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-TNFα 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 $P = 3.3\times10^{-9}$ 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.iscacon sortium.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 ∼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 samples 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 ∼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, nutricional 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 lauched 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.