

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].

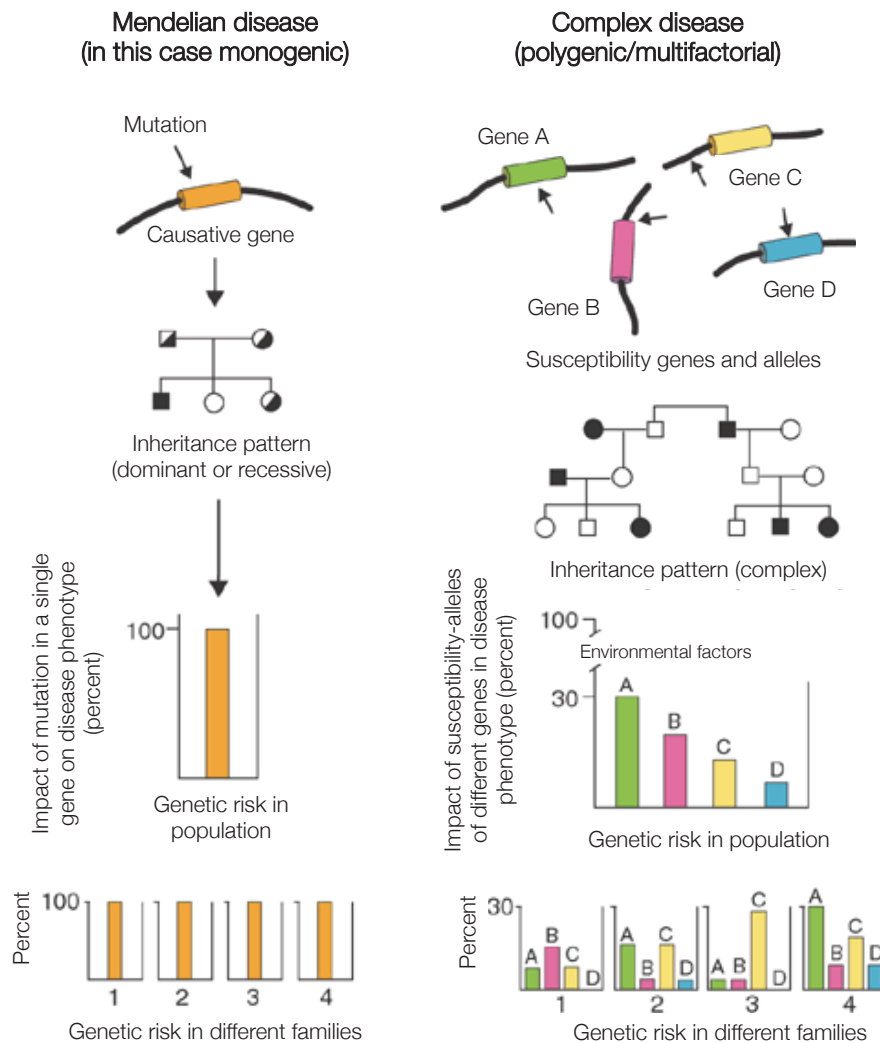


Figure 1.1 Inheritance of monogenic and complex disorders.

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**).

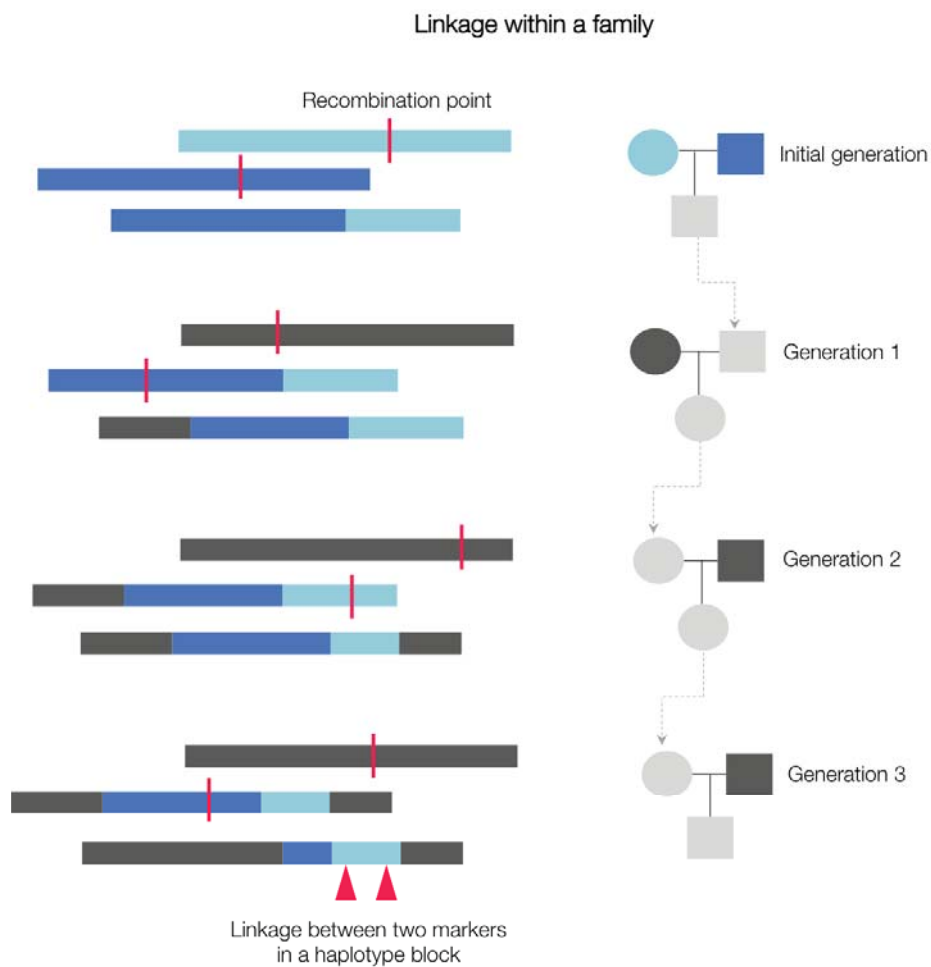


Figure 1.2 Linkage within a family

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.

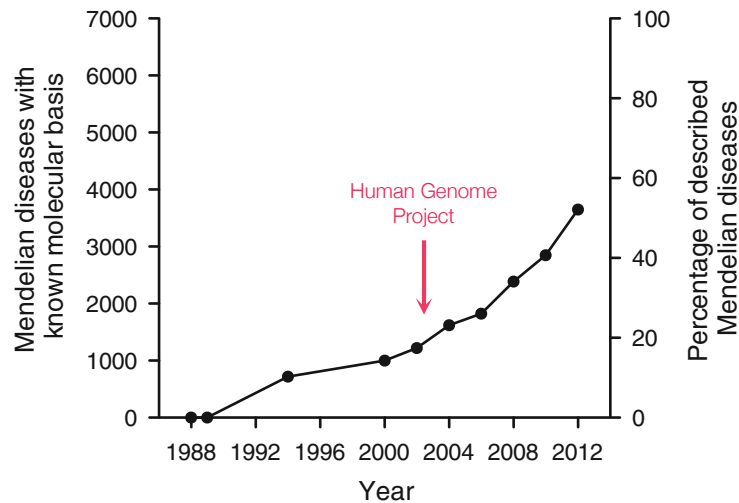


Figure 1.3 Mendelian disease genes of known molecular basis. The left-hand y-axis indicates the cumulative number of diseases for which a molecular basis is identified. The right-hand y-axis expresses that as a percentage of the $\sim 7,000$ Mendelian disorders that have been described and are present 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.

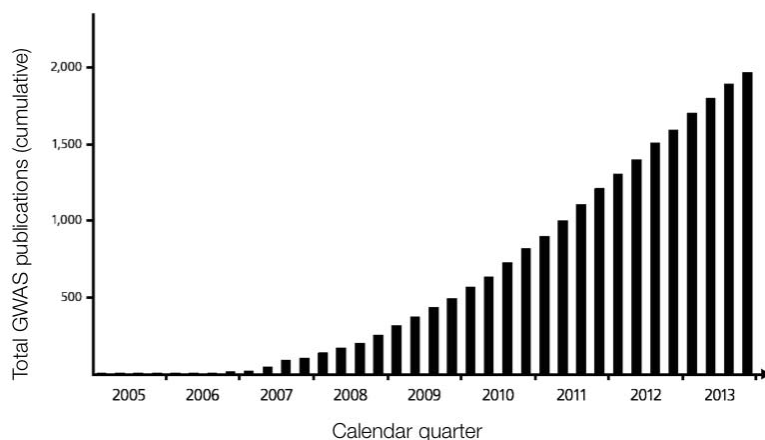


Figure 1.4 Number of genome-wide association studies published between 2005 and 2013. Image credit: Genome Research Limited.

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

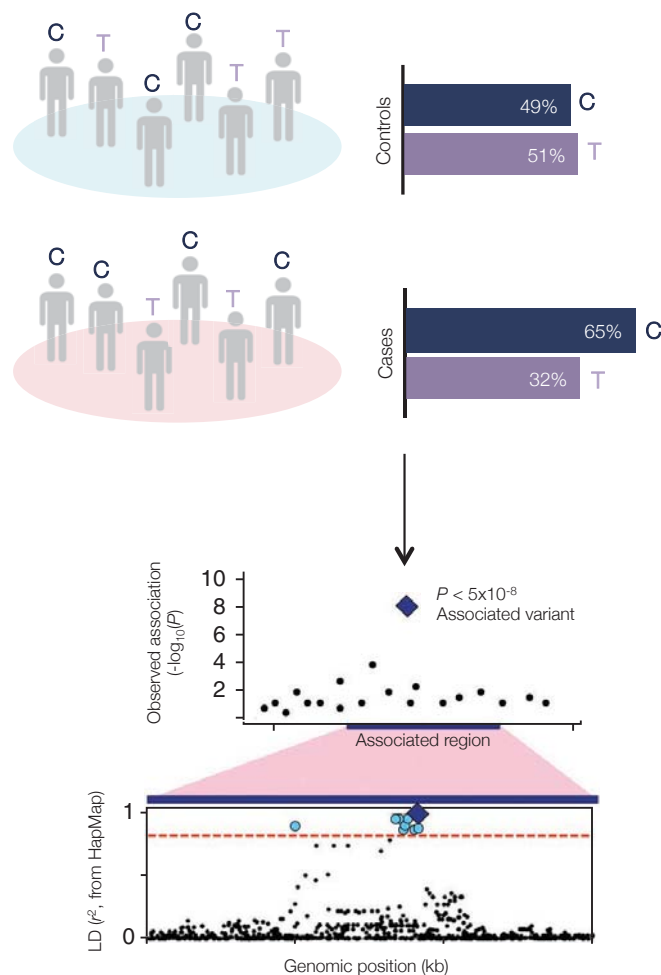


Figure 1.5 Schematic representation of a case-control (or binary) GWAS study.

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).

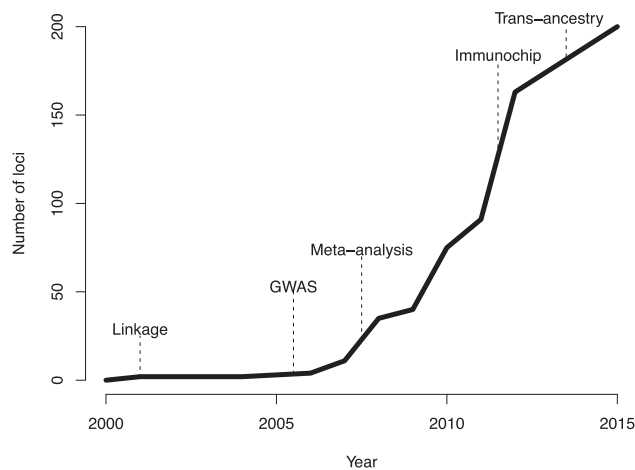


Figure 1.6 The number of IBD-associated loci identified using various study designs over the past fifteen years. Other than meta-analyses, IBD researchers also used a custom genotyping array (Immunochip) to aid replication and fine-mapping strategies, and to allow more cost-efficient genotyping in larger 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

a human genome (at high depth, 30x) down to \$1,000 in 10 years [436]. Since then, many NGS technologies have been developed (**Figure 1.7**), with the rate of throughput continually climbing [418]. The Illumina/Solexa platforms have constantly dominated the market, and have offered diverse systems ranging from small, low-cost "desktop sequencers" such as the MiSeq machine, to population-scale sequencers (HiSeq X Ten). Most of the NGS data generated for my dissertation was produced between 2010-2015, using the then state-of-the-art Illumina's HiSeq 2000 system (**Figure 1.7**).

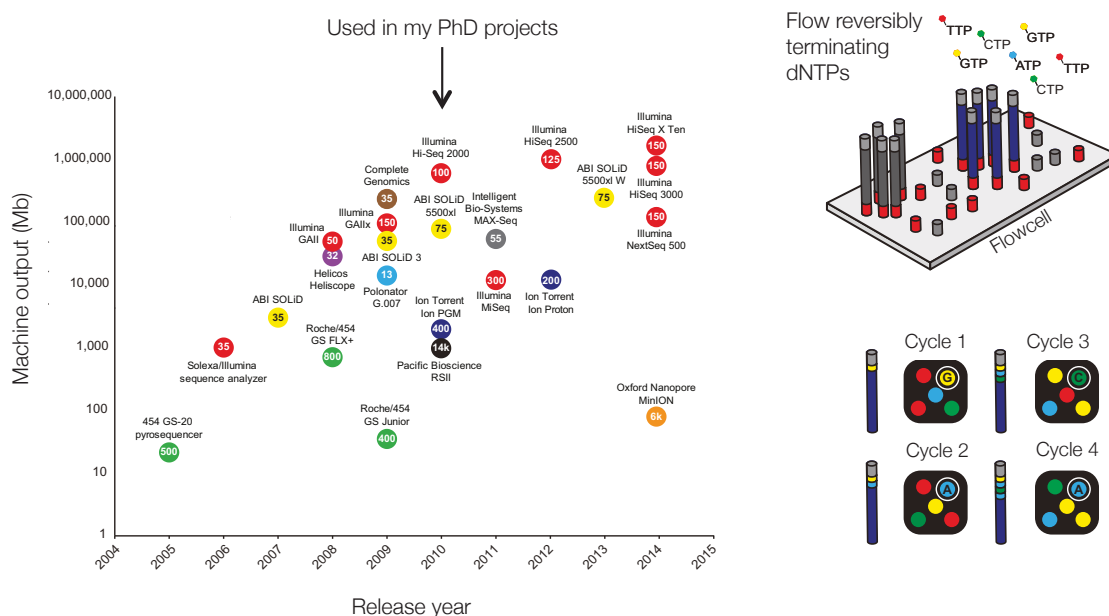


Figure 1.7 **A)** Timeline and comparison of NGS instruments released to date. Release date versus machine outputs per platform are shown. Numbers inside data points denote current read lengths. Sequencing platforms are colour coded according to manufacturer. **B)** Illumina's HiSeq 2000 sequencing method. The sequencing process includes clonal *in-situ*-amplification of DNA fragments (or templates) that are ligated to adaptors on the surface of a glass slide. Nucleotide bases are read using a "cyclic reversible termination" strategy, which sequences the template strand one base at the time through successive rounds of incorporation of fluorescently-labeled complementary bases (dNTPs), washing to remove unincorporated dNTPs and fluorescent imaging to determine the added bases. Images adapted from Reuter *et al* [418].

In 2007, Nimblegen released a sequence capture technology that was able to select specific DNA sequences by microarray hybridization [12], now known as "targeted capture". Using this method, any subset of the genome, from a handful of genes, to virtually all protein-coding regions (the "exome"), could be sequenced much quicker and

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.

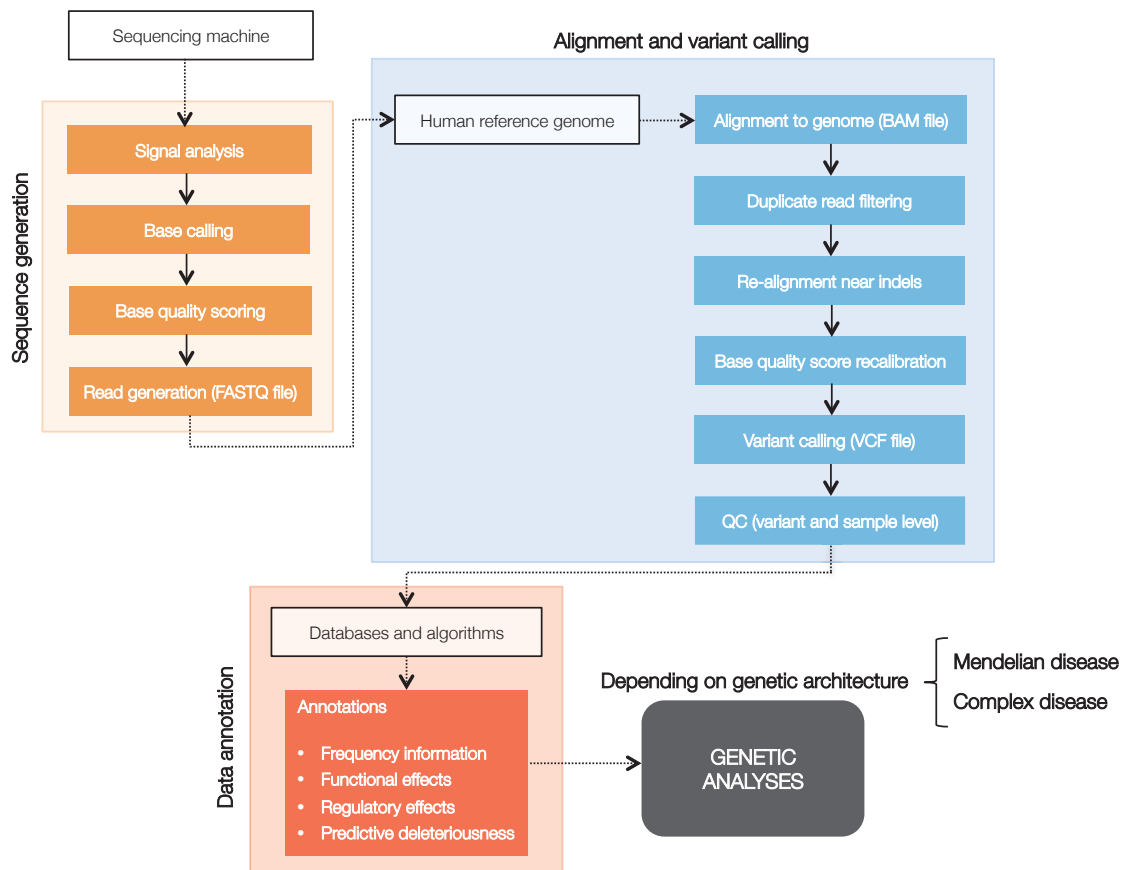


Figure 1.8 Flow diagram of the major computational steps involved in NGS data generation. The first step occurs inside the sequencing machine and involves the conversion of the raw imaging signal into sequencing reads. The second step is the alignment of the reads to the human genome, followed by several 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.

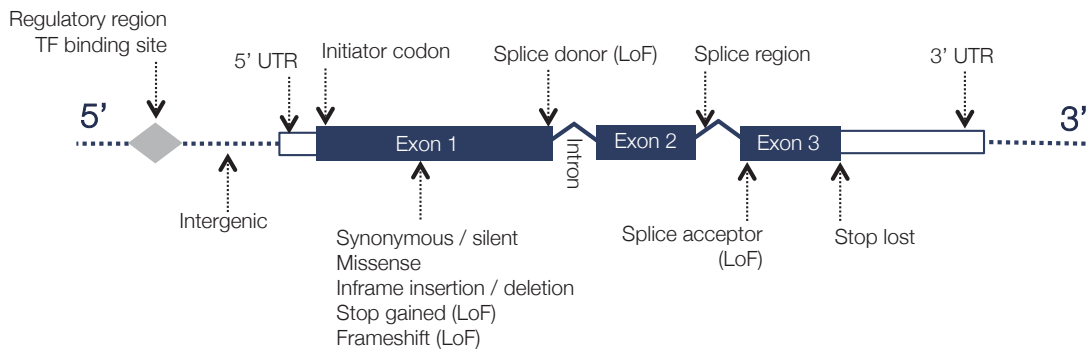


Figure 1.9 The impact of variants at the protein level. The diagram illustrates the set of functional consequence terms given by the Ensembl Variant Effect Predictor (VEP) tool [322]. A splice donor is splice variant that changes the invariable 2-base region at the 5' end of an intron. A splice acceptor 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.

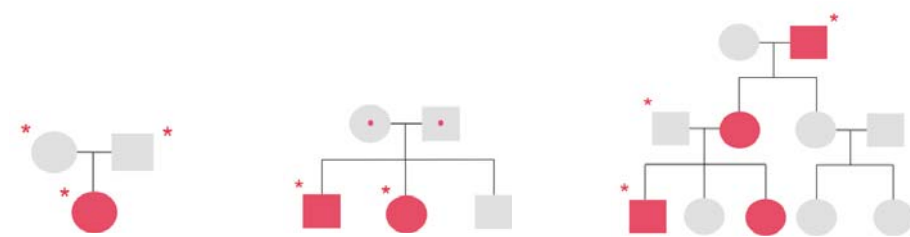
1.7 NGS genetic analyses in Mendelian diseases

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.

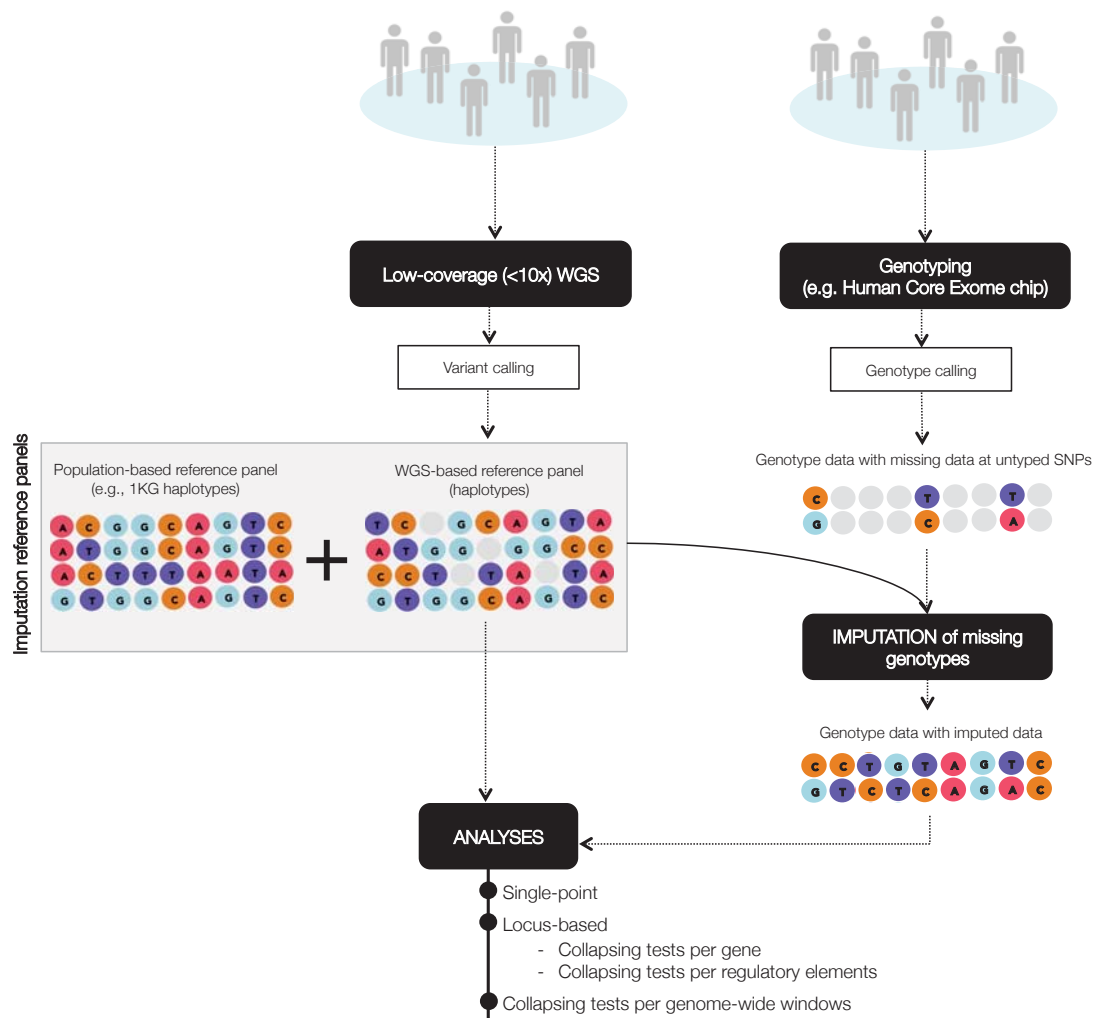


Figure 1.10 Diagram illustrating a popular study design now used in NGS-based complex disease studies. Low-coverage whole-exome sequencing is performed in as many cases and controls as possible. In parallel, additional cases and controls are genotyped for an ordinary GWAS array. The low-coverage sequences 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.

