Chapter 4

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

4.1 Introduction

4.1.1 What is inflammatory bowel disease?

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

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

	Crohn's disease (CD)	Ulcerative colitis (UC)
Incidence figures		
European rate	322 per 100,000 persons	505 per 100,00 persons
Gender	More common in woman	Equal rates in men and woman
Geographical prevalence	More prevalent in developed countries	More prevalent in still-developing countries
Disease pattern		
Location	Any part of the GI tract	Limited to the colon
Inflammation pattern	Patchy, discontinous inflammation	Continuous inflammation in the affected area
Histopathology		
Penetrance	Transmural inflammation - all layers of the bowel wall	Superficial inflammation - mucosa and submucosa
Appearance	Thickened colon wall with granulomas	Projecting masses of scar tissue - pseudopolyps
Complications	Abscess formation, fistula, strictures & colon cancer	Severe bleeding, toxic megacolon, rupture of the bowel, passage of mucus & colon cancer
Serological markers	Anti-*Saccaromyces cervisiae* antibodies	Anti-neutrophil cytoplasmic antibodies
Schematic		

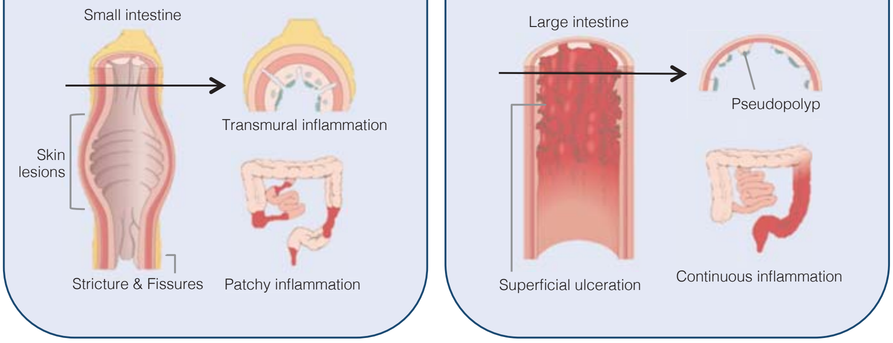


Figure 4.1 Epidemiological and clinical features of the two inflammatory bowel disease subtypes: Crohn's disease and ulcerative colitis [109] [35]. GI: gastrointestinal tract.

4.1.2 The genetics of IBD

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

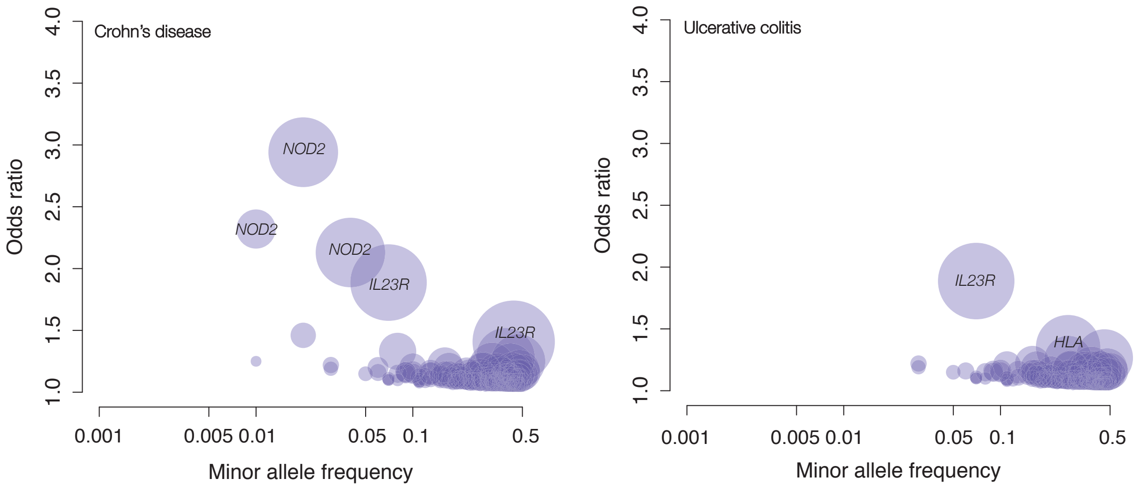


Figure 4.2 The genetic architecture of Crohn's disease and ulcerative colitis. Known CD/UC-associated variants are plotted according to their minor allele frequency and odds ratio (OR) [290]. The OR of protective alleles were inverted for illustrative purposes. The size of the circles represents the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

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

Biology	Context	Genes
Innate immunity	Provides an initial and quick response to microbes through pattern-recognition receptors	
Epithelial barrier function and repair	Maintains a physical and chemical barrier to commensal and pathogenic microorganisms	*HNF4A, CDH1, LAMB1, OSMR, ERRFI1, GNA12, ITLN1, MUC19, PLA2G2E, PTGER4, REL, STAT3, NKX2-3*
Innate mucosal defence	Cell-surface receptors or adaptor proteins involved in mediating innate immune response signalling	*NKX2-3, CARD9, FCGR2A, IL18RAP, ITLN1, NOD2, REL, SLC11A1*
Autophagy pathway	Intracellular degradation and recycling system for clearing intracellular organelles, proteins and macromolecular complexes. Crucial for cellular activity and protein recycling	*ATG16L1, CUL2, DAP, IRGM, LRRK2, NOD2, ATG4B, PARK7*
Apoptosis	Important mechanism of peripheral immune tolerance that controls programmed cell death of mucosa cells	*DAP, FASLG, MST1, PUS10, THDA, TNFSF15*
Adaptive immunity	Highly specialised cells and processes tailored to eliminate or prevent specific pathogens, while also developing immunological memory (through memory B and memory T cells)	
Activation		
IL23R pathway	Important for Th17 proliferation and survival. Crucial for host defence against foreign pathogens at the mucosal surface	*CCR6, IL12B, IL21, IL23R, JAK2, STAT3, STAT4, TYK2, PTPN2*
NF-kB pathway	Involved in cellular inflammatory responses in the mucosal environment, which leads to subsequent production of cytokines (TNF and IL-1β) and antimicrobial peptides (defensins) by T helper cells, subsequently activating the adaptive immune system	*NFKB1, REL, TNFAIP3, TNIP1, NFKBIZ, TNFSF15, TNFRSF9, RIPK2*
Aminopeptidases	Involved in the generation of HLA class I-binding peptides. HLA-I proteins display short peptides derived from pathogens in their cell surface for recognition by the appropriate T-cells	*ERAP1, ERAP2*
IL2 and IL21 dependent T-cell activation	Cytokine growth factors that optimise T-cell responses. Produced by CD4+ T-helper cells and CD8+ cytotoxic T-cells upon antigen-induced activation	*IL2, IL21, IL2RA*
Regulation		
Th17 cell differentiation	These cells are abundant in the intestine, especially in the terminal ileum and control epithelial cell proliferation, wound healing and the production of defensins	*AHR, CCR6, IL2, IL22, IL23R, IRF4, JAK2, RORC, STAT3, TNFSF15, TYK2*
T-cell regulation		*ICOSLG, IFNG, IL12B, IL2, IL21, IL23R, IL2RA, IL7R, NDFIP1, PIM3, PRDM1, TAGAP, TNFRSF9, TNFSF8, LY75, CD28, NFATC1, CCL20, IL10*
B-cell regulation		*BACH2, IKZF1, IL5, IL7R, IRF5, NFATC1, IL10*

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

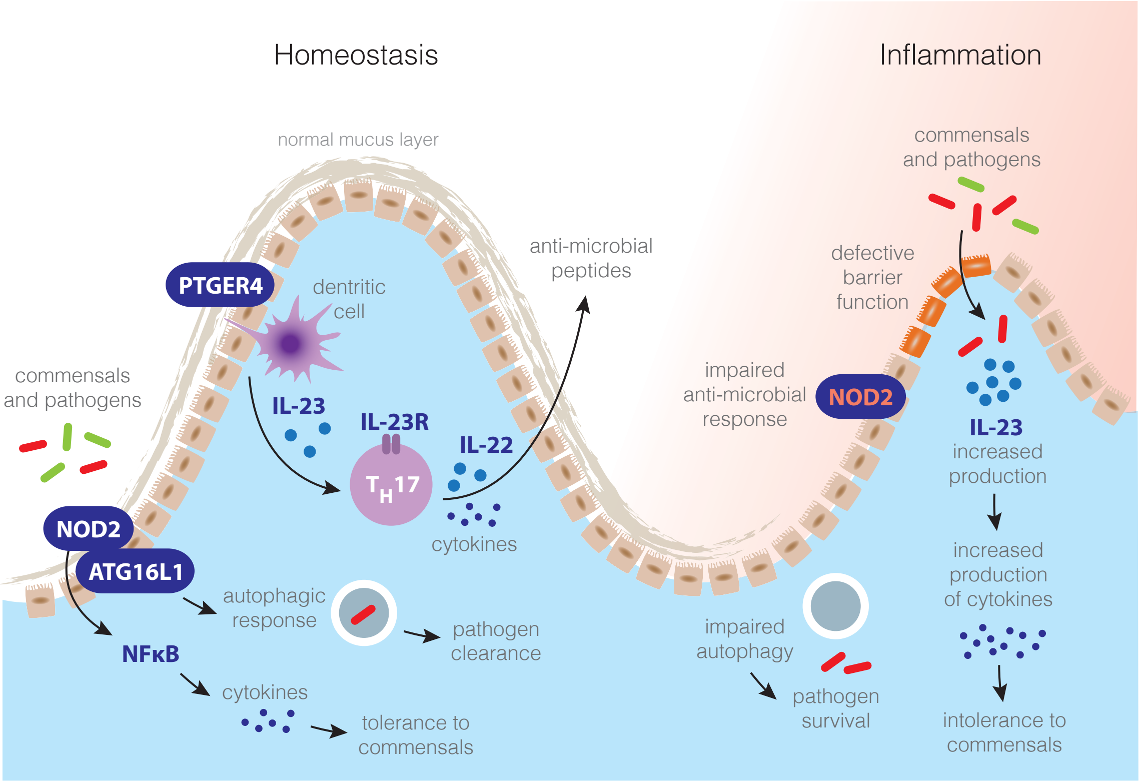
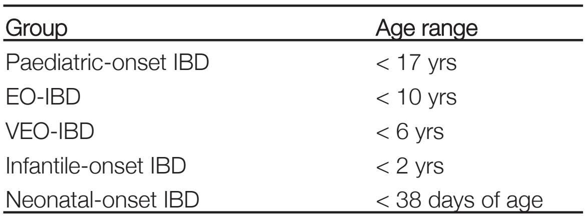


Figure 4.3 A simplified overview of the intestinal immunity in health and disease.

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

4.1.3 Paediatric-IBD

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



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

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

4.1.4 Very-early-onset IBD

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Monogenic disorders with IBD-like inflammation

Inspired by stories such as *XIAP*, *TTC7A* and *FOXP3*, there is now increasing awareness that several monogenic disorders present with overlapping pathology with CD or UC and, most frequently, their histological and endoscopic information does not allow a clear distinction to IBD [5]. These conditions have been termed to exhibit an 'IBD-like inflammation' with varying levels of penetrance of the IBD phenotype, which has been estimated to range from 2 to 30% [495]. A total of 59 IBD-like conditions have been identified and associated to IBD-like inflammation [495, 496], the majority of which represent PIDs caused by familial defects in key components of the immune system. Most of these abnormalities are recessively inherited (~72%) and can be divided into distinct subtypes depending on the biological mechanisms by which they affect intestinal immune homeostasis (**Figure** 4.4). Collectively, they disturb multiple layers of immune competence that severely compromise intestinal immunity.

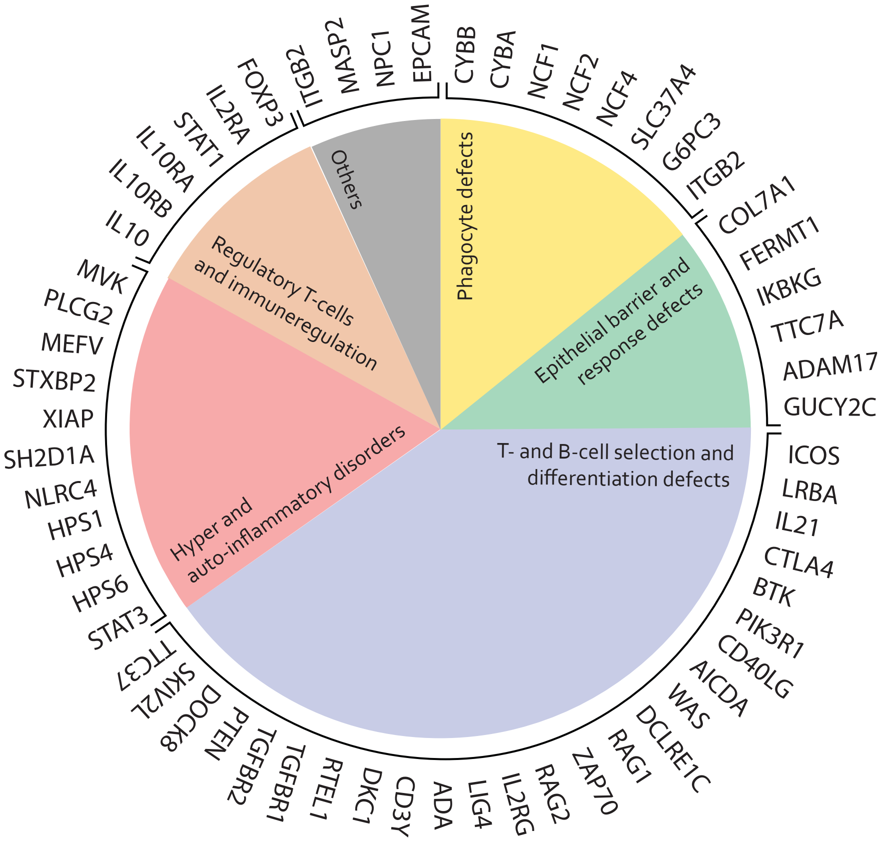


Figure 4.4 Monogenic defects (n=59) associated with VEO-IBD and IBD-like immunopathology stratified by biological category [495, 496].

It has been suggested that the intestinal immune dysfunction seen in these 59 conditions has similarities to those seen in VEO-IBD and that rare, penetrating defects within any of these loci may underlie the many VEO-IBD cases that still await a genetic diagnosis [495]. Following this hypothesis, recent guidelines on the diagnostic approach to VEO-IBD recommended excluding other possible causes of early-onset inflammation, such as these immune function defects [496], which would also allow for more targeted treatment strategies. Given that adult-IBD associated loci are enriched 4.9-fold for PID genes [232], the importance of these 59 loci may not be limited to the VEO-IBD phenotype, but may extend to the biology of IBD in general, albeit with a different magnitude of effect. Moreover, a substantial proportion of proteins encoded by the genes mutated in these disorders directly, or indirectly, interact with loci that confer susceptibility to IBD, suggesting common signalling pathways predisposing to colitis [495]. Examples include *XIAP* and *IKBKG*, which interact with IBD-loci such as *RIPK2* and *NOD2* [256], as well as *CYBA* and *CYBB* that interact with *RELA* and *NFBK1* [196].

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

4.1.6 Another hypothesis for the aetiology of VEO-IBD

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

Two subsequent GWAS, focusing solely on children with a mean age of 12 years but less than 18 years, confirmed the association of *NOD2* and *IL23R* and identified children-specific loci overlapping with *TNFRSF6B*, *PSMG1* and *IL27* [222, 261]. However, all of these associations have been subsequently replicated in a larger meta-analysis of adult IBD cohorts [153], leaving no paediatric-specific loci behind.

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

4.2 Aims

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

4.3 Colleagues

All the work presented in this chapter is my own work, unless otherwise stated. This research was carried out under the supervision and guidance of Dr Carl Anderson at the Wellcome Trust Sanger Institute (WTSI). This work was done in close collaboration with other colleagues at the University of Oxford, namely Professor Holm Uhlig and Dr Tobias Schwerd.

4.4 Methods

4.4.1 Patients

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

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

4.4.2 Controls

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

4.4.3 Exome sequencing and variant calling

Whole-exome sequencing of both cases and controls was performed and processed at the WTSI by the Sanger Institute Core Sequencing pipeline. Genomic DNA (1-3µg) extracted from blood was sheared to 100-400bp using a Covaris E210 or LE220 (Covaris, Woburn, Massachusetts, USA). Sheared DNA was subjected to Illumina paired-end DNA library preparation and enriched for targeted sequencing (Agilent Technologies, Santa Clara, CA, USA; Human All Exon 50 Mb – ELID S04380110) according to the manufacturer's recommendations (Agilent Technologies, Santa Clara, CA, USA; SureSelectXT Automated Target Enrichment for Illumina Paired-End Multiplexed Sequencing). Enriched libraries were sequenced (eight samples over two lanes) using the HiSeq 2000 platform (Illumina) as paired-end 75 base reads according to the manufacture's protocol. The Burrows-Wheeler Aligner [280] was used for alignment to the human reference genome build UCSC hg19/Grch37. Variants were first called at the single sample level using GATK Haplotype Caller (version 3.4) [116] and then joint-called across all cases and controls using GATK CombineVCFs and GenotypeVCFs at default settings. These steps were performed by the Human Genetics Informatics team at the WTSI.

4.4.4 Data quality control

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

Individual-level QC

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

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

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

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

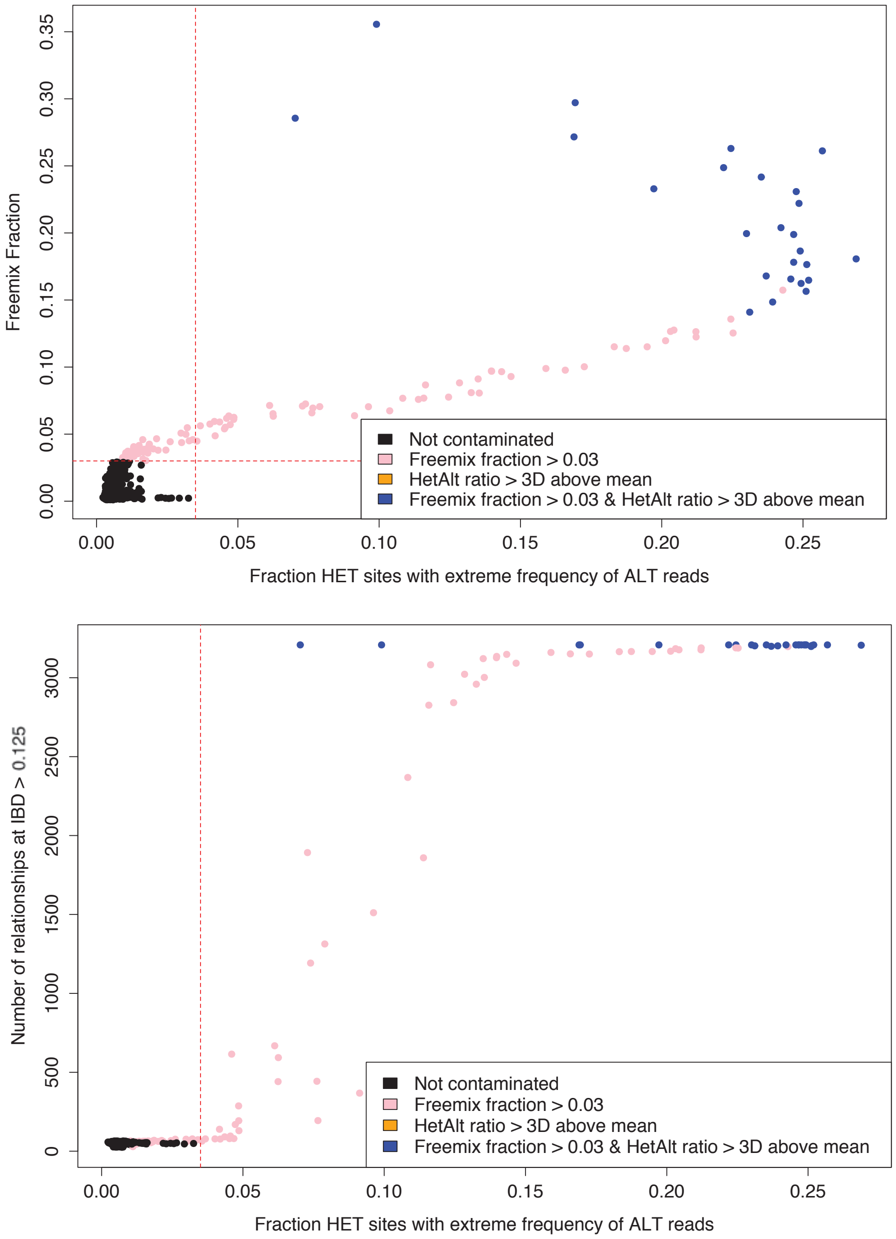


Figure 4.5 Contamination metrics. **A)** Fraction of heterozygous sites with extreme frequency of alternative reads vs. the freemix fraction. A heterozygous (HET) site was termed to have an extreme frequency of alternative (ALT) reads if the frequency of ALT reads were greater than 0.8 or lower than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

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

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

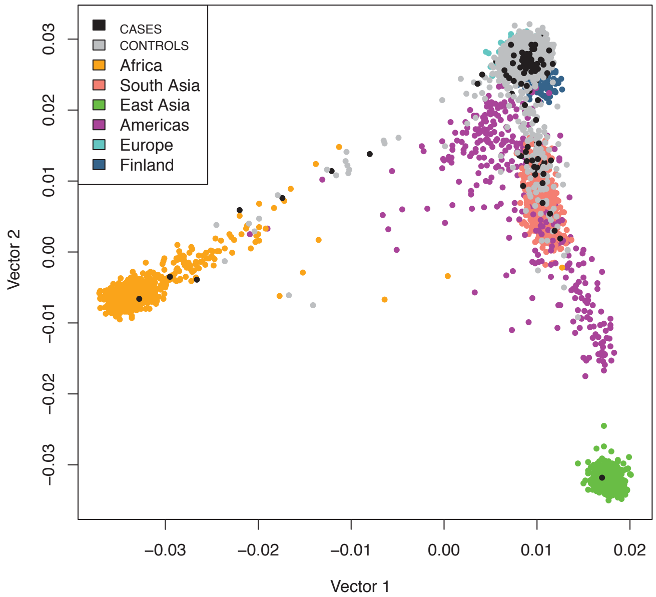


Figure 4.6 Principal component analysis (PCA) of VEO-IBD cases and INTERVAL controls with 1KG Phase 3 reference populations.

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

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

A total of 291 controls were outliers for at least one of these metrics, and one case had a considerably high rate (~0.07) of non-called genotypes (data not shown). All of these samples were removed from the dataset. Importantly, this analysis also revealed cases and controls were sequenced at mean depths of 69x and 53x, respectively, which represented a significant difference (P-value<2.2x10^{-16}) that needed to be corrected, if possible, with further variant-QC (explained below), or accounted for in downstream analyses.

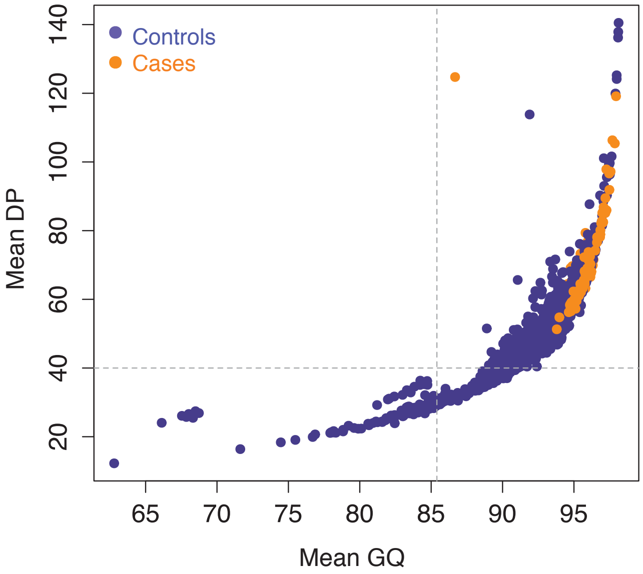


Figure 4.7 Mean genotype quality (GQ) and mean depth (DP) per sample. Grey dashed lines represent the applied thresholds of GQ=85.4 and DP=40. The GQ threshold represented 3SD from the mean and the DP threshold was empirically derived.

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