier, F. P., Brudno, M., Knoppers, B., Marcadier, J., Dymont, D., Adam, S., Bulman, D. E., Jones, S. J. M., Avard, D., Nguyen, M. T., Rousseau, F., Marshall, C., Wintle, R. F., Shen, Y., Scherer, S. W., Friedman, J. M., Michaud, J. L., and Boycott, K. M. (2014). FORGE Canada Consortium: Outcomes of a 2-year national rare-disease gene-discovery project. *American Journal of Human Genetics*, 94(6):809–817.
- [39] Begue, B., Verdier, J., Rieux-Laucat, F., Goulet, O., Morali, A., Canioni, D., Hugot, J.-P., Daussy, C., Verkarre, V., Pigneur, B., Fischer, A., Klein, C., Cerf-Bensussan, N., and Ruemmele, F. M. (2011). Defective IL10 signaling defining a subgroup of patients with inflammatory bowel disease. *The American Journal of Gastroenterology*, 106(8):1544–55.
- [40] Bell, C. J., Dinwiddie, D. L., Miller, N. a., Hateley, S. L., Ganusova, E. E., Mudge, J., Langley, R. J., Zhang, L., Lee, C. C., Schilkey, F. D., Sheth, V., Woodward, J. E., Peckham, H. E., Schroth, G. P., Kim, R. W., Kingsmore, S. F., Elena, E., and Ryan, W. (2011). Carrier Testing for Severe Childhood Recessive Disease by Next-Generation Sequencing. *Science Translational Medicine*, 3(65):65ra4.
- [41] Belsky, D. and Israel, S. (2014). Integrating Genetics and Social Science: Genetic Risk Scores. *Biodemographic Soc Biol*, 60:137–155.
- [42] Berg, J. S., Khoury, M. J., and Evans, J. P. (2011). Deploying whole genome sequencing in clinical practice and public health: meeting the challenge one bin at a time. *Genetics in medicine: official journal of the American College of Medical Genetics*, 13(6):499–504.
- [43] Bhattacharya, S., Das, A., Ghosh, S., Dasgupta, R., and Bagchi, A. (2014). Hypoglycosylation of dystroglycan due to T192M mutation: A molecular insight behind the fact. *Gene*, 537(1):108–114.

- [44] Bhoj, E., Halbach, S., McDonald-McGinn, D., Tan, C., Lande, R., Waggoner, D., and Zackai, E. (2013). Expanding the spectrum of microdeletion 4q21 syndrome: A partial phenotype with incomplete deletion of the minimal critical region and a new association with cleft palate and pierre robin sequence. *American Journal of Medical Genetics, Part A*, 161(9):2327–2333.
- [45] Biank, V., Broeckel, U., and Kugathasan, S. (2007). Pediatric inflammatory bowel disease: clinical and molecular genetics. *Inflammatory Bowel Diseases*, 13(11):1430–1438.
- [46] Biben, C., Weber, R., Kesteven, S., Stanley, E., McDonald, L., Elliott, D. A., Barnett, L., Ko, F., Robb, L., Feneley, M., and Harvey, R. P. (2000). Cardiac Septal and Valvular Dymorphogenesis in Mice Heterozygous for Mutations in the Homeobox Gene *Nkx2-5*. *Circulation Research*, 10:888–896.
- [47] Biebermann, H., Schöneberg, T., Krude, H., Schultz, G., Gudermann, T., and Grüters, a. (1997). Mutations of the human thyrotropin receptor gene causing thyroid hypoplasia and persistent congenital hypothyroidism. *The Journal of Clinical Endocrinology and Metabolism*, 82(10):3471–80.
- [48] Bittles (2001). Consanguinity and its relevance to clinical genetics. *Clinical Genetics*, 60:89–98.
- [49] Blair, D. R., Lyttle, C. S., Mortensen, J. M., Bearden, C. F., Jensen, A. B., Khiabani, H., Melamed, R., Rabadan, R., Bernstam, E. V., Brunak, S., Jensen, L. J., Nicolae, D., Shah, N. H., Grossman, R. L., Cox, N. J., White, K. P., and Rzhetsky, A. (2013). A nondegenerate code of deleterious variants in mendelian loci contributes to complex disease risk. *Cell*, 155(1):70–80.
- [50] Blau, N., Van Spronsen, F. J., and Levy, H. L. (2010). Phenylketonuria. *The Lancet*, 376(9750):1417–1427.
- [51] Bonnet, C., Andrieux, J., Béri-Dexheimer, M., Leheup, B., Boute, O., Manouvrier, S., Delobel, B., Copin, H., Receveur, a., Mathieu, M., Thiriez, G., Le Caignec, C., David, a., de Blois, M. C., Malan, V., Philippe, a., Cormier-Daire, V., Colleaux, L., Flori, E., Dollfus, H., Pelletier, V., Thauvin-Robinet, C., Masurel-Paulet, a., Faivre, L., Tardieu, M., Bahi-Buisson, N., Callier, P., Mugneret, F., Edery, P., Jonveaux, P., and Sanlaville, D. (2010). Microdeletion at chromosome 4q21 defines a new emerging syndrome with marked growth restriction, mental retardation and absent or severely delayed speech. *Journal of Medical Genetics*, 47(6):377–84.
- [52] Booth, C., Gilmour, K. C., Veys, P., Gennery, A. R., Slatter, M. A., Chapel, H., Heath, P. T., Steward, C. G., Smith, O., Meara, A. O., Kerrigan, H., Mahlaoui, N., Cavazzana-calvo, M., Fischer, A., Moshous, D., Blanche, S., Schmid, J. P., Latour, S., Saint-basile, G. D., Albert, M., Notheis, G., Rieber, N., Strahm, B., Ritterbusch, H., Sedlacek, P., Jazbec, J., Kanegane, H., Nichols, K. E., Hanson, I. C., and Kapoor, N. (2011). X-linked lymphoproliferative disease due to SAP / SH2D1A deficiency: a multicenter study on the manifestations , management and outcome of the disease. *Blood*, 117(1):53–62.

- [53] Botstein, D. and Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genetics*, 33 Suppl(march):228–37.
- [54] Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, 32(3):314–31.
- [55] Boycott, K. M., Vanstone, M. R., Bulman, D. E., and MacKenzie, A. E. (2013). Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nature Reviews Genetics*, 14(10):681–91.
- [56] Brandau, O., Schuster, V., Weiss, M., Hellebrand, H., Fink, F. M., Kreczy, A., Friedrich, W., Strahm, B., Niemeyer, C., Belohradsky, B. H., and Meindl, A. (1999). Epstein-Barr virus-negative boys with non-Hodgkin lymphoma are mutated in the SH2D1A gene, as are patients with X-linked lymphoproliferative disease (XLP). *Human Molecular Genetics*, 8(13):2407–2413.
- [57] BROAD Institute (2013). White Paper: Creating a Global Alliance to Enable Responsible Sharing of Genomic and Clinical Data. *BROAD Communications*.
- [58] Brommage, R., Liu, J., Hansen, G. M., Kirkpatrick, L. L., Potter, D. G., Sands, A. T., Zambrowicz, B., Powell, D. R., and Vogel, P. (2014). High-throughput screening of mouse gene knockouts identifies established and novel skeletal phenotypes. *Bone Research*, 2:14034.
- [59] Brown, R. S. and Demmer, L. a. (2002). The etiology of thyroid dysgenesis-still an enigma after all these years. *The Journal of Clinical Endocrinology and Metabolism*, 87(9):4069–71.
- [60] Brunham, L. R. and Hayden, M. R. (2013). Hunting human disease genes: Lessons from the past, challenges for the future. *Human Genetics*, 132(6):603–617.
- [61] Brust, E. S., Beltrao, C. B., Chammas, M. C., Watanabe, T., Sapienza, M. T., and Marui, S. (2012). Absence of mutations in PAX8, NKX2.5, and TSH receptor genes in patients with thyroid dysgenesis. *Arq Bras Endocrinol Metabol*, 56(4):8–12.
- [62] Buijsse, B., Simmons, R. K., Griffin, S. J., and Schulze, M. B. (2011). Risk assessment tools for identifying individuals at risk of developing type 2 diabetes. *Epidemiologic Reviews*, 33(1):46–62.
- [63] Burniat, A., Pirson, I., Vilain, C., Kulik, W., Afink, G., Moreno-Reyes, R., Corvilain, B., and Abramowicz, M. (2012). Iodotyrosine deiodinase defect identified via genome-wide approach. *Journal of Clinical Endocrinology and Metabolism*, 97(July 2012):1276–1283.
- [64] Bush, W. S. and Moore, J. H. (2012). Chapter 11: Genome-Wide Association Studies. *PLoS Computational Biology*, 8(12).

- [65] Buysse, K., Riemersma, M., Powell, G., Van reeuwijk, J., Chitayat, D., Roscioli, T., Kamsteeg, E. J., Van den elzen, C., Van beusekom, E., Blaser, S., Babul-Hirji, R., Halliday, W., Wright, G. J., Stemple, D. L., Lin, Y. Y., Lefeber, D. J., and Van bokhoven, H. (2013). Missense mutations in beta-1,3-N-acetylglucosaminyltransferase 1 (B3GNT1) cause Walker-Warburg syndrome. *Human Molecular Genetics*, 22(9):1746–1754.
- [66] Cangul, H., Aycaan, Z., Olivera-Nappa, A., Saglam, H., Schoenmakers, N. a., Boelaert, K., Cetinkaya, S., Tarim, O., Bober, E., Darendeliler, F., Bas, V., Demir, K., Aydin, B. K., Kendall, M., Cole, T., Högler, W., Chatterjee, V. K. K., Barrett, T. G., and Maher, E. R. (2013). Thyroid dysmorphogenesis is mainly caused by TPO mutations in consanguineous community. *Clinical Endocrinology*, 79(2):275–81.
- [67] Cangul, H., Morgan, N. V., Forman, J. R., Saglam, H., Aycaan, Z., Yakut, T., Gulden, T., Tarim, O., Bober, E., Cesur, Y., Kirby, G. A., Pasha, S., Karkucak, M., Eren, E., Cetinkaya, S., Bas, V., Demir, K., Yuca, S. A., Meyer, E., Kendall, M., Hogler, W., Barrett, T. G., and Maher, E. R. (2010). Novel TSHR mutations in consanguineous families with congenital nongoitrous hypothyroidism. *Clinical Endocrinology*, 73(5):671–677.
- [68] Cannioto, Z., Berti, I., Martelossi, S., Bruno, I., Giurici, N., Crovella, S., and Ventura, a. (2009). IBD and IBD mimicking enterocolitis in children younger than 2 years of age. *European Journal of Pediatrics*, 168(2):149–55.
- [69] Cao, X.-Y., Jiang, X.-M., Dou, Z.-H., Rakeman, M. A., Zhang, M.-L., O'Donnell, K., Ma, T., Amette, K., DeLong, N., and DeLong, G. R. (1994). Timing of Vulnerability of the Brain to Iodine Deficiency in Endemic Cretinism. *The New England Journal of Medicine*, 331:1739–1744.
- [70] Carre, A., Rachdi, L., Tron, E., Richard, B., Castanet, M., Schlumberger, M., Bidart, J. M., Szinnai, G., and Polak, M. (2011). Hes1 is required for appropriate morphogenesis and differentiation during mouse thyroid gland development. *PLoS ONE*, 6(2).
- [71] Carson, A. R., Smith, E. N., Matsui, H., Brækkan, S. K., Jepsen, K., Hansen, J.-b., and Frazer, K. A. (2014). Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. *BMC Bioinformatics*, 15:1–15.
- [72] Carss, K. J., Hillman, S. C., Parthiban, V., McMullan, D. J., Maher, E. R., Kilby, M. D., and Hurles, M. E. (2014). Exome sequencing improves genetic diagnosis of structural fetal abnormalities revealed by ultrasound. *Human Molecular Genetics*, 23(12):3269–77.
- [73] Carter, M. J. (2004). Guidelines for the management of inflammatory bowel disease in adults. *Gut*, 53(suppl_5):v1–v16.
- [74] Castanet, M., Polak, M., Bonai, C., Lyonnet, S., Czernichow, P., and Le, J. (2001). Nineteen Years of National Screening for Congenital Hypothyroidism: Familial Cases with Thyroid Dysgenesis. *The Journal of Clinical Endocrinology & Metabolism*, 86(5):2009–2014.

- [75] Castanet, M., Sura-Trueba, S., Chauty, A., Carré, A., de Roux, N., Heath, S., Léger, J., Lyonnet, S., Czernichow, P., and Polak, M. (2005). Linkage and mutational analysis of familial thyroid dysgenesis demonstrate genetic heterogeneity implicating novel genes. *European Journal of Human Genetics: EJHG*, 13(2):232–239.
- [76] Cheng, A. Y., Teo, Y. Y., and Ong, R. T. H. (2014). Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals. *Bioinformatics*, 30(12):1707–1713.
- [77] Chong, J. X., Buckingham, K. J., Jhangiani, S. N., Boehm, C., Sobreira, N., Smith, J. D., Harrell, T. M., McMillin, M. J., Wiszniewski, W., Gambin, T., Coban Akdemir, Z. H., Doheny, K., Scott, A. F., Avramopoulos, D., Chakravarti, A., Hoover-Fong, J., Mathews, D., Witmer, P. D., Ling, H., Hetrick, K., Watkins, L., Patterson, K. E., Reinier, F., Blue, E., Muzny, D., Kircher, M., Bilguvar, K., Lopez-Giraldez, F., Sutton, V. R., Tabor, H. K., Leal, S. M., Gunel, M., Mane, S., Gibbs, R. A., Boerwinkle, E., Hamosh, A., Shendure, J., Lupski, J. R., Lifton, R. P., Valle, D., Nickerson, D. A., and Bamshad, M. J. (2015). The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *American Journal of Human Genetics*, 97(2):199–215.
- [78] Christodoulou, K., Wiskin, A. E., Gibson, J., Tapper, W., Willis, C., Afzal, N. a., Upstill-Goddard, R., Holloway, J. W., Simpson, M. a., Beattie, R. M., Collins, A., and Ennis, S. (2013). Next generation exome sequencing of paediatric inflammatory bowel disease patients identifies rare and novel variants in candidate genes. *Gut*, 62(7):977–984.
- [79] Chun, S. and Fay, J. C. (2009). Identification of deleterious mutations within three human genomes. *Genome research*, 19:1553–1561.
- [80] Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., and Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w 1118; iso-2; iso-3. *Fly*, 6(2):80–92.
- [81] Ciofani, M., Madar, A., Galan, C., Sellars, M., MacE, K., Pauli, F., Agarwal, A., Huang, W., Parkurst, C. N., Muratet, M., Newberry, K. M., Meadows, S., Greenfield, A., Yang, Y., Jain, P., Kirigin, F. K., Birchmeier, C., Wagner, E. F., Murphy, K. M., Myers, R. M., Bonneau, R., and Littman, D. R. (2012). A validated regulatory network for Th17 cell specification. *Cell*, 151(2):289–303.
- [82] Citterio, C. E., Machiavelli, G. A., Miras, M. B., Gruñeiro-Papendieck, L., Lachlan, K., Sobrero, G., Chiesa, A., Walker, J., Muñoz, L., Testa, G., Belforte, F. S., González-Sarmiento, R., Rivolta, C. M., and Targovnik, H. M. (2013). New insights into thyroglobulin gene: Molecular analysis of seven novel mutations associated with goiter and hypothyroidism. *Molecular and Cellular Endocrinology*, 365(2):277–291.
- [83] Civitareale, D., Filippis, V. D. E., Cisternino, C., and Tassi, V. (1997). Absence of Mutations in the Gene Transcription Thyroid Dysgenesis in Patients with EncodingThyroid. *Thyroid*, 7(3).

- [84] Clark, M. J., Chen, R. R., Lam, H. Y. K., Karczewski, K. J., Chen, R. R., Euskirchen, G., Butte, A. J., and Snyder, M. (2011). Performance comparison of exome DNA sequencing technologies. *Nature Biotechnology*, 29(10):908–914.
- [85] Clarke, G. M., Anderson, C. a., Pettersson, F. H., Cardon, L. R., and Andrew, P. (2011). Basic statistical analysis in genetic case-control studies. *Nature Protocols*, 6(2):121–133.
- [86] Clément, K., Vaisse, C., Lahlou, N., Cabrol, S., Pelloux, V., Cassuto, D., Gourmelon, M., Dina, C., Chambaz, J., Lacorte, J. M., Basdevant, a., Bougnères, P., Lebouc, Y., Froguel, P., and Guy-Grand, B. (1998). A mutation in the human leptin receptor gene causes obesity and pituitary dysfunction. *Nature*, 392(6674):398–401.
- [87] Cleynen, I., Boucher, G., Jostins, L., Schumm, L. P., Zeissig, S., Ahmad, T., Andersen, V., Andrews, J. M., Annese, V., Brand, S., Brant, S. R., Cho, J. H., Daly, M. J., Dubinsky, M., Duerr, R. H., Ferguson, L. R., Franke, A., Kupcinskis, L., Lawrance, I. C., Lee, J. C., Satsangi, J., Schreiber, S., Théâtre, E., Jong, A. E. V. D. M.-d., Mansfi, J., Silverberg, M. S., Radford-smith, G., MCGovern, D. P. B., Barrett, J. C., Lees, C. W., and Adelaide, R. (2015). Inherited determinants of Crohn ’ s disease and ulcerative colitis phenotypes: a genetic association study. *The Lancet*, 6736(15):1–12.
- [88] Clifton-Bligh, R. J., Wentworth, J. M., Heinz, P., Crisp, M. S., John, R., Lazarus, J. H., Ludgate, M., and Chatterjee, V. K. (1998). Mutation of the gene encoding human TTF-2 associated with thyroid agenesis, cleft palate and choanal atresia. *Nature Genetics*, 19(4):399–401.
- [89] Coffey, a. J., Brooksbank, R. a., Brandau, O., Oohashi, T., Howell, G. R., Bye, J. M., Cahn, a. P., Durham, J., Heath, P., Wray, P., Pavitt, R., Wilkinson, J., Leversha, M., Huckle, E., Shaw-Smith, C. J., Dunham, a., Rhodes, S., Schuster, V., Porta, G., Yin, L., Serafini, P., Sylla, B., Zollo, M., Franco, B., Bolino, a., Seri, M., Lanyi, a., Davis, J. R., Webster, D., Harris, a., Lenoir, G., de St Basile, G., Jones, a., Behloradsky, B. H., Achatz, H., Murken, J., Fassler, R., Sumegi, J., Romeo, G., Vaudin, M., Ross, M. T., Meindl, a., and Bentley, D. R. (1998). Host response to EBV infection in X-linked lymphoproliferative disease results from mutations in an SH2-domain encoding gene. *Nature Genetics*, 20(2):129–135.
- [90] Cognard, N., Scerbo, M. J., Obringer, C., Yu, X., Costa, F., Haser, E., Le, D., Stoetzel, C., Roux, M. J., Moulin, B., Dollfus, H., and Marion, V. (2015). Comparing the Bbs10 complete knockout phenotype with a specific renal epithelial knockout one highlights the link between renal defects and systemic inactivation in mice. *Cilia*, 4(1):10.
- [91] Colombel, J. and Grandbastien, B. (1996). Clinical characteristics of Crohn’s disease in 72 families. *Gastroenterology*, 111(3):604–607.
- [92] Congdon, T., Nguyen, L. Q., Nogueira, C. R., Habiby, R. L., Medeiros-neto, G., Kopp, P., Endocrinology, D., C, M. M. T., and H, P. E. R. L. (2001). A Novel Mutation (Q40P) in PAX8 Associated with Congenital Hypothyroidism and Thyroid Hypoplasia: Evidence for Phenotypic Variability in Mother and Child. *The Journal of Clinical Endocrinology and Metabolism*, 86(August):3962–3967.

- [93] Connelly, T. M., Berg, A. S., Iii, L. H., Brinton, D., Deiling, S., and Koltun, W. A. (2015). Genetic determinants associated with early age of diagnosis of IBD. *Diseases of the Colon and Rectum*, 58(3):321–327.
- [94] Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T. D., Barnes, C., Campbell, P., Fitzgerald, T., Hu, M., Ihm, C. H., Kristiansson, K., Macarthur, D. G., Macdonald, J. R., Onyiah, I., Pang, A. W. C., Robson, S., Stirrups, K., Valsesia, A., Walter, K., Wei, J., Tyler-Smith, C., Carter, N. P., Lee, C., Scherer, S. W., and Hurles, M. E. (2010). Origins and functional impact of copy number variation in the human genome. *Nature*, 464(7289):704–12.
- [95] CONVERGE Consortium (2015). Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature*, 523(7562):588.
- [96] Cooley, M. A., Kern, C. B., Fresco, V. M., Wessels, A., Thompson, R. P., McQuinn, T. C., Twal, W. O., Mjaatvedt, C. H., Drake, C. J., and Argraves, W. S. (2008). Fibulin-1 is required for morphogenesis of neural crest-derived structures. *Developmental Biology*, 319(2):336–345.
- [97] Cooper, D. N., Krawczak, M., Polychronakos, C., Tyler-Smith, C., and Kehrer-Sawatzki, H. (2013). Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Human genetics*, 132(10):1077–130.
- [98] Cooper, G. M., Stone, E. a., Asimenos, G., Green, E. D., Batzoglou, S., and Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. *Genome research*, 15(7):901–13.
- [99] Corbetta, C., Weber, G., Cortinovis, F., Calebiro, D., Passoni, A., Vigone, M. C., Beck-Peccoz, P., Chiumello, G., and Persani, L. (2009). A 7-year experience with low blood TSH cutoff levels for neonatal screening reveals an unsuspected frequency of congenital hypothyroidism (CH). *Clinical Endocrinology*, 71(5):739–45.
- [100] Corder, E. H., Saunders, a. M., Strittmatter, W. J., Schmechel, D. E., Gaskell, P. C., Small, G. W., Roses, a. D., Haines, J. L., and Pericak-Vance, M. a. (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer’s disease in late onset families. *Science (New York, N.Y.)*, 261(5123):921–923.
- [101] Crow, J. F. and Weinberg, W. (2000). The origins, patterns and implications of human spontaneous mutation. *Nature Reviews Genetics*, 1(October):40–47.
- [102] Cutler DJ, Zwick ME, Okou DT, Prahalad S, Walters T, G. S. (2015). Dissecting allele architecture of early onset IBD using high-density genotyping. *PLoS ONE*, 10(6):1–12.
- [103] D’Alessandro, L. C. A., Al Turki, S., Manickaraj, A. K., Manase, D., Mulder, B. J. M., Bergin, L., Rosenberg, H. C., Mondal, T., Gordon, E., Loughheed, J., Smythe, J., Devriendt, K., Bhattacharya, S., Watkins, H., Bentham, J., Bowdin, S., Hurles, M. E., and Mital, S. (2016). Exome sequencing identifies rare variants in multiple genes in atrioventricular septal defect. *Genetics in Medicine*, 18(2):189–198.

- [104] Damgaard, R. B., Fiil, B. K., Speckmann, C., Yabal, M., zur Stadt, U., Bekker-Jensen, S., Jost, P. J., Ehl, S., Mailand, N., and Gyrd-Hansen, M. (2013). Disease-causing mutations in the XIAP BIR2 domain impair NOD2-dependent immune signalling. *EMBO Molecular Medicine*, 5(8):1278–1295.
- [105] Danecek, P., Auton, A., Abecasis, G., Albers, C. a., Banks, E., DePristo, M. a., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., and Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics (Oxford, England)*, 27(15):2156–8.
- [106] Danjou, F., Zoledziewska, M., Sidore, C., Steri, M., Busonero, F., Maschio, A., Mulas, A., Perseu, L., Barella, S., Porcu, E., Pistis, G., Pitzalis, M., Pala, M., Menzel, S., Metrustry, S., Spector, T. D., Leoni, L., Angius, A., Uda, M., Moi, P., Thein, S. L., Galanello, R., Abecasis, G. R., Schlessinger, D., Sanna, S., and Cucca, F. (2015). Genome-wide association analyses based on whole-genome sequencing in Sardinia provide insights into regulation of hemoglobin levels. *Nature Genetics*, 47(11):1264–71.
- [107] Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Computational Biology*, 6(12).
- [108] de Bakker, P. I. W., Ferreira, M. A. R., Jia, X., Neale, B. M., Raychaudhuri, S., and Voight, B. F. (2008). Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Human Molecular Genetics*, 17(R2):122–128.
- [109] de Lange, K. M. and Barrett, J. C. (2015). Understanding inflammatory bowel disease via immunogenetics. *Journal of Autoimmunity*, 64:91–100.
- [110] de Lange, K. M., Moutsianas, L., Lee, J. C., Lamb, C. A., Luo, Y., Kennedy, N. A., Jostins, L., Rice, D. L., Gutierrez-Achury, J., Ji, S.-G., Heap, G., Nimmo, E. R., Edwards, C., Henderson, P., Mowat, C., Sanderson, J., Satsangi, J., Simmons, A., Wilson, D. C., Tremelling, M., Hart, A., Mathew, C. G., Newman, W. G., Parkes, M., Lees, C. W., Uhlig, H., Hawkey, C., Prescott, N. J., Ahmad, T., Mansfield, J., Anderson, C. A., and Barrett, J. (2016). Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *BioRxiv*, pages 1–19.
- [111] de Ligt, J., Willemsen, M. H., van Bon, B. W., Kleefstra, T., Yntema, H. G., Kroes, T., Vulto-van Silfhout, A. T., Koolen, D. a., de Vries, P., Gilissen, C., del Rosario, M., Hoischen, A., Scheffer, H., de Vries, B. B., Brunner, H. G., Veltman, J. a., and Vissers, L. E. (2012). Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability. *New England Journal of Medicine*, 367(20):1921–1929.
- [112] de Luca, A., Smeekens, S. P., Casagrande, A., Iannitti, R., Conway, K. L., Gresnigt, M. S., Begun, J., Plantinga, T. S., Joosten, L. a. B., van der Meer, J. W. M., Chamilos, G., Netea, M. G., Xavier, R. J., Dinarello, C. a., Romani, L., and van de Veerdonk, F. L. (2014). IL-1 receptor blockade restores autophagy and reduces inflammation in chronic granulomatous disease in mice and in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 111(9):3526–31.

- [113] DeBaun, M. R., Niemitz, E. L., McNeil, D. E., Brandenburg, S. A., Lee, M. P., and Feinberg, A. P. (2002). Epigenetic alterations of H19 and LIT1 distinguish patients with Beckwith-Wiedemann syndrome with cancer and birth defects. *American Journal of Human Genetics*, 70(3):604–11.
- [114] Deladoey, J. (2012). Congenital Hypothyroidism due to Thyroid Dysgenesis: From Epidemiology to Molecular Mechanisms. In Dr. Drahomira Springer (Ed.), editor, *A New Look at Hypothyroidism*, pages 229–242. InTech.
- [115] Dentice, M., Cordeddu, V., Rosica, A., Ferrara, A. M., Santarpia, L., Salvatore, D., Chiovato, L., Perri, A., Moschini, L., Fazzini, C., Olivieri, A., Costa, P., Stoppioni, V., Baserga, M., De Felice, M., Sorcini, M., Fenzi, G., Di Lauro, R., Tartaglia, M., and Macchia, P. E. (2006). Missense mutation in the transcription factor NKX2-5: a novel molecular event in the pathogenesis of thyroid dysgenesis. *The Journal of Clinical Endocrinology and Metabolism*, 91(4):1428–33.
- [116] DePristo, M. a., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. a., del Angel, G., Rivas, M. a., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., and Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491–8.
- [117] Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Merkel, A., Gonzalez, D., Lagarde, J., Jb, B., Lipovich, L., Gonzalez, J., Ca, D., Tr, G., Hubbard, T., Notredame, C., Harrow, J., and Guigó, R. (2012). The GENCODE v7 catalogue of human long non-coding RNAs: Analysis of their structure , evolution and expression. *Genome Research*, 22:1775–1789.
- [118] Desmet, F. O., Hamroun, D., Collod-Bérout, G., Claustres, M., and Bérout, C. (2010). Bioinformatics identification of splice site signals and prediction of mutation effects. *Research Advances in Nucleic Acids Research*, (September 2015):1–14.
- [119] Di Palma, T., Conti, A., de Cristofaro, T., Scala, S., Nitsch, L., and Zannini, M. (2011). Identification of novel Pax8 targets in FRTL-5 thyroid cells by gene silencing and expression microarray analysis. *PloS one*, 6(9):e25162.
- [120] Dietrich, J. W., Landgrafe, G., and Fotiadou, E. H. (2012). TSH and Thyrotropic Agonists: Key Actors in Thyroid Homeostasis. *Journal of Thyroid Research*, 2012:351864.
- [121] Dinwiddie, D. L., Bracken, J. M., Bass, J. a., Christenson, K., Soden, S. E., Saunders, C. J., Miller, N. a., Singh, V., Zwick, D. L., Roberts, C. C., Dalal, J., and Kingsmore, S. F. (2013). Molecular diagnosis of infantile onset inflammatory bowel disease by exome sequencing. *Genomics*, 102(5-6):442–7.
- [122] Do, R., Kathiresan, S., and Abecasis, G. R. (2012). Exome sequencing and complex disease: practical aspects of rare variant association studies. *Human Molecular Genetics*, 21(R1):R1–9.

- [123] Do, R., Stitzziel, N. O., Won, H.-H., Jørgensen, A. B., Duga, S., Angelica Merlini, P., Kiezun, A., Farrall, M., Goel, A., Zuk, O., Guella, I., Asselta, R., Lange, L. a., Peloso, G. M., Auer, P. L., Girelli, D., Martinelli, N., Farlow, D. N., DePristo, M. a., Roberts, R., Stewart, A. F. R., Saleheen, D., Danesh, J., Epstein, S. E., Sivapalaratnam, S., Hovingh, G. K., Kastelein, J. J., Samani, N. J., Schunkert, H., Erdmann, J., Shah, S. H., Kraus, W. E., Davies, R., Nikpay, M., Johansen, C. T., Wang, J., Hegele, R. a., Hechter, E., Marz, W., Kleber, M. E., Huang, J., Johnson, A. D., Li, M., Burke, G. L., Gross, M., Liu, Y., Assimes, T. L., Heiss, G., Lange, E. M., Folsom, A. R., Taylor, H. a., Olivieri, O., Hamsten, A., Clarke, R., Reilly, D. F., Yin, W., Rivas, M. a., Donnelly, P., Rossouw, J. E., Psaty, B. M., Herrington, D. M., Wilson, J. G., Rich, S. S., Bamshad, M. J., Tracy, R. P., Cupples, L. A., Rader, D. J., Reilly, M. P., Spertus, J. a., Cresci, S., Hartiala, J., Tang, W. H. W., Hazen, S. L., Allayee, H., Reiner, A. P., Carlson, C. S., Kooperberg, C., Jackson, R. D., Boerwinkle, E., Lander, E. S., Schwartz, S. M., Siscovick, D. S., McPherson, R., Tybjaerg-Hansen, A., Abecasis, G. R., Watkins, H., Nickerson, D. a., Ardissono, D., Sunyaev, S. R., O'Donnell, C. J., Altshuler, D., Gabriel, S., and Kathiresan, S. (2015). Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. *Nature*, 518(7537):102–6.
- [124] Donaldson, M. and Jones, J. (2013). Optimising Outcome in Congenital Hypothyroidism: Current Opinions on Best Practice in Initial Assessment and Subsequent Management. *J Clin Res Pediatr Endocrinol*, 5(Suppl 1):13–22.
- [125] Dudas, P. L., Mentone, S., Greineder, C. F., Biemesderfer, D., Aronson, P. S., and Paul, L. (2006). Immunolocalization of anion transporter Slc26a7 in mouse kidney. *American journal of Physiology*, 8029:937–945.
- [126] Duerr, R. H., Taylor, K. D., Brant, S. R., Rioux, J. D., Silverberg, M. S., Daly, M. J., Steinhardt, a. H., Abraham, C., Regueiro, M., Griffiths, a., Dassopoulos, T., Bitton, a., Yang, H., Targan, S. R., Datta, L. W., Kistner, E. O., Schumm, L. P., Lee, a. T., Gregersen, P. K., Barmada, M. M., Rotter, J. I., Nicolae, D. L., and Cho, J. H. (2006). A Genome-Wide Association Study Identifies IL23R as an Inflammatory Bowel Disease Gene. *Science*, 314(5804):1461–1463.
- [127] El Omari, K., de Mesmaeker, J., Karia, D., Ginn, H., Bhattacharya, S., and Mancini, E. J. (2012). Structure of the DNA-bound T-box domain of human TBX1, a transcription factor associated with the DiGeorge syndrome. *Proteins: Structure, Function and Bioinformatics*, 80(2):655–660.
- [128] Eng, L., Coutinho, G., Nahas, S., Yeo, G., Tanouye, R., Babaei, M., Dork, T., Burge, C., and Gatti, R. A. (2004). Nonclassical Splicing Mutations in the Coding and Noncoding Regions of the ATM Gene: Maximum Entropy Estimates of Splice Junction Strengths. *Human Mutation*, 23(1):67–76.
- [129] Essers, J., Lee, J., Kugathasan, S., Stevens, C. R., Consortium, N. I. B. D. G., Grant, R., and Daly, M. J. (2009). Established genetic risk factors do not distinguish early and later onset Crohn's Disease. *Inflammatory bowel disease*, 15(10):1508–1514.
- [130] Evangelou, E. and Ioannidis, J. P. A. (2013). Meta-analysis methods for genome-wide association studies and beyond. *Nature Reviews Genetics*, 14(6):379–89.

- [131] Evans, D. M. and Cardon, L. R. (2004). Guidelines for genotyping in genomewide linkage studies: single-nucleotide-polymorphism maps versus microsatellite maps. *American Journal of Human Genetics*, 75:687–692.
- [132] Evans, D. M., Spencer, C. C. a., Pointon, J. J., Su, Z., Harvey, D., Kochan, G., Oppermann, U., Dilthey, A., Pirinen, M., Stone, M. a., Appleton, L., Moutsianas, L., Leslie, S., Wordsworth, T., Kenna, T. J., Karaderi, T., Thomas, G. P., Ward, M. M., Weisman, M. H., Farrar, C., Bradbury, L. a., Danoy, P., Inman, R. D., Maksymowych, W., Gladman, D., Rahman, P., Morgan, A., Marzo-Ortega, H., Bowness, P., Gaffney, K., Gaston, J. S. H., Smith, M., Bruges-Armas, J., Couto, A.-R., Sorrentino, R., Paladini, F., Ferreira, M. a., Xu, H., Liu, Y., Jiang, L., Lopez-Larrea, C., Díaz-Peña, R., López-Vázquez, A., Zayats, T., Band, G., Bellenguez, C., Blackburn, H., Blackwell, J. M., Bramon, E., Bumpstead, S. J., Casas, J. P., Corvin, A., Craddock, N., Deloukas, P., Dronov, S., Duncanson, A., Edkins, S., Freeman, C., Gillman, M., Gray, E., Gwilliam, R., Hammond, N., Hunt, S. E., Jankowski, J., Jayakumar, A., Langford, C., Liddle, J., Markus, H. S., Mathew, C. G., McCann, O. T., McCarthy, M. I., Palmer, C. N. a., Peltonen, L., Plomin, R., Potter, S. C., Rautanen, A., Ravindrarajah, R., Ricketts, M., Samani, N., Sawcer, S. J., Strange, A., Trembath, R. C., Viswanathan, A. C., Waller, M., Weston, P., Whittaker, P., Widaa, S., Wood, N. W., McVean, G., Reveille, J. D., Wordsworth, B. P., Brown, M. a., and Donnelly, P. (2011). Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. *Nature Genetics*, 43(8):761–767.
- [133] Evans, D. M., Visscher, P. M., and Wray, N. R. (2009). Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Human Molecular Genetics*, 18(18):3525–3531.
- [134] Evans, W. E. and Relling, M. V. (2007). Pharmacogenomics: Translating Functional Genomics into Rational Therapeutics. *Science*, 487(1999).
- [135] Exome Aggregation Consortium (2015). Analysis of protein-coding genetic variation in 60,706 humans. *bioRxiv*, pages 1–26.
- [136] Eyre-Walker, A. and Keightley, P. D. (2007). The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, 8(8):610–8.
- [137] Fagman, H., Amendola, E., Parrillo, L., Zoppoli, P., Marotta, P., Scarfò, M., De Luca, P., de Carvalho, D. P., Ceccarelli, M., De Felice, M., and Di Lauro, R. (2011). Gene expression profiling at early organogenesis reveals both common and diverse mechanisms in foregut patterning. *Developmental biology*, 359(2):163–75.
- [138] Fagman, H., Grände, M., Gritli-Linde, A., and Nilsson, M. (2004). Genetic deletion of sonic hedgehog causes hemiagenesis and ectopic development of the thyroid in mouse. *The American journal of pathology*, 164(5):1865–1872.
- [139] Fagman, H., Liao, J., Westerlund, J., Andersson, L., Morrow, B. E., and Nilsson, M. (2007). The 22q11 deletion syndrome candidate gene *Tbx1* determines thyroid size and positioning. *Human Molecular Genetics*, 16(3):276–85.

- [140] Fagman, H. and Nilsson, M. (2010a). Morphogenesis of the thyroid gland. *Molecular and cellular endocrinology*, 323(1):35–54.
- [141] Fagman, H. and Nilsson, M. (2010b). Morphogenetics of early thyroid development. *Journal of Molecular Endocrinology*, 46(1):R33–R42.
- [142] Fankhauser, C., Chory, J., Kircher, S., Ringli, C., Boylan, M. T., Quail, P. H., Deng, X. W., Puente, P., Whitelam, G. C., Harberd, N. P., Lisch, D. R., Quail, P. H., and Wang, H. (2007). A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*, 318(November):1305–1309.
- [143] Felice, M. D., Ovitt, C., Biffali, E., Rodriguez-mallon, A., Arra, C., Anastassiadis, K., Macchia, P. E., Mattei, M.-g., Mariano, A., Schöler, H., Macchia, V., and Lauro, R. D. (1998). A mouse model for hereditary thyroid dysgenesis and cleft palate. *Nature Genetics*, 19(august):3–6.
- [144] Felice, M. D. E., Lauro, R. D. I., Zoologica, S., Dohrn, A., and Biology, M. (2004). Thyroid Development and Its Disorders: Genetics and Molecular Mechanisms. *Endocrine Reviews*, 25(August):722–746.
- [145] Fellows, A., Griffin, M. E., Petrella, B. L., Zhong, L., Parvin-nejad, F. P., and Matera, A. G. (2012). AUF1/hnRNP D represses expression of VEGF in macrophages. *Molecular Biology of the Cell*, 23:1414–1422.
- [146] Fernández, L. P., López-Márquez, A., Martínez, A. M., Gómez-López, G., and Santisteban, P. (2013). New insights into FoxE1 functions: identification of direct FoxE1 targets in thyroid cells. *PloS one*, 8(5):e62849.
- [147] Fernández, L. P., López-Márquez, A., and Santisteban, P. (2015). Thyroid transcription factors in development, differentiation and disease. *Nature Reviews. Endocrinology*, 11(1):29–42.
- [148] Filipovich, A. H., Zhang, K., Snow, A. L., and Marsh, R. A. (2010). X-linked lymphoproliferative syndromes: Brothers or distant cousins? *Blood*, 116(18):3398–3408.
- [149] Firth, H. V., Richards, S. M., Bevan, A. P., Clayton, S., Corpas, M., Rajan, D., Vooren, S. V., Moreau, Y., Pettett, R. M., and Carter, N. P. (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *American Journal of Human Genetics*, 84(4):524–533.
- [150] Fitzgerald, T. W., Gerety, S. S., Jones, W. D., van Kogelenberg, M., King, D. A., McRae, J., Morley, K. I., Parthiban, V., Al-Turki, S., Ambridge, K., Barrett, D. M., Bayzatinova, T., Clayton, S., Coomber, E. L., Gribble, S., Jones, P., Krishnappa, N., Mason, L. E., Middleton, A., Miller, R., Prigmore, E., Rajan, D., Sifrim, A., Tivey, A. R., Ahmed, M., Akawi, N., Andrews, R., Anjum, U., Archer, H., Armstrong, R., Balasubramanian, M., Banerjee, R., Baralle, D., Batstone, P., Baty, D., Bennett, C., Berg, J., Bernhard, B., Bevan, A. P., Blair, E., Blyth, M., Bohanna, D., Bourdon, L., Bourn, D., Brady, A., Bragin, E., Brewer, C., Brueton, L., Brunstrom, K., Bumpstead, S. J., Bunyan, D. J., Burn, J., Burton, J., Canham, N., Castle, B.,

- Chandler, K., Clasper, S., Clayton-Smith, J., Cole, T., Collins, A., Collinson, M. N., Connell, F., Cooper, N., Cox, H., Cresswell, L., Cross, G., Crow, Y., D'Alessandro, M., Dabir, T., Davidson, R., Davies, S., Dean, J., Deshpande, C., Devlin, G., Dixit, A., Dominiczak, A., Donnelly, C., Donnelly, D., Douglas, A., Duncan, A., Eason, J., Edkins, S., Ellard, S., Ellis, P., Elmslie, F., Evans, K., Everest, S., Fendick, T., Fisher, R., Flinter, F., Foulds, N., Fryer, A., Fu, B., Gardiner, C., Gaunt, L., Ghali, N., Gibbons, R., Gomes Pereira, S. L., Goodship, J., Goudie, D., Gray, E., Greene, P., Greenhalgh, L., Harrison, L., Hawkins, R., Hellens, S., Henderson, A., Hobson, E., Holden, S., Holder, S., Hollingsworth, G., Homfray, T., Humphreys, M., Hurst, J., Ingram, S., Irving, M., Jarvis, J., Jenkins, L., Johnson, D., Jones, D., Jones, E., Josifova, D., Joss, S., Kaemba, B., Kazembe, S., Kerr, B., Kini, U., Kinning, E., Kirby, G., Kirk, C., Kivuva, E., Kraus, A., Kumar, D., Lachlan, K., Lam, W., Lampe, A., Langman, C., Lees, M., Lim, D., Lowther, G., Lynch, S. A., Magee, A., Maher, E., Mansour, S., Marks, K., Martin, K., Maye, U., McCann, E., McConnell, V., McEntagart, M., McGowan, R., McKay, K., McKee, S., McMullan, D. J., McNerlan, S., Mehta, S., Metcalfe, K., Miles, E., Mohammed, S., Montgomery, T., Moore, D., Morgan, S., Morris, A., Morton, J., Mugalaasi, H., Murday, V., Nevitt, L., Newbury-Ecob, R., Norman, A., O'Shea, R., Ogilvie, C., Park, S., Parker, M. J., Patel, C., Paterson, J., Payne, S., Phipps, J., Pilz, D. T., Porteous, D., Pratt, N., Prescott, K., Price, S., Pridham, A., Procter, A., Purnell, H., Ragge, N., Rankin, J., Raymond, L., Rice, D., Robert, L., Roberts, E., Roberts, G., Roberts, J., Roberts, P., Ross, A., Rosser, E., Saggar, A., Samant, S., Sandford, R., Sarkar, A., Schweiger, S., Scott, C., Scott, R., Selby, A., Seller, A., Sequeira, C., Shannon, N., Sharif, S., Shaw-Smith, C., Shearing, E., Shears, D., Simonic, I., Simpkin, D., Singzon, R., Skitt, Z., Smith, A., Smith, B., Smith, K., Smithson, S., Sneddon, L., Splitt, M., Squires, M., Stewart, F., Stewart, H., Suri, M., Sutton, V., Swaminathan, G. J., Sweeney, E., Tatton-Brown, K., Taylor, C., Taylor, R., Tein, M., Temple, I. K., Thomson, J., Tolmie, J., Torokwa, A., Treacy, B., Turner, C., Turnpenny, P., Tysoe, C., Vandersteen, A., Vasudevan, P., Vogt, J., Wakeling, E., Walker, D., Waters, J., Weber, A., Wellesley, D., Whiteford, M., Widaa, S., Wilcox, S., Williams, D., Williams, N., Woods, G., Wragg, C., Wright, M., Yang, F., Yau, M., Carter, N. P., Parker, M., Firth, H. V., FitzPatrick, D. R., Wright, C. F., Barrett, J. C., and Hurles, M. E. (2014). Large-scale discovery of novel genetic causes of developmental disorders. *Nature*, 519(7542):223–228.
- [151] Forrest, A. R. R., Kawaji, H., Rehli, M., Kenneth Baillie, J., de Hoon, M. J. L., Haberle, V., Lassmann, T., Kulakovskiy, I. V., Lizio, M., Itoh, M., Andersson, R., Mungall, C. J., Meehan, T. F., Schmeier, S., Bertin, N., Jørgensen, M., Dimont, E., Arner, E., Schmidl, C., Schaefer, U., Medvedeva, Y. a., Plessy, C., Vitezic, M., Severin, J., Semple, C. a., Ishizu, Y., Young, R. S., Francescato, M., Alam, I., Albanese, D., Altschuler, G. M., Arakawa, T., Archer, J. a. C., Arner, P., Babina, M., Rennie, S., Balwiercz, P. J., Beckhouse, A. G., Pradhan-Bhatt, S., Blake, J. a., Blumenthal, A., Bodega, B., Bonetti, A., Briggs, J., Brombacher, F., Maxwell Burroughs, a., Califano, A., Cannistraci, C. V., Carbajo, D., Chen, Y., Chierici, M., Ciani, Y., Clevers, H. C., Dalla, E., Davis, C. a., Detmar, M., Diehl, A. D., Dohi, T., Drabløs, F., Edge, A. S. B., Edinger, M., Ekwall, K., Endoh, M., Enomoto, H., Fagiolini, M., Fairbairn, L., Fang, H., Farach-Carson, M. C., Faulkner, G. J., Favorov, A. V., Fisher, M. E., Frith, M. C., Fujita, R., Fukuda, S., Furlanello, C., Furuno, M., Furusawa, J.-i., Geijtenbeek, T. B., Gibson, A. P., Gingeras, T.,

- Goldowitz, D., Gough, J., Guhl, S., Guler, R., Gustincich, S., Ha, T. J., Hamaguchi, M., Hara, M., Harbers, M., Harshbarger, J., Hasegawa, A., Hasegawa, Y., Hashimoto, T., Herlyn, M., Hitchens, K. J., Ho Sui, S. J., Hofmann, O. M., Hoof, I., Hori, F., Huminiecki, L., Iida, K., Ikawa, T., Jankovic, B. R., Jia, H., Joshi, A., Jurman, G., Kaczkowski, B., Kai, C., Kaida, K., Kaiho, A., Kajiyama, K., Kanamori-Katayama, M., Kasianov, A. S., Kasukawa, T., Katayama, S., Kato, S., Kawaguchi, S., Kawamoto, H., Kawamura, Y. I., Kawashima, T., Kempfle, J. S., Kenna, T. J., Kere, J., Khachigian, L. M., Kitamura, T., Peter Klinken, S., Knox, A. J., Kojima, M., Kojima, S., Kondo, N., Koseki, H., Koyasu, S., Krampitz, S., Kubosaki, A., Kwon, A. T., Laros, J. F. J., Lee, W., Lennartsson, A., Li, K., Lilje, B., Lipovich, L., Mackay-sim, A., Manabe, R.-i., Mar, J. C., Marchand, B., Mathelier, A., Mejhert, N., Meynert, A., Mizuno, Y., de Lima Morais, D. a., Morikawa, H., Morimoto, M., Moro, K., Motakis, E., Motohashi, H., Mummery, C. L., Murata, M., Nagao-Sato, S., Nakachi, Y., Nakahara, F., Nakamura, T., Nakamura, Y., Nakazato, K., van Nimwegen, E., Ninomiya, N., Nishiyori, H., Noma, S., Nozaki, T., Ogishima, S., Ohkura, N., Ohmiya, H., Ohno, H., Ohshima, M., Okada-Hatakeyama, M., Okazaki, Y., Orlando, V., Ovchinnikov, D. a., Pain, A., Passier, R., Patrikakis, M., Persson, H., Piazza, S., Prendergast, J. G. D., Rackham, O. J. L., Ramilowski, J. a., Rashid, M., Ravasi, T., Rizzu, P., Roncador, M., Roy, S., Rye, M. B., Saijyo, E., Sajantila, A., Saka, A., Sakaguchi, S., Sakai, M., Sato, H., Satoh, H., Savvi, S., Saxena, A., Schneider, C., Schultes, E. a., Schulze-Tanzil, G. G., Schwegmann, A., Sengstag, T., Sheng, G., Shimoji, H., Shimoni, Y., Shin, J. W., Simon, C., Sugiyama, D., Sugiyama, T., Suzuki, M., Suzuki, N., Swoboda, R. K., 't Hoen, P. a. C., Tagami, M., Takahashi, N., Takai, J., Tanaka, H., Tatsukawa, H., Tatum, Z., Thompson, M., Toyoda, H., Toyoda, T., Valen, E., van de Wetering, M., van den Berg, L. M., Verardo, R., Vijayan, D., Vorontsov, I. E., Wasserman, W. W., Watanabe, S., Wells, C. a., Winteringham, L. N., Wolvetang, E., Wood, E. J., Yamaguchi, Y., Yamamoto, M., Yoneda, M., Yonekura, Y., Yoshida, S., Zabierowski, S. E., Zhang, P. G., Zhao, X., Zucchelli, S., Summers, K. M., Suzuki, H., Daub, C. O., Kawai, J., Heutink, P., Hide, W., Freeman, T. C., Lenhard, B., Bajic, V. B., Taylor, M. S., Makeev, V. J., Sandelin, A., Hume, D. a., Carninci, P., Hayashizaki, Y., and Baillie, J. K. (2014). A promoter-level mammalian expression atlas. *Nature*, 507(7493):462–470.
- [152] Fortune, M. D., Guo, H., Burren, O., Schofield, E., Walker, N. M., Ban, M., Sawcer, S. J., Bowes, J., Worthington, J., Barton, A., Eyre, S., Todd, J. a., and Wallace, C. (2015). Statistical colocalization of genetic risk variants for related autoimmune diseases in the context of common controls. *Nature Genetics*, 47(7):839–846.
- [153] Franke, A., McGovern, D. P. B., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., Lees, C. W., Balschun, T., Lee, J., Roberts, R., Anderson, C. A., Bis, J. C., Bumpstead, S., Ellinghaus, D., Festen, E. M., Georges, M., Green, T., Haritunians, T., Jostins, L., Latiano, A., Mathew, C. G., Montgomery, G. W., Prescott, N. J., Raychaudhuri, S., Rotter, J. I., Schumm, P., Sharma, Y., Simms, L. A., Taylor, K. D., Whiteman, D., Wijmenga, C., Baldassano, R. N., Barclay, M., Bayless, T. M., Brand, S., Büning, C., Cohen, A., Colombel, J.-F., Cottone, M., Stronati, L., Denson, T., De Vos, M., D’Inca, R., Dubinsky, M., Edwards, C., Florin, T., Franchimont, D., Gearry, R., Glas, J., Van Gossom, A., Guthery, S. L., Halfvarson, J., Verspaget, H. W., Hugot, J.-P., Karban, A., Laukens, D., Lawrance,

- I., Lemann, M., Levine, A., Libioulle, C., Louis, E., Mowat, C., Newman, W., Panés, J., Phillips, A., Proctor, D. D., Regueiro, M., Russell, R., Rutgeerts, P., Sanderson, J., Sans, M., Seibold, F., Steinhart, A. H., Stokkers, P. C. F., Torkvist, L., Kullak-Ublick, G., Wilson, D., Walters, T., Targan, S. R., Brant, S. R., Rioux, J. D., D'Amato, M., Weersma, R. K., Kugathasan, S., Griffiths, A. M., Mansfield, J. C., Vermeire, S., Duerr, R. H., Silverberg, M. S., Satsangi, J., Schreiber, S., Cho, J. H., Annese, V., Hakonarson, H., Daly, M. J., and Parkes, M. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genetics*, 42(12):1118–25.
- [154] Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., Hu, W., Li, C., Lin, W., Liu, S., Pan, H., Tang, X., Wang, J., Wang, W., Yu, J., Zhang, B., Zhang, Q., Zhao, H., Zhao, H., Zhou, J., Gabriel, S. B., Barry, R., Blumenstiel, B., Camargo, A., Defelice, M., Faggart, M., Goyette, M., Gupta, S., Moore, J., Nguyen, H., Onofrio, R. C., Parkin, M., Roy, J., Stahl, E., Winchester, E., Ziaugra, L., Altshuler, D., Shen, Y., Yao, Z., Huang, W., Chu, X., He, Y., Jin, L., Liu, Y., Shen, Y., Sun, W., Wang, H., Wang, Y., Wang, Y., Xiong, X., Xu, L., Waye, M. M. Y., Tsui, S. K. W., Xue, H., Wong, J. T.-F., Galver, L. M., Fan, J.-B., Gunderson, K., Murray, S. S., Oliphant, A. R., Chee, M. S., Montpetit, A., Chagnon, F., Ferretti, V., Leboeuf, M., Olivier, J.-F., Phillips, M. S., Roumy, S., Sallée, C., Verner, A., Hudson, T. J., Kwok, P.-Y., Cai, D., Koboldt, D. C., Miller, R. D., Pawlikowska, L., Taillon-Miller, P., Xiao, M., Tsui, L.-C., Mak, W., Song, Y. Q., Tam, P. K. H., Nakamura, Y., Kawaguchi, T., Kitamoto, T., Morizono, T., Nagashima, A., Ohnishi, Y., Sekine, A., Tanaka, T., Tsunoda, T., Deloukas, P., Bird, C. P., Delgado, M., Dermitzakis, E. T., Gwilliam, R., Hunt, S., Morrison, J., Powell, D., Stranger, B. E., Whittaker, P., Bentley, D. R., Daly, M. J., de Bakker, P. I. W., Barrett, J., Chretien, Y. R., Maller, J., McCarroll, S., Patterson, N., Pe'er, I., Price, A., Purcell, S., Richter, D. J., Sabeti, P., Saxena, R., Schaffner, S. F., Sham, P. C., Varilly, P., Stein, L. D., Krishnan, L., Smith, A. V., Tello-Ruiz, M. K., Thorisson, G. A., Chakravarti, A., Chen, P. E., Cutler, D. J., Kashuk, C. S., Lin, S., Abecasis, G. R., Guan, W., Li, Y., Munro, H. M., Qin, Z. S., Thomas, D. J., McVean, G., Auton, A., Bottolo, L., Cardin, N., Eyheramendy, S., Freeman, C., Marchini, J., Myers, S., Spencer, C., Stephens, M., Donnelly, P., Cardon, L. R., Clarke, G., Evans, D. M., Morris, A. P., Weir, B. S., Mullikin, J. C., Sherry, S. T., Feolo, M., Skol, A., Zhang, H., Matsuda, I., Fukushima, Y., Macer, D. R., Suda, E., Rotimi, C. N., Adebamowo, C. A., Ajayi, I., Aniagwu, T., Marshall, P. A., Nkwodimmah, C., Royal, C. D. M., Leppert, M. F., Dixon, M., Peiffer, A., Qiu, R., Kent, A., Kato, K., Niikawa, N., Adewole, I. F., Knoppers, B. M., Foster, M. W., Clayton, E. W., Watkin, J., Muzny, D., Nazareth, L., Sodergren, E., Weinstock, G. M., Yakub, I., Birren, B. W., Wilson, R. K., Fulton, L. L., Rogers, J., Burton, J., Carter, N. P., Clee, C. M., Griffiths, M., Jones, M. C., McLay, K., Plumb, R. W., Ross, M. T., Sims, S. K., Willey, D. L., Chen, Z., Han, H., Kang, L., Godbout, M., Wallenburg, J. C., L'Archevêque, P., Bellemare, G., Saeki, K., Wang, H., An, D., Fu, H., Li, Q., Wang, Z., Wang, R., Holden, A. L., Brooks, L. D., McEwen, J. E., Guyer, M. S., Wang, V. O., Peterson, J. L., Shi, M., Spiegel, J., Sung, L. M., Zacharia, L. F., Collins, F. S., Kennedy, K., Jamieson, R., and Stewart, J. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–61.

- [155] Fromer, M., Pocklington, A. J., Kavanagh, D. H., Williams, H. J., Dwyer, S., Gormley, P., Georgieva, L., Rees, E., Palta, P., Ruderfer, D. M., Carrera, N., Humphreys, I., Johnson, J. S., Roussos, P., Barker, D. D., Banks, E., Milanova, V., Grant, S. G., Hannon, E., Rose, S. A., Chambert, K., Mahajan, M., Scolnick, E. M., Moran, J. L., Kirov, G., Palotie, A., McCarroll, S. A., Holmans, P., Sklar, P., Owen, M. J., Purcell, S. M., and O'Donovan, M. C. (2014). De novo mutations in schizophrenia implicate synaptic networks. *Nature*, 506(7487):179–184.
- [156] Fuchsberger, C., Flannick, J., Teslovich, T. M., Mahajan, A., Agarwala, V., Gaulton, K. J., Ma, C., Fontanillas, P., Moutsianas, L., McCarthy, D. J., Rivas, M. A., Perry, J. R. B., Sim, X., Blackwell, T. W., Robertson, N. R., Rayner, N. W., Cingolani, P., Locke, A. E., Tajes, J. F., Highland, H. M., Dupuis, J., Chines, P. S., Lindgren, C. M., Hartl, C., Jackson, A. U., Chen, H., Huyghe, J. R., van de Bunt, M., Pearson, R. D., Kumar, A., Müller-Nurasyid, M., Grarup, N., Stringham, H. M., Gamazon, E. R., Lee, J., Chen, Y., Scott, R. A., Below, J. E., Chen, P., Huang, J., Go, M. J., Stitzel, M. L., Pasko, D., Parker, S. C. J., Varga, T. V., Green, T., Beer, N. L., Day-Williams, A. G., Ferreira, T., Fingerlin, T., Horikoshi, M., Hu, C., Huh, I., Ikram, M. K., Kim, B.-J., Kim, Y., Kim, Y. J., Kwon, M.-S., Lee, J., Lee, S., Lin, K.-H., Maxwell, T. J., Nagai, Y., Wang, X., Welch, R. P., Yoon, J., Zhang, W., Barzilai, N., Voight, B. F., Han, B.-G., Jenkinson, C. P., Kuulasmaa, T., Kuusisto, J., Manning, A., Ng, M. C. Y., Palmer, N. D., Balkau, B., Stančáková, A., Abboud, H. E., Boeing, H., Giedraitis, V., Prabhakaran, D., Gottesman, O., Scott, J., Carey, J., Kwan, P., Grant, G., Smith, J. D., Neale, B. M., Purcell, S., Butterworth, A. S., Howson, J. M. M., Lee, H. M., Lu, Y., Kwak, S.-H., Zhao, W., Danesh, J., Lam, V. K. L., Park, K. S., Saleheen, D., So, W. Y., Tam, C. H. T., Afzal, U., Aguilar, D., Arya, R., Aung, T., Chan, E., Navarro, C., Cheng, C.-Y., Palli, D., Correa, A., Curran, J. E., Rybin, D., Farook, V. S., Fowler, S. P., Freedman, B. I., Griswold, M., Hale, D. E., Hicks, P. J., Khor, C.-C., Kumar, S., Lehne, B., Thuillier, D., Lim, W. Y., Liu, J., van der Schouw, Y. T., Loh, M., Musani, S. K., Puppala, S., Scott, W. R., Yengo, L., Tan, S.-T., Jr., H. A. T., Thameem, F., Wilson, G., Wong, T. Y., Njølstad, P. R., Levy, J. C., Mangino, M., Bonnycastle, L. L., Schwarzmayr, T., Fadista, J., Surdulescu, G. L., Herder, C., Groves, C. J., Wieland, T., Bork-Jensen, J., Brandslund, I., Christensen, C., Koistinen, H. A., Doney, A. S. F., Kinnunen, L., Esko, T., Farmer, A. J., Hakaste, L., Hodgkiss, D., Kravic, J., Lyssenko, V., Hollensted, M., Jørgensen, M. E., Jørgensen, T., Ladenvall, C., Justesen, J. M., Käräjämäki, A., Kriebel, J., Rathmann, W., Lannfelt, L., Lauritzen, T., Narisu, N., Linneberg, A., Melander, O., Milani, L., Neville, M., Orho-Melander, M., Qi, L., Qi, Q., Roden, M., Rolandsson, O., Swift, A., Rosengren, A. H., Stirrups, K., Wood, A. R., Mihailov, E., Blancher, C., Carneiro, M. O., Maguire, J., Poplin, R., Shakir, K., Fennell, T., DePristo, M., de Angelis, M. H., Deloukas, P., Gjesing, A. P., Jun, G., Nilsson, P., Murphy, J., Onofrio, R., Thorand, B., Hansen, T., Meisinger, C., Hu, F. B., Isomaa, B., Karpe, F., Liang, L., Peters, A., Huth, C., O'Rahilly, S. P., Palmer, C. N. A., Pedersen, O., Rauramaa, R., Tuomilehto, J., Salomaa, V., Watanabe, R. M., Syvänen, A.-C., Bergman, R. N., Bharadwaj, D., Bottinger, E. P., Cho, Y. S., Chandak, G. R., Chan, J. C. N., Chia, K. S., Daly, M. J., Ebrahim, S. B., Langenberg, C., Elliott, P., Jablonski, K. A., Lehman, D. M., Jia, W., Ma, R. C. W., Pollin, T. I., Sandhu, M., Tandon, N., Froguel, P., Barroso, I., Teo, Y. Y., Zeggini, E., Loos, R. J. F., Small, K. S., Ried, J. S., DeFronzo, R. A., Grallert, H., Glaser, B., Metspalu, A., Wareham, N. J., Walker, M., Banks, E., Gieger, C., Ingelsson, E.,

- Im, H. K., Illig, T., Franks, P. W., Buck, G., Trakalo, J., Buck, D., Prokopenko, I., Mägi, R., Lind, L., Farjoun, Y., Owen, K. R., Gloyn, A. L., Strauch, K., Tuomi, T., Kooner, J. S., Lee, J.-Y., Park, T., Donnelly, P., Morris, A. D., Hattersley, A. T., Bowden, D. W., Collins, F. S., Atzmon, G., Chambers, J. C., Spector, T. D., Laakso, M., Strom, T. M., Bell, G. I., Blangero, J., Duggirala, R., Tai, E. S., McVean, G., Hanis, C. L., Wilson, J. G., Seielstad, M., Frayling, T. M., Meigs, J. B., Cox, N. J., Sladek, R., Lander, E. S., Gabriel, S., Burt, N. P., Mohlke, K. L., Meitinger, T., Groop, L., Abecasis, G., Florez, J. C., Scott, L. J., Morris, A. P., Kang, H. M., Boehnke, M., Altshuler, D., and McCarthy, M. I. (2016). The genetic architecture of type 2 diabetes. *Nature*, 536:41–47.
- [157] Fugazzola, L., Cerutti, N., Mannavola, D., Vannucchi, G., Fallini, C., Persani, L., and Beck-Peccoz, P. (2003). Monoallelic expression of mutant thyroid peroxidase allele causing total iodide organification defect. *Journal of Clinical Endocrinology and Metabolism*, 88(7):3264–3271.
- [158] Fujimoto, A., Nakagawa, H., Hosono, N., Nakano, K., Abe, T., Boroevich, K. a., Nagasaki, M., Yamaguchi, R., Shibuya, T., Kubo, M., Miyano, S., Nakamura, Y., and Tsunoda, T. (2010). Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. *Nature Genetics*, 42(11):931–936.
- [159] Funato, N. (2016). 22q11.2 Deletion Syndrome: Unmasking the Role of Tbx1 in Craniofacial Development. *Austin Med Sciences*, 1(1):22–25.
- [160] Futema, M., Plagnol, V., Li, K., Whittall, R. a., Neil, H. A. W., Seed, M., Bertolini, S., Calandra, S., Descamps, O. S., Graham, C. a., Hegele, R. a., Karpe, F., Durst, R., Leitersdorf, E., Lench, N., Nair, D. R., Soran, H., Van Bockxmeer, F. M., and Humphries, S. E. (2014). Whole exome sequencing of familial hypercholesterolaemia patients negative for LDLR/APOB/PCSK9 mutations. *Journal of Medical Genetics*, 51(8):537–44.
- [161] Gajecka, M. (2016). Unrevealed mosaicism in the next-generation sequencing era. *Molecular Genetics and Genomics*, 291(2):513–530.
- [162] Gálvez, J. (2014). Role of Th17 Cells in the Pathogenesis of Human IBD. *ISRN inflammation*, 2014:928461.
- [163] García-Carpizo, V., Ruiz-Llorente, L., Fraga, M., and Aranda, A. (2011). The growing role of gene methylation on endocrine function. *Journal of Molecular Endocrinology*, 47(2).
- [164] Gazouli, M., Pachoula, I., Panayotou, I., Mantzaris, G., Chrousos, G., Anagnou, N. P., and Roma-giannikou, E. (2010). NOD2/CARD15, ATG16L1 and IL23R gene polymorphisms and childhood-onset of Crohn’s Disease. *World journal of gastroenterology: WJG*, 16(14):1753–1758.
- [165] Genetic Analysis of Psoriasis Consortium & the Wellcome Trust Case Control Consortium 2, Strange, A., Capon, F., Spencer, C. C., Knight, J., Weale, M. E., Allen, M. H., Barton, A., Band, G., Bellenguez, C., Bergboer, J. G. M., Blackwell, J. M., Bramon, E., Bumpstead, S. J., Casas, J. P., Cork, M. J., Corvin, A., Deloukas,

- P., Dilthey, A., Duncanson, A., Edkins, S., Estivill, X., Fitzgerald, O., Freeman, C., Giardina, E., Gray, E., Hofer, A., Hüffmeier, U., Hunt, S. E., Irvine, A. D., Jankowski, J., Kirby, B., Langford, C., Lascorz, J., Leman, J., Leslie, S., Mallbris, L., Markus, H. S., Mathew, C. G., McLean, W. H. I., McManus, R., Mössner, R., Moutsianas, L., Naluai, A. T., Nestle, F. O., Novelli, G., Onoufriadis, A., Palmer, C. N. A., Perricone, C., Pirinen, M., Plomin, R., Potter, S. C., Pujol, R. M., Rautanen, A., Riveira-Munoz, E., Ryan, A. W., Salmhofer, W., Samuelsson, L., Sawcer, S. J., Schalkwijk, J., Smith, C. H., Stähle, M., Su, Z., Tazi-Ahnini, R., Traupe, H., Viswanathan, A. C., Warren, R. B., Weger, W., Wolk, K., Wood, N., Worthington, J., Young, H. S., Zeeuwen, P. L. J. M., Hayday, A., Burden, A. D., Griffiths, C. E. M., Kere, J., Reis, A., McVean, G., Evans, D. M., Brown, M. A., Barker, J. N., Peltonen, L., Donnelly, P., and Trembath, R. C. (2010). A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. *Nature Genetics*, 42(11):985–90.
- [166] Génin, E., Feingold, J., and Clerget-Darpoux, F. (2008). Identifying modifier genes of monogenic disease: Strategies and difficulties. *Human Genetics*, 124(4):357–368.
- [167] Genomics England (2015). Genomes Project Protocol. *White paper Genomics England*, (February).
- [168] Gibbs, W. W. (2014). Medicine gets up close and personal. *Nature*, 506(7487):144–5.
- [169] Gilissen, C., Hoischen, A., Brunner, H. G., and Veltman, J. A. (2012). Disease gene identification strategies for exome sequencing. *European Journal of Human Genetics: EJHG*, 20(5):490–7.
- [170] Gilissen C, Hehir-Kwa JY, Thung DT, van de Vorst M, van Bon BWM, Willemsen MH, Kwint M, Janssen IM, Hoischen A, Schenck A, Leach R, Klein R, Tearle R, Bo T, Pfundt R, Yntema HG, de Vries BBA, Kleefstra T, Brunner HG, Vissers LELM, Veltman JA (2014). Genome sequencing identifies major causes of severe intellectual disability. *Nature*, 511(7509):344–347.
- [171] Gillam, M. P. and Kopp, P. (2001). Genetic defects in thyroid hormone synthesis. *Current opinion in Pediatrics*, 13:364–372.
- [172] Girard, S. L., Gauthier, J., Noreau, A., Xiong, L., Zhou, S., Jouan, L., Dionne-Laporte, A., Spiegelman, D., Henrion, E., Diallo, O., Thibodeau, P., Bachand, I., Bao, J. Y. J., Tong, A. H. Y., Lin, C.-H., Millet, B., Jaafari, N., Joobar, R., Dion, P. A., Lok, S., Krebs, M.-O., and Rouleau, G. A. (2011). Increased exonic de novo mutation rate in individuals with schizophrenia. *Nature Genetics*, 43(9):860–3.
- [173] Glickman, J. (2005). Pathology of inflammatory bowel disease in children. *Current Diagnostic Pathology*, 11(2):117–124.
- [174] Glocker, E.-O., Kotlarz, D., Boztug, K., Gertz, E. M., Schäffer, A. a., Noyan, F., Perro, M., Diestelhorst, J., Allroth, A., Murugan, D., Hätscher, N., Pfeifer, D., Sykora, K.-W., Sauer, M., Kreipe, H., Lacher, M., Nustede, R., Woellner, C., Baumann, U., Salzer, U., Koletzko, S., Shah, N., Segal, A. W., Sauerbrey, A.,

- Buderus, S., Snapper, S. B., Grimbacher, B., and Klein, C. (2009). Inflammatory bowel disease and mutations affecting the interleukin-10 receptor. *The New England Journal of Medicine*, 361(21):2033–45.
- [175] Gong, W., Gottlieb, S., Collins, J., Blescia, a., Dietz, H., Goldmuntz, E., McDonald-McGinn, D. M., Zackai, E. H., Emanuel, B. S., Driscoll, D. a., and Budarf, M. L. (2001). Mutation analysis of TBX1 in non-deleted patients with features of DGS/VCFS or isolated cardiovascular defects. *Journal of Medical Genetics*, 38:E45.
- [176] Gonzaga-Jauregui, C. (2012). Human genome sequencing in health and disease. *Annual review of Medicine*, 63:35–61.
- [177] Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351.
- [178] Grasberger, H. (2010). Defects of thyroidal hydrogen peroxide generation in congenital hypothyroidism. *Molecular and cellular endocrinology*, 322(1-2):99–106.
- [179] Grasberger, H. and Rofetoff, S. (2012). Genetic causes of congenital hypothyroidism due to dysshormonogenesis. *Current opinion in Pediatrics*, 23(4):421–428.
- [180] Greene, D., Richardson, S., and Turro, E. (2016). Phenotype Similarity Regression for Identifying the Genetic Determinants of Rare Diseases. *American Journal of Human Genetics*, 98(3):490–499.
- [181] Griffiths, A. M. (2004). Specificities of inflammatory bowel disease in childhood. *Best practice & research. Clinical gastroenterology*, 18(3):509–23.
- [182] Groza, T., Kohler, S., Moldenhauer, D., Vasilevsky, N., Baynam, G., Zemojtel, T., Schriml, L. M., Kibbe, W. A., Schofield, P. N., Beck, T., Vasant, D., Brookes, A. J., Zankl, A., Washington, N. L., Mungall, C. J., Lewis, S. E., Haendel, M. A., Parkinson, H., and Robinson, P. N. (2015). The Human Phenotype Ontology: Semantic Unification of Common and Rare Disease. *American Journal of Human Genetics*, 97(1):111–124.
- [183] Grozeva, D., Carss, K., Spasic-Boskovic, O., Parker, M. J., Archer, H., Firth, H. V., Park, S.-M., Canham, N., Holder, S. E., Wilson, M., Hackett, A., Field, M., Floyd, J. a. B., Hurles, M., and Raymond, F. L. (2014). De novo loss-of-function mutations in SETD5, encoding a methyltransferase in a 3p25 microdeletion syndrome critical region, cause intellectual disability. *American Journal of Human Genetics*, 94(4):618–24.
- [184] Grozeva, D., Carss, K., Spasic-Boskovic, O., Tejada, M.-I., Gecz, J., Shaw, M., Corbett, M., Haan, E., Thompson, E., Friend, K., Hussain, Z., Hackett, A., Field, M., Renieri, A., Stevenson, R., Schwartz, C., Floyd, J. A. B., Bentham, J., Cosgrove, C., Keavney, B., Bhattacharya, S., Hurles, M., and Raymond, F. L. (2015). Targeted Next-Generation Sequencing Analysis of 1,000 Individuals with Intellectual Disability. *Human mutation*, 36(12):1197–204.

- [185] Guipponi, M., Santoni, F. A., Setola, V., Gehrig, C., Rotharmel, M., Cuenca, M., Guillin, O., Dikeos, D., Georgantopoulos, G., Papadimitriou, G., Curtis, L., Méary, A., Schürhoff, F., Jamain, S., Avramopoulos, D., Leboyer, M., Rujescu, D., Pulver, A., Campion, D., Siderovski, D. P., and Antonarakis, S. E. (2014). Exome sequencing in 53 sporadic cases of schizophrenia identifies 18 putative candidate genes. *PLoS ONE*, 9(11):1–12.
- [186] Gulsuner, S., Walsh, T., Watts, A. C., Lee, M. K., Thornton, A. M., Casadei, S., Rippey, C., Shahin, H., Nimgaonkar, V. L., Go, R. C. P., Savage, R. M., Swerdlow, N. R., Gur, R. E., Braff, D. L., King, M. C., and McClellan, J. M. (2013). Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell*, 154(3):518–529.
- [187] Guo, Y., Ye, F., Sheng, Q., and Samuels, D. C. (2013). Three-stage quality control strategies for DNA re-sequencing data. *Briefings in Bioinformatics*, page 1–11.
- [188] Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., Karthikeyan, S., Iles, L., Pollard, M. O., Choudhury, A., Ritchie, G. R. S., Xue, Y., Asimit, J., Nsubuga, R. N., Young, E. H., Pomilla, C., Kivinen, K., Rockett, K., Kamali, A., Doumatey, A. P., Asiki, G., Seeley, J., Sisay-Joof, F., Jallow, M., Tollman, S., Mekonnen, E., Ekong, R., Oljira, T., Bradman, N., Bojang, K., Ramsay, M., Adeyemo, A., Bekele, E., Motala, A., Norris, S. A., Pirie, F., Kaleebu, P., Kwiatkowski, D., Tyler-Smith, C., Rotimi, C., Zeggini, E., and Sandhu, M. S. (2015). The African Genome Variation Project shapes medical genetics in Africa. *Nature*, 517(7534):327–332.
- [189] Gusella, J. (1983). A polymorphic DNA marker genetically linked to Huntington’s disease. *Nature*, 53(9):1689–1699.
- [190] Halme, L., Paavola-sakki, P., Turunen, U., Lappalainen, M., Färkkilä, M., and Kontula, K. (2006). Family and twin studies in inflammatory bowel disease. *World journal of gastroenterology: WJG*, 12(23):3668–3672.
- [191] Hamosh, A., Scott, A. F., Amberger, J., Bocchini, C., Valle, D., and McKusick, V. a. (2002). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 30(1):52–5.
- [192] Hampe, J., Franke, A., Rosenstiel, P., Till, A., Teuber, M., Huse, K., Albrecht, M., Mayr, G., De La Vega, F. M., Briggs, J., Günther, S., Prescott, N. J., Onnie, C. M., Häslér, R., Sipos, B., Fölsch, U. R., Lengauer, T., Platzer, M., Mathew, C. G., Krawczak, M., and Schreiber, S. (2007). A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. *Nature Genetics*, 39(2):207–211.
- [193] Harbour, S. N., Maynard, C. L., Zindl, C. L., Schoeb, T. R., and Weaver, C. T. (2015). Th17 cells give rise to Th1 cells that are required for the pathogenesis of colitis. *Proceedings of the National Academy of Sciences*, 112(22):201415675.

- [194] Harris, K. B. and Pass, K. A. (2007). Increase in congenital hypothyroidism in New York State and in the United States. *Molecular Genetics and Metabolism*, 91(3):268–277.
- [195] Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N., Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J. M., Ezkurdia, I., Van Baren, J., Brent, M., Haussler, D., Kellis, M., Valencia, A., Reymond, A., Gerstein, M., Guigo, R., and Hubbard, T. J. (2012). GENCODE: The reference human genome annotation for the ENCODE project. *Genome Research*, 22(9):1760–1774.
- [196] Hayden, M. S., Ghosh, S., Hayden, M. S., and Ghosh, S. (2012). NF- κ B, the first quarter-century: remarkable progress and outstanding questions. *Genes & Development*, pages 203–234.
- [197] Heinrich, J., Proepper, C., Schmidt, T., Linta, L., Liebau, S., and Boeckers, T. M. (2012a). The postsynaptic density protein Abelson interactor protein 1 interacts with the motor protein Kinesin family member 26B in hippocampal neurons. *Neuroscience*, 221:86–95.
- [198] Heinrich, V., Stange, J., Dickhaus, T., Imkeller, P., Krüger, U., Bauer, S., Mundlos, S., Robinson, P. N., Hecht, J., and Krawitz, P. M. (2012b). The allele distribution in next-generation sequencing data sets is accurately described as the result of a stochastic branching process. *Nucleic Acids Research*, 40(6):2426–31.
- [199] Heinzen, E. L., Depondt, C., Cavalleri, G. L., Ruzzo, E. K., Walley, N. M., Need, A. C., Ge, D., He, M., Cirulli, E. T., Zhao, Q., Cronin, K. D., Gumbs, C. E., Campbell, C. R., Hong, L. K., Maia, J. M., Shianna, K. V., McCormack, M., Radtke, R. A., O’Conner, G. D., Mikati, M. A., Gallentine, W. B., Husain, A. M., Sinha, S. R., Chinthapalli, K., Puranam, R. S., McNamara, J. O., Ottman, R., Sisodiya, S. M., Delanty, N., and Goldstein, D. B. (2012). Exome sequencing followed by large-scale genotyping fails to identify single rare variants of large effect in idiopathic generalized epilepsy. *American Journal of Human Genetics*, 91(2):293–302.
- [200] Helsmoortel, C., Vulto-van Silfhout, A. T., Coe, B. P., Vandeweyer, G., Rooms, L., van den Ende, J., Schuurs-Hoeijmakers, J. H. M., Marcelis, C. L., Willemsen, M. H., Vissers, L. E. L. M., Yntema, H. G., Bakshi, M., Wilson, M., Witherspoon, K. T., Malmgren, H., Nordgren, A., Annerén, G., Fichera, M., Bosco, P., Romano, C., de Vries, B. B. A., Kleefstra, T., Kooy, R. F., Eichler, E. E., and Van der Aa, N. (2014). A SWI/SNF-related autism syndrome caused by de novo mutations in ADNP. *Nature Genetics*, 46(4):380–4.
- [201] Hermanns, P., Grasberger, H., Refetoff, S., and Pohlenz, J. (2011). Mutations in the NKX2.5 gene and the PAX8 promoter in a girl with thyroid dysgenesis. *The Journal of Clinical Endocrinology and Metabolism*, 96(6):E977–81.

- [202] Heyman, M. B., Kirschner, B. S., Goldstein, B. A., and Ferry, G. D. (2005). Children with early-onset inflammatory bowel disease (IBD): analysis of a pediatric ibd consortium registry. *The Journal of pediatrics*, 146:35–40.
- [203] Heyworth, P. G., Cross, A. R., and Curnutte, J. T. (2003). Chronic granulomatous disease. *Current Opinion in Immunology*, 15(5):578–584.
- [204] Hieter, P. and Boycott, K. M. (2014). Understanding rare disease pathogenesis: A grand challenge for model organisms. *Genetics*, 198(2):443–445.
- [205] Higgins, J. P. T., Thompson, S. G., Deeks, J. J., and Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ: British Medical Journal*, 327(7414):557–560.
- [206] Hilton, I. B., D’Ippolito, A. M., Vockley, C. M., Thakore, P. I., Crawford, G. E., Reddy, T. E., and Gersbach, C. A. (2015). Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nature Biotechnology*, 33(5):510–7.
- [207] Hinds, D. a., Stuve, L. L., Nilsen, G. B., Halperin, E., Eskin, E., Ballinger, D. G., Frazer, K. a., and Cox, D. R. (2005). Whole-genome patterns of common DNA variation in three human populations. *Science*, 307(5712):1072–1079.
- [208] Hirschhorn, J. N. and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2):95–108.
- [209] Hishinuma, A., Onigata, K., Kuribayashi, T., and Ieiri, T. (1998). Sequence Analysis of Thyroid Transcription Reveals Absence of Mutations in Patients Dysgenesis but Presence of Polymorphisms Flanking Region and Intron Factor-1 Gene with Thyroid in the 5’. *Endocrinology*, 45(4):563–567.
- [210] Hoischen, A., Krumm, N., and Eichler, E. E. (2014). Prioritization of neurodevelopmental disease genes by discovery of new mutations. *Nature neuroscience*, 17(6):764–72.
- [211] Holm, H., Gudbjartsson, D. F., Sulem, P., Masson, G., Helgadóttir, H. T., Zanon, C., Magnusson, O. T., Helgason, A., Saemundsdóttir, J., Gylfason, A., Stefansdóttir, H., Gretarsdóttir, S., Matthiasson, S. E., Thorgeirsson, G. M., Jonasdóttir, A., Sigurdsson, A., Stefansson, H., Werge, T., Rafnar, T., Kiemeny, L. A., Parvez, B., Muhammad, R., Roden, D. M., Darbar, D., Thorleifsson, G., Walters, G. B., Kong, A., Thorsteinsdóttir, U., Arnar, D. O., and Stefansson, K. (2011). A rare variant in MYH6 is associated with high risk of sick sinus syndrome. *Nature Genetics*, 43(4):316–320.
- [212] Hood, L. and Price, N. D. (2014). Promoting Wellness & Demystifying Disease: The 100K Project. *Clinical Omics*, (3).
- [213] Houdayer, C., Caux-Moncoutier, V., Krieger, S., Barrois, M., Bonnet, F., Bourdon, V., Bronner, M., Buisson, M., Coulet, F., Gaildrat, P., Lefol, C., Leone, M., Mazoyer, S., Muller, D., Remenieras, A., Revillion, F., Rouleau, E., Sokolowska, J., Vert, J. P., Lidereau, R., Soubrier, F., Sobol, H., Sevenet, N., Bressac-de Paillerets,

- B., Hardouin, A., Tosi, M., Sinilnikova, O. M., and Stoppa-Lyonnet, D. (2012). Guidelines for splicing analysis in molecular diagnosis derived from a set of 327 combined in silico/in vitro studies on BRCA1 and BRCA2 variants. *Human Mutation*, 33(8):1228–1238.
- [214] Howie, B. N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, 5(6).
- [215] Howrigan, D. P., Simonson, M. a., and Keller, M. C. (2011). Detecting autozygosity through runs of homozygosity: a comparison of three autozygosity detection algorithms. *BMC genomics*, 12(1):460.
- [216] Huang, D. W., Sherman, B. T., and Lempicki, R. a. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57.
- [217] Huang, H., Fang, M., Jostins, L., Anderson, C. A., Andersen, V., Cleynen, I., Cortes, A., Amato, M. D., Dimitrieva, J., Elansary, M., Farh, K. K.-h., Franke, A., Gori, S., Goyette, P., Halfvarson, J., Haritunians, T., Knight, J., Lawrance, I. C., Lees, C. W., Louis, E., Mariman, R., Mni, M., Momozawa, Y., Parkes, M., Trynka, G., Satsangi, J., Sommeren, S. V., Vermeire, S., Xavier, R. J., Weersma, R. K., Duerr, R. H., Mathew, C. G., Rioux, J. D., Cho, J. H., Georges, M., and Daly, M. J. (2015). Association mapping of inflammatory Bowel Disease loci to single variant resolution. *BioRxiv*.
- [218] Hugot, J. P., Chamaillard, M., Zouali, H., Lesage, S., Cézard, J. P., Belaiche, J., Almer, S., Tysk, C., O’Morain, C. a., Gassull, M., Binder, V., Finkel, Y., Cortot, a., Modigliani, R., Laurent-Puig, P., Gower-Rousseau, C., Macry, J., Colombel, J. F., Sahbatou, M., and Thomas, G. (2001). Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn’s disease. *Nature*, 411(6837):599–603.
- [219] Hugot, J. P., Laurent-Puig, P., Gower-Rousseau, C., Olson, J. M., Lee, J. C., Beaugerie, L., Naom, I., Dupas, J. L., Van Gossum, a., Orholm, M., Bonaiti-Pellie, C., Weissenbach, J., Mathew, C. G., Lennard-Jones, J. E., Cortot, a., Colombel, J. F., and Thomas, G. (1996). Mapping of a susceptibility locus for Crohn’s disease on chromosome 16.
- [220] Hular, I., Hermanns, P., Nestoris, C., Heger, S., Refetoff, S., Pohlenz, J., and Grasberger, H. (2011). A single copy of the recently identified dual oxidase maturation factor (DUOXA) 1 gene produces only mild transient hypothyroidism in a patient with a novel biallelic DUOXA2 mutation and monoallelic DUOXA1 deletion. *The Journal of Clinical Endocrinology and Metabolism*, 96(5):E841–5.
- [221] Hutchison, C. A. (2007). DNA sequencing: Bench to bedside and beyond. *Nucleic Acids Research*, 35(18):6227–6237.
- [222] Imielinski, M., Baldassano, R. N., Griffiths, A. M., Russell, R. K., Annese, V., Dubinsky, M., Kugathasan, S., Bradfield, J. P., Walters, T. D., Sleiman, P., Kim, C. E., Muise, A., Wang, K., Glessner, J. T., Saeed, S., Zhang, H., Frackelton, E. C., Hou, C., Flory, J. H., Otieno, G., Chiavacci, R. M., Grundmeier, R., Castro, M.,

- Latiano, A., Dallapiccola, B., Stempak, J., Abrams, D. J., Taylor, K. D., McGovern, D., Silber, G., Wrobel, I., Quiros, A., Barrett, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H., Duerr, R. H., Rioux, J. D., Brant, S. R., Silverberg, M. S., Taylor, K. D., Barmuda, M. M., Bitton, A., Dassopoulos, T., Datta, L. W., Green, T., Griffiths, A. M., Kistner, E. O., Murtha, M. T., Regueiro, M. D., Rotter, J. I., Schumm, L. P., Steinhart, A. H., Targan, S. R., Xavier, R. J., Libioulle, C., Sandor, C., Lathrop, M., Belaiche, J., Dewit, O., Gut, I., Heath, S., Laukens, D., Mni, M., Rutgeerts, P., Van Gossum, A., Zelenika, D., Franchimont, D., Hugot, J. P., de Vos, M., Vermeire, S., Louis, E., Cardon, L. R., Anderson, C. A., Drummond, H., Nimmo, E., Ahmad, T., Prescott, N. J., Onnie, C. M., Fisher, S. a., Marchini, J., Ghori, J., Bumpstead, S., Gwillam, R., Tremelling, M., Delukas, P., Mansfield, J., Jewell, D., Satsangi, J., Mathew, C. G., Parkes, M., Georges, M., Daly, M. J., Heyman, M. B., Ferry, G. D., Kirschner, B., Lee, J., Essers, J., Grand, R., Stephens, M., Levine, A., Piccoli, D., Van Limbergen, J., Cucchiara, S., Monos, D. S., Guthery, S. L., Denson, L., Wilson, D. C., Grant, S. F. a., Daly, M. J., and Hakonarson, H. (2009). Common variants at five new loci associated with early-onset inflammatory bowel disease. *Nature Genetics*, 41(12):1335–1340.
- [223] Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., Yamrom, B., Lee, Y.-H., Narzisi, G., Leotta, A., Kendall, J., Grabowska, E., Ma, B., Marks, S., Rodgers, L., Stepansky, A., Troge, J., Andrews, P., Bekritsky, M., Pradhan, K., Ghiban, E., Kramer, M., Parla, J., Demeter, R., Fulton, L. L., Fulton, R. S., Magrini, V. J., Ye, K., Darnell, J. C., Darnell, R. B., Mardis, E. R., Wilson, R. K., Schatz, M. C., McCombie, W. R., and Wigler, M. (2012). De novo gene disruptions in children on the autistic spectrum. *Neuron*, 74(2):285–99.
- [224] Ishiguro, Y., Yamagata, K., Sakuraba, H., Munakata, A., Nakane, A., Morita, T., and Nishihira, J. (2004). Macrophage migration inhibitory factor and activator protein-1 in ulcerative colitis. *Annals of the New York Academy of Sciences*, 1029:348–349.
- [225] Itan, Y. and Casanova, J.-L. (2015). Novel primary immunodeficiency candidate genes predicted by the human gene connectome. *Frontiers in immunology*, 6(April):142.
- [226] Iwatani, N., Mabe, H., Devriendt, K., Kodama, M., and Miike, T. (2000). Deletion of NKX2.1 gene encoding thyroid transcription factor-1 in two siblings with hypothyroidism and respiratory failure. *The Journal of pediatrics*, 137(2):272–6.
- [227] J. Gustav Smith and Christopher Newton-Cheh (2009). *Genome-wide association study in humans*, volume 573. Humana Press.
- [228] Jian, X., Boerwinkle, E., and Liu, X. (2014). In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Research*, 42(22):13534–44.
- [229] Jin, H. Y., Heo, S.-H., Kim, Y.-M., Kim, G.-H., Choi, J.-H., Lee, B.-H., and Yoo, H.-W. (2014). High frequency of DUOX2 mutations in transient or permanent congenital hypothyroidism with eutopic thyroid glands. *Hormone research in paediatrics*, 82(4):252–60.

- [230] Johnston, J. J., Rubinstein, W. S., Facio, F. M., Ng, D., Singh, L. N., Teer, J. K., Mullikin, J. C., and Biesecker, L. G. (2012). Secondary variants in individuals undergoing exome sequencing: screening of 572 individuals identifies high-penetrance mutations in cancer-susceptibility genes. *American Journal of Human Genetics*, 91(1):97–108.
- [231] Jostins, L. and Barrett, J. C. (2011). Genetic risk prediction in complex disease. *Human Molecular Genetics*, 20(R2):R182–8.
- [232] Jostins, L., Ripke, S., Weersma, R. K., Duerr, R. H., McGovern, D. P., Hui, K. Y., Lee, J. C., Philip Schumm, L., Sharma, Y., Anderson, C. a., Essers, J., Mitrovic, M., Ning, K., Cleynen, I., Theatre, E., Spain, S. L., Raychaudhuri, S., Goyette, P., Wei, Z., Abraham, C., Achkar, J.-P., Ahmad, T., Amininejad, L., Ananthakrishnan, A. N., Andersen, V., Andrews, J. M., Baidoo, L., Balschun, T., Bampton, P. a., Bitton, A., Boucher, G., Brand, S., Büning, C., Cohain, A., Cichon, S., D’Amato, M., De Jong, D., Devaney, K. L., Dubinsky, M., Edwards, C., Ellinghaus, D., Ferguson, L. R., Franchimont, D., Fransen, K., Gearry, R., Georges, M., Gieger, C., Glas, J., Haritunians, T., Hart, A., Hawkey, C., Hedl, M., Hu, X., Karlsen, T. H., Kupcinskis, L., Kugathasan, S., Latiano, A., Laukens, D., Lawrance, I. C., Lees, C. W., Louis, E., Mahy, G., Mansfield, J., Morgan, A. R., Mowat, C., Newman, W., Palmieri, O., Ponsioen, C. Y., Potocnik, U., Prescott, N. J., Regueiro, M., Rotter, J. I., Russell, R. K., Sanderson, J. D., Sans, M., Satsangi, J., Schreiber, S., Simms, L. a., Sventoraityte, J., Targan, S. R., Taylor, K. D., Tremelling, M., Verspaget, H. W., De Vos, M., Wijmenga, C., Wilson, D. C., Winkelmann, J., Xavier, R. J., Zeissig, S., Zhang, B., Zhang, C. K., Zhao, H., Silverberg, M. S., Annese, V., Hakonarson, H., Brant, S. R., Radford-Smith, G., Mathew, C. G., Rioux, J. D., Schadt, E. E., Daly, M. J., Franke, A., Parkes, M., Vermeire, S., Barrett, J. C., and Cho, J. H. (2012). Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, 491(7422):119–124.
- [233] Jun, G., Flickinger, M., Hetrick, K. N., Romm, J. M., Doheny, K. F., Abecasis, G. R., Boehnke, M., and Kang, H. M. (2012). Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *American Journal of Human Genetics*, 91(5):839–48.
- [234] K Lichti-Kaiser, ZeRuth, G., and Anton, J. (2014). Transcription factor Gli-similar 3 (GLIS3): Implications for the development of congenital hypothyroidism. *Journal of Endocrinology and Diabetes and Obesity*, 2(2).
- [235] Kamboh, M. I., Barmada, M. M., Demirci, F. Y., Minster, R. L., Carrasquillo, M. M., Pankratz, V. S., Younkin, S. G., Saykin, A. J., Sweet, R. A., Feingold, E., DeKosky, S. T., and Lopez, O. L. (2012). Genome-wide association analysis of age-at-onset in Alzheimer’s disease. *Molecular Psychiatry*, 17(12):1340–1346.
- [236] Kammermeier, J., Drury, S., James, C. T., Dziubak, R., Ocaña, L., Elawad, M., Beales, P., Lench, N., Uhlig, H. H., Bacchelli, C., and Shah, N. (2014). Targeted gene panel sequencing in children with very early onset inflammatory bowel disease—evaluation and prospective analysis. *Journal of Medical Genetics*, 51(11):748–55.

- [237] Kang, H. S., Kim, Y.-S., ZeRuth, G., Beak, J. Y., Gerrish, K., Kilic, G., Sosa-Pineda, B., Jensen, J., Pierreux, C. E., Lemaigre, F. P., Foley, J., and Jetten, A. M. (2009). Transcription factor Glis3, a novel critical player in the regulation of pancreatic beta-cell development and insulin gene expression. *Molecular and cellular biology*, 29(24):6366–79.
- [238] Kasahara, H. and Benson, D. W. (2004). Biochemical analyses of eight NKX2.5 homeodomain missense mutations causing atrioventricular block and cardiac anomalies. *Cardiovascular Research*, 64(1):40–51.
- [239] Kathiresan, S., Melander, O., Anevski, D., and C, G. (2008). Polymorphisms Associated with Cholesterol and Risk of Cardiovascular Events. *The New England Journal of Medicine*, 358(12):1240–1249.
- [240] Katsanis, S. H. and Katsanis, N. (2013). Molecular genetic testing and the future of clinical genomics. *Nature Reviews Genetics*, 14(6):415–26.
- [241] Kelly, L., Mezulis, S., Yates, C., Wass, M., and Sternberg, M. (2015). The Phyre2 web portal for protein modelling, prediction, and analysis. *Nature Protocols*, 10(6):845–858.
- [242] Kelly, T. N., Takeuchi, F., Tabara, Y., Edwards, T. L., Kim, Y. J., Chen, P., Li, H., Wu, Y., Yang, C. F., Zhang, Y., Gu, D., Katsuya, T., Ohkubo, T., Gao, Y. T., Go, M. J., Teo, Y. Y., Lu, L., Lee, N. R., Chang, L. C., Peng, H., Zhao, Q., Nakashima, E., Kita, Y., Shu, X. O., Kim, N. H., Tai, E. S., Wang, Y., Adair, L. S., Chen, C. H., Zhang, S., Li, C., Nabika, T., Umemura, S., Cai, Q., Cho, Y. S., Wong, T. Y., Zhu, J., Wu, J. Y., Gao, X., Hixson, J. E., Cai, H., Lee, J., Cheng, C. Y., Rao, D. C., Xiang, Y. B., Cho, M. C., Han, B. G., Wang, A., Tsai, F. J., Mohlke, K., Lin, X., Ikram, M. K., Lee, J. Y., Zheng, W., Tetsuro, M., Kato, N., and He, J. (2013). Genome-wide association study meta-analysis reveals transethnic replication of mean arterial and pulse pressure loci. *Hypertension*, 62(5):853–859.
- [243] Kelsen, J. R., Baldassano, R. N., Artis, D., and Sonnenberg, G. F. (2015a). Maintaining Intestinal Health: The Genetics and Immunology of Very Early Onset Inflammatory Bowel Disease. *CMGH Cellular and Molecular Gastroenterology and Hepatology*, 1(5):462–476.
- [244] Kelsen, J. R., Dawany, N., Moran, C. J., Petersen, B.-S., Sarmady, M., Sasson, A., Pauly-Hubbard, H., Martinez, A., Maurer, K., Soong, J., Rappaport, E., Franke, A., Keller, A., Winter, H. S., Mamula, P., Piccoli, D., Artis, D., Sonnenberg, G. F., Daly, M., Sullivan, K. E., Baldassano, R. N., and Devoto, M. (2015b). Exome Sequencing Analysis Reveals Variants in Primary Immunodeficiency Genes in Patients With Very Early Onset Inflammatory Bowel Disease. *Gastroenterology*, 149(6):1415–1424.
- [245] Kiezun, A., Garimella, K., Do, R., Stitzel, N. O., Neale, B. M., McLaren, P. J., Gupta, N., Sklar, P., Sullivan, P. F., Moran, J. L., Hultman, C. M., Lichtenstein, P., Magnusson, P., Lehner, T., Shugart, Y. Y., Price, A. L., de Bakker, P. I. W., Purcell, S. M., and Sunyaev, S. R. (2012). Exome sequencing and the genetic basis of complex traits. *Nature Genetics*, 44(6):623–30.

- [246] Kim, K., Seong, M.-w., Chung, W.-h., Park, S. S., Leem, S., Park, W., Kim, J., Lee, K., Park, R. W., and Kim, N. (2015). Effect of Next-Generation Exome Sequencing Depth for Discovery of Diagnostic Variants. *Genomics and Informatics*, 13(2):31–39.
- [247] Kim, K. X., Sanneman, J. D., Kim, H.-M., Harbidge, D. G., Xu, J., Soleimani, M., Wangemann, P., and Marcus, D. C. (2014). Slc26a7 chloride channel activity and localization in mouse Reissner’s membrane epithelium. *PLoS one*, 9(5):e97191.
- [248] Kimura, S., Ward, J. M., and Minoo, P. (1999). Thyroid-specific enhancer-binding protein/thyroid transcription factor 1 is not required for the initial specification of the thyroid and lung primordia. *Biochimie*, 81(4):321–327.
- [249] King, D. A., Fitzgerald, T. W., Miller, R., Canham, N., Clayton-Smith, J., Johnson, D., Mansour, S., Stewart, F., Vasudevan, P., and Hurles, M. E. (2014). A novel method for detecting uniparental disomy from trio genotypes identifies a significant excess in children with developmental disorders. *Genome Research*, 24(4):673–687.
- [250] Kinkorová, J. (2015). Biobanks in the era of personalized medicine: objectives, challenges, and innovation: Overview. *The EPMA journal*, 7:4.
- [251] Kircher, M., Witten, D. M., Jain, P., O’Roak, B. J., Cooper, G. M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3):310–315.
- [252] Klein Rj Fau - Zeiss, C., Zeiss C Fau - Chew, E. Y., Chew Ey Fau - Tsai, J.-Y., Tsai Jy Fau - Sackler, R. S., Sackler Rs Fau - Haynes, C., Haynes C Fau - Henning, A. K., Henning Ak Fau - SanGiovanni, J. P., SanGiovanni Jp Fau - Mane, S. M., Mane Sm Fau - Mayne, S. T., Mayne St Fau - Bracken, M. B., Bracken Mb Fau - Ferris, F. L., Ferris Fl Fau - Ott, J., Ott J Fau - Barnstable, C., Barnstable C Fau - Hoh, J., and Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science*, 308(5720):385–389.
- [253] Kohler, S., Doelken, S. C., Mungall, C. J., Bauer, S., Firth, H. V., Bailleul-Forestier, I., Black, G. C. M., Brown, D. L., Brudno, M., Campbell, J., Fitzpatrick, D. R., Eppig, J. T., Jackson, A. P., Freson, K., Girdea, M., Helbig, I., Hurst, J. A., John, J., Jackson, L. G., Kelly, A. M., Ledbetter, D. H., Mansour, S., Martin, C. L., Moss, C., Mumford, A., Ouwehand, W. H., Park, S. M., Riggs, E. R., Scott, R. H., Sisodiya, S., Vooren, S. V., Wapner, R. J., Wilkie, A. O. M., Wright, C. F., Vulto-Van Silfhout, A. T., Leeuw, N. D., De Vries, B. B. A., Washington, N. L., Smith, C. L., Westerfield, M., Schofield, P., Ruef, B. J., Gkoutos, G. V., Haendel, M., Smedley, D., Lewis, S. E., and Robinson, P. N. (2014). The Human Phenotype Ontology project: Linking molecular biology and disease through phenotype data. *Nucleic Acids Research*, 42(D1):966–974.
- [254] Kondo, T., Ezzat, S., and Asa, S. L. (2006). Pathogenetic mechanisms in thyroid follicular-cell neoplasia. *Nature Reviews. Cancer*, 6(4):292–306.
- [255] Kotlarz, D., Beier, R., Murugan, D., Diestelhorst, J., Jensen, O., Boztug, K., Pfeifer, D., Kreipe, H., Pfister, E.-D., Baumann, U., Puchalka, J., Bohne, J., Egritas,

- O., Dalgic, B., Kolho, K.-L., Sauerbrey, A., Buderus, S., Güngör, T., Enninger, A., Koda, Y. K. L., Guariso, G., Weiss, B., Corbacioglu, S., Socha, P., Uslu, N., Metin, A., Wahbeh, G. T., Husain, K., Ramadan, D., Al-Herz, W., Grimbacher, B., Sauer, M., Sykora, K.-W., Koletzko, S., and Klein, C. (2012). Loss of interleukin-10 signaling and infantile inflammatory bowel disease: implications for diagnosis and therapy. *Gastroenterology*, 143(2):347–55.
- [256] Krieg, A., Correa, R. G., Garrison, J. B., Le Negrato, G., Welsh, K., Huang, Z., Knoefel, W. T., and Reed, J. C. (2009). XIAP mediates NOD signaling via interaction with RIP2. *Proceedings of the National Academy of Sciences of the United States of America*, 106(34):14524–14529.
- [257] Krude, H., Schütz, B., Biebermann, H., von Moers, A., Schnabel, D., Neitzel, H., Tönnies, H., Weise, D., Lafferty, A., Schwarz, S., DeFelice, M., von Deimling, A., van Landeghem, F., DiLauro, R., and Grüters, A. (2002). Choreoathetosis, hypothyroidism, and pulmonary alterations due to human NKX2-1 haploinsufficiency. *Journal of Clinical Investigation*, 109(4):475–480.
- [258] Krumm, N., Turner, T. N., Baker, C., Vives, L., Mohajeri, K., Witherspoon, K., Raja, A., Coe, B. P., Stessman, H. a., He, Z.-X., Leal, S. M., Bernier, R., and Eichler, E. E. (2015). Excess of rare, inherited truncating mutations in autism. *Nature Genetics*, 47(6):582–588.
- [259] Kryukov, G. V., Pennacchio, L. a., and Sunyaev, S. R. (2007). Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *American Journal of Human Genetics*, 80(4):727–39.
- [260] Ku, C.-s., Cooper, D. N., Polychronakos, C., Naidoo, N., Wu, M., and Soong, R. (2012). Exome Sequencing: Dual Role as a Discovery and Diagnostic Tool. *Annals of Neurology*, 71(1):5–14.
- [261] Kugathasan, S., Baldassano, R. N., Bradfield, J. P., Sleiman, P. M. A., Imielinski, M., Guthery, S. L., Cucchiara, S., Kim, C. E., Frackelton, E. C., Annaiah, K., Glessner, J. T., Santa, E., Willson, T., Eckert, A. W., Bonkowski, E., Shaner, J. L., Smith, R. M., Otieno, F. G., Peterson, N., Abrams, D. J., Chiavacci, R. M., Grundmeier, R., Mamula, P., Tomer, G., Piccoli, D. A., Monos, D. S., Annesse, V., Denson, L. A., Grant, S. F. A., and Hakonarson, H. (2008). Loci on 20q13 and 21q22 are associated with pediatric-onset inflammatory bowel disease. *Nat Genet*, 40(10):1211–1215.
- [262] Kühn, R., Löhler, J., Rennick, D., Rajewsky, K., and Müller, W. (1993). Interleukin-10-deficient mice develop chronic enterocolitis. *Cell*, 75(2):263–274.
- [263] Kühnen, P., Turan, S., Fröhler, S., Güran, T., Abali, S., Biebermann, H., Bereket, A., Grüters, A., Chen, W., and Krude, H. (2014). Identification of PEN-DRIN (SLC26A4) mutations in patients with congenital hypothyroidism and "apparent" thyroid dysgenesis. *The Journal of Clinical Endocrinology and Metabolism*, 99(1):E169–76.

- [264] Kvikstad, E. M., Tyekucheva, S., Chiaromonte, F., and Makova, K. D. (2007). A macaque’s-eye view of human insertions and deletions: differences in mechanisms. *PLoS computational biology*, 3(9):1772–82.
- [265] Kyogoku, C., Langefeld, C. D., Ortmann, W. A., Lee, A., Selby, S., Carlton, V. E. H., Chang, M., Ramos, P., Baechler, E. C., Batliwalla, F. M., Novitzke, J., Williams, A. H., Gillett, C., Rodine, P., Graham, R. R., Ardlie, K. G., Gaffney, P. M., Moser, K. L., Petri, M., Begovich, A. B., Gregersen, P. K., and Behrens, T. W. (2004). Genetic association of the R620W polymorphism of protein tyrosine phosphatase PTPN22 with human SLE. *American Journal of Human Genetics*, 75(3):504–7.
- [266] LA. Everett, B. Glasser, J. B. (1997). Pendred syndrome is caused by mutations in a putative sulphate transporter gene (PDS). *Nature Genetics*, 17:411–422.
- [267] Ladinsky, H. T., Perez, E. E., and Dorsey, M. J. (2013). Chronic granulomatous disease associated colitis leading to profound zinc deficiency. *The journal of allergy and clinical immunology. In practice*, 2(2):217–9.
- [268] Lalande, M. (2001). Imprints of disease at GNAS1. *Journal of Clinical Investigation*, 107(7):793–794.
- [269] Lange, L. A., Hu, Y., Zhang, H., Xue, C., Schmidt, E. M., Tang, Z. Z., Bizon, C., Lange, E. M., Smith, J. D., Turner, E. H., Jun, G., Kang, H. M., Peloso, G., Auer, P., Li, K. P., Flannick, J., Zhang, J., Fuchsberger, C., Gaulton, K., Lindgren, C., Locke, A., Manning, A., Sim, X., Rivas, M. A., Holmen, O. L., Gottesman, O., Lu, Y., Ruderfer, D., Stahl, E. A., Duan, Q., Li, Y., Durda, P., Jiao, S., Isaacs, A., Hofman, A., Bis, J. C., Correa, A., Griswold, M. E., Jakobsdottir, J., Smith, A. V., Schreiner, P. J., Feitosa, M. F., Zhang, Q., Huffman, J. E., Crosby, J., Wassel, C. L., Do, R., Franceschini, N., Martin, L. W., Robinson, J. G., Assimes, T. L., Crosslin, D. R., Rosenthal, E. A., Tsai, M., Rieder, M. J., Farlow, D. N., Folsom, A. R., Lumley, T., Fox, E. R., Carlson, C. S., Peters, U., Jackson, R. D., Van Duijn, C. M., Uitterlinden, A. G., Levy, D., Rotter, J. I., Taylor, H. A., Gudnason, V., Siscovick, D. S., Fornage, M., Borecki, I. B., Hayward, C., Rudan, I., Chen, Y. E., Bottinger, E. P., Loos, R. J. F., Sætrom, P., Hveem, K., Boehnke, M., Groop, L., McCarthy, M., Meitinger, T., Ballantyne, C. M., Gabriel, S. B., O’Donnell, C. J., Post, W. S., North, K. E., Reiner, A. P., Boerwinkle, E., Psaty, B. M., Altshuler, D., Kathiresan, S., Lin, D. Y., Jarvik, G. P., Cupples, L. A., Kooperberg, C., Wilson, J. G., Nickerson, D. A., Abecasis, G. R., Rich, S. S., Tracy, R. P., and Willer, C. J. (2014). Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. *American Journal of Human Genetics*, 94(2):233–245.
- [270] Lania, G., Zhang, Z., Huynh, T., Caprio, C., Moon, A. M., Vitelli, F., and Baldini, A. (2009). Early thyroid development requires a Tbx1-Fgf8 pathway. *Developmental biology*, 328(1):109–117.
- [271] Lappalainen, I., Thusberg, J., Shen, B., and Vihinen, M. (2008). Genome wide analysis of pathogenic SH2 domain mutations. *Proteins: Structure, Function and Genetics*, 72(2):779–792.

- [272] Lee, H., Graham, J. M., Rimoin, D. L., Lachman, R. S., Krejci, P., Tompson, S. W., Nelson, S. F., Krakow, D., and Cohn, D. H. (2012). Exome sequencing identifies PDE4D mutations in acrodysostosis. *American Journal of Human Genetics*, 90(4):746–51.
- [273] Lee, J., Jeso, B. D., and Arvan, P. (2008). The cholinesterase-like domain of thyroglobulin functions as an intramolecular chaperone. *The Journal of clinical investigation*, 118(8).
- [274] Lee, J., Jeso, B. D., and Arvan, P. (2011). Maturation of Thyroglobulin Protein Region I*. *The Journal of biological chemistry*, 286(38):33045–33052.
- [275] Lee, J. C., Espéli, M., Anderson, C. A., Linterman, M. A., Pocock, J. M., Williams, N. J., Roberts, R., Viatte, S., Fu, B., Peshu, N., Hien, T. T., Phu, N. H., Wesley, E., Edwards, C., Ahmad, T., Mansfield, J. C., Gearry, R., Dunstan, S., Williams, T. N., Barton, A., Vinuesa, C. G., Parkes, M., Lyons, P. A., and Smith, K. G. C. (2013). Human SNP links differential outcomes in inflammatory and infectious disease to a FOXO3-regulated pathway. *Cell*, 155(1):57–69.
- [276] Lee, S., Abecasis, G. R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: Study designs and statistical tests. *American Journal of Human Genetics*, 95(1):5–23.
- [277] Léger, J., Olivieri, A., Donaldson, M., Torresani, T., Krude, H., van Vliet, G., Polak, M., and Butler, G. (2014). European Society for Paediatric Endocrinology consensus guidelines on screening, diagnosis, and management of congenital hypothyroidism. *Hormone research in paediatrics*, 81(2):80–103.
- [278] Lelieveld, S. H., Veltman, J. A., and Gilissen, C. (2016). Novel bioinformatic developments for exome sequencing. *Human Genetics*, 135(6):1–12.
- [279] Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., Axelrod, N., Huang, J., Kirkness, E. F., Denisov, G., Lin, Y., MacDonald, J. R., Pang, A. W. C., Shago, M., Stockwell, T. B., Tsiamouri, A., Bafna, V., Bansal, V., Kravitz, S. A., Busam, D. A., Beeson, K. Y., McIntosh, T. C., Remington, K. A., Abril, J. F., Gill, J., Borman, J., Rogers, Y. H., Frazier, M. E., Scherer, S. W., Strausberg, R. L., and Venter, J. C. (2007). The diploid genome sequence of an individual human. *PLoS Biology*, 5(10):2113–2144.
- [280] Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14):1754–60.
- [281] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–9.
- [282] Li, L.-h., Ho, S.-f., Chen, C.-h., Wei, C.-y., Wong, W.-c., Li, L.-y., Hung, S.-i., Chung, W.-h., Pan, W.-h., Lee, M.-t. M., Tsai, F.-j., Chang, C.-f., Wu, J.-y., and Chen, Y.-t. (2006). Long Contiguous Stretches of Homozygosity in the Human Genome. *Human mutation*, 27(11):1115–1121.

- [283] Li, Y., Sidore, C., Kang, H. M., Boehnke, M., and Abecasis, G. R. (2011). Low-coverage sequencing: implications for design of complex trait association studies. *Genome . . .*, 21:940–951.
- [284] Li, Y. J., Scott, W. K., Hedges, D. J., Zhang, F., Gaskell, P. C., Nance, M. A., Watts, R. L., Hubble, J. P., Koller, W. C., Pahwa, R., Stern, M. B., Hiner, B. C., Jankovic, J., Allen, F. A., Goetz, C. G., Mastaglia, F., Stajich, J. M., Gibson, R. A., Middleton, L. T., Saunders, A. M., Scott, B. L., Small, G. W., Nicodemus, K. K., Reed, A. D., Schmechel, D. E., Welsh-Bohmer, K. A., Conneally, P. M., Roses, A. D., Gilbert, J. R., Vance, J. M., Haines, J. L., and Pericak-Vance, M. A. (2002). Age at onset in two common neurodegenerative diseases is genetically controlled. *American Journal of Human Genetics*, 70(4):985–993.
- [285] Liebert, M. A., Yeo, G., and Burge, C. B. (2004). Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. *Journal of Computational Biology*, 11:377–394.
- [286] Lindgren, C. M., Heid, I. M., Randall, J. C., Lamina, C., Steinthorsdottir, V., Qi, L., Speliotes, E. K., Thorleifsson, G., Willer, C. J., Herrera, B. M., Jackson, A. U., Lim, N., Scheet, P., Soranzo, N., Amin, N., Aulchenko, Y. S., Chambers, J. C., Drong, A., an Luan, J. N., Lyon, H. N., Rivadeneira, F., Sanna, S., Timpson, N. J., Zillikens, M. C., Jing, H. Z., Almgren, P., Bandinelli, S., Bennett, A. J., Bergman, R. N., Bonnycastle, L. L., Bumpstead, S. J., Chanoock, S. J., Cherkas, L., Chines, P., Coin, L., Cooper, C., Crawford, G., Doering, A., Dominiczak, A., Doney, A. S. F., Ebrahim, S., Elliott, P., Erdos, M. R., Estrada, K., Ferrucci, L., Fischer, G., Forouhi, N. G., Gieger, C., Grallert, H., Groves, C. J., Grundy, S., Guiducci, C., Hadley, D., Hamsten, A., Havulinna, A. S., Hofman, A., Holle, R., Holloway, J. W., Illig, T., Isomaa, B., Jacobs, L. C., Jameson, K., Jousilahti, P., Karpe, F., Kuusisto, J., Laitinen, J., Lathrop, G. M., Lawlor, D. A., Mangino, M., McArdle, W. L., Meitinger, T., Morken, M. A., Morris, A. P., Munroe, P., Narisu, N., Nordstrom, A., Nordstrom, P., Oostra, B. A., Palmer, C. N. A., Payne, F., Peden, J. F., Prokopenko, I., Renstr??m, F., Ruukonen, A., Salomaa, V., Sandhu, M. S., Scott, L. J., Scuteri, A., Silander, K., Song, K., Yuan, X., Stringham, H. M., Swift, A. J., Tuomi, T., Uda, M., Vollenweider, P., Waeber, G., Wallace, C., Walters, G. B., Weedon, M. N., Witteman, J. C. M., Zhang, C., Zhang, W., Caulfield, M. J., Collins, F. S., Smith, G. D., Day, I. N. M., Franks, P. W., Hattersley, A. T., Hu, F. B., Jarvelin, M. R., Kong, A., Kooner, J. S., Laakso, M., Lakatta, E., Mooser, V., Morris, A. D., Peltonen, L., Samani, N. J., Spector, T. D., Strachan, D. P., Tanaka, T., Tuomilehto, J., Uitterlinden, A. G., Van Duijn, C. M., Wareham, N. J., Watkins, H., Waterworth, D. M., Boehnke, M., Deloukas, P., Groop, L., Hunter, D. J., Thorsteinsdottir, U., Schlessinger, D., Wichmann, H. E., Frayling, T. M., Abecasis, G. R., Hirschhorn, J. N., Loos, R. J. F., Stefansson, K., Mohlke, K. L., Barroso, I., and McCarthy, M. I. (2009). Genome-wide association scan meta-analysis identifies three loci influencing adiposity and fat distribution. *PLoS Genetics*, 5(6).
- [287] Liu, C.-t., Raghavan, S., Maruthur, N., Kabagambe, E. K., Hong, J., Ng, M. C. Y., Hivert, M.-f., Lu, Y., An, P., Bentley, A. R., Drolet, A. M., Gaulton, K. J., Guo, X., Armstrong, L. L., Irvin, M. R., Li, M., Lipovich, L., Rybin, D. V., Taylor, K. D., Agyemang, C., Palmer, N. D., Cade, B. E., Chen, W.-m., Dauriz, M., Delaney,

- J. A. C., Edwards, T. L., Evans, D. S., Evans, M. K., Lange, L. A., Leong, A., Liu, J., Liu, Y., Nayak, U., Patel, S. R., Porneala, B. C., Rasmussen-torvik, L. J., Snijder, M. B., Stallings, S. C., Tanaka, T., Yanek, L. R., Zhao, W., Becker, D. M., Bielak, L. F., Biggs, M. L., Bottinger, E. P., Bowden, D. W., Chen, G., Correa, A., Couper, D. J., Crawford, D. C., Cushman, M., Eicher, J. D., Fornage, M., Franceschini, N., Fu, Y.-p., Goodarzi, M. O., Gottesman, O., Hara, K., Harris, T. B., Jensen, R. A., Johnson, A. D., Jhun, M. A., Karter, A. J., Keller, M. F., Kho, A. N., Kizer, J. R., Krauss, R. M., Langefeld, C. D., and Li, X. (2016). Trans-ethnic Meta-Analysis and Functional Annotation Illuminates the Genetic Architecture of Fasting Glucose and Insulin. *American Journal of Human Genetics*, 99:1–20.
- [288] Liu, J. Z. and Anderson, C. A. (2014). Genetic studies of Crohn’s disease: Past, present and future. *Best Practice and Research: Clinical Gastroenterology*, 28(3):373–386.
- [289] Liu, J. Z., Hov, J. R., Folseraas, T., Ellinghaus, E., Rushbrook, S. M., Doncheva, N. T., Andreassen, O. a., Weersma, R. K., Weismüller, T. J., Eksteen, B., Invernizzi, P., Hirschfield, G. M., Gotthardt, D. N., Pares, A., Ellinghaus, D., Shah, T., Juran, B. D., Milkiewicz, P., Rust, C., Schramm, C., Müller, T., Srivastava, B., Dalekos, G., Nöthen, M. M., Herms, S., Winkelmann, J., Mitrovic, M., Braun, F., Ponsioen, C. Y., Croucher, P. J. P., Sterneck, M., Teufel, A., Mason, A. L., Saarela, J., Leppa, V., Dorfman, R., Alvaro, D., Floreani, A., Onengut-Gumuscu, S., Rich, S. S., Thompson, W. K., Schork, A. J., Næss, S., Thomsen, I., Mayr, G., König, I. R., Hveem, K., Cleynen, I., Gutierrez-Achury, J., Ricaño-Ponce, I., van Heel, D., Björnsson, E., Sandford, R. N., Durie, P. R., Melum, E., Vatn, M. H., Silverberg, M. S., Duerr, R. H., Padyukov, L., Brand, S., Sans, M., Annese, V., Achkar, J.-P., Boberg, K. M., Marschall, H.-U., Chazouillères, O., Bowlus, C. L., Wijmenga, C., Schrupf, E., Vermeire, S., Albrecht, M., Rioux, J. D., Alexander, G., Bergquist, A., Cho, J., Schreiber, S., Manns, M. P., Färkkilä, M., Dale, A. M., Chapman, R. W., Lazaridis, K. N., Franke, A., Anderson, C. a., and Karlsen, T. H. (2013). Dense genotyping of immune-related disease regions identifies nine new risk loci for primary sclerosing cholangitis. *Nature Genetics*, 45(6):670–5.
- [290] Liu, J. Z., van Sommeren, S., Huang, H., Ng, S. C., Alberts, R., Takahashi, A., Ripke, S., Lee, J. C., Jostins, L., Shah, T., Abedian, S., Cheon, J. H., Cho, J., Daryani, N. E., Franke, L., Fuyuno, Y., Hart, A., Juyal, R. C., Juyal, G., Kim, W. H., Morris, A. P., Poustchi, H., Newman, W. G., Midha, V., Orchard, T. R., Vahedi, H., Sood, A., Sung, J. J. Y., Malekzadeh, R., Westra, H.-J., Yamazaki, K., Yang, S.-K., Barrett, J. C., Franke, A., Alizadeh, B. Z., Parkes, M., B K, T., Daly, M. J., Kubo, M., Anderson, C. A., and Weersma, R. K. (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature Genetics*, advance on(9).
- [291] Lloyd-Jones, D. M., Leip, E. P., Larson, M. G., D’Agostino, R. B., Beiser, A., Wilson, P. W. F., Wolf, P. A., and Levy, D. (2006). Prediction of lifetime risk for cardiovascular disease by risk factor burden at 50 years of age. *Circulation*, 113(6):791–798.
- [292] Lobo, I. (2008). Same Genetic Mutation, Different Genetic Disease Phenotype. *Nature Education*, 1(1):64.

- [293] Locke, A., Kahali, B., Berndt, S., Justice, A., and Pers, T. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206.
- [294] Loomis, E. W., Eid, J. S., Peluso, P., Yin, J., Hickey, L., Rank, D., Mccalmon, S., Hagerman, R. J., Tassone, F., and Hagerman, P. J. (2013). Sequencing the unsequenceable: Expanded CGG-repeat alleles of the fragile X gene Sequencing the unsequenceable: Expanded CGG-repeat alleles of the fragile X gene. (530):121–128.
- [295] Luo, Y., de Lange, K. M., Jostins, L., Moutsianas, L., Randall, J., Kennedy, N. A., Lamb, C. A., McCarthy, S., Ahmad, T., Edwards, C., Serra, E. G., Hart, A., Hawkey, C., Mansfield, J. C., Mowat, C., Newman, W. G., Nichols, S., Pollard, M., Satsangi, J., Simmons, A., Tremelling, M., Uhlig, H., Wilson, D. W., Lee, J. C., Prescott, N. J., Lees, C. W., Mathew, C. G., Parkes, M., Barrett, J. C., and Anderson, C. A. (2016). Exploring the genetic architecture of inflammatory bowel disease by whole genome sequencing identifies association at ADCY7. *BioRxiv*, pages 1–24.
- [296] Lupski, J. R., Reid, J. G., Gonzaga-Jauregui, C., Rio Deiros, D., Chen, D. C. Y., Nazareth, L., Bainbridge, M., Dinh, H., Jing, C., Wheeler, D. a., McGuire, A. L., Zhang, F., Stankiewicz, P., Halperin, J. J., Yang, C., Gehman, C., Guo, D., Irikat, R. K., Tom, W., Fantin, N. J., Muzny, D. M., and Gibbs, R. a. (2010). Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *The New England Journal of Medicine*, 362(13):1181–1191.
- [297] Lynch, M. (2010). Rate, molecular spectrum, and consequences of human mutation. *Proceedings of the National Academy of Sciences of the United States of America*, 107(3):961–8.
- [298] Lyons, I., Parsons, L. M., Hartley, L., Li, R., Andrews, J. E., Robb, L., and Harvey, R. P. (1995). Myogenic and morphogenetic defects in the heart tubes of murine embryos lacking the homeo box gene Nkx2-5. *Genes and Development*, 9(13):1654–1666.
- [299] MacArthur, D. G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J. K., Montgomery, S. B., Albers, C. A., Zhang, Z. D., Conrad, D. F., Lunter, G., Zheng, H., Ayub, Q., DePristo, M. A., Banks, E., Hu, M., Handsaker, R. E., Rosenfeld, J. A., Fromer, M., Jin, M., Mu, X. J., Khurana, E., Ye, K., Kay, M., Saunders, G. I., Suner, M.-M., Hunt, T., Barnes, I. H. A., Amid, C., Carvalho-Silva, D. R., Bignell, A. H., Snow, C., Yngvadottir, B., Bumpstead, S., Cooper, D. N., Xue, Y., Romero, I. G., Wang, J., Li, Y., Gibbs, R. A., McCarroll, S. A., Dermitzakis, E. T., Pritchard, J. K., Barrett, J. C., Harrow, J., Hurler, M. E., Gerstein, M. B., and Tyler-Smith, C. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. *Science (New York, N.Y.)*, 335(6070):823–8.
- [300] MacArthur, D. G., Manolio, T. A., Dimmock, D. P., Rehm, H. L., Shendure, J., Abecasis, G. R., Adams, D. R., and Altman, R. B. (2014). Guidelines for investigating causality of sequence variants in human disease. *Nature*, 508(7497):469–476.
- [301] Macchia, P. E., Krude, H., and Pirro, M. T. (1998). PAX8 mutations associated with congenital hypothyroidism caused by thyroid dysgenesis. *Nat Genet*, 19:83–86.

- [302] Madsen, B. E. and Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics*, 5(2).
- [303] Mahajan, A., Go, M. J., Zhang, W., Below, J. E., Gaulton, K. J., Ferreira, T., Horikoshi, M., Johnson, A. D., Ng, M. C. Y., Prokopenko, I., Saleheen, D., Wang, X., Zeggini, E., Abecasis, G. R., Adair, L. S., Almgren, P., Atalay, M., Aung, T., Baldassarre, D., Balkau, B., Bao, Y., Barnett, A. H., Barroso, I., Basit, A., Been, L. F., Beilby, J., Bell, G. I., Benediktsson, R., Bergman, R. N., Boehm, B. O., Boerwinkle, E., Bonnycastle, L. L., Burt, N., Cai, Q., Campbell, H., Carey, J., Cauchi, S., Caulfield, M., Chan, J. C. N., Chang, L.-C., Chang, T.-J., Chang, Y.-C., Charpentier, G., Chen, C.-H., Chen, H., Chen, Y.-T., Chia, K.-S., Chidambaram, M., Chines, P. S., Cho, N. H., Cho, Y. M., Chuang, L.-M., Collins, F. S., Cornelis, M. C., Couper, D. J., Crenshaw, A. T., van Dam, R. M., Danesh, J., Das, D., de Faire, U., Dedoussis, G., Deloukas, P., Dimas, A. S., Dina, C., Doney, A. S., Donnelly, P. J., Dorkhan, M., van Duijn, C., Dupuis, J., Edkins, S., Elliott, P., Emilsson, V., Erbel, R., Eriksson, J. G., Escobedo, J., Esko, T., Eury, E., Florez, J. C., Fontanillas, P., Forouhi, N. G., Forsen, T., Fox, C., Fraser, R. M., Frayling, T. M., Froguel, P., Frossard, P., Gao, Y., Gertow, K., Gieger, C., Gigante, B., Grallert, H., Grant, G. B., Grrop, L. C., Groves, C. J., Grundberg, E., Guiducci, C., Hamsten, A., Han, B.-G., Hara, K., Hassanali, N., Hattersley, A. T., Hayward, C., Hedman, A. K., Herder, C., Hofman, A., Holmen, O. L., Hovingh, K., Hreidarsson, A. B., Hu, C., Hu, F. B., Hui, J., Humphries, S. E., Hunt, S. E., Hunter, D. J., Hveem, K., Hydrie, Z. I., Ikegami, H., Illig, T., Ingelsson, E., Islam, M., Isomaa, B., Jackson, A. U., Jafar, T., James, A., Jia, W., Jöckel, K.-H., Jonsson, A., Jowett, J. B. M., Kadowaki, T., Kang, H. M., Kanoni, S., Kao, W. H. L., Kathiresan, S., Kato, N., Katulanda, P., Keinanen-Kiukkaanniemi, K. M., Kelly, A. M., Khan, H., Khaw, K.-T., Khor, C.-C., Kim, H.-L., Kim, S., Kim, Y. J., Kinnunen, L., Klopp, N., Kong, A., Korpi-Hyövälti, E., Kowlessur, S., Kraft, P., Kravic, J., Kristensen, M. M., Krithika, S., Kumar, A., Kumate, J., Kuusisto, J., Kwak, S. H., Laakso, M., Lagou, V., Lakka, T. A., Langenberg, C., Langford, C., Lawrence, R., Leander, K., Lee, J.-M., Lee, N. R., Li, M., Li, X., Li, Y., Liang, J., Liju, S., Lim, W.-Y., Lind, L., Lindgren, C. M., Lindholm, E., Liu, C.-T., Liu, J. J., Lobbens, S., Long, J., Loos, R. J. F., Lu, W., Luan, J., Lyssenko, V., Ma, R. C. W., Maeda, S., Mägi, R., Männistö, S., Matthews, D. R., Meigs, J. B., Melander, O., Metspalu, A., Meyer, J., Mirza, G., Mihailov, E., Moebus, S., Mohan, V., Mohlke, K. L., Morris, A. D., Mühleisen, T. W., Müller-Nurasyid, M., Musk, B., Nakamura, J., Nakashima, E., Navarro, P., Ng, P.-K., Nica, A. C., Nilsson, P. M., Njølstad, I., Nöthen, M. M., Ohnaka, K., Ong, T. H., Owen, K. R., Palmer, C. N. A., Pankow, J. S., Park, K. S., Parkin, M., Pechlivanis, S., Pedersen, N. L., Peltonen, L., Perry, J. R. B., Peters, A., Pinidiyapathirage, J. M., Platou, C. G., Potter, S., Price, J. F., Qi, L., Radha, V., Rallidis, L., Rasheed, A., Rathman, W., Rauramaa, R., Raychaudhuri, S., Rayner, N. W., Rees, S. D., Rehnberg, E., Ripatti, S., Robertson, N., Roden, M., Rossin, E. J., Rudan, I., Rybin, D., Saaristo, T. E., Salomaa, V., Saltevo, J., Samuel, M., Sanghera, D. K., Saramies, J., Scott, J., Scott, L. J., Scott, R. A., Segrè, A. V., Sehmi, J., Sennblad, B., Shah, N., Shah, S., Shera, A. S., Shu, X. O., Shuldiner, A. R., Sigurdsson, G., Sijbrands, E., Silveira, A., Sim, X., Sivapalaratnam, S., Small, K. S., So, W. Y., Stančáková, A., Stefansson, K., Steinbach, G., Steinthorsdóttir, V., Stirrups, K., Strawbridge, R. J., Stringham, H. M., Sun, Q., Suo, C., Syvänen, A.-C., Takayanagi, R., Takeuchi, F., Tay, W. T., Teslovich, T. M., Thorand, B.,

- Thorleifsson, G., Thorsteinsdottir, U., Tikkanen, E., Trakalo, J., Tremoli, E., Trip, M. D., Tsai, F. J., Tuomi, T., Tuomilehto, J., Uitterlinden, A. G., Valladares-Salgado, A., Vedantam, S., Veglia, F., Voight, B. F., Wang, C., Wareham, N. J., Wennauer, R., Wickremasinghe, A. R., Wilsgaard, T., Wilson, J. F., Wiltshire, S., Winckler, W., Wong, T. Y., Wood, A. R., Wu, J.-Y., Wu, Y., Yamamoto, K., Yamauchi, T., Yang, M., Yengo, L., Yokota, M., Young, R., Zabaneh, D., Zhang, F., Zhang, R., Zheng, W., Zimmet, P. Z., Altshuler, D., Bowden, D. W., Cho, Y. S., Cox, N. J., Cruz, M., Hanis, C. L., Kooner, J., Lee, J.-Y., Seielstad, M., Teo, Y. Y., Boehnke, M., Parra, E. J., Chambers, J. C., Tai, E. S., McCarthy, M. I., and Morris, A. P. (2014). Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature Genetics*, 46(3):234–44.
- [304] Manase, D., Alessandro, L. C. A. D., Manickaraj, A. K., Turki, S. A., and Hurles, M. E. (2014). High throughput exome coverage of clinically relevant cardiac genes. *BMC medical genomics*, 7:1–10.
- [305] Manley, N. R. and Capecchi, M. R. (1995). The role of Hoxa-3 in mouse thymus and thyroid development. *Development (Cambridge, England)*, 121(7):1989–2003.
- [306] Manley, N. R. and Capecchi, M. R. (1998). Hox group 3 paralogs regulate the development and migration of the thymus, thyroid, and parathyroid glands. *Developmental biology*, 195(1):1–15.
- [307] Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F. C., McCarroll, S. A., and Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–53.
- [308] Manrai, A. K., Funke, B. H., Rehm, H. L., Olesen, M. S., Maron, B. A., Szolovits, P., Margulies, D. M., Loscalzo, J., and Kohane, I. S. (2016). Genetic Misdiagnoses and the Potential for Health Disparities. *New England Journal of Medicine*, 375(7):655–665.
- [309] Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7):499–511.
- [310] Marco, G. D., Agretti, P., Montanelli, L., Cosmo, C. D., Bagattini, B., Servi, M. D., Ferrarini, E., Dimida, A., Claudia, A., Ferreira, F., Molinaro, A., Ceccarelli, C., Brozzi, F., Pinchera, A., Vitti, P., and Tonacchera, M. (2011). Identification and Functional Analysis of Novel Dual Oxidase 2 (DUOX2) Mutations in Children with Congenital or Subclinical Hypothyroidism. *Journal of Clinical Endocrinology and Metabolism*, 96(8):1335–1339.
- [311] Marinovic, D., Garel, C., Bonai, C., Polak, M., Czernichow, P., Unit, P. E., and Inserm, U. J. L. (2002). Thyroid Developmental Anomalies in First Degree Relatives of Children with Congenital Hypothyroidism. *The Journal of Clinical Endocrinology and Metabolism*, 87(August):575–580.

- [312] Maruo, Y., Takahashi, H., Soeda, I., Nishikura, N., Matsui, K., Ota, Y., Mimura, Y., Mori, A., Sato, H., and Takeuchi, Y. (2008). Transient congenital hypothyroidism caused by biallelic mutations of the dual oxidase 2 gene in Japanese patients detected by a neonatal screening program. *The Journal of Clinical Endocrinology and Metabolism*, 93(11):4261–7.
- [313] Mathew, C. G. and Lewis, C. M. (2004). Genetics of inflammatory bowel disease: progress and prospects. *Human Molecular Genetics*, 13 Spec No(1):R161—R168.
- [314] Mathieson, I. and McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. *Nature Genetics*, 44(3):243–6.
- [315] Matthews, A. G., Finkelstein, D. M., and Betensky, R. A. (2008). Analysis of familial aggregation studies with complex ascertainment schemes. *Statistics in Medicine*, 27(24):5076–5092.
- [316] Matute, J. D., Arias, A. A., Wright, N. A. M., Wrobel, I., Waterhouse, C. C. M., Li, X. J., Marchal, C. C., Stull, N. D., Lewis, D. B., Steele, M., Kellner, J. D., Yu, W., Meroueh, S. O., Nauseef, W. M., and Dinauer, M. C. (2009). A new genetic subgroup of chronic granulomatous disease with autosomal recessive mutations in p40 phox and selective defects in neutrophil NADPH oxidase activity. *Blood*, 114(15):3309–3315.
- [317] Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutuyavin, T., Stehling-Sun, S., Johnson, A. K., Canfield, T. K., Giste, E., Diegel, M., Bates, D., Hansen, R. S., Neph, S., Sabo, P. J., Heimfeld, S., Raubitschek, A., Ziegler, S., Cotsapas, C., Sotoodehnia, N., Glass, I., Sunyaev, S. R., Kaul, R., and Stamatoyannopoulos, J. A. (2012). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science*, 337(6099):1190–1195.
- [318] Max, V. (2015). The DNA of a Nation. *Nature*, 524:503–505.
- [319] McCarthy, S., Das, S., Kretzschmar, W., Durbin, R., Abecasis, G., and Marchini, J. (2015). A reference panel of 64,976 haplotypes for genotype imputation. *bioRxiv*.
- [320] McCarthy, S. E., Gillis, J., Kramer, M., Lihm, J., Yoon, S., Berstein, Y., Mistry, M., Pavlidis, P., Solomon, R., Ghiban, E., Antoniou, E., Kelleher, E., O’Brien, C., Donohoe, G., Gill, M., Morris, D. W., McCombie, W. R., and Corvin, A. (2014). De novo mutations in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and intellectual disability. *Molecular psychiatry*, 19(6):652–8.
- [321] McKernan, K. J., Peckham, H. E., Costa, G., McLaughlin, S., Tsung, E., Fu, Y., Clouser, C., Dunkan, C., Ichikawa, J., Lee, C., Zhang, Z., Sheridan, A., Fu, H., Ranade, S., Dimilanta, E., Sokolsky, T., Zhang, L., Hendrickson, C., Bin Li1, L. K., Stuart, J., Malek, J., Manning, J., Antipova, A., Perez, D., Moore, M., Hayashibara, K., Lyons, M., Beaudoin, R., Coleman, B., Laptewicz, M., Sanicandro, A., Rhodes, M., Vega, F. D. L., Gottimukkala, R. K., Hyland, F., Reese, M., Yang, S., Bafna, V., Bashir, A., MacBride, A., Aklan, C., Kidd, J. M., Eichler, E. E., and Blanchard,

- A. P. (2009). Sequence and Structural Variation in a Human Genome Uncovered by Short-Read , Massively Parallel Ligation. *Genome Research*, 19:1527–1541.
- [322] McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., and Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics (Oxford, England)*, 26(16):2069–70.
- [323] McMillan, H. J., Worthylake, T., Schwartzentruber, J., Gottlieb, C. C., Lawrence, S. E., Mackenzie, A., Beaulieu, C. L., Mooyer, P. a. W., FORGE Canada Consortium, Wanders, R. J. a., Majewski, J., Bulman, D. E., Geraghty, M. T., Ferdinandusse, S., and Boycott, K. M. (2012). Specific combination of compound heterozygous mutations in 17 β -hydroxysteroid dehydrogenase type 4 (HSD17B4) defines a new subtype of D-bifunctional protein deficiency. *Orphanet Journal of Rare Diseases*, 7(1):90.
- [324] McRae, J. F., Clayton, S., Fitzgerald, T. W., Kaplanis, J., Prigmore, E., Rajan, D., Sifrim, A., Aitken, S., Akawi, N., Alvi, M., Ambridge, K., Barrett, D. M., Bayzatinova, T., Jones, P., Jones, W. D., King, D., Krishnappa, N., Mason, L. E., Singh, T., Tivey, A. R., Ahmed, M., Anjum, U., Archer, H., Armstrong, R., Awada, J., Balasubramanian, M., Banka, S., Baralle, D., Barnicoat, A., Batstone, P., Baty, D., Bennett, C., Berg, J., Bernhard, B., Bevan, A. P., Bitner-Glindzicz, M., Blair, E., Blyth, M., Bohanna, D., Bourdon, L., Bourn, D., Bradley, L., Brady, A., Brent, S., Brewer, C., Brunstrom, K., Bunyan, D. J., Burn, J., Canham, N., Castle, B., Chandler, K., Chatzimichali, E., Cilliers, D., Clarke, A., Clasper, S., Clayton-Smith, J., Clowes, V., Coates, A., Cole, T., Colgiu, I., Collins, A., Collinson, M. N., Connell, F., Cooper, N., Cox, H., Cresswell, L., Cross, G., Crow, Y., D\textquoterightAlessandro, M., Dabir, T., Davidson, R., Davies, S., de Vries, D., Dean, J., Deshpande, C., Devlin, G., Dixit, A., Dobbie, A., Donaldson, A., Donnai, D., Donnelly, D., Donnelly, C., Douglas, A., Douzgou, S., Duncan, A., Eason, J., Ellard, S., Ellis, I., Elmslie, F., Evans, K., Everest, S., Fendick, T., Fisher, R., Flinter, F., Foulds, N., Fry, A., Fryer, A., Gardiner, C., Gaunt, L., Ghali, N., Gibbons, R., Gill, H., Goodship, J., Goudie, D., Gray, E., Green, A., Greene, P., Greenhalgh, L., Gribble, S., Harrison, R., Harrison, L., Harrison, V., Hawkins, R., He, L., Hellens, S., Henderson, A., Hewitt, S., Hildyard, L., Hobson, E., Holden, S., Holder, M., Holder, S., Hollingsworth, G., Homfray, T., Humphreys, M., Hurst, J., Hutton, B., Ingram, S., Irving, M., Islam, L., Jackson, A., Jarvis, J., Jenkins, L., Johnson, D., Jones, E., Josifova, D., Joss, S., Kaemba, B., Kazembe, S., Kellsell, R., Kerr, B., Kingston, H., Kini, U., Kinning, E., Kirby, G., Kirk, C., Kivuva, E., Kraus, A., Kumar, D., Kumar, V. K. A., Lachlan, K., Lam, W., Lampe, A., Langman, C., Lees, M., Lim, D., Longman, C., Lowther, G., Lynch, S. A., Magee, A., Maher, E., Male, A., Mansour, S., Marks, K., Martin, K., Maye, U., McCann, E., McConnell, V., McEntagart, M., McGowan, R., McKay, K., McKee, S., McMullan, D. J., McNerlan, S., McWilliam, C., Mehta, S., Metcalfe, K., Middleton, A., Miedzybrodzka, Z., Miles, E., Mohammed, S., Montgomery, T., Moore, D., Morgan, S., Morton, J., Mugalaasi, H., Murday, V., Murphy, H., Naik, S., Nemeth, A., Nevitt, L., Newbury-Ecob, R., Norman, A., O\textquoterightShea, R., Ogilvie, C., Ong, K.-R., Park, S.-M., Parker, M. J., Patel, C., Paterson, J., Payne, S., Perrett, D., Phipps, J., Pilz, D. T., Pollard, M., Pottinger, C., Poulton, J., Pratt, N., Prescott, K., Price, S., Pridham, A., Procter, A., Purnell, H., Quarrell, O., Ragge, N., Rahbari, R., Randall, J.,

- Rankin, J., Raymond, L., Rice, D., Robert, L., Roberts, E., Roberts, J., Roberts, P., Roberts, G., Ross, A., Rosser, E., Saggar, A., Samant, S., Sampson, J., Sandford, R., Sarkar, A., Schweiger, S., Scott, R., Scurr, I., Selby, A., Seller, A., Sequeira, C., Shannon, N., Sharif, S., Shaw-Smith, C., Shearing, E., Shears, D., Sheridan, E., Simonic, I., Singzon, R., Skitt, Z., Smith, A., Smith, K., Smithson, S., Sneddon, L., Splitt, M., Squires, M., Stewart, F., Stewart, H., Straub, V., Suri, M., Sutton, V., Swaminathan, G. J., Sweeney, E., Tatton-Brown, K., Taylor, C., Taylor, R., Tein, M., Temple, I. K., Thomson, J., Tischkowitz, M., Tomkins, S., Torokwa, A., Treacy, B., Turner, C., Turnpenny, P., Tysoe, C., Vandersteen, A., Varghese, V., Vasudevan, P., Vijayarangakannan, P., Vogt, J., Wakeling, E., Wallwark, S., Waters, J., Weber, A., Wellesley, D., Whiteford, M., Widaa, S., Wilcox, S., Wilkinson, E., Williams, D., Williams, N., Wilson, L., Woods, G., Wragg, C., Wright, M., Yates, L., Yau, M., Nellaker, C., Parker, M. J., Firth, H. V., Wright, C. F., FitzPatrick, D. R., Barrett, J. C., and Hurles, M. E. (2016). Prevalence, phenotype and architecture of developmental disorders caused by de novo mutation. *bioRxiv*.
- [325] McVean, G. A. (2004). The Fine-Scale Structure of Recombination Rate Variation in the Human Genome. *Science*, 304(5670):581–584.
- [326] Mesut Erzurumluoglu, A., Rodriguez, S., Shihab, H. A., Baird, D., Richardson, T. G., Day, I. N. M., and Gaunt, T. R. (2015). Identifying highly penetrant disease causal mutations using next generation sequencing: Guide to whole process. *BioMed Research International*, 2015:1–16.
- [327] Meunier, D., Aubin, J., and Jeannotte, L. (2003). Perturbed thyroid morphology and transient hypothyroidism symptoms in *Hoxa5* mutant mice. *Developmental Dynamics*, 227(3):367–378.
- [328] Minoo, P., Su, G., Drum, H., Bringas, P., and Kimura, S. (1999). Defects in tracheoesophageal and lung morphogenesis in *Nkx2.1(-/-)* mouse embryos. *Developmental biology*, 209(1):60–71.
- [329] Mitchell, K. J. (2012). What is complex about complex disorders? *Genome biology*, 13(1):237.
- [330] Mizoguchi, A. and Mizoguchi, E. (2010). Animal models of IBD: Linkage to human disease. *Current Opinion in Pharmacology*, 10(5):578–587.
- [331] Molodecky, N. A., Soon, I. S., Rabi, D. M., Ghali, W. A., Ferris, M., Chernoff, G., Benchimol, E. I., Panaccione, R., Ghosh, S., Barkema, H. W., and Kaplan, G. G. (2012). Increasing Incidence and Prevalence of the Inflammatory Bowel Diseases With Time, Based on Systematic Review. *Gastroenterology*, 142(1):46–54.e42.
- [332] Monteleone, I., Pallone, F., and Monteleone, G. (2011). Th17-related cytokines: new players in the control of chronic intestinal inflammation. *BMC Medicine*, 9(1):122.
- [333] Moore, K. W., Malefyt, R. D. W., Robert, L., and Garra, A. O. (2001). Interleukin -10 and the Interleukin -10 receptor. *Molecular and Cellular Biology*, 1(1):683–765.

- [334] Moreno JC, Bikker H, Kempers MJ, van Trotsenburg AS, Baas F, de Vijlder JJ, Vulsma T, R.-S. C. (2002). Inactivating mutations in the gene for thyroid oxidase 2 (THOX2) and Congenital Hypothyroidism. *New England Journal of Medicine*, 347(2):95–102.
- [335] Morgenthaler, S. and Thilly, W. G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, 615(1-2):28–56.
- [336] Moriyama, I., Ishihara, S., Rumi, M. A., Aziz, M. D., Mishima, Y., Oshima, N., Kadota, C., Kadowaki, Y., Amano, Y., and Kinoshita, Y. (2008). Decoy oligodeoxynucleotide targeting activator protein-1 (AP-1) attenuates intestinal inflammation in murine experimental colitis. *Laboratory investigation; a journal of technical methods and pathology*, 88(6):652–663.
- [337] Morris, J. A., Randall, J. C., Maller, J. B., and Barrett, J. C. (2010). Evoker: A visualization tool for genotype intensity data. *Bioinformatics*, 26(14):1786–1787.
- [338] Morrison, A. C., Bare, L. A., Chambless, L. E., Ellis, S. G., Malloy, M., Kane, J. P., Pankow, J. S., Devlin, J. J., Willerson, J. T., and Boerwinkle, E. (2007). Prediction of Coronary Heart Disease Risk using a Genetic Risk Score: The Atherosclerosis Risk in Communities Study. *American Journal of Epidemiology*, 166(1):28–35.
- [339] Mullur, R., Liu, Y.-Y., and Brent, G. a. (2014). Thyroid hormone regulation of metabolism. *Physiological reviews*, 94(2):355–82.
- [340] Murray, P. J. (2006). STAT3-mediated anti-inflammatory signalling. *Biochemical Society transactions*, 34(Pt 6):1028–31.
- [341] Murugan, D., Albert, M. H., Langemeier, J., Bohne, J., Puchalka, J., Järvinen, P. M., Hauck, F., Klenk, A. K., Prell, C., Schatz, S., Diestelhorst, J., Sciskala, B., Kohistani, N., Belohradsky, B. H., Müller, S., Kirchner, T., Walter, M. R., Bufler, P., Muise, A. M., Snapper, S. B., Koletzko, S., Klein, C., and Kotlarz, D. (2014). Very early onset inflammatory bowel disease associated with aberrant trafficking of IL-10R1 and cure by T cell replete haploidentical bone marrow transplantation. *Journal of clinical immunology*, 34(3):331–9.
- [342] Muzza, M., Rabbiosi, S., Vigone, M. C., Zamproni, I., Cirello, V., Maffini, M. a., Maruca, K., Schoenmakers, N., Beccaria, L., Gallo, F., Park, S.-M., Beck-Peccoz, P., Persani, L., Weber, G., and Fugazzola, L. (2014). The clinical and molecular characterization of patients with dys hormonogenic congenital hypothyroidism reveals specific diagnostic clues for DUOX2 defects. *The Journal of Clinical Endocrinology and Metabolism*, 99(3):E544–53.
- [343] Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M. C., Hirai, A., Takahashi, H., Altaf-Ul-Amin, M., Ogasawara, N., and Kanaya, S. (2011). Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research*, 39(13).

- [344] Narasimhan, V., Danecek, P., Scally, A., Xue, Y., Tyler-smith, C., and Durbin, R. (2016). BCFtools / RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics (Oxford, England)*, (January):2–3.
- [345] Narumi, S., Muroya, K., Asakura, Y., Aachi, M., and Hasegawa, T. (2011). Molecular Basis of Thyroid Dysmorphogenesis: Genetic Screening in Population-Based Japanese. *Journal of Clinical Endocrinology and Metabolism*, 96:1838–1842.
- [346] Narumi, S., Muroya, K., Asakura, Y., Adachi, M., and Hasegawa, T. (2010). Transcription factor mutations and congenital hypothyroidism: Systematic genetic screening of a population-based cohort of Japanese patients. *Journal of Clinical Endocrinology and Metabolism*, 95(4):1981–1985.
- [347] Neale, B. M., Kou, Y., Liu, L., Ma’ayan, A., Samocha, K. E., Sabo, A., Lin, C.-F., Stevens, C., Wang, L.-S., Makarov, V., Polak, P., Yoon, S., Maguire, J., Crawford, E. L., Campbell, N. G., Geller, E. T., Valladares, O., Schafer, C., Liu, H., Zhao, T., Cai, G., Lihm, J., Dannenfelser, R., Jabado, O., Peralta, Z., Nagaswamy, U., Muzny, D., Reid, J. G., Newsham, I., Wu, Y., Lewis, L., Han, Y., Voight, B. F., Lim, E., Rossin, E., Kirby, A., Flannick, J., Fromer, M., Shakir, K., Fennell, T., Garimella, K., Banks, E., Poplin, R., Gabriel, S., DePristo, M., Wimbish, J. R., Boone, B. E., Levy, S. E., Betancur, C., Sunyaev, S., Boerwinkle, E., Buxbaum, J. D., Cook, E. H., Devlin, B., Gibbs, R. a., Roeder, K., Schellenberg, G. D., Sutcliffe, J. S., and Daly, M. J. (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature*, 485(7397):242–5.
- [348] Neurath, M. F. (2014). Cytokines in inflammatory bowel disease. *Nature Reviews Immunology*, 14(5):329–342.
- [349] Ng, P. C. (2003). SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31(13):3812–3814.
- [350] Ng, P. C., Levy, S., Huang, J., Stockwell, T. B., Walenz, B. P., Li, K., Axelrod, N., Busam, D. A., Strausberg, R. L., and Venter, J. C. (2008). Genetic variation in an individual human exome. *PLoS Genetics*, 4(8).
- [351] Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., Huff, C. D., Shannon, P. T., Jabs, E. W., Nickerson, D. a., Shendure, J., and Bamshad, M. J. (2010a). Exome sequencing identifies the cause of a mendelian disorder. *Nature Genetics*, 42(1):30–35.
- [352] Ng, S. B., Nickerson, D. a., Bamshad, M. J., and Shendure, J. (2010b). Massively parallel sequencing and rare disease. *Human Molecular Genetics*, 19(R2):R119–24.
- [353] Ng, S. C., Tang, W., Ching, J. Y., Wong, M., Chow, C. M., Hui, A. J., Wong, T. C., Leung, V. K., Tsang, S. W., Yu, H. H., Li, M. F., Ng, K. K., Kamm, M. A., Studd, C., Bell, S., Leong, R., de Silva, H. J., Kasturiratne, A., Mufeen, M. N. F., Ling, K. L., Ooi, C. J., Tan, P. S., Ong, D., Goh, K. L., Hilmi, I., Pisespongsa, P., Manatsathit, S., Rerknimitr, R., Aniwaniwan, S., Wang, Y. F., Ouyang, Q., Zeng, Z., Zhu, Z., Chen, M. H., Hu, P. J., Wu, K., Wang, X., Simadibrata, M., Abdullah, M., Wu, J. C., Sung, J. J. Y., and Chan, F. K. L. (2013). Incidence and Phenotype of

- Inflammatory Bowel Disease Based on Results From the Asia-Pacific Crohn's and Colitis Epidemiology Study. *Gastroenterology*, 145(1):158—165.e2.
- [354] Nichols, K. E., Harkin, D. P., Levitz, S., Krainer, M., Kolquist, K. a., Genovese, C., Bernard, a., Ferguson, M., Zuo, L., Snyder, E., Buckler, a. J., Wise, C., Ashley, J., Lovett, M., Valentine, M. B., Look, a. T., Gerald, W., Housman, D. E., and Haber, D. a. (1998). Inactivating mutations in an SH2 domain-encoding gene in X-linked lymphoproliferative syndrome. *Proceedings of the National Academy of Sciences of the United States of America*, 95(23):13765–70.
- [355] Nielsen, R., Paul, J. S., Albrechtsen, A., and Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6):443–51.
- [356] Nieuwenhuis, E. E. S. and Escher, J. C. (2008). Early onset IBD: what's the difference? *Digestive and liver disease: official journal of the Italian Society of Gastroenterology and the Italian Association for the Study of the Liver*, 40(1):12–5.
- [357] Nisticò, L., Buzzetti, R., Pritchard, L. E., Van der Auwera, B., Giovannini, C., Bosi, E., Larrad, M. T., Rios, M. S., Chow, C. C., Cockram, C. S., Jacobs, K., Mijovic, C., Bain, S. C., Barnett, a. H., Vandewalle, C. L., Schuit, F., Gorus, F. K., Tosi, R., Pozzilli, P., and Todd, J. a. (1996). The CTLA-4 gene region of chromosome 2q33 is linked to, and associated with, type 1 diabetes. Belgian Diabetes Registry. *Human Molecular Genetics*, 5(7):1075–1080.
- [358] O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., Traglia, M., Huang, J., Huffman, J. E., Rudan, I., McQuillan, R., Fraser, R. M., Campbell, H., Polasek, O., Asiki, G., Ekoru, K., Hayward, C., Wright, A. F., Vitart, V., Navarro, P., Zagury, J. F., Wilson, J. F., Toniolo, D., Gasparini, P., Soranzo, N., Sandhu, M. S., and Marchini, J. (2014). A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. *PLoS Genetics*, 10(4).
- [359] Ogura, Y., Bonen, D. K., Inohara, N., Nicolae, D. L., Chen, F. F., Ramos, R., Britton, H., Moran, T., Karaliuskas, R., Duerr, R. H., Achkar, J. P., Brant, S. R., Bayless, T. M., Kirschner, B. S., Hanauer, S. B., Nuñez, G., and Cho, J. H. (2001). A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature*, 411(6837):603–606.
- [360] Okou, D. T., Mondal, K., Faubion, W. A., Kobrynski, L. J., Denson, L. A., Mulle, J. G., Ramachandran, D., Xiong, Y., Svingen, P., Patel, V., Bose, P., Waters, J. P., Prahalad, S., Cutler, D. J., Zwick, M. E., and Kugathasan, S. (2014). Exome Sequencing Identifies a Novel FOXP3 Mutation in a 2-Generation Family With Inflammatory Bowel Disease. *Journal of Pediatric Gastroenterology and Nutrition*, 58(5):561–568.
- [361] Oliver, G. R., Hart, S. N., and Klee, E. W. (2015). Bioinformatics for clinical next generation sequencing. *Clinical Chemistry*, 61(1):124–135.
- [362] Olivieri, A., Stazi, M. A., Mastroiacovo, P., Fazzini, C., Medda, E., Spagnolo, A., De Angelis, S., Grandolfo, M. E., Taruscio, D., Cordeddu, V., and Sorcini, M. (2002). A population-based study on the frequency of additional congenital malformations in

- infants with congenital hypothyroidism: Data from the Italian registry for congenital hypothyroidism (1991-1998). *Journal of Clinical Endocrinology and Metabolism*, 87(2):557–562.
- [363] Opitz, R., Hitz, M.-P., Vandernoot, I., Trubiroha, A., Abu-Khudir, R., Samuels, M., Désilets, V., Costagliola, S., Andelfinger, G., and Deladoëy, J. (2015). Functional zebrafish studies based on human genotyping point to netrin-1 as a link between aberrant cardiovascular development and thyroid dysgenesis. *Endocrinology*, 156(1):377–88.
- [364] O’Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., Bodily, P., Tian, L., Hakonarson, H., Johnson, W. E., Wei, Z., Wang, K., and Lyon, G. J. (2013). Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome medicine*, 5(3):28.
- [365] O’Roak, B. J., Deriziotis, P., Lee, C., Vives, L., Schwartz, J. J., Girirajan, S., Karakoc, E., Mackenzie, A. P., Ng, S. B., Baker, C., Rieder, M. J., Nickerson, D. a., Bernier, R., Fisher, S. E., Shendure, J., and Eichler, E. E. (2011). Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nature Genetics*, 43(6):585–9.
- [366] O’Roak, B. J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B. P., Levy, R., Ko, A., Lee, C., Smith, J. D., Turner, E. H., Stanaway, I. B., Vernot, B., Malig, M., Baker, C., Reilly, B., Akey, J. M., Borenstein, E., Rieder, M. J., Nickerson, D. a., Bernier, R., Shendure, J., and Eichler, E. E. (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*, 485(7397):246–50.
- [367] Ou, Z., Berg, J. S., Yonath, H., Enciso, V. B., Miller, D. T., Picker, J., Lenzi, T., Keegan, C. E., Sutton, V. R., Belmont, J., Chinault, a. C., Lupski, J. R., Cheung, S. W., Roeder, E., and Patel, A. (2008). Microduplications of 22q11.2 are frequently inherited and are associated with variable phenotypes. *Genetics in medicine: official journal of the American College of Medical Genetics*, 10(4):267–77.
- [368] Owaga, E., Hsieh, R.-H., Mugendi, B., Masuku, S., Shih, C.-K., and Chang, J.-S. (2015). Th17 Cells as Potential Probiotic Therapeutic Targets in Inflammatory Bowel Diseases. *International journal of molecular sciences*, 16(9):20841–58.
- [369] Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M. R., Zschocke, J., and Trajanoski, Z. (2014). A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in Bioinformatics*, 15(2):256–278.
- [370] Panagiotou, O. A., Ioannidis, J. P. A., Hirschhorn, J. N., Abecasis, G. R., Frayling, T. M., McCarthy, M. I., Lindgren, C. M., Beaty, T. H., Eriksson, N., Polychronakos, C., Kathirensan, S., Plenge, R. M., Spritz, R., Payami, H., Martin, E. R., Vance, J., Su, W. H., Chang, Y. S., Bei, J. X., Zeng, Y. X., Paré, G., Faraone, S. V., Neale, B., Anney, R. J., Traynor, B. J., Scherag, A., Hebebrand, J., Hinney, A., Froguel, P., Meyre, D., Chanock, S. J., and Kesheng, W. (2012). What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *International Journal of Epidemiology*, 41(1):273–286.

- [371] Panoutsopoulou, K., Tachmazidou, I., and Zeggini, E. (2013). In search of low-frequency and rare variants affecting complex traits. *Human Molecular Genetics*, 22(R1):16–21.
- [372] Park, K.-j., Park, H.-k. H.-d., Kim, Y.-j., Lee, K.-r., Park, J.-h. J.-h., Park, J.-h. J.-h., Park, H.-k. H.-d., Lee, S.-y., and Kim, J.-w. (2016). DUOX2 Mutations Are Frequently Associated With Congenital Hypothyroidism in the Korean Population. *Annals of laboratory medicine*, 36:145–153.
- [373] Park, S. M. and Chatterjee, V. K. K. (2005). Genetics of congenital hypothyroidism. *Journal of Medical Genetics*, 42(5):379–89.
- [374] Parkes, M., Barrett, J. C., Prescott, N. J., Tremelling, M., Anderson, C. A., Fisher, S. A., Roberts, R. G., Nimmo, E. R., Cummings, F. R., Soars, D., Drummond, H., Lees, C. W., Khawaja, S. A., Bagnall, R., Burke, D. A., Todhunter, C. E., Ahmad, T., Onnie, C. M., McArdle, W., Strachan, D., Bethel, G., Bryan, C., Lewis, C. M., Deloukas, P., Forbes, A., Sanderson, J., Jewell, D. P., Satsangi, J., Mansfield, J. C., Cardon, L., and Mathew, C. G. (2007). Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn’s disease susceptibility. *Nature Genetics*, 39(7):830–2.
- [375] Parkes, M., Cortes, A., van Heel, D. A., and Brown, M. A. (2013). Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nature Reviews Genetics*, 14(9):661–673.
- [376] Parks, J. S., Lin, M., Grosse, S. D., Hinton, C. F., Drummond-Borg, M., Borgfeld, L., and Sullivan, K. M. (2010). The impact of transient hypothyroidism on the increasing rate of congenital hypothyroidism in the United States. *Pediatrics*, 125 Suppl:S54–63.
- [377] Parlato, R., Rosica, A., Rodriguez-Mallon, A., Affuso, A., Postiglione, M. P., Arra, C., Mansouri, A., Kimura, S., Di Lauro, R., and De Felice, M. (2004). An integrated regulatory network controlling survival and migration in thyroid organogenesis. *Developmental Biology*, 276(2):464–475.
- [378] Passeri, E., Frigerio, M., De Filippis, T., Valaperta, R., Capelli, P., Costa, E., Fugazzola, L., Marelli, F., Porazzi, P., Arcidiacono, C., Carminati, M., Ambrosi, B., Persani, L., and Corbetta, S. (2011). Increased risk for non-autoimmune hypothyroidism in young patients with congenital heart defects. *Journal of Clinical Endocrinology and Metabolism*, 96(7):1115–1119.
- [379] Paul, D. S., Soranzo, N., and Beck, S. (2014). Functional interpretation of non-coding sequence variation: Concepts and challenges. *BioEssays*, 36(2):191–199.
- [380] Peeters, M., Nevens, H., Baert, F., Hiele, M., Vlietinck, R., Rutgeerts, P., and Collection, D. (1996). Adjusted Risk and Concordance in Clinical Characteristics. *Gastroenterology*, 111:597–603.
- [381] Pelak, K., Shianna, K. V., Ge, D., Maia, J. M., Zhu, M., Smith, J. P., Cirulli, E. T., Fellay, J., Dickson, S. P., Gumbs, C. E., Heinzen, E. L., Need, A. C., Ruzzo, E. K., Singh, A., Campbell, C. R., Hong, L. K., Lornsen, K. A., McKenzie, A. M.,

- Sobreira, N. L. M., Hoover-Fong, J. E., Milner, J. D., Ottman, R., Haynes, B. F., Goedert, J. J., and Goldstein, D. B. (2010). The characterization of twenty sequenced human genomes. *PLoS Genetics*, 6(9).
- [382] Peltonen, L. and McKusick, V. a. (2001). Dissecting Human Disease in the Postgenomic Era. *Science*, 291(February):1224–1229.
- [383] Perez-Pinera, P., Kocak, D. D., Vockley, C. M., Adler, A. F., Kabadi, A. M., Polstein, L. R., Thakore, P. I., Glass, K. A., Ousterout, D. G., Leong, K. W., Guilak, F., Crawford, G. E., Reddy, T. E., and Gersbach, C. A. (2013). RNA-guided gene activation by CRISPR-Cas9-based transcription factors. *Nature Methods*, 10(10):973–6.
- [384] Perry, R., Heinrichs, C., Bourdoux, P., Khoury, K., Dussault, J. H., Vassart, G., Vliet, G. U. Y. V. A. N., and Szo, O. I. S. (2002). Discordance of Monozygotic Twins for Thyroid Dysgenesis: Implications for Screening and for Molecular Pathophysiology. *The Journal of Clinical Endocrinology and Metabolism*, 87(August):4072–4077.
- [385] Persani, L. (2012). Congenital Hypothyroidism with Gland in situ is More Frequent than Previously Thought. *Frontiers in endocrinology*, 3(February):18.
- [386] Persani, L., Calebiro, D., Cordella, D., Weber, G., Gelmini, G., Libri, D., de Filippis, T., and Bonomi, M. (2010). Genetics and phenomics of hypothyroidism due to TSH resistance. *Molecular and cellular endocrinology*, 322(1-2):72–82.
- [387] Peterson, N., Guthery, S., Denson, L., Lee, J., Saeed, S., Prahalad, S., Biank, V., Ehlert, R., Tomer, G., Grand, R., Rudolph, C., and Kugathasan, S. (2008). Genetic variants in the autophagy pathway contribute to paediatric Crohn’s disease. *Gut*, 57(9):1336–7; author reply 1337.
- [388] Petrovic, S., Ju, X., Barone, S., Seidler, U., Alper, S. L., Lohi, H., Kere, J., and Soleimani, M. (2003). Identification of a basolateral Cl⁻ / HCO³⁻ exchanger specific to gastric parietal cells. *American journal of Physiology and Gastrointestinal Liver Physiology*, 0585:1093–1103.
- [389] Pfarr, N., Korsch, E., Kaspers, S., Herbst, A., Stach, A., Zimmer, C., and Pohlenz, J. (2006). Congenital hypothyroidism caused by new mutations in the thyroid oxidase 2 (THOX2) gene. *Clinical Endocrinology*, 65(6):810–5.
- [390] Pfeiffer, R. M., Gail, M. H., and Pee, D. (2009). On Combining Data From Genome-Wide Association Studies to Discover Disease-Associated SNPs. *Statistical Science*, 24(4):547–560.
- [391] Philippakis, A. A., Azzariti, D. R., Beltran, S., Brookes, A. J., Brownstein, C. A., Brudno, M., Brunner, H. G., Buske, O. J., Carey, K., Doll, C., Dumitriu, S., Dyke, S. O. M., den Dunnen, J. T., Firth, H. V., Gibbs, R. A., Girdea, M., Gonzalez, M., Haendel, M. A., Hamosh, A., Holm, I. A., Huang, L., Hurles, M. E., Hutton, B., Krier, J. B., Misyura, A., Mungall, C. J., Paschall, J., Paten, B., Robinson, P. N., Schiettecatte, F., Sobreira, N. L., Swaminathan, G. J., Taschner, P. E., Terry, S. F., Washington, N. L., Züchner, S., Boycott, K. M., and Rehm, H. L. (2015). The Matchmaker Exchange: A Platform for Rare Disease Gene Discovery. *Human Mutation*, 36(10):915–921.

- [392] Pidasheva, S., Trifari, S., Phillips, A., Hackney, J. A., Ma, Y., Smith, A., Sohn, S. J., Spits, H., Little, R. D., Behrens, T. W., Honigberg, L., Ghilardi, N., and Clark, H. F. (2011). Functional studies on the IBD susceptibility gene IL23R implicate reduced receptor function in the protective genetic variant R381Q. *PLoS one*, 6(10):e25038.
- [393] Pirinen, M., Donnelly, P., and Spencer, C. C. A. (2012). Including known covariates can reduce power to detect genetic effects in case-control studies. *Nature Genetics*, 44(8):848–51.
- [394] Pitceathly, R. D. S., Rahman, S., Wedatilake, Y., Polke, J. M., Cirak, S., Foley, a. R., Sailer, A., Hurler, M. E., Stalker, J., Hargreaves, I., Woodward, C. E., Sweeney, M. G., Muntoni, F., Houlden, H., Taanman, J.-W., and Hanna, M. G. (2013). NDUFA4 mutations underlie dysfunction of a cytochrome c oxidase subunit linked to human neurological disease. *Cell reports*, 3(6):1795–805.
- [395] Plenge, R. M., Padyukov, L., Remmers, E. F., Purcell, S., Lee, A. T., Karlson, E. W., Wolfe, F., Kastner, D. L., Alfredsson, L., Altshuler, D., Gregersen, P. K., Klareskog, L., and Rioux, J. D. (2005). Replication of putative candidate-gene associations with rheumatoid arthritis in >4,000 samples from North America and Sweden: association of susceptibility with PTPN22, CTLA4, and PADI4. *American Journal of Human Genetics*, 77(6):1044–1060.
- [396] Pn, R., Krawitz, P., Strategies, M. S., Robinson, P. N., Krawitz, P., and Mundlos, S. (2011). Strategies for exome and genome sequence data analysis in disease-gene discovery projects. *Clinical Genetics*, 80(2):127–32.
- [397] Polak, M., Sura-Trueba, S., Chauty, A., Szinnai, G., Carré, A., and Castanet, M. (2004). Molecular mechanisms of thyroid dysgenesis. *Hormone research*, 62 Suppl 3(suppl 3):14–21.
- [398] Porcu, E., Medici, M., Pistis, G., Volpato, C. B., Wilson, S. G., Cappola, A. R., Bos, S. D., Deelen, J., den Heijer, M., Freathy, R. M., Lahti, J., Liu, C., Lopez, L. M., Nolte, I. M., O’Connell, J. R., Tanaka, T., Trompet, S., Arnold, A., Bandinelli, S., Beekman, M., Böhringer, S., Brown, S. J., Buckley, B. M., Camaschella, C., de Craen, A. J. M., Davies, G., de Visser, M. C. H., Ford, I., Forsen, T., Frayling, T. M., Fugazzola, L., Gögele, M., Hattersley, A. T., Hermus, A. R., Hofman, A., Houwing-Duistermaat, J. J., Jensen, R. a., Kajantie, E., Kloppenburg, M., Lim, E. M., Masciullo, C., Mariotti, S., Minelli, C., Mitchell, B. D., Nagaraja, R., Netea-Maier, R. T., Palotie, A., Persani, L., Piras, M. G., Psaty, B. M., Rääkkönen, K., Richards, J. B., Rivadeneira, F., Sala, C., Sabra, M. M., Sattar, N., Shields, B. M., Soranzo, N., Starr, J. M., Stott, D. J., Sweep, F. C. G. J., Usala, G., van der Klauw, M. M., van Heemst, D., van Mullem, A., H Vermeulen, S., Visser, W. E., Walsh, J. P., Westendorp, R. G. J., Widen, E., Zhai, G., Cucca, F., Deary, I. J., Eriksson, J. G., Ferrucci, L., Fox, C. S., Jukema, J. W., Kiemeny, L. a., Pramstaller, P. P., Schlessinger, D., Shuldiner, A. R., Slagboom, E. P., Uitterlinden, A. G., Vaidya, B., Visser, T. J., Wolfenbuttel, B. H. R., Meulenbelt, I., Rotter, J. I., Spector, T. D., Hicks, A. a., Toniolo, D., Sanna, S., Peeters, R. P., and Naitza, S. (2013). A meta-analysis of thyroid-related traits reveals novel Loci and gender-specific differences in the regulation of thyroid function. *PLoS genetics*, 9(2):e1003266.

- [399] Prenzel, F. and Uhlig, H. H. (2009). Frequency of indeterminate colitis in children and adults with IBD - a metaanalysis. *Journal of Crohn's & colitis*, 3(4):277–81.
- [400] Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. a., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–9.
- [401] Pritchard, J. K. (2001). Are rare variants responsible for susceptibility to complex diseases? *American Journal of Human Genetics*, 69(1):124–137.
- [402] Project, G., Asia, E., Africa, S., Figs, S., and Tables, S. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 135(V):0–9.
- [403] Project, T. T., of the Exome Sequencing, H. W. G., National Heart, Lung, and Institute, B. (2014). Loss-of-Function Mutations in APOC3, Triglycerides, and Coronary Disease. *The New England Journal of Medicine*, 371:1–10.
- [404] Pruim, R. J., Welch, R. P., Sanna, S., Teslovich, T. M., Chines, P. S., Gliedt, T. P., Boehnke, M., Abecasis, G. R., Willer, C. J., and Frishman, D. (2011). LocusZoom: Regional visualization of genome-wide association scan results. *Bioinformatics*, 27(13):2336–2337.
- [405] Pugh, T. J., Delaney, a. D., Farnoud, N., Flibotte, S., Griffith, M., Li, H. I., Qian, H., Farinha, P., Gascoyne, R. D., and Marra, M. a. (2008). Impact of whole genome amplification on analysis of copy number variants. *Nucleic Acids Research*, 36(13):e80.
- [406] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. a. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., and Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3):559–75.
- [407] Purcell, S. M., Moran, J. L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., O'Dushlaine, C., Chambert, K., Bergen, S. E., Kähler, A., Duncan, L., Stahl, E., Genovese, G., Fernández, E., Collins, M. O., Komiyama, N. H., Choudhary, J. S., Magnusson, P. K. E., Banks, E., Shakir, K., Garimella, K., Fennell, T., DePristo, M., Grant, S. G. N., Haggarty, S. J., Gabriel, S., Scolnick, E. M., Lander, E. S., Hultman, C. M., Sullivan, P. F., McCarroll, S. A., and Sklar, P. (2014). A polygenic burden of rare disruptive mutations in schizophrenia. *Nature*, 506(7487):185–90.
- [408] Rabbani, B., Tekin, M., and Mahdieh, N. (2014). The promise of whole-exome sequencing in medical genetics. *Journal of human genetics*, 59(1):5–15.
- [409] Rabbiosi, S., Vigone, M. C., Cortinovis, F., Zamproni, I., Fugazzola, L., Persani, L., Corbetta, C., Chiumello, G., and Weber, G. (2013). Congenital hypothyroidism with eutopic thyroid gland: analysis of clinical and biochemical features at diagnosis and after re-evaluation. *The Journal of Clinical Endocrinology and Metabolism*, 98(4):1395–402.
- [410] Rae, J., Noack, D., Heyworth, P. G., Ellis, B. A., Curnutte, J. T., and Cross, A. R. (2000). Molecular analysis of 9 new families with chronic granulomatous disease caused by mutations in CYBA, the gene encoding p22(phox). *Blood*, 96(3):1106–12.

- [411] Ramasundara, M., Leach, S. T., Lemberg, D. A., and Day, A. S. (2009). Defensins and inflammation: The role of defensins in inflammatory bowel disease. *Journal of Gastroenterology and Hepatology (Australia)*, 24(2):202–208.
- [412] Ramos, H. E., Carre, A., Chevrier, L., Szinnai, G., Tron, E., Cerqueira, T. L. O., Leger, J., Cabrol, S., Puel, O., Queindec, C., De Roux, N., Guillot, L., Castanet, M., and Polak, M. (2014). Extreme phenotypic variability of thyroid dysgenesis in six new cases of congenital hypothyroidism due to PAX8 gene loss-of-function mutations. *European Journal of Endocrinology*, 171(4):499–507.
- [413] Ramu, A., Noordam, M. J., Schwartz, R. S., Wuster, A., Hurles, M. E., Cartwright, R. A., and Conrad, D. F. (2013). Indel and point mutation discovery and phasing. *Nature Methods*, 10(10):3–7.
- [414] Rastogi, M. V. and LaFranchi, S. H. (2010). Congenital hypothyroidism. *Orphanet Journal of Rare Diseases*, 5:17.
- [415] Rauch, A., Wiczorek, D., Graf, E., Wieland, T., Endeke, S., Schwarzmayr, T., Albrecht, B., Bartholdi, D., Beygo, J., Di Donato, N., Dufke, A., Cremer, K., Hempel, M., Horn, D., Hoyer, J., Joset, P., Röpke, A., Moog, U., Riess, A., Thiel, C. T., Tzschach, A., Wiesener, A., Wohlleber, E., Zweier, C., Ekici, A. B., Zink, A. M., Rump, A., Meisinger, C., Grallert, H., Sticht, H., Schenck, A., Engels, H., Rappold, G., Schröck, E., Wieacker, P., Riess, O., Meitinger, T., Reis, A., and Strom, T. M. (2012). Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet*, 380(9854):1674–82.
- [416] Reamon-Buettner, S. M. and Borlak, J. (2010). NKX2-5: An update on this hypermutable homeodomain protein and its role in human congenital heart disease (CHD). *Human Mutation*, 31(11):1185–1194.
- [417] Retterer, K., Juusola, J., Cho, M. T., Vitazka, P., Millan, F., Gibellini, F., Vertino-Bell, A., Smaoui, N., Neidich, J., Monaghan, K. G., McKnight, D., Bai, R., Suchy, S., Friedman, B., Tahiliani, J., Pineda-Alvarez, D., Richard, G., Brandt, T., Haverfield, E., Chung, W. K., and Bale, S. (2015). Clinical application of whole-exome sequencing across clinical indications. *Genetics in medicine: official journal of the American College of Medical Genetics*, 18(March):1–9.
- [418] Reuter, J. A., Spacek, D. V., and Snyder, M. P. (2015). High-Throughput Sequencing Technologies. *Molecular Cell*, 58(4):586–597.
- [419] Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., and Rehm, H. L. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*, 17(5):405–424.
- [420] Rigoli, L. and Caruso, R. A. (2014). Inflammatory bowel disease in pediatric and adolescent patients: A biomolecular and histopathological review. *World journal of gastroenterology: WJG*, 20(30):10262–10278.

- [421] Riise, R., Andréasson, S., Borgaström, M. K., Wright, A. F., Tommerup, N., Rosenberg, T., and Tornqvist, K. (1997). Intrafamilial variation of the phenotype in Bardet-Biedl syndrome. *The British journal of ophthalmology*, 81(5):378–85.
- [422] Ripatti, S., Tikkanen, E., Orho-Melander, M., Havulinna, A. S., Silander, K., Sharma, A., Guiducci, C., Perola, M., Jula, A., Sinisalo, J., Lokki, M.-L., Nieminen, M. S., Melander, O., Salomaa, V., Peltonen, L., and Kathiresan, S. (2010). A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. *The Lancet*, 376(9750):1393–1400.
- [423] Ris-Stalpers, C. and Bikker, H. (2010). Genetics and phenomics of hypothyroidism and goiter due to TPO mutations. *Molecular and cellular endocrinology*, 322(1-2):38–43.
- [424] Ritchie, G. R. S., Dunham, I., Zeggini, E., and Flicek, P. (2014). Functional annotation of noncoding sequence variants. *Nature Methods*, 11(3):294–6.
- [425] Roadmap Epigenomics Consortium (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330.
- [426] Roak, B. J. O., Vives, L., Fu, W., Egertson, J. D., Stanaway, I. B., Phelps, I. G., Carvill, G., Kumar, A., Lee, C., Ankenman, K., Munson, J., Hiatt, J. B., Turner, E. H., Levy, R., Day, D. R. O., Krumm, N., Coe, B. P., Martin, B. K., Borenstein, E., Nickerson, D. A., Mefford, H. C., Doherty, D., Akey, J. M., Bernier, R., Eichler, E. E., and Shendure, J. (2012). Multiplex Targeted Sequencing Identifies Recurrently Mutated Genes in Autism Spectrum Disorders. *Science*, 23(December).
- [427] Robinson, A. N. N. (2016). Genomics – the future of healthcare and medicine. *Prescriber*, 2003(April 2003):51–55.
- [428] Roos, D. and de Boer, M. (2014). Molecular diagnosis of chronic granulomatous disease. *Clinical and Experimental Immunology*, 175(2):139–149.
- [429] Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., and Feldman, M. W. (2002). Genetic Structure of Human Populations. *Science*, 298(December):2381–2385.
- [430] Sabbagh, A., Pasmant, E., Laurendeau, I., Parfait, B., Barbarot, S., Guillot, B., Combemale, P., Ferkal, S., Vidaud, M., Aubourg, P., Vidaud, D., and Wolkenstein, P. (2009). Unravelling the genetic basis of variable clinical expression in neurofibromatosis 1. *Human Molecular Genetics*, 18(15):2768–2778.
- [431] Samocha, K. E., Robinson, E. B., Sanders, S. J., Stevens, C., Sabo, A., McGrath, L. M., Kosmicki, J. A., Rehnström, K., Mallick, S., Kirby, A., Wall, D. P., MacArthur, D. G., Gabriel, S. B., DePristo, M., Purcell, S. M., Palotie, A., Boerwinkle, E., Buxbaum, J. D., Cook, E. H., Gibbs, R. A., Schellenberg, G. D., Sutcliffe, J. S., Devlin, B., Roeder, K., Neale, B. M., and Daly, M. J. (2014). A framework for the interpretation of de novo mutation in human disease. *Nature Genetics*, 46(9):944–50.

- [432] Sanders, S. J., Murtha, M. T., Gupta, A. R., Murdoch, J. D., Raubeson, M. J., Willsey, a. J., Ercan-Sencicek, a. G., DiLullo, N. M., Parikshak, N. N., Stein, J. L., Walker, M. F., Ober, G. T., Teran, N. a., Song, Y., El-Fishawy, P., Murtha, R. C., Choi, M., Overton, J. D., Bjornson, R. D., Carriero, N. J., Meyer, K. a., Bilguvar, K., Mane, S. M., Sestan, N., Lifton, R. P., Günel, M., Roeder, K., Geschwind, D. H., Devlin, B., and State, M. W. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature*, 485(7397):237–41.
- [433] Sawczenko, A. and Sandhu, B. K. (2006). Presenting features of inflammatory bowel disease in Great Britain and Ireland. *Arch Dis Child*, 88:124–148.
- [434] Schäffer, A. a. (2013). Digenic inheritance in medical genetics. *Journal of Medical Genetics*, 50(10):641–52.
- [435] Schäppi, M. G., Smith, V. V., Goldblatt, D., Lindley, K. J., and Milla, P. J. (2001). Colitis in chronic granulomatous disease. *Arch Dis Child*, 84:147–151.
- [436] Schloss, J. a. (2008). How to get genomes at one ten-thousandth the cost. *Nat. Biotechnol.*, 26(10):1113–1115.
- [437] Schork, N. J. (2015). Personalized medicine: Time for one-person trials. *Nature*, 520(7549):609–611.
- [438] Schott, J., Benson, D. W., Basson, C. T., Pease, W., Silberbach, G. M., Moak, J. P., Maron, B. J., Seidman, C. E., and Seidman, J. G. (1998). Congenital Heart Disease Caused by Mutations in the Transcription Factor NKX2-5. *Science*, 281(5373):108–111.
- [439] Scinicariello, F., Murray, H. E., Smith, L., Wilbur, S., and Fowler, B. A. (2005). Genetic factors that might lead to different responses in individuals exposed to perchlorate. *Environmental Health Perspectives*, 113(11):1479–1484.
- [440] Seah, Y. F. S., El Farran, C. A., Warriar, T., Xu, J., and Loh, Y. H. (2015). Induced pluripotency and gene editing in disease modelling: Perspectives and challenges. *International Journal of Molecular Sciences*, 16(12):28614–28634.
- [441] Senée, V., Chelala, C., Duchatelet, S., Feng, D., Blanc, H., Cossec, J.-C., Charon, C., Nicolino, M., Boileau, P., Cavener, D. R., Bougnères, P., Taha, D., and Julier, C. (2006). Mutations in GLIS3 are responsible for a rare syndrome with neonatal diabetes mellitus and congenital hypothyroidism. *Nature Genetics*, 38(6):682–687.
- [442] Shaikh, T. H., Gai, X., Perin, J. C., Glessner, J. T., Xie, H., Murphy, K., O’Hara, R., Casalunovo, T., Conlin, L. K., D’Arcy, M., Frackelton, E. C., Geiger, E. a., Haldeman-Englert, C., Imielinski, M., Kim, C. E., Medne, L., Annaiah, K., Bradfield, J. P., Dabaghyan, E., Eckert, A., Onyiah, C. C., Ostapenko, S., Otieno, F. G., Santa, E., Shaner, J. L., Skraban, R., Smith, R. M., Elia, J., Goldmuntz, E., Spinner, N. B., Zackai, E. H., Chiavacci, R. M., Grundmeier, R., Rappaport, E. F., Grant, S. F. a., White, P. S., and Hakonarson, H. (2009). High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications. *Genome research*, 19(9):1682–90.

- [443] Sham, P. C. and Purcell, S. M. (2014). Statistical power and significance testing in large-scale genetic studies. *Nature Reviews Genetics*, 15(5):335–346.
- [444] Shaulian, E. and Karin, M. (2002). AP-1 as a regulator of cell life and death. *Nat Cell Biol*, 4(5):E131–6.
- [445] Shawky, R. M. (2014). Reduced penetrance in human inherited disease. *Egyptian Journal of Medical Human Genetics*, 15(2):103–111.
- [446] Sherry, S. T., Ward, M., and Sirotkin, K. (1999). dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Research*, 9(8):677–679.
- [447] Shim, J. O. and Seo, J. K. (2014). Very early-onset inflammatory bowel disease (IBD) in infancy is a different disease entity from adult-onset IBD; one form of interleukin-10 receptor mutations. *Journal of human genetics*, 59(6):337–41.
- [448] Shyr, C., Tarailo-Graovac, M., Gottlieb, M., Lee, J. J. Y., van Karnebeek, C., and Wasserman, W. W. (2014). FLAGS, frequently mutated genes in public exomes. *BMC medical genomics*, 7:64.
- [449] Singh, T., Kurki, M. I., Curtis, D., Purcell, S. M., Crooks, L., McRae, J., Suvisaari, J., Chheda, H., Blackwood, D., Breen, G., Pietilainen, O., Gerety, S. S., Ayub, M., Blyth, M., Cole, T., Collier, D., Coomber, E. L., Craddock, N., Daly, M. J., Danesh, J., DiForti, M., Foster, A., Freimer, N. B., Geschwind, D., Johnstone, M., Joss, S., Kirov, G., Korkko, J., Kuismin, O., Holmans, P., Hultman, C. M., Iyegbe, C., Lonnqvist, J., Mannikko, M., McCarroll, S. A., McGuffin, P., McIntosh, A. M., McQuillin, A., Moilanen, J. S., Moore, C., Murray, R. M., Newbury-Ecob, R., Ouwehand, W., Paunio, T., Prigmore, E., Rees, E., Roberts, D., Sambrook, J., Sklar, P., St. Clair, D., Veijola, J., Walters, J. T. R., Williams, H., Sullivan, P. F., Hurles, M. E., O'Donovan, M. C., Palotie, A., Owen, M. J., and Barrett, J. C. (2016). Rare loss-of-function variants in KMT2F are associated with schizophrenia and developmental disorders. *bioRxiv*, 19(4).
- [450] Singleton, A. B. (2011). Exome sequencing: a transformative technology. *Lancet neurology*, 10(10):942–6.
- [451] Siva, N. (2015). UK gears up to decode 100 000 genomes from NHS patients. *The Lancet*, 385(9963):103–104.
- [452] Smith, G. D., Ebrahim, S., Lewis, S., Hansell, A. L., Palmer, L. J., and Burton, P. R. (2005). Genetic epidemiology and public health: Hope, hype, and future prospects. *Lancet*, 366(9495):1484–1498.
- [453] Smoller, J. W. (2013). Identification of risk loci with shared effects on five major psychiatric disorders: A genome-wide analysis. *The Lancet*, 381(9875):1371–1379.
- [454] Sobreira, N., Schiettecatte, F., Boehm, C., Valle, D., and Hamosh, A. (2015). New tools for mendelian disease gene identification: PhenoDB variant analysis module; and genematcher, a web-based tool for linking investigators with an interest in the same gene. *Human Mutation*, 36(4):425–431.

- [455] Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M., and Smoller, J. W. (2013). Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics*, 14(7):483–95.
- [456] Spain, S. L. and Barrett, J. C. (2015). Strategies for fine-mapping complex traits. *Human Molecular Genetics*, 24(R1):R111–R119.
- [457] Speckmann, C., Lehmborg, K., Albert, M. H., Damgaard, R. B., Fritsch, M., Gyrd-Hansen, M., Rensing-Ehl, A., Vraetz, T., Grimbacher, B., Salzer, U., Fuchs, I., Ufheil, H., Belohradsky, B. H., Hassan, A., Cale, C. M., Elawad, M., Strahm, B., Schibli, S., Lauten, M., Kohl, M., Meerpohl, J. J., Rodeck, B., Kolb, R., Eberl, W., Soerensen, J., von Bernuth, H., Lorenz, M., Schwarz, K., zur Stadt, U., and Ehl, S. (2013). X-linked inhibitor of apoptosis (XIAP) deficiency: The spectrum of presenting manifestations beyond hemophagocytic lymphohistiocytosis. *Clinical Immunology*, 149(1):133–141.
- [458] Spielman, R. S., McGinnis, R. E., and Ewenst, W. J. (1993). Transmission Test for Linkage Disequilibrium: The Insulin Gene Region and Insulin-dependent Diabetes Mellitus (IDDM). *American Journal of Human Genetics*, 52:506–516.
- [459] Spitzweg, C. and Morris, J. C. (2010). Genetics and phenomics of hypothyroidism and goiter due to NIS mutations. *Molecular and cellular endocrinology*, 322(1-2):56–63.
- [460] Sriprapradang, C., Tenenbaum-Rakover, Y., Weiss, M., Barkoff, M. S., Admoni, O., Kawthar, D., Caltabiano, G., Pardo, L., Dumitrescu, A. M., and Refetoff, S. (2011). The coexistence of a novel inactivating mutant thyrotropin receptor allele with two thyroid peroxidase mutations: a genotype-phenotype correlation. *The Journal of Clinical Endocrinology and Metabolism*, 96(6):E1001–6.
- [461] St George-Hyslop, P., Tanzi, R., Polinsky, R., Haines, J., Nee, L., Watkins, P., Myers, R., Feldman, R., Pollen, D., Drachman, D., and Al., E. (1987). The genetic defect causing familial Alzheimer’s disease maps on chromosome 21. *Science*, 235(4791):885–890.
- [462] Steinthorsdottir, V., Thorleifsson, G., Sulem, P., Helgason, H., Grarup, N., Sigurdsson, a., Helgadottir, H. T., Johannsdottir, H., Magnusson, O. T., Gudjonsson, S. a., Justesen, J. M., Harder, M. N., Jorgensen, M. E., Christensen, C., Brandslund, I., Sandbaek, a., Lauritzen, T., Vestergaard, H., Linneberg, a., Jorgensen, T., Hansen, T., Daneshpour, M. S., Fallah, M. S., Hreidarsson, a. B., Sigurdsson, G., Azizi, F., Benediktsson, R., Masson, G., Helgason, a., Kong, a., Gudbjartsson, D. F., Pedersen, O., Thorsteinsdottir, U., and Stefansson, K. (2014). Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nat Genet*, 46(3):294–298.
- [463] Sternecker, J. L., Reinhardt, P., and Schöler, H. R. (2014). Investigating human disease using stem cell models. *Nature Reviews Genetics*, 15(9):625–639.
- [464] Stitzel, N. O., Kiezun, A., and Sunyaev, S. (2011). Computational and statistical approaches to analyzing variants identified by exome sequencing. *Genome biology*, 12(9):227.

- [465] Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., and Collins, R. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Medicine*, 12(3):1–10.
- [466] Sumegi, J., Huang, D., Lanyi, a., Davis, J. D., Seemayer, T. a., Maeda, a., Klein, G., Seri, M., Wakiguchi, H., Purtilo, D. T., and Gross, T. G. (2000). Correlation of mutations of the SH2D1A gene and epstein-barr virus infection with clinical phenotype and outcome in X-linked lymphoproliferative disease. *Blood*, 96(9):3118–3125.
- [467] Surakka, I., Horikoshi, M., Magi, R., Sarin, A. P., Mahajan, A., Lagou, V., Marullo, L., Ferreira, T., Miraglio, B., Timonen, S., Kettunen, J., Pirinen, M., Karjalainen, J., Thorleifsson, G., Hagg, S., Hottenga, J. J., Isaacs, A., Ladenvall, C., Beekman, M., Esko, T., Ried, J. S., Nelson, C. P., Willenborg, C., Gustafsson, S., Westra, H. J., Blades, M., de Craen, A. J., de Geus, E. J., Deelen, J., Grallert, H., Hamsten, A., Havulinna, A. S., Hengstenberg, C., Houwing-Duistermaat, J. J., Hypponen, E., Karssen, L. C., Lehtimäki, T., Lyssenko, V., Magnusson, P. K., Mihailov, E., Müller-Nurasyid, M., Mpindi, J. P., Pedersen, N. L., Penninx, B. W., Perola, M., Pers, T. H., Peters, A., Rung, J., Smit, J. H., Steinthorsdóttir, V., Tobin, M. D., Tsernikova, N., van Leeuwen, E. M., Viikari, J. S., Willems, S. M., Willemsen, G., Schunkert, H., Erdmann, J., Samani, N. J., Kaprio, J., Lind, L., Gieger, C., Metspalu, A., Slagboom, P. E., Groop, L., van Duijn, C. M., Eriksson, J. G., Jula, A., Salomaa, V., Boomsma, D. I., Power, C., Raitakari, O. T., Ingelsson, E., Jarvelin, M. R., Thorsteinsdóttir, U., Franke, L., Ikonen, E., Kallioniemi, O., Pietiäinen, V., Lindgren, C. M., Stefansson, K., Palotie, A., McCarthy, M. I., Morris, A. P., Prokopenko, I., and Ripatti, S. f. t. E. C. (2015). The impact of low-frequency and rare variants on lipid levels. *Nat Genet*, 47(6):589–597.
- [468] Syvänen, A.-C. (2005). Toward genome-wide SNP genotyping. *Nature Genetics*, 37(June):S5–S10.
- [469] Takata, A., Xu, B., Ionita-Laza, I., Roos, J. L., Gogos, J. A., and Karayiorgou, M. (2014). Loss-of-Function Variants in Schizophrenia Risk and SETD1A as a Candidate Susceptibility Gene. *Neuron*, 82(4):773–780.
- [470] Talmud, P. J., Shah, S., Whittall, R., Futema, M., Howard, P., Cooper, J. a., Harrison, S. C., Li, K., Drenos, F., Karpe, F., Neil, H. A. W., Descamps, O. S., Langenberg, C., Lench, N., Kivimäki, M., Whittaker, J., Hingorani, A. D., Kumari, M., and Humphries, S. E. (2013). Use of low-density lipoprotein cholesterol gene score to distinguish patients with polygenic and monogenic familial hypercholesterolaemia: a case-control study. *Lancet*, 381(9874):1293–301.
- [471] Targovnik, H. M., Citterio, C. E., and Rivolta, C. M. (2011). Thyroglobulin Gene Mutations in Congenital Hypothyroidism. *Hormone Research in Paediatrics*, 75:311–321.
- [472] Targovnik, H. M., Esperante, S. a., and Rivolta, C. M. (2010). Genetics and phenomics of hypothyroidism and goiter due to thyroglobulin mutations. *Molecular and cellular endocrinology*, 322(1-2):44–55.

- [473] Tattini, L., D'Aurizio, R., and Magi, A. (2015). Detection of Genomic Structural Variants from Next-Generation Sequencing Data. *Frontiers in bioengineering and biotechnology*, 3(June):92.
- [474] Taylor, P. N., Porcu, E., Chew, S., Campbell, P. J., Traglia, M., Brown, S. J., Mullin, B. H., Shihab, H. a., Min, J., Walter, K., Memari, Y., Huang, J., Barnes, M. R., Beilby, J. P., Charoen, P., Danecek, P., Dudbridge, F., Forgetta, V., Greenwood, C., Grundberg, E., Johnson, A. D., Hui, J., Lim, E. M., McCarthy, S., Muddyman, D., Panicker, V., Perry, J. R. B., Bell, J. T., Yuan, W., Relton, C., Gaunt, T., Schlessinger, D., Abecasis, G., Cucca, F., Surdulescu, G. L., Woltersdorf, W., Zeggini, E., Zheng, H.-F., Toniolo, D., Dayan, C. M., Naitza, S., Walsh, J. P., Spector, T., Davey Smith, G., Durbin, R., Richards, J. B., Sanna, S., Soranzo, N., Timpson, N. J., and Wilson, S. G. (2015). Whole-genome sequence-based analysis of thyroid function. *Nature communications*, 6(May):5681.
- [475] Tenenbaum-Rakover, Y., Grasberger, H., Mamanasiri, S., Ringkananont, U., Montanelli, L., Barkoff, M. S., Dahood, A. M. H., and Refetoff, S. (2009). Loss-of-function mutations in the thyrotropin receptor gene as a major determinant of hyperthyrotropinemia in a consanguineous community. *Journal of Clinical Endocrinology and Metabolism*, 94(5):1706–1712.
- [476] Tennessen, J. a., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., Kang, H. M., Jordan, D., Leal, S. M., Gabriel, S., Rieder, M. J., Abecasis, G., Altshuler, D., Nickerson, D. a., Boerwinkle, E., Sunyaev, S., Bustamante, C. D., Bamshad, M. J., and Akey, J. M. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science (New York, N.Y.)*, 337(6090):64–9.
- [477] Teo, S. M., Pawitan, Y., Ku, C. S., Chia, K. S., and Salim, A. (2012). Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics*, 28(21):2711–2718.
- [478] The Encode Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.
- [479] The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*, 437(October):1299–1320.
- [480] The International Schizophrenia Consortium (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460(9):748–752.
- [481] The International SNP Map Working Group (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409(6822):928–933.
- [482] Thompson, E. A. (2013). Identity by descent: Variation in meiosis, across genomes, and in populations. *Genetics*, 194(2):301–326.
- [483] Thornhill, P., Bassett, D., Lochmüller, H., Bushby, K., and Straub, V. (2008). Developmental defects in a zebrafish model for muscular dystrophies associated with the loss of fukutin-related protein (FKRP). *Brain*, 131(6):1551–1561.

- [484] Thorvaldsdóttir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, 14(2):178–92.
- [485] Thorwarth, a., Mueller, I., Biebermann, H., Ropers, H. H., Grueters, a., Krude, H., and Ullmann, R. (2010). Screening chromosomal aberrations by array comparative genomic hybridization in 80 patients with congenital hypothyroidism and thyroid dysgenesis. *The Journal of Clinical Endocrinology and Metabolism*, 95(7):3446–52.
- [486] Thorwarth, A., Schnittert-Hübener, S., Schruppf, P., Müller, I., Jyrch, S., Dame, C., Biebermann, H., Kleinau, G., Katchanov, J., Schuelke, M., Ebert, G., Steininger, A., Bönnemann, C., Brockmann, K., Christen, H.-J., Crock, P., DeZegher, F., Griese, M., Hewitt, J., Ivarsson, S., Hübner, C., Kapelari, K., Plecko, B., Rating, D., Stoeva, I., Ropers, H.-H., Grüters, A., Ullmann, R., and Krude, H. (2014). Comprehensive genotyping and clinical characterisation reveal 27 novel NKX2-1 mutations and expand the phenotypic spectrum. *Journal of Medical Genetics*, 51(6):375–87.
- [487] Tipney, H., Painter, J. L., Shen, J., Nicoletti, P., Shen, Y., Floratos, A., Sham, P. C., Li, M. J., Wang, J., Cardon, L. R., Whittaker, J. C., Sanseau, P., and Nelson, M. R. (2015). The support of human genetic evidence for approved drug indications. *Nature Genetics*, 47(8):1–7.
- [488] Todd, J. a., Walker, N. M., Cooper, J. D., Smyth, D. J., Downes, K., Plagnol, V., Bailey, R., Nejentsev, S., Field, S. F., Payne, F., Lowe, C. E., Szeszko, J. S., Hafler, J. P., Zeitels, L., Yang, J. H. M., Vella, A., Nutland, S., Stevens, H. E., Schuilenburg, H., Coleman, G., Maisuria, M., Meadows, W., Smink, L. J., Healy, B., Burren, O. S., Lam, A. a. C., Ovington, N. R., Allen, J., Adlem, E., Leung, H.-T., Wallace, C., Howson, J. M. M., Guja, C., Ionescu-Tîrgoviște, C., Simmonds, M. J., Heward, J. M., Gough, S. C. L., Dunger, D. B., Wicker, L. S., and Clayton, D. G. (2007). Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nature Genetics*, 39(7):857–64.
- [489] Torres-ruiz, R. and Rodriguez-perales, S. (2016). CRISPR-Cas9 technology: applications and human disease modelling. *Briefings in Functional Genomics*, pages 1–9.
- [490] Trojanowicz, B., Brodauf, L., Sekulla, C., Lorenz, K., Finke, R., Dralle, H., and Hoang-vu, C. (2009). The role of AUF1 in thyroid carcinoma progression. *Endocrine-related Cancer*, 16:857–871.
- [491] Trojanowicz, B., Dralle, H., and Hoang-vu, C. (2011). AUF1 and HuR: possible implications of mRNA stability in thyroid function and disorders. *Thyroid Research*, 4(Suppl 1):S5.
- [492] Tsui, L., Buchwald, M., Barker, D., Braman, J., Knowlton, R., Schumm, H., Eiberg, J., Mohr, D., Kennedy, N., and Plavsic (1985). Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker. *Science*, 230:1054—1057.
- [493] Tsuma, Y., Imamura, T., Ichise, E., Sakamoto, K., Ouchi, K., Osone, S., Ishida, H., Wada, T., and Hosoi, H. (2015). Successful treatment of idiopathic colitis related

- to XIAP deficiency with allo-HSCT using reduced-intensity conditioning. *Pediatric Transplantation*, 19(1):E25–E28.
- [494] Uccellatore, F., Sava, L., Giuffrida, D., Fazio, T., Calaciura, F., Regalbuto, C., Vigneri, R., Costituzionale, P., Catania, U., and Garibaldi, O. (1990). Cytogenetic analysis in congenital hypothyroidism. *Journal of Clinical Investigation*, 13:605–607.
- [495] Uhlig, H. H. (2013). Monogenic diseases associated with intestinal inflammation: implications for the understanding of inflammatory bowel disease. *Gut*, 62:1795–1805.
- [496] Uhlig, H. H., Schwerd, T., Koletzko, S., Shah, N., Kammermeier, J., Elkadri, A., Ouahed, J., Wilson, D. C., Travis, S. P., Turner, D., Klein, C., Snapper, S. B., and Muise, A. M. (2014). The Diagnostic Approach to Monogenic Very Early Onset Inflammatory Bowel Disease. *Gastroenterology*, 147(5):990–1007.e3.
- [497] Uzel, G., Orange, J. S., Poliak, N., Marciano, B. E., Heller, T., and Holland, S. M. (2010). Complications of tumor necrosis factor- α blockade in chronic granulomatous disease-related colitis. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America*, 51(12):1429–34.
- [498] van Dongen, J., Slagboom, P. E., Draisma, H. H. M., Martin, N. G., and Boomsma, D. I. (2012). The continuing value of twin studies in the omics era. *Nature Publishing Group*, 13(9):640–653.
- [499] van Engelen, K., Mommersteeg, M. T. M., Baars, M. J. H., Lam, J., Ilgun, A., van Trotsenburg, a. S. P., Smets, A. M. J. B., Christoffels, V. M., Mulder, B. J. M., and Postma, A. V. (2012). The ambiguous role of NKX2-5 mutations in thyroid dysgenesis. *PloS one*, 7(12):e52685.
- [500] Van Limbergen, J., Russell, R. K., Drummond, H. E., Aldhous, M. C., Round, N. K., Nimmo, E. R., Smith, L., Gillett, P. M., McGrogan, P., Weaver, L. T., Bisset, W. M., Mahdi, G., Arnott, I. D., Satsangi, J., and Wilson, D. C. (2008). Definition of Phenotypic Characteristics of Childhood-Onset Inflammatory Bowel Disease. *Gastroenterology*, 135(4):1114–1122.
- [501] Vansteelandt, S., Goetgeluk, S., Lutz, S., Waldman, I., Lyon, H., Schadt, E. E., Weiss, S. T., and Lange, C. (2009). On the adjustment for covariates in genetic association analysis: A novel, simple principle to infer direct causal effects. *Genetic Epidemiology*, 33(5):394–405.
- [502] Vigone, M. C., Fugazzola, L., Zamproni, I., Passoni, A., Di Candia, S., Chiumello, G., Persani, L., and Weber, G. (2005). Persistent mild hypothyroidism associated with novel sequence variants of the DUOX2 gene in two siblings. *Human mutation*, 26(4):395.
- [503] Vissers, L. E. L. M., de Ligt, J., Gilissen, C., Janssen, I., Steehouwer, M., de Vries, P., van Lier, B., Arts, P., Wieskamp, N., del Rosario, M., van Bon, B. W. M., Hoischen, A., de Vries, B. B. a., Brunner, H. G., and Veltman, J. a. (2010). A de novo paradigm for mental retardation. *Nature Genetics*, 42(12):1109–1112.
- [504] Vitelli, F., Huynh, T., and Baldini, A. (2009). Gain of function of Tbx1 affects pharyngeal and heart development in the mouse. *Genesis*, 47(3):188–195.

- [505] Vojta, A., Dobrinić, P., Tadić, V., Bočkor, L., Korać, P., Julg, B., Klasić, M., and Zoldoš, V. (2016). Repurposing the CRISPR-Cas9 system for targeted DNA methylation. *Nucleic Acids Research*, (24):1–14.
- [506] Wada, T., Kanegane, H., Ohta, K., Katoh, F., Imamura, T., Nakazawa, Y., Miyashita, R., Hara, J., Hamamoto, K., Yang, X., Filipovich, A. H., Marsh, R. A., and Yachie, A. (2014). Sustained elevation of serum interleukin-18 and its association with hemophagocytic lymphohistiocytosis in XIAP deficiency. *Cytokine*, 65(1):74–78.
- [507] Walter, K., Min, J. L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J. R. B., Xu, C., Futema, M., Lawson, D., Iotchkova, V., Schiffels, S., Hendricks, A. E., Danecek, P., Li, R., Floyd, J., Wain, L. V., Barroso, I., Humphries, S. E., Hurles, M. E., Zeggini, E., Barrett, J. C., Plagnol, V., Brent Richards, J., Greenwood, C. M. T., Timpson, N. J., Durbin, R., Soranzo, N., Bala, S., Clapham, P., Coates, G., Cox, T., Daly, A., Danecek, P., Du, Y., Durbin, R., Edkins, S., Ellis, P., Flicek, P., Guo, X., Guo, X., Huang, L., Jackson, D. K., Joyce, C., Keane, T., Kolb-Kokocinski, A., Langford, C., Li, Y., Liang, J., Lin, H., Liu, R., Maslen, J., McCarthy, S., Muddyman, D., Quail, M. A., Stalker, J., Sun, J., Tian, J., Wang, G., Wang, J., Wang, Y., Wong, K., Zhang, P., Barroso, I., Birney, E., Bousted, C., Chen, L., Clement, G., Cocca, M., Danecek, P., Davey Smith, G., Day, I. N. M., Day-Williams, A., Down, T., Dunham, I., Durbin, R., Evans, D. M., Gaunt, T. R., Geijs, M., Greenwood, C. M. T., Hart, D., Hendricks, A. E., Howie, B., Huang, J., Hubbard, T., Hysi, P., Iotchkova, V., Jamshidi, Y., Karczewski, K. J., Kemp, J. P., Lachance, G., Lawson, D., Lek, M., Lopes, M., MacArthur, D. G., Marchini, J., Mangino, M., Mathieson, I., McCarthy, S., Memari, Y., Metrustry, S., Min, J. L., Moayyeri, A., Muddyman, D., Northstone, K., Panoutsopoulou, K., Paternoster, L., Perry, J. R. B., Quayle, L., Brent Richards, J., Ring, S., Ritchie, G. R. S., Schiffels, S., Shihab, H. A., Shin, S.-Y., Small, K. S., Soler Artigas, M., Soranzo, N., Southam, L., Spector, T. D., St Pourcain, B., Surdulescu, G., Tachmazidou, I., Timpson, N. J., Tobin, M. D., Valdes, A. M., Visscher, P. M., Wain, L. V., Walter, K., Ward, K., Wilson, S. G., Wong, K., Yang, J., Zeggini, E., Zhang, F., Zheng, H.-F., Anney, R., Ayub, M., Barrett, J. C., Blackwood, D., Bolton, P. F., Breen, G., Collier, D. A., Craddock, N., Crooks, L., Curran, S., Curtis, D., Durbin, R., Gallagher, L., Geschwind, D., Gurling, H., Holmans, P., Lee, I., Lönnqvist, J., McCarthy, S., McGuffin, P., McIntosh, A. M., McKechnie, A. G., McQuillin, A., Morris, J., Muddyman, D., O'Donovan, M. C., Owen, M. J., Palotie, A., Parr, J. R., Paunio, T., Pietilainen, O., Rehnström, K., Sharp, S. I., Skuse, D., St Clair, D., Suvisaari, J., Walters, J. T. R., Williams, H. J., Barroso, I., Bochukova, E., Bounds, R., Dominiczak, A., Durbin, R., Farooqi, I. S., Hendricks, A. E., Keogh, J., Marenne, G., McCarthy, S., Morris, A., Muddyman, D., O'Rahilly, S., Porteous, D. J., Smith, B. H., Tachmazidou, I., Wheeler, E., Zeggini, E., Al Turki, S., Anderson, C. A., Antony, D., Barroso, I., Beales, P., Bentham, J., Bhattacharya, S., Calissano, M., Carss, K., Chatterjee, K., Cirak, S., Cosgrove, C., Durbin, R., Fitzpatrick, D. R., Floyd, J., Reghan Foley, A., Franklin, C. S., Futema, M., Grozeva, D., Humphries, S. E., Hurles, M. E., McCarthy, S., Mitchison, H. M., Muddyman, D., Muntoni, F., O'Rahilly, S., Onoufriadis, A., Parker, V., Payne, F., Plagnol, V., Lucy Raymond, F., Roberts, N., Savage, D. B., Scambler, P., Schmidts, M., Schoenmakers, N., Semple, R. K., Serra, E., Spasic-Boskovic, O., Stevens, E., van Kogelenberg, M., Vijayarangakannan, P., Walter, K., Williamson, K. A., Wilson, C., Whyte, T., Ciampi, A., Greenwood, C. M. T., Hendricks, A. E.,

- Li, R., Metrustry, S., Oualkacha, K., Tachmazidou, I., Xu, C., Zeggini, E., Bobrow, M., Bolton, P. F., Durbin, R., Fitzpatrick, D. R., Griffin, H., Hurles, M. E., Kaye, J., Kennedy, K., Kent, A., Muddyman, D., Muntoni, F., Lucy Raymond, F., Semple, R. K., Smee, C., Spector, T. D., Timpson, N. J., Charlton, R., Ekong, R., Futema, M., Humphries, S. E., Khawaja, F., Lopes, L. R., Migone, N., Payne, S. J., Plagnol, V., Pollitt, R. C., Povey, S., Ridout, C. K., Robinson, R. L., Scott, R. H., Shaw, A., Syrris, P., Taylor, R., Vandersteen, A. M., Barrett, J. C., Barroso, I., Davey Smith, G., Durbin, R., Farooqi, I. S., Fitzpatrick, D. R., Hurles, M. E., Kaye, J., Kennedy, K., Langford, C., McCarthy, S., Muddyman, D., Owen, M. J., Palotie, A., Brent Richards, J., Soranzo, N., Spector, T. D., Stalker, J., Timpson, N. J., Zeggini, E., Amuzu, A., Pablo Casas, J., Chambers, J. C., Cocca, M., Dedoussis, G., Gambaro, G., Gasparini, P., Gaunt, T. R., Huang, J., Iotchkova, V., Isaacs, A., Johnson, J., Kleber, M. E., Kooner, J. S., Langenberg, C., Luan, J., Malerba, G., März, W., Matchan, A., Min, J. L., Morris, R., Nordestgaard, B. G., Benn, M., Ring, S., Scott, R. A., Soranzo, N., Southam, L., Timpson, N. J., Toniolo, D., Traglia, M., Tybjaerg-Hansen, A., van Duijn, C. M., van Leeuwen, E. M., Varbo, A., Whincup, P., Zaza, G., Zeggini, E., and Zhang, W. (2015). The UK10K project identifies rare variants in health and disease. *Nature*, 526(7571):82–90.
- [508] Wang, F., Lu, K., Yang, Z., Zhang, S., Lu, W., Zhang, L., and Liu, S. (2014). Genotypes and phenotypes of congenital goitre and hypothyroidism caused by mutations in dual oxidase 2 genes. *Clinical Endocrinology*, 81:452–457.
- [509] Wang, L., Jia, P., Wolfinger, R. D., Chen, X., and Zhao, Z. (2011). Gene set analysis of genome-wide association studies: Methodological issues and perspectives. *Genomics*, 98(1):1–8.
- [510] Wang, X., Chua, H.-x., Chen, P., Ong, R. T.-h., Sim, X., Zhang, W., Takeuchi, F., Liu, X., Khor, C.-c., Tay, W.-t., Cheng, C.-y., Suo, C., Liu, J., Aung, T., Chia, K.-s., Kato, N., and Teo, Y.-y. (2013). Comparing methods for performing trans-ethnic meta-analysis of genome-wide association studies. *Human Molecular Genetics*, 22(11):2303–2311.
- [511] Watanabe, N., Hiramatsu, K., Miyamoto, R., Yasuda, K., Suzuki, N., Oshima, N., Kiyonari, H., Shiba, D., Nishio, S., Mochizuki, T., Yokoyama, T., Maruyama, S., Matsuo, S., Wakamatsu, Y., and Hashimoto, H. (2009). A murine model of neonatal diabetes mellitus in Glis3-deficient mice. *FEBS letters*, 583(12):2108–13.
- [512] Weber, J. L. and May, P. E. (1989). Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *American Journal of Human Genetics*, 44:388–396.
- [513] Weinzimer, S. a. (2001). Endocrine aspects of the 22q11.2 deletion syndrome. *Genetics in medicine: official journal of the American College of Medical Genetics*, 3(1):19–22.
- [514] Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., and Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 42(D1):1001–1006.

- [515] Wentzel, C., Fernström, M., Öhrner, Y., Annerén, G., and Thuresson, A. C. (2008). Clinical variability of the 22q11.2 duplication syndrome. *European Journal of Medical Genetics*, 51(6):501–510.
- [516] Wessels, M. W. and Willems, P. J. (2010). Genetic factors in non-syndromic congenital heart malformations. *Clinical Genetics*, 78(2):103–123.
- [517] Westbury, S. K., Turro, E., Greene, D., Lentaingne, C., Kelly, A. M., Bariana, T. K., Simeoni, I., Pillois, X., Attwood, A., Austin, S., Jansen, S. B. G., Bakchoul, T., Crisp-Hihn, A., Erber, W. N., Favier, R., Foad, N., Gattens, M., Jolley, J. D., Liesner, R., Meacham, S., Millar, C. M., Nurden, A. T., Peerlinck, K., Perry, D. J., Poudel, P., Schulman, S., Schulze, H., Stephens, J. C., Furie, B., Van Geet, C., Rendon, A., Gomez, K., Laffan, M. A., Lambert, M. P., Nurden, P., Ouwehand, W. H., Richardson, S., Mumford, A. D., and Robinson, P. N. (2015). Human phenotype ontology annotation and cluster analysis to unravel genetic defects in 707 cases with unexplained bleeding and platelet disorders. *Genome Medicine*, 7:36.
- [518] Westerlund, J., Andersson, L., Carlsson, T., Zoppoli, P., Fagman, H., and Nilsson, M. (2008). Expression of *islet1* in thyroid development related to budding, migration, and fusion of primordia. *Developmental Dynamics*, 237(12):3820–3829.
- [519] Wheeler, D. a., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.-J., Makhijani, V., Roth, G. T., Gomes, X., Tartaro, K., Niazi, F., Turcotte, C. L., Irzyk, G. P., Lupski, J. R., Chinault, C., Song, X.-z., Liu, Y., Yuan, Y., Nazareth, L., Qin, X., Muzny, D. M., Margulies, M., Weinstock, G. M., Gibbs, R. a., and Rothberg, J. M. (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 452(7189):872–6.
- [520] Willer, C. J., Li, Y., and Abecasis, G. R. (2010). METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, 26(17):2190–2191.
- [521] Willer, C. J. and Mohlke, K. L. (2012). Finding genes and variants for lipid levels after genome-wide association analysis. *Current opinion in lipidology*, 23(2):98–103.
- [522] Williamson, K. a., Rainger, J., Floyd, J. a. B., Ansari, M., Meynert, A., Aldridge, K. V., Rainger, J. K., Anderson, C. a., Moore, A. T., Hurles, M. E., Clarke, A., van Heyningen, V., Verloes, A., Taylor, M. S., Wilkie, A. O. M., and Fitzpatrick, D. R. (2014). Heterozygous loss-of-function mutations in *YAP1* cause both isolated and syndromic optic fissure closure defects. *American Journal of Human Genetics*, 94(2):295–302.
- [523] Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., Chu, A. Y., Estrada, K., Luan, J., Kutalik, Z., Amin, N., Buchkovich, M. L., Croteau-Chonka, D. C., Day, F. R., Duan, Y., Fall, T., Fehrmann, R., Ferreira, T., Jackson, A. U., Karjalainen, J., Lo, K. S., Locke, A. E., Mägi, R., Mihailov, E., Porcu, E., Randall, J. C., Scherag, A., Vinkhuyzen, A. A. E., Westra, H.-J., Winkler, T. W., Workalemahu, T., Zhao, J. H., Absher, D., Albrecht, E., Anderson, D., Baron, J., Beekman, M., Demirkan, A., Ehret, G. B., Feenstra, B., Feitosa, M. F., Fischer, K., Fraser, R. M., Goel, A., Gong, J., Justice, A. E., Kanoni, S., Kleber, M. E., Kristiansson, K., Lim, U., Lotay, V., Lui, J. C., Mangino, M., Leach,

- I. M., Medina-Gomez, C., Nalls, M. A., Nyholt, D. R., Palmer, C. D., Pasko, D., Pechlivanis, S., Prokopenko, I., Ried, J. S., Ripke, S., Shungin, D., Stancáková, A., Strawbridge, R. J., Sung, Y. J., Tanaka, T., Teumer, A., Trompet, S., van der Laan, S. W., van Setten, J., Van Vliet-Ostaptchouk, J. V., Wang, Z., Yengo, L., Zhang, W., Afzal, U., Arnlöv, J., Arscott, G. M., Bandinelli, S., Barrett, A., Bellis, C., Bennett, A. J., Berne, C., Blüher, M., Bolton, J. L., Böttcher, Y., Boyd, H. A., Bruinenberg, M., Buckley, B. M., Buyske, S., Caspersen, I. H., Chines, P. S., Clarke, R., Claudi-Boehm, S., Cooper, M., Daw, E. W., De Jong, P. A., Deelen, J., Delgado, G., Denny, J. C., Dhonukshe-Rutten, R., Dimitriou, M., Doney, A. S. F., Dörr, M., Eklund, N., Eury, E., Folkersen, L., Garcia, M. E., Geller, F., Giedraitis, V., Go, A. S., Grallert, H., Grammer, T. B., Gräßler, J., Grönberg, H., de Groot, L. C. P. G. M., Groves, C. J., Haessler, J., Hall, P., Haller, T., Hallmans, G., Hannemann, A., Hartman, C. A., Hassinen, M., Hayward, C., Heard-Costa, N. L., Helmer, Q., Hemani, G., Henders, A. K., Hillege, H. L., Hlatky, M. A., Hoffmann, W., Hoffmann, P., Holmen, O., Houwing-Duistermaat, J. J., Illig, T., Isaacs, A., James, A. L., Jeff, J., Johansen, B., Johansson, A., Jolley, J., Juliusdottir, T., Junttila, J., Kho, A. N., Kinnunen, L., Klopp, N., Kocher, T., Kratzer, W., Lichtner, P., Lind, L., Lindström, J., Lobbens, S., Lorentzon, M., Lu, Y., Lyssenko, V., Magnusson, P. K. E., Mahajan, A., Maillard, M., McArdle, W. L., McKenzie, C. A., McLachlan, S., McLaren, P. J., Menni, C., Merger, S., Milani, L., Moayyeri, A., Monda, K. L., Morken, M. A., Müller, G., Müller-Nurasyid, M., Musk, A. W., Narisu, N., Nauck, M., Nolte, I. M., Nöthen, M. M., Oozageer, L., Pilz, S., Rayner, N. W., Renstrom, F., Robertson, N. R., Rose, L. M., Roussel, R., Sanna, S., Schernagl, H., Scholtens, S., Schumacher, F. R., Schunkert, H., Scott, R. A., Sehmi, J., Seufferlein, T., Shi, J., Silventoinen, K., Smit, J. H., Smith, A. V., Smolonska, J., Stanton, A. V., Stirrups, K., Stott, D. J., Stringham, H. M., Sundström, J., Swertz, M. A., Syvänen, A.-C., Tayo, B. O., Thorleifsson, G., Tyrer, J. P., van Dijk, S., van Schoor, N. M., van der Velde, N., van Heemst, D., van Oort, F. V. A., Vermeulen, S. H., Verweij, N., Vonk, J. M., Waite, L. L., Waldenberger, M., Wennauer, R., Wilkens, L. R., Willenborg, C., Wilsgaard, T., Wojczynski, M. K., Wong, A., Wright, A. F., Zhang, Q., Arveiler, D., Bakker, S. J. L., Beilby, J., Bergman, R. N., Bergmann, S., Biffar, R., Blangero, J., Boomsma, D. I., Bornstein, S. R., Bovet, P., Brambilla, P., Brown, M. J., Campbell, H., Caulfield, M. J., Chakravarti, A., Collins, R., Collins, F. S., Crawford, D. C., Cupples, L. A., Danesh, J., de Faire, U., den Ruijter, H. M., Erbel, R., Erdmann, J., Eriksson, J. G., Farrall, M., Ferrannini, E., Ferrières, J., Ford, I., Forouhi, N. G., Forrester, T., Gansevoort, R. T., Gejman, P. V., Gieger, C., Golay, A., Gottesman, O., Gudnason, V., Gyllensten, U., Haas, D. W., Hall, A. S., Harris, T. B., Hattersley, A. T., Heath, A. C., Hengstenberg, C., Hicks, A. A., Hindorff, L. A., Hingorani, A. D., Hofman, A., Hovingh, G. K., Humphries, S. E., Hunt, S. C., Hyppönen, E., Jacobs, K. B., Jarvelin, M.-R., Jousilahti, P., Jula, A. M., Kaprio, J., Kastelein, J. J. P., Kayser, M., Kee, F., Keinanen-Kiukaanniemi, S. M., Kiemeny, L. A., Kooner, J. S., Kooperberg, C., Koskinen, S., Kovacs, P., Kraja, A. T., Kumari, M., Kuusisto, J., Lakka, T. A., Langenberg, C., Le Marchand, L., Lehtimäki, T., Lupoli, S., Madden, P. A. F., Männistö, S., Manunta, P., Marette, A., Matise, T. C., McKnight, B., Meitinger, T., Moll, F. L., Montgomery, G. W., Morris, A. D., Morris, A. P., Murray, J. C., Nelis, M., Ohlsson, C., Oldehinkel, A. J., Ong, K. K., Ouwehand, W. H., Pasterkamp, G., Peters, A., Pramstaller, P. P., Price, J. F., Qi, L., Raitakari, O. T., Rankinen, T., Rao, D. C., Rice, T. K., Ritchie, M., Rudan, I., Salomaa,

- V., Samani, N. J., Saramies, J., Sarzynski, M. A., Schwarz, P. E. H., Sebert, S., Sever, P., Shuldiner, A. R., Sinisalo, J., Steinthorsdottir, V., Stolk, R. P., Tardif, J.-C., Tönjes, A., Tremblay, A., Tremoli, E., Virtamo, J., Vohl, M.-C., Amouyel, P., Asselbergs, F. W., Assimes, T. L., Bochud, M., Boehm, B. O., Boerwinkle, E., Bottinger, E. P., Bouchard, C., Cauchi, S., Chambers, J. C., Chanock, S. J., Cooper, R. S., de Bakker, P. I. W., Dedoussis, G., Ferrucci, L., Franks, P. W., Froguel, P., Groop, L. C., Haiman, C. A., Hamsten, A., Hayes, M. G., Hui, J., Hunter, D. J., Hveem, K., Jukema, J. W., Kaplan, R. C., Kivimaki, M., Kuh, D., Laakso, M., Liu, Y., Martin, N. G., März, W., Melbye, M., Moebus, S., Munroe, P. B., Njølstad, I., Oostra, B. A., Palmer, C. N. A., Pedersen, N. L., Perola, M., Pérusse, L., Peters, U., Powell, J. E., Power, C., Quertermous, T., Rauramaa, R., Reinmaa, E., Ridker, P. M., Rivadeneira, F., Rotter, J. I., Saaristo, T. E., Saleheen, D., Schlessinger, D., Slagboom, P. E., Snieder, H., Spector, T. D., Strauch, K., Stumvoll, M., Tuomilehto, J., Uusitupa, M., van der Harst, P., Völzke, H., Walker, M., Wareham, N. J., Watkins, H., Wichmann, H.-E., Wilson, J. F., Zanen, P., Deloukas, P., Heid, I. M., Lindgren, C. M., Mohlke, K. L., Speliotes, E. K., Thorsteinsdottir, U., Barroso, I., Fox, C. S., North, K. E., Strachan, D. P., Beckmann, J. S., Berndt, S. I., Boehnke, M., Borecki, I. B., McCarthy, M. I., Metspalu, A., Stefansson, K., Uitterlinden, A. G., van Duijn, C. M., Franke, L., Willer, C. J., Price, A. L., Lettre, G., Loos, R. J. F., Weedon, M. N., Ingelsson, E., O'Connell, J. R., Abecasis, G. R., Chasman, D. I., Goddard, M. E., Visscher, P. M., Hirschhorn, J. N., and Frayling, T. M. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics*, 46(11):1173–86.
- [524] Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., Collins, N., Gregory, S., Gumbs, C., and Micklem, G. (1995). Identification of the breast cancer susceptibility gene BRCA2. *Nature*, 378(6559):789–92.
- [525] Worthey, E. a., Mayer, A. N., Syverson, G. D., Helbling, D., Bonacci, B. B., Decker, B., Serpe, J. M., Dasu, T., Tschannen, M. R., Veith, R. L., Basehore, M. J., Broeckel, U., Tomita-Mitchell, A., Arca, M. J., Casper, J. T., Margolis, D. a., Bick, D. P., Hessner, M. J., Routes, J. M., Verbsky, J. W., Jacob, H. J., and Dimmock, D. P. (2011). Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genetics in medicine: official journal of the American College of Medical Genetics*, 13(3):255–62.
- [526] Wray, N., Goddard, M., and Visscher, P. (2007). Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Research*, 17:1520–1528.
- [527] Wright, C. F., Fitzgerald, T. W., Jones, W. D., Clayton, S., McRae, J. F., Van Kogelenberg, M., King, D. A., Ambridge, K., Barrett, D. M., Bayzatinova, T., Bevan, A. P., Bragin, E., Chatzimichali, E. A., Gribble, S., Jones, P., Krishnappa, N., Mason, L. E., Miller, R., Morley, K. I., Parthiban, V., Prigmore, E., Rajan, D., Sifrim, A., Swaminathan, G. J., Tivey, A. R., Middleton, A., Parker, M., Carter, N. P., Barrett, J. C., Hurles, M. E., Fitzpatrick, D. R., and Firth, H. V. (2015). Genetic diagnosis of developmental disorders in the DDD study: A scalable analysis of genome-wide research data. *The Lancet*, 385(9975):1305–1314.

- [528] WTCCC (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–78.
- [529] Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics*, 89(1):82–93.
- [530] Xu, B., Ionita-Laza, I., Roos, J. L., Boone, B., Woodrick, S., Sun, Y., Levy, S., Gogos, J. A., and Karayiorgou, M. (2012). De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nature Genetics*, 44(12):1365–9.
- [531] Xue, Y., Chen, Y., Ayub, Q., Huang, N., Ball, E. V., Mort, M., Phillips, A. D., Shaw, K., Stenson, P. D., Cooper, D. N., and Tyler-Smith, C. (2012). Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. *American Journal of Human Genetics*, 91(6):1022–32.
- [532] Yagi, H., Furutani, Y., Hamada, H., Sasaki, T., Asakawa, S., Minoshima, S., Ichida, F., Joo, K., Kimura, M., ichiro Imamura, S., Kamatani, N., Momma, K., Takao, A., Nakazawa, M., Shimizu, N., and Matsuoka, R. (2003). Role of TBX1 in human del22q11.2 syndrome. *Lancet*, 362(9393):1366–1373.
- [533] Yamagata, K., Oda, N., Kaisaki, P. J., Menzel, S., Furuta, H., Vaxillaire, M., Southam, L., Cox, R. D., Lathrop, G. M., Boriraj, V. V., Chen, X., Cox, N. J., Oda, Y., Yano, H., Le Beau, M. M., Yamada, S., Nishigori, H., Takeda, J., Fajans, S. S., Hattersley, a. T., Iwasaki, N., Hansen, T., Pedersen, O., Polonsky, K. S., and Bell, G. I. (1996). Mutations in the hepatocyte nuclear factor-1alpha gene in maturity-onset diabetes of the young (MODY3). *Nature*, 384(6608):455–458.
- [534] Yang, X., Kanegane, H., Nishida, N., Imamura, T., Hamamoto, K., Miyashita, R., Imai, K., Nonoyama, S., Sanayama, K., Yamaide, A., Kato, F., Nagai, K., Ishii, E., van Zelm, M. C., Latour, S., Zhao, X.-D., and Miyawaki, T. (2012). Clinical and genetic characteristics of XIAP deficiency in Japan. *Journal of clinical immunology*, 32(3):411–20.
- [535] Yang, Y., Muzny, D. M., Reid, J. G., Bainbridge, M. N., Willis, A., Ward, P. a., Braxton, A., Beuten, J., Xia, F., Niu, Z., Hardison, M., Person, R., Bekheirnia, M. R., Leduc, M. S., Kirby, A., Pham, P., Scull, J., Wang, M., Ding, Y., Plon, S. E., Lupski, J. R., Beaudet, A. L., Gibbs, R. a., and Eng, C. M. (2013). Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *The New England Journal of Medicine*, 369(16):1502–11.
- [536] Yang, Y., Muzny, D. M., Xia, F., Niu, Z., Person, R., Ding, Y., Ward, P., Braxton, A., Wang, M., Buhay, C., Veeraraghavan, N., Hawes, A., Chiang, T., Leduc, M., Beuten, J., Zhang, J., He, W., Scull, J., Willis, A., Landsverk, M., Craigen, W. J., Bekheirnia, M. R., Stray-Pedersen, A., Liu, P., Wen, S., Alcaraz, W., Cui, H., Walkiewicz, M., Reid, J., Bainbridge, M., Patel, A., Boerwinkle, E., Beaudet, A. L., Lupski, J. R., Plon, S. E., Gibbs, R. a., and Eng, C. M. (2014). Molecular Findings Among Patients Referred for Clinical Whole-Exome Sequencing. *Jama*, 312(18):1870.

- [537] Zamproni, I., Grasberger, H., Cortinovis, F., Vigone, M. C., Chiumello, G., Mora, S., Onigata, K., Fugazzola, L., Refetoff, S., Persani, L., and Weber, G. (2008). Biallelic inactivation of the dual oxidase maturation factor 2 (DUOXA2) gene as a novel cause of congenital hypothyroidism. *The Journal of Clinical Endocrinology and Metabolism*, 93(2):605–610.
- [538] Zeggini, E., Panoutsopoulou, K., Southam, L., Rayner, N. W., Day-Williams, A. G., Lopes, M. C., Boraska, V., Esko, T., Evangelou, E., Hofman, A., Houwing-Duistermaat, J. J., Ingvarsson, T., Jonsdottir, I., Jonsson, H., Kerkhof, H. J. M., Kloppenburg, M., Bos, S. D., Mangino, M., Metrustry, S., Slagboom, P. E., Thorleifsson, G., Raine, E. V. A., Ratnayake, M., Ricketts, M., Beazley, C., Blackburn, H., Bumpstead, S., Elliott, K. S., Hunt, S. E., Potter, S. C., Shin, S. Y., Yadav, V. K., Zhai, G., Sherburn, K., Dixon, K., Arden, E., Aslam, N., Battley, P. K., Carluke, I., Doherty, S., Gordon, A., Joseph, J., Keen, R., Koller, N. C., Mitchell, S., O'Neill, F., Paling, E., Reed, M. R., Rivadeneira, F., Swift, D., Walker, K., Watkins, B., Wheeler, M., Birrell, F., Ioannidis, J. P. A., Meulenberg, I., Metspalu, A., Rai, A., Salter, D., Stefansson, K., Styrkarsdottir, U., Uitterlinden, A. G., Van Meurs, J. B. J., Chapman, K., Deloukas, P., Ollier, W. E. R., Wallis, G. A., Arden, N., Carr, A., Doherty, M., McCaskie, A., Wilkinson, J. M., Ralston, S. H., Valdes, A. M., Spector, T. D., and Loughlin, J. (2012). Identification of new susceptibility loci for osteoarthritis (arcOGEN): A genome-wide association study. *The Lancet*, 380(9844):815–823.
- [539] Zeggini, E., Scott, L. J., Saxena, R., Voight, B. F., Marchini, J. L., Hu, T., de Bakker, P. I. W., Abecasis, G. R., Almgren, P., Andersen, G., Ardlie, K., Boström, K. B., Bergman, R. N., Bonnycastle, L. L., Borch-Johnsen, K., Burt, N. P., Chen, H., Chines, P. S., Daly, M. J., Deodhar, P., Ding, C.-J., Doney, A. S. F., Duren, W. L., Elliott, K. S., Erdos, M. R., Frayling, T. M., Freathy, R. M., Gianniny, L., Grallert, H., Grarup, N., Groves, C. J., Guiducci, C., Hansen, T., Herder, C., Hitman, G. A., Hughes, T. E., Isomaa, B., Jackson, A. U., Jørgensen, T., Kong, A., Kubalanza, K., Kuruvilla, F. G., Kuusisto, J., Langenberg, C., Lango, H., Lauritzen, T., Li, Y., Lindgren, C. M., Lyssenko, V., Marvelle, A. F., Meisinger, C., Midtjell, K., Mohlke, K. L., Morken, M. A., Morris, A. D., Narisu, N., Nilsson, P., Owen, K. R., Palmer, C. N. A., Payne, F., Perry, J. R. B., Pettersen, E., Platou, C., Prokopenko, I., Qi, L., Qin, L., Rayner, N. W., Rees, M., Roix, J. J., Sandbaek, A., Shields, B., Sjögren, M., Steinthorsdottir, V., Stringham, H. M., Swift, A. J., Thorleifsson, G., Thorsteinsdottir, U., Timpson, N. J., Tuomi, T., Tuomilehto, J., Walker, M., Watanabe, R. M., Weedon, M. N., Willer, C. J., Illig, T., Hveem, K., Hu, F. B., Laakso, M., Stefansson, K., Pedersen, O., Wareham, N. J., Barroso, I., Hattersley, A. T., Collins, F. S., Groop, L., McCarthy, M. I., Boehnke, M., and Altshuler, D. (2008). Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature Genetics*, 40(5):638–45.
- [540] Zeggini, E., Weedon, M. N., Lindgren, C. M., Frayling, T. M., Elliott, K. S., Lango, H., Timpson, N. J., Perry, J. R. B., Rayner, N. W., Freathy, R. M., Barrett, J. C., Shields, B., Morris, A. P., Ellard, S., Groves, C. J., Harries, L. W., Marchini, J. L., Owen, K. R., Knight, B., Cardon, L. R., Walker, M., Hitman, G. A., Morris, A. D., Doney, A. S. F., McCarthy, M. I., and Hattersley, A. T. (2007). Replication of

- genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science (New York, N.Y.)*, 316(5829):1336–41.
- [541] Zhang, F., Gu, W., Hurles, M. E., and Lupski, J. R. (2009). Copy number variation in human health, disease, and evolution. *Annual review of genomics and human genetics*, 10:451–81.
- [542] Zheng, H. F., Rong, J. J., Liu, M., Han, F., Zhang, X. W., Richards, J. B., and Wang, L. (2015). Performance of genotype imputation for low frequency and rare variants from the 1000 genomes. *PLoS ONE*, 10(1):1–10.
- [543] Zhu, W., Shiojima, I., Hiroi, Y., Zou, Y., Akazawa, H., Mizukami, M., Toko, H., Yazaki, Y., Nagai, R., and Komuro, I. (2000). Functional analyses of three Csx/Nkx-2.5 mutations that cause human congenital heart disease. *Journal of Biological Chemistry*, 275(45):35291–35296.
- [544] Zhu, X., Petrovski, S., Xie, P., Ruzzo, E. K., Lu, Y.-F., McSweeney, K. M., Ben-Zeev, B., Nissenkorn, A., Anikster, Y., Oz-Levi, D., Dhindsa, R. S., Hitomi, Y., Schoch, K., Spillmann, R. C., Heimer, G., Marek-Yagel, D., Tzadok, M., Han, Y., Worley, G., Goldstein, J., Jiang, Y.-H., Lancet, D., Pras, E., Shashi, V., McHale, D., Need, A. C., and Goldstein, D. B. (2015). Whole-exome sequencing in undiagnosed genetic diseases: interpreting 119 trios. *Genetics in medicine: official journal of the American College of Medical Genetics*, 17(10):774–81.
- [545] Zuk, O., Hechter, E., Sunyaev, S. R., and Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences of the United States of America*, 109(4):1193–8.
- [546] Zuk, O., Schaffner, S. F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M. J., Neale, B. M., Sunyaev, S. R., and Lander, E. S. (2014). Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences of the United States of America*, 111(4):E455–64.
- [547] Zweier, C., Sticht, H., Aydin-Yaylagül, I., Campbell, C. E., and Rauch, A. (2007). Human TBX1 missense mutations cause gain of function resulting in the same phenotype as 22q11.2 deletions. *American Journal of Human Genetics*, 80(3):510–7.

Appendix A

Appendix

VQSR inputs	Variant type	
	SNVs	Indels
Annotations	QD, MQ, MQRankSum, ReadPosRankSum, FS, InbreedingCoeff	QD, FS, ReadPosRankSum, MQRankSum, InbreedingCoeff
Training set	HapMap 3.3: hapmap_3.3.b37.sites.vcf.gz	Mills-Devine and 1000GP Phase I:
	Omni 2.5M chip: 1000G_omni2.5.b37.sites.vcf.gz	Mills_and_1000G_gold_standard.indels.b37.sites.vcf.gz
	1000GP Phase I: 1000G_phase1.snps.high_confidence.b37.vcf	.
Truth set	HapMap 3.3: hapmap_3.3.b37.sites.vcf.gz	Mills-Devine and 1000GP Phase I:
	Omni 2.5M chip: 1000G_omni2.5.b37.sites.vcf.gz	Mills_and_1000G_gold_standard.indels.b37.sites.vcf.gz
Known set	dbSNP build 137: dbsnp_137.b37.excluding_sites_after_129.vcf	dbSNP build 137: dbsnp_138.vcf.gz

Table A.1 Annotations and training sets used for VQSR variant QC.

Variant caller	Filter	Description	Variant type	
			SNPs	Indels
SAMtools	StrandBias	One DNA strand being favored over the other	0.0001	0.0001
	EndDistBias	One DNA strand being favored over the other at the end of reads	0.0001	0.0001
	MaxDP	Maximum depth allowed	2000	2000
	MinDP	Minimum depth allowed	4	4
	MinMQ	Minimum mapping quality allowed	10	10
	MinAB	Minimum number of alternate bases	2	2
	Qual	Minimum value of the overall quality field	10	10
	GapWin	Window size to filter adjacent gaps	3	3
	MapQualBias	Minimum P-value for Mapping quality bias	0	0
	SnpGap	SNP within a certain distance of an Indel to be filtered out	10	10
GATK	MinDP	Minimum depth allowed	4	4
	MinQD	Minimum quality over depth allowed	2	2
	MinMQ	Minimum mapping quality allowed	10	.
	MaxFS	Maximum Fishers P-value	60	200
	MaxHS	Maximum haplotype score	13	.
	MinMQRank	Minimum Z-score	-12.5	.
	MinPosRank	Minimum Z-score	-8	-20
	MinInbreed	Minimum inbreeding coefficient	.	-0.8

Table A.2 Filters and thresholds applied on variants from the TS experiment.

ID	Country of origin	Gender	TSH	FT4 (pmol/L)	Thyroid gland	Comment	Mutations
Monogenic families							
F2a	Saudi Arabia (C)	M	>100	16	Goitre, normal on L-T4 (u)		TG R159* (hom)
F2b	Saudi Arabia (C)	M	150	6.2	Normal on L-T4 (u)		TG R159* (hom)
F1a	Pakistan (C)	M	>375	3.9	Avid uptake (i)	Siblings	TG R451* (hom)
F1b	Pakistan (C)	M	NA	NA	NA	Siblings	TG R451* (hom)
F3a	Turkish (C)	F	>75	6.4	Nodular, enlarged (u)	Siblings	TG S528* (hom)
F3b	Turkish (C)	F	>51.4	6	Nodular goitre (u)		TG S528* (hom)
F4	Oman	M	>100	NA	2 siblings with goitrous CH and the same TG genotype	2 affected siblings	TG S2121Afs*32 (hom)
F5a	JK	M	>150	6	Normal (i)	Siblings	TG R296* (het), TG C160S (het)
F5b	JK	F	>150	11	NA		TG R296* (het), TG C160S (het)
F7a	Iraq (C)	M	16 (bs)	NA	Normal (u)	Siblings	TG c.638+5G>A (hom)
F7b	Iraq (C)	F	43.6	16.8	Avid uptake (i)		TG c.638+5G>A (hom)
F17	JAE (C)	M	27	NA	Avid uptake, normally sited (i)		TPO R665Q (hom)
F18	JK	M	920	1.2	Normal (i)		TPO R291H (het), TPO G331V (het)
F16	JK	F	NA	NA	Nodular goitre (u)		TPO R491H (het), TPO A397Pfs*76 (het)
F20	Oman (C)	M	7.1 on L-T4	13.5 on L-T4	Avid uptake (i)	2 affected siblings	TPO C808Afs*24 (hom)
F24	Oman	M	55	NA	Normal (u)		DUOX2 L1028Afs*3 (hom)
F23	Bangladesh (C)	F	47.6	15.8	Good uptake, normally sited (i)	Transient CH in sibling	DUOX2 F966Sfs*29 (het)
F26	JAE (C)	M	21	NA	Normal (u)	Cousin with CH	TSHR P68S (het)
Digenic families							
F8a	Turkey (C)	M	NA	NA	Goitre (u)	Siblings	DUOX2 Q686* (het), TG 1493Y (hom)
F8b	Turkey (C)	M	79.6	1	Goitre (u)		DUOX2 Q686* (het), TG 1493Y (hom)
F6a	Turkey (C)	F	123.3	8.9	Nodular goitre (u)	Siblings	TG W1050L (hom), TG C726Y (hom)
F6b	Turkey (C)	F	NA	NA	Nodular goitre (u)		TG W1050L (hom), TG C726Y (hom), DUOX2 Q686* (het)
F9a	Turkey (C)	M	>75	TT4 <0.5	Normal (u)	Siblings	TG W2685L (hom), DUOX2 R354W (het)
F9b	Turkey (C)	M	>75	<1.5	Normal (i)		TG W2685L (hom), DUOX2 R354W (het)
F11	JAE (C)	M	250	2	Normal (u, b)		TPO R491H (hom), TG Q164E (het)
F21	JK	M	>100	3.8	Avid uptake, enlarged (i, u)		TPO E17Dfs*77 (het), TPO Y453D (het), SLC26A4 E384G (het)
F10a	Pakistan (C)	M	>150	NA	Normal (u)	Cousins with transient CH	DUOX2 G570L (hom), TG R1691C (het), TG L2647Q (het)
F19a	JK/African-Caribbean	F	400	0.8	Normal (i)		TPO R684Q (hom), SLC26A4 N324Y (het)
F19b	JK/African-Caribbean	M	620	0.7	Normal (i)	Siblings	TPO R684Q (hom), SLC26A4 I713M (het)

Table A.3 Detailed genotype and phenotype information for all CH patients with causative mutations in the eight known *gland-in-situ* genes.

ID	Country of origin	Gender	TSH	fT4 (pmol/L)	Thyroid gland	Other features	Mutations
Ambiguous cases							
F22	UK	F	10	14	Normal (u)		TPO E510Afs*14 (het)
F12a	India/UK	M	11.3	13.6	Left lobe smaller than right (u)	Siblings	TG Q870H (het)
F12b	India/UK	M	10.05	12.3	Left lobe smaller than right (u)		TG Q870H (het)
F13	Wales	M	96	12	Normal (u)	Cleft palate	TG Q771* (het)
F15a	Somalia	F	Mildly elevated	NA	Normal (u)		TG c.3433+3_3433+6delGAGT (het)
F15b	Somalia	M	Mildly elevated	NA	Normal (u)	Siblings	TG c.3433+3_3433+6delGAGT (het)
F15c	Somalia	M	40.3	17.8	Avid uptake (i)		TG c.3433+3_3433+6delGAGT (het)
F14a	Yemen (C)	F	NA	NA	Normal (u)		TG Y759C (het)
F14b	Yemen (C)	F	8	16	Normal (u)	Identical twins	TG Y759C (het)
F25*	Sri Lanka	M	NA	NA	Avid uptake (i)	2 affected siblings	DUOX2 R764W (het)
F27	Bangladesh (C)	F	>50	4.3	Normal (u)		DUOX2 c. 555-5G>A (hom)
Unsolved cases							
F28	UAE (C)	M	200	NA	Normal (u)		
F29a	Saudi Arabia (C)	M	>100	1.4	Goitre		
F29b	Saudi Arabia (C)	M	>100	0.49	NA	Siblings	
F30	Oman (C)	M	NA	NA	Normal (u)	1 affected sibling	
F31	Australia	M	20.7	11.5	Normally sited, decreased uptake (i)	1 affected sibling	
F32	Poland	F	31.2	12.4	Normal (i)		
F33a	UK	F	12.3 (bs)	NA	Normal (i)	Siblings	
F33b	UK	F	22	NA	NA		
F34	UAE (C)	M	153	5.7	Normally sized and sited (u)		

Table A.4 Detailed genotype and phenotype information for all CH patients that were considered 'ambiguous' or 'unsolved' due to a lack of convincing causative variants in *gland-in-situ* genes.

Mouse gene	Human gene	Mouse gene	Human gene
Mouse models of TD		Genes enriched in the mouse thyroid bud at E10.5	
<i>Shh</i>	<i>SHH</i>	<i>Adrbk2</i>	<i>ADRBK2</i>
<i>Foxe1</i>	<i>FOXE1</i>	<i>Atap1L2</i>	<i>AFAP1L2</i>
<i>Chrd</i>	<i>CHRD</i>	<i>Atp10a</i>	<i>ATP10A</i>
<i>Edn1</i>	<i>EDN1</i>	<i>Bcl11b</i>	<i>BCL11B</i>
<i>Eya1</i>	<i>EYA1</i>	<i>Bcl2</i>	<i>BCL2</i>
<i>Fbln1</i>	<i>FBLN1</i>	<i>Calml3</i>	<i>CALML3</i>
<i>Hes1</i>	<i>HES1</i>	<i>Capg</i>	<i>CAPG</i>
<i>Hoxa5</i>	<i>HOXA5</i>	<i>Cckar</i>	<i>CCKAR</i>
<i>Isl1</i>	<i>ISL1</i>	<i>Cd44</i>	<i>CD44</i>
<i>Nkx2-5</i>	<i>NKX2-5</i>	<i>Chdh</i>	<i>CHDH</i>
<i>Frs2</i>	<i>FRS2</i>	<i>Clstn2</i>	<i>CLSTN2</i>
<i>Hoxa3</i>	<i>HOXA3</i>	<i>Cpne4</i>	<i>CPNE4</i>
<i>Hoxb3</i>	<i>HOXB3</i>	<i>Cpxm2</i>	<i>CPXM2</i>
<i>Hoxd3</i>	<i>HOXD3</i>	<i>Ctnn3</i>	<i>CTNN3</i>
<i>Pax3</i>	<i>PAX3</i>	<i>Cxcl12</i>	<i>CXCL12</i>
<i>Fgfr2</i>	<i>FGFR2</i>	<i>Elfn1</i>	<i>ELFN1</i>
<i>Fgf10</i>	<i>FGF10</i>	<i>Galns</i>	<i>GALNS</i>
<i>Hhex</i>	<i>HHEX</i>	<i>Gcgr</i>	<i>GCGR</i>
<i>Nkx2-1</i>	<i>NKX2-1</i>	<i>Hexb</i>	<i>HEXB</i>
<i>Pax8</i>	<i>PAX8</i>	<i>Hivep3</i>	<i>HIVEP3</i>
<i>Twsg1</i>	<i>TWSG1</i>	<i>Htra1</i>	<i>HTRA1</i>
<i>Tbx1</i>	<i>TBX1</i>	<i>Irs4</i>	<i>IRS4</i>
Zebrafish models of TD		<i>Klhl14</i>	<i>KLHL14</i>
<i>ace</i>	<i>ACE</i>	<i>Lypd6b</i>	<i>LYPD6B</i>
<i>cyc</i>	<i>CYC1</i>	<i>Matn2</i>	<i>MATN2</i>
<i>fau</i>	<i>FAU</i>	<i>Nbeal2</i>	<i>NBEAL2</i>
<i>hand2</i>	<i>HAND2</i>	<i>Nptx1</i>	<i>NPTX1</i>
		<i>Pla2g7</i>	<i>PLA2G7</i>
		<i>Prlr</i>	<i>PRLR</i>
		<i>Ptpre</i>	<i>PTPRE</i>
		<i>Ryr3</i>	<i>RYR3</i>
		<i>Scara5</i>	<i>SCARA5</i>
		<i>Slc16a2</i>	<i>SLC16A2</i>
		<i>Slc44a3</i>	<i>SLC44A3</i>
		<i>Slc4a4</i>	<i>SLC4A4</i>
		<i>Slc4A5</i>	<i>SLC4A5</i>
		<i>Sorbs2</i>	<i>SORBS2</i>
		<i>Stc2</i>	<i>STC2</i>
		<i>Tbx3</i>	<i>TBX3</i>
		<i>Tcfcp2l1</i>	<i>TFCP2L1</i>
		<i>Zbtb20</i>	<i>ZBTB20</i>
		<i>Zbtb4</i>	<i>ZBTB4</i>

Table A.5 List of CH candidate genes, part 1.

Mouse gene	Human gene	Foxe1 targets	Pax8 targets
Genes enriched in both the mouse thyroid bud and lung at E10.5		AHCY	CDH16
<i>Alcam</i>	ALCAM	AMIGO3	CITED2
<i>Ap1m2</i>	AP1M2	ANKRD37	EGR1
<i>Arhgef16</i>	ARHGEF16	ATMIN	IGFBP7
<i>Cdcp1</i>	CDCP1	BET1	KCNJ15
<i>Cdh1</i>	CDH1	CASP4	KCNJ16
<i>Cdh16</i>	CDH16	COQ10B	NFKB1
<i>Cldn3</i>	CLDN3	CRELD2	RAB17
<i>Cldn6</i>	CLDN6	CTGF	RUNX2
<i>Cldn7</i>	CLDN7	DDIT3	SPARC
<i>Clu</i>	CLU	DERL3	TRIB1
<i>Crb3</i>	CRB3	DNAJB11	WBP2
<i>Ct14</i>	SAGE1	DNAJB9	WNT4
<i>Dsg2</i>	DSG2	DNAJC3	
<i>Epcam</i>	EPCAM	DUSP5	
<i>Eppk1</i>	EPPK1	ENGASE	
<i>Esrp2</i>	ESRP2	ERO1LB	
<i>Inadl</i>	INADL	ETV5	
<i>Kcnk1</i>	KCNK1	GGCT	
<i>Mapk13</i>	MAPK13	GMPPB	
<i>Marveld2</i>	MARVELD2	HSP90B1	
<i>Marveld3</i>	MARVELD3	HSPA5	
<i>Mbip</i>	MBIP	HYOU1	
<i>Meg3</i>	MEG3	IGF2BP2	
<i>Mfsd6</i>	MFS6	IL23A	
<i>Npnt</i>	NPNT	MANF	
<i>Pitpnm1</i>	PITPNM1	MFS2	
<i>Prss8</i>	PRSS8	NR4A2	
<i>Pygl</i>	PYGL	NUPR1	
<i>Rassf10</i>	RASSF10	PDIA4	
<i>Ripk4</i>	RIPK4	RIOK3	
<i>Sh3gl2</i>	SH3GL2	SDF2L1	
<i>Slco2a1</i>	SLCO2A1	SEC23B	
<i>Sorcs2</i>	SORCS2	SEL1L	
<i>Sorl1</i>	SORL1	TM4SF1	
<i>Spint1</i>	SPINT1	TMEM66	
<i>Spint2</i>	SPINT2	ZFAND2A	
<i>Tie2</i>	TEK	ADAMTS9	
<i>Tmem176a</i>	TMEM176A	BCAM	
<i>Tnk1</i>	TNK1	CDH1	
		CRIP2	
		DYNLRB2	
		ELOVL2	
		FGF18	
		FOLR1	
		KRT20	
		PRIMA1	
		PRSS8	
		RIL	
		S100A4	
		SLIT1	
		TMEM140	

Table A.6 List of CH candidate genes, part 2.

Baseline patient characteristics		N = 145	
Gender (<i>female/male</i>)		63 (43%) / 82 (57%)	
Ethnicity (<i>Caucasian / Black / Asian / Jewish / Others / unknown</i>)		99 (68%) / 2 / 21 / 1 / 11 / 11	
Age at diagnosis, years			
Mean \pm SD		3.6 \pm 1.8	
Median (range)		3.5 (range 4 weeks to 7 years)	
Diagnosis (<i>CD / UC / IBDU</i>)		66 (46%) / 51 (35%) / 28 (19%)	
Positive family history		29/137 (21.2%)	
Paris Crohn's Classification (n=66) *			
Disease location	L1 Ileum	2	
	L2 Colon	20	
	L3 Ileocolonic	42	
	+ L4 (upper GI tract)	34	
Disease behavior	B1 nonstricturing-nonpenetrating	54	
	B2 stricturing	6	
	B3 penetrating	2	
	B2B3 penetrating and stricturing	3	
	Perianal involvement	17 (24.2%)	
Paris UC classification		UC (n=49) **	IBDU (n=25) ***
Disease location	E1 Ulcerative proctitis	1	0
	E2 Left sided UC to splenic flexure	6	0
	E3 Extensive UC to hepatic flexure	4	1
	E4 Pancolitis	38	24
GI surgery			
Colectomy		23/144 (16.0%)	
Associate medical therapy			
Steroids		93/144 (64.6%)	
Azathioprine/6-MP		102/144 (70.8%)	
MTX		11/144 (7.6%)	
CYA		9/144 (6.3%)	
Anti-TNF α		51/144 (35.4%)	

Table kindly generated and provided by Dr Tobias Schwerd.
One data set incomplete (David Wilson), "unknown" patient details excluded for analysis;

* CD: 2/66 patients with oral and perianal CD only

** UC: 1/51 patient unknown location, 1/51 data set incomplete

*** IBDU: 3/28 patients without macroscopic inflammation

Table A.7 Disease phenotype and therapy characteristics for the VEO-IBD cohort.

