**UNIVERSITY OF SOUTHAMPTON**

FACULTY OF MEDICINE

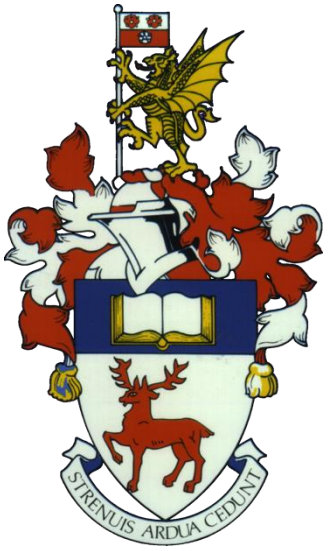
Human Development and Health

# GENOMIC ANALYSES OF PAEDIATRIC INFLAMMATORY BOWEL DISEASE

Thesis for the degree of Doctor of Philosophy

by

Gaia Andreoletti




Supervisory team: Prof. Sarah Ennis, Dr Mark R Beattie, Dr Jane Gibson and Prof. Andrew Collins

May 2016

# University of Southampton

## Abstract

FACULTY OF MEDICINE

Human Genetics

Doctor of Philosophy

**GENOMIC ANALYSES OF PAEDIATRIC INFLAMMATORY BOWEL DISEASE**

Gaia Andreoletti

Ulcerative colitis, Crohn's disease and indeterminate colitis are forms of inflammatory bowel disease (IBD) an inflammatory autoimmune disorder of the gastrointestinal tract. It is believed the disease arises from interaction of environmental and genetics triggers in genetically susceptible individuals. Children present with a more severe phenotype compared with adults and are postulated to harbour a stronger genetic component to their disease with less environmental influence compared to adults.

Next generation sequencing (NGS) has become a feasible method for studying the missing heritability of complex diseases not explained by previous genetic studies. Analysis of exome data in complex disease is not yet routine; however, assessing the pathogenic variants specific to individual patients has the potential to uncover generic immune inadequacy and may, in the future, help guide personalised treatments.

The work presented herein is based on one of the largest reported cohort of children with IBD with whole exome sequencing analysis (n=147). All children (aged < 18) were recruited following diagnosis by the paediatric gastroenterology service at University Hospital Southampton. Despite a modest sample size, this paediatric cohort enabled us to conduct adequately-powered association analysis and identify novel disease genes.

After reviewing the current state of art of IBD, I herein present four analytical chapters. The first study investigates the incidence of comorbidities within the Southampton paediatric IBD (pIBD) cohort. This analysis aimed to provide insight into the relationship between paediatric IBD and concurrent autoimmune diseases in children. From the analysis, forty-nine (28.3%) pIBD (18.49%CD, 8.6% UC and 2 1.15% IBDU patients) had a concurrent clinical diagnosis of at least one other autoimmune

disorder. Asthma was the most prevalent, affecting 16.2% of the pIBD cohort. For a subset of patients with pIBD and concurrent asthma (n=18), exome data was interrogated to ascertain the burden of pathogenic variants within the 49 genes implicated in asthma. Association testing was conducted between cases and population controls using the SKAT-O test. Rare and common variant association testing revealed six significant genes ($p< 0.05$) prior to Bonferroni adjustment. Three of these genes were previously implicated in both asthma and IBD (*ZPBP2 IL1R1* and *IL18R1*) and three in asthma only (*PYHIN1*, *IL2RB* and *GSTP1*). By interrogating the exome data for a subset of children we were able to show that for a group of patients the relationship between concurrent pIBD and asthma could be caused by a systemic immune dysregulation rather than organ specific immune dysfunctions.

The second research project describes the application of astatistical test of rare variation within the 41 genes involved in the NOD signalling pathway. In this analysis we used a discovery cohort composed of 136 pIBD and 106 control samples. We compared the burden of common, rare and private mutation between these two groups using the SKAT-O test. An independent replication cohort of 33 cases and 111 controls was used to validate significant findings. Variation was observed in 40 of 41 genes comprising the NOD signalling pathway. Four genes were significantly associated with disease in the discovery cohort (*BIRC2* p=0.004, *NFKB1* p= 0.005, *NOD2* p=0.029 and *SUGT1* p=0.047). Statistical significance was replicated for *BIRC2* (0.041) and *NOD2* (p=0.045) in the independent validation cohort. Six variants within *BIRC2* were observed in the discovery cohort across 15 cases and 4 controls. The evidence contributing to the association signal for these genes is primarily driven by rare variants that would not have been assessed in array-based studies. The identification of *BIRC2* as a novel pIBD gene provides a wider role for the inhibitor of apoptosis gene family in IBD pathogenesis and overall this data demonstrates the potential utility of rare genetic variation to stratify complex diseases.

The third project details the analysis conducted between the Southampton Informatics group and the University of Stanford (USA) in order to comprehensively investigate the role of *HSPA1L* in IBD. The Stanford group performed a family-based whole-exome sequencing analysis on an index family (Family A) and identified a potential causal

mutation within *HSPA1L*. We subsequently analysed exome data from the Southampton paediatric cohort (136 patients and 106 controls) to further investigate mutations in the candidate gene *HSPA1L*. Biochemical assays on *de novo* and rare (MAF<0.01) mutation variant proteins further validated the predicted deleterious effects of the identified alleles. In the proband of Family A, a heterozygous *de novo* mutation (c.830C>T; p.Ser277Leu) in *HSPA1L* was found. Through analysis of exome data from our cohort of 136 patients, we identified five additional rare *HSPA1L* mutations (p.Gly77Ser, p.Leu172del, p.Thr267Ile, p.Ala268Thr, p.Glu558Asp) in six patients. In contrast, rare *HSPA1L* mutations were not observed in controls, and were significantly enriched in patients (p=0.02). Biochemical assays revealed that all six rare *HSPA1L* variants proteins showed decreased chaperone activity *in vitro*. These results indicated that *de novo* and rare mutations in *HSPA1L* might be associated with IBD and provide insights into the pathogenesis of IBD, as well as expand our understanding of the roles of heat shock proteins in human disease.

The fourth project discuss the application of the American college of medical genetics guidelines to efficiently detect pathogenic mutations across the 51 genes known to cause a monogenic form of IBD. Whole exome data for all 147 children with IBD was interrogated for extract variation within the monogenic genes. 574 variants were identified across 51 genes in all 147 patients. Subsequently, variants were categorised in line with ACMG guidance to remove benign variants and to identify 'pathogenic' and 'likely pathogenic' variants. In six patients we observed six pathogenic variants of which *CYBA*(c.287+2T>C), *COL7A1*(c.6501+1G>C), *LIG4*(p.R814X), and *XIAP*(p.T470S) were known causative mutations and *FERMT1*(p.R271Q ) and *SKIV2L*(c.354+5G>A ) were novel. In the three patients with *XIAP*, *SKIV2L* and *FERMT1* variants, individuals' disease features resembled the monogenic phenotype. This was despite apparent heterozygous carriage of pathogenic variation for the latter two genes. The *XIAP* variant was observed in a hemizygous male. This research project demonstrates the power of whole-exome sequencing to identify known and *novel* potentially causative mutations in genes associated with monogenic IBD. Whilst these are rare conditions it is important to identify causative mutations early in order to improve prognosis. We postulate that in a subset of IBD, heterozygous mutations (in genes thought to manifest IBD through autosomal recessive inheritance) may contribute to clinical

presentation.

In conclusion, with the rapid increase of the application of the NGS technologies, and consequently the increasing number of genomic data produced, there is the need to better functionally annotate variants and to conduct functional analysis in order to truly assess the causality of mutations on the phenotype. Ultimately, the combination of genomic, transcriptomic and functional information on a case-by-case basis will lead provide further insight into disease mechanism and bring us closer to more effective personalised treatment plan.

# Table of contents

## Abbreviations

| | |
|---|---|
| 1KG | 1000 Genome project |
| 46CG | 46 complete genomics |
| 5-ASAs | Aminosalicylic acids |
| ACMG | The American College of Medical Genetics and Genomics |
| AIH | Autoimmune hepatitis |
| AJs | Adherens junction |
| AMPs | Antimicrobial peptides |
| ATP | triphosphate |
| AZA | Azathioprine |
| BAM | Binary Alignment/Map files |
| BIR | Baculovirus inhibitior of apoptosis protein repeat |
| CADD | Combined Annotation Dependent Depletion |
| CAGI | Critical Assessment of Genome Interpretation |
| CD | Crohn's disease |
| CD/CV | Common Disease/Common Variant |
| CNVs | Copy number variations |
| Condel | CONsensus DELeteriousness |
| DAVID | Database for Annotation, Visualization and Integrated Discovery |
| DDD | Deciphering Developmental Disorders |
| ddNTP's | Dideoxynucleotides |
| EC-IBD | Europe a prospective study |
| EDTA | Ethylenediamine tetraacetic acid |
| ELB | Erythrocyte lysis buffer |
| EO-IBD | Early-onset IBD |
| EVS | Exome Variants Server |
| FATHMM | The Functional Analysis through Hidden Markov Models |
| Fi/fd | Frameshift isnertion/deletion |
| G | Genome |
| GATK | The Genome Analysis Toolkit |
| GERP | Genomic evolutionary rate profiling |
| GI | Gastrointestinal tract |

| | |
|---|---|
| GWAS | Genome wide association studies |
| Hg19 | Human genome 19 |
| HGP | The human genome project |
| HLA | human leucocyte antigen |
| HPMR | Hyperphosphatasia mental retardation |
| HSF1 | transcription factor heat shock factor 1 |
| HSP70 | 70-kD heat shock protein family |
| HSPs | Heat shock proteins |
| IBD | Inflammatory Bowel disease |
| IBDU | Inflammatory bowel disease unclassified |
| IC | Indeterminate colitis |
| IGV | Integrative Genomic Viewer |
| IL10 | Interleukyn 10 |
| IL10RA/B | Interleukin recepto A/B |
| indels | Insertions/deletions |
| KEGG | Kyoto Encyclopedia of Genes and Genomes |
| LOD | logarithm of the odds |
| LRT | Likelihood ratio test |
| MAF | Minor allele frequency |
| MAPK | Mitogen-activated protein kinases |
| MAPP | Multivariate Analysis of Protein Polymorphism |
| miRNA | Micro RNA |
| MP | Mercaptopurine |
| MS | Multiple sclerosis |
| N | Number |
| NGS | Next generation sequencing |
| NICE | National Institute for Health and Care Excellence |
| NLR | NOD-like receptors |
| Ns | Non-synonymous |
| OH | Hydroxyl group |
| PCA | Principle component analysis |
| PhD-SNP | Predictor of human Deleterious Single Nucleotide Polymorphisms |
| PhyloP | phylogenetic P-values |

| | |
|---|---|
| pIBD | Paediatric IBD |
| PNG | Peptidoglycan |
| PolyPhen | Polymorphism Phenotyping |
| PSC | Primary sclerosing cholangitis |
| RA | Rheumatoid arthritis |
| RAPID | Resource of Asian Primary Immunodeficiency Diseases |
| REC | Research Ethics Committee |
| rpm | Rotation per minute |
| SAM | Sequence Alignment/Map |
| SIFT | Sorting Intolerant From Tolerant |
| SKAT | Sequence Kernel Association Test |
| SKAT-O | Sequence kernel association testing optimal unified test |
| SLE | Systemic lupus erythematosus |
| SLR | Sitewise likelihood-ratio |
| Sn | Synonymous |
| SNPs | Single nucleotide polymorphisms |
| Sp | Splicing |
| TLR | Toll-like receptor |
| TNFα | Tumor Necrosis Factor alpha |
| UC | Ulcerative colitis |
| UCSC | University of Santa Cruz |
| UHS | University Hospital Southampton |
| UTRs | Untranslated regions |
| VCF | Variant Call Format |
| VEO-IBD | Very early onset |
| VUS | Variants of unknown significance |
| WES | Whole exome sequencing |
| WGS | Whole genome sequencing |
| WT | Wild type |
| WTCCC | Wellcome Trust Case Control Consortium |
| XIAP | X-linked inhibitor of apoptosis |

## List of figures

## List of tables

## Published papers

- Andreoletti G, Ashton JJ, Coelho T, *et al.* **Identification of novel and known variants in genes associated with monogenic inflammatory bowel disease by whole exome sequencing in a paediatric IBD cohort.** *Inflammatory Bowel Diseases.* (2016)
- Seaby E, Gilbert R, Pengelly R, Andreoletti G, *et al* **Exome sequencing reveals the genetic cause of myoclonic epilepsy associated with Fanconi syndrome** *Journal of the Royal Society of Medicine* (2016, *accepted*)
- Andreoletti G, Ashton JJ, Coelho T *et al*. **Exome analysis of patients with concurrent pediatric inflammatory bowel disease (PIBD) and autoimmune disease** *Inflammatory Bowel Disease.* (2015)
- Coelho T, Andreoletti G, Ashton JJ *et al.* **Immuno- Genomic Profiling of Patients with Inflammatory Bowel Disease: A Systematic Review of Genetic and Functional in vivo Studies of Implicated Genes**; *Inflammatory Bowel Diseases.* (2014)
- Pengelly RJ, Gibson J, Andreoletti G, Collins A, Mattocks CJ, Ennis S. **A SNP profiling panel for sample tracking in whole-exome sequencing studies.** *Genome Med*. (2013)

## Submitted papers

- Andreoletti G, Shakhnovich V, Christenson K, *et al.* **Exome Analysis of Rare and Common Variants within the NOD Signaling Pathway.** *Submitted to Scientific reports*
- Andreoletti G, Takahashi S, Chen R, *et al.* **De Novo and Rare Mutations in the HSPA1L Heat Shock Gene Associated with Inflammatory Bowel Disease.** *Submitted to Genome Medicine*
- Seaby EG, Gilbert RD, Andreoletti G, *et al.* **Pre-clinical cancer: an example of how genomics could reshape the clinical prognostic paradigm** *review to BMJ case report*
- Andreoletti G, Seaby EG, Dewing JM, *et al.* ***AMMECR1*: a single point mutation causes mental retardation, midface hypoplasia, and elliptocytosis.** *Submitted to J Med Genet*
- Andreoletti G, Coelho T, Ashton JJ *et al.* **Genes implicated in thiopurine-induced toxicity: Comparison of TPMT phenotype with clinical phenotype and exome data in a paediatric IBD cohort.** *Submitted to Scientific reports*

## Papers in progress

- Andreoletti G, Mercer CL, Carroll A, *et al.* **Familial Ebstein's anomaly: whole exome sequencing identifies novel phenotype associated with *FLNA*.**
- Andreoletti G, Douglas AGL, Talbot K, *et al.* ***ADCY5*-related dyskinesia presenting as familial myoclonic dystonia**

## Acknowledgements

I would like to acknowledge all those who helped me with my research; first of all Prof Sarah Ennis, Dr Jane Gibson, Prof Andrew Collins and Dr William Tapper for the help and support given me over these three year.

I would like to express my gratitude towards all the patients and their families that have taken part to the research study. I would like to thank the University Hospital Southampton NHS Foundation Trust in specific Dr Mark Beattie, Dr Tracy Coelho and Liz Blake, Senior Paediatric Research Sister, and Rachel Haggarty, Senior Children's Research Nurse for their help in recruiting the patients and families.

Special thanks also to Nikki Graham, Sylvia Diaper and the Iridis team at Southampton University for their support in the wet lab analysis and computational support.

I also wish to thank the many friends I have made at the University of Southampton for their support and for successfully distracting me from my work on many occasions during the last years.

## Ethics approval

This study was approved by the Southampton & South West Hampshire Research Ethics Committee (REC) (09/H0504/125).

## Funders

Heartfelt thanks to Crohn's in Childhood Research Association (CICRA) and the Gerald Kerkut Charitable Trust for funding my PhD project.

*Alla mia mamma e al mio papà*

# DECLARATION OF AUTHORSHIP

I, .....[please print name]

declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

[title of thesis] .....

.....

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. [Delete as appropriate] None of this work has been published before submission [or] Parts of this work have been published as: [please list references below]:

Signed: .....

Date:

# Chapter 1 **Introduction**

This chapter will introduce concepts regarding the aetiology of inflammatory bowel disease and the analysis methods used within this thesis to explore whole-exome sequencing data. The analyses presented in this thesis are based on the largest reported cohort of children with IBD with whole exome sequencing analysis. Using this cutting edge sequencing data, it is possible to further understand the genetic background of this complex autoimmune condition.

The first section represents a discussion of the three major forms of IBD, their phenotypic classification, the epidemiology of IBD, clinical presentation, current treatments and environmental and genetics triggers of IBD. The second section of this introductory chapter deals with genetic variation in human disease and the past and present methods for analysing genomic data.

## 1.1 Inflammatory bowel disease

Inflammatory Bowel disease (IBD) is the umbrella term that encompasses a group of diseases including ulcerative colitis (UC), Crohn's disease (CD) and indeterminate colitis (IC) which are inflammatory auto-immune diseases of the gastrointestinal tract. The aetiology of IBD is still unknown but it is accepted that the disease occurs in genetically predisposed individuals with an inappropriate immune response to the normal gut flora[1] (Figure 1.1). CD and UC, the two major forms of IBD, are characterised by differences in location of the lesions within the gastro-intestinal tract; in the type of pattern of inflammation (continuous or discontinuous); the macroscopic histological differences of the lesions. Clinical, endoscopic and histological caracteristics of CD and UC are given in Table 1.1.

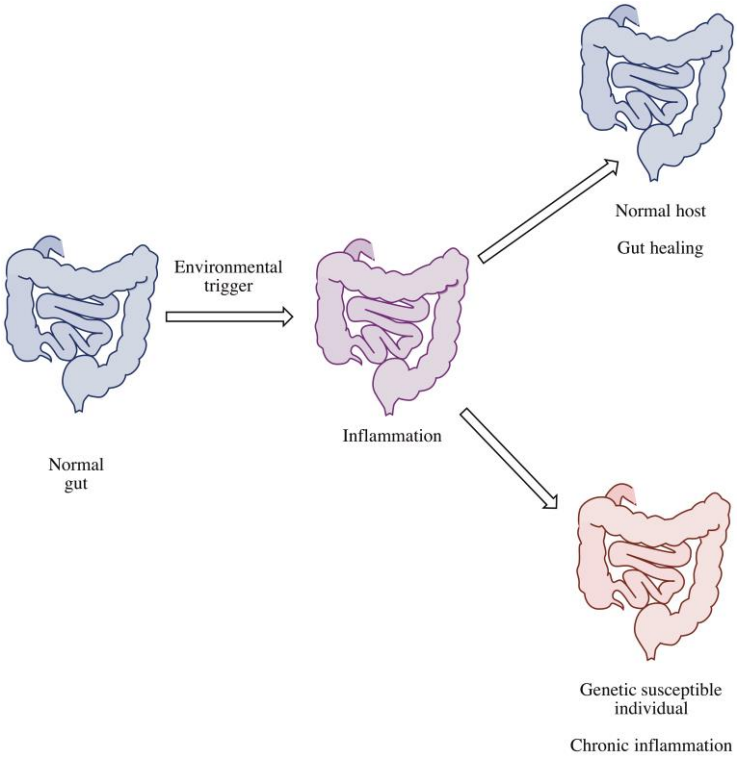



**Figure 1.1 Overview of the Pathogenesis of IBD**. The aetiology of IBD is still unknown but it is accepted that the disease arises from the contribution of genetics and environmental triggers (for example diet and smoking) in genetically predisposed individuals.

**Table 1.1 Endoscopy and histology in inflammatory bowel disease according to the Porto Criteria**

| | **Crohn's disease** | **Ulcerative colitis** |
|---|---|---|
| Endoscopy and visualization of oral and perianal regions | Ulcers | Ulcers |
| | Cobblestoning | Erythema |
| | Skip lesions | Loss of vascular pattern |
| | Strictures | Granularity |
| | Fistulas | Spontaneous bleeding |
| | Abnormalities in oral or perianal regions | Pseudopolyps |
| | Any part of the gastrointestinal tract may be involved, most frequently the terminal ileum and colon | Inflammation is limited to rectal (proctitis) and colonic mucosal layers |
| | Segmental distribution | Continuous with variable proximal extension from rectum |
| Histology | Submucosal or transmural involvement | Mucosal involvement |
| | Ulcers: crypt distortion | Crupt distortion |
| | Crypt abscess | Crypt abscess |
| | Granulomas | Goblet cell depletion |
| | Focal changes | Mucin granulomas |
| | Patchy distribution | Continous distribution |

### 1.1.1 Crohn's disease

Crohn's disease can occur in any part of the digestive system, from the mouth to the anus and is characterised clinically by abdominal pain, loss of weight and diarrhoea[2]. The inflammation may occur in patches in one or several organs in the digestive system and is characterised by macroscopic lesions on the mucosa, strictures, fistulas and ulcers. Fistulai are defined as abnormal communications between the lumen of the gut and/or another organ. Fistulas are often associated with strictures which are luminal narrowing and bowel wall thickening.

Currently there are five classifications of CD based on the anatomical location affected. The most common form of CD in the adult population is ileocolitis which affects the terminal ileum and the colon in 47% of cases[3]. For the second most common type of CD, 28% of the cases are due to ileitis which affects only the ileum. 21% of CD diagnoses are of gastroduodenal Crohn's disease involving the stomach and duodenum. Rarer forms of CD are jejunoileitis, affecting the jejunum, the upper half of the small intestine, and Crohn's (granulomatous) colitis affecting the colon only for 3% of cases[4]. In the paediatric population the disease is commonly colonic (>50% of cases) and affects the upper gastrointestinal tract: stomach (67% of the cases), oesophagus (54% of the cases) and duoduodenum (22% of the cases)[3]. CD can also be distinguished by the behaviour of the disease and classified as non-stricturing and non-penetrating (70% of cases), stricturing (17% of cases), and penetrating in 13% of the patients diagnosed[5]. A study conducted on the Southampton paediatric IBD cohort dealing with the comparison of histological and endoscopic findings revealed that across 107 CD cases the most common location for disease was the descending colon (69%) and ascending colon (69%)(Unpublished data). The ileum was involved in 49% of cases, oesophagus, stomach and duodenum in 18.4%, 42.9% and 27.6% of cases respectively. With regards to the histological findings 85.9% of cases had stomach involvement, 72% ileal involvement and 3% of the cases had isolated upper GI disease (Unpublished data).

### 1.1.2 Ulcerative colitis

Ulcerative colitis is an idiopathic, chronic inflammatory disorder of the colonic mucosa, which starts in the rectum and generally extends in a continuous manner through part

of, or the entire, colon[6]. UC is characterised microscopically by diffuse inflammatory lesions and distortion of the villi and lamina propria (Figure 1.2).

In the adult population the most common form of UC is ulcerative proctitis which is limited to the rectum affecting 40-50% of the cases. In 30-40% of the patients the disease affects the rectum and the sigmoid colon, called proctosigmoiditis, while in a further 20% the disease affects the rectum and extends to the descending colon, defined as left-sided colitis. Contrary to CD, the extension of UC is continuous and can affect the whole colon, pancolitis, in 10-20% of the patients[7]. In contrast to the adult population at the time of presentation 44-49% of the paediatric cases present with UC in the rectum, 36-41% in the colon and the rectum and 14%-37% in the ascending colon[8]. Histological examination of UC biopsies shows alteration of the shape of the crypt with an increase in the lamina propria of inflammatory cells. The lamina propria underlies the epithelium with a rich vascular and lymphatic network. The crypts are tubular invaginations of the epithelium around the villi involved primarily in secretion. Stem cells are at the base of the crypts providing the source of all the epithelial cells in the crypts and on the villi[7]. An unpublished study conducted on the Southampton paediatric IBD cohort revealed that across 50 UC patients the most common disease location was the stomach (40%) and duodenum (2.2%). A more extensive histological disease was observed in the rectum, descending colon, transverse colon and ascending colon in 93.6%, 95.7%, 93.3% and 86% of the patients respectively.

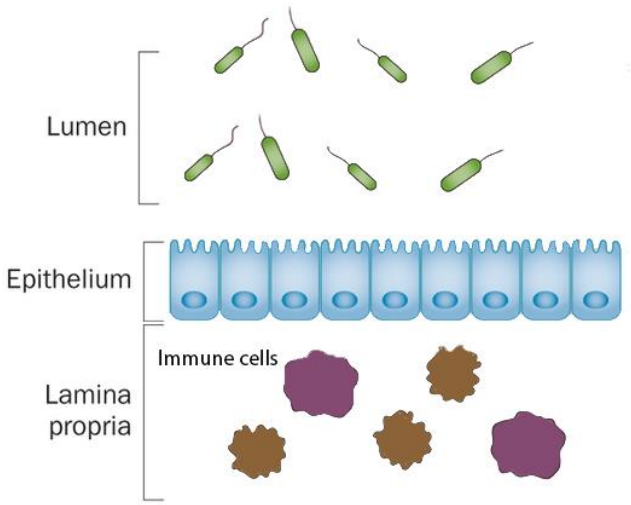



**Figure 1.2 The intestinal epithelium**. The intestinal epithelium is organized into crypts and villi. The epithelium is composed of two cell types: enterocytes and globet cells. The lamina propria is the connective tissue that underlies the epithelium and is where relies the immune cells which protect from host-bacteria[7].

### 1.1.3 Inflammatory bowel disease unclassified

A clear distinction between CD and UC is not always possible, for this reason some patients with colonic inflammation and features of both Crohn's disease and ulcerative colitis are given a diagnosis of indeterminate colitis (IC) also called colonic inflammatory bowel disease unclassified (IBDU)[9]. Patients with IBDU present with colonic disease and endoscopic/histological features that make the diagnosis of UC or CD uncertain. By definition IC is a disease limited to the colon but it also includes architectural distortion with inflammatory features of the mucosa, which makes the distinction difficult between UC and CD. A study performed on the Southampton cohort of paediatric IBD revealed that across 15 IBDU patients the most common location of disease was the sole rectum (86.7%) whereas the ileum was the location of disease in only in five cases (33%). Histologically 93.3% of cases presented pancolitis and no patients had granulomas. These results indicate the challenge in confirming CD or UC in these patients, as there is a large variation of disease location at an endoscopic level and the high rate of pancolitis at histology. Because a standard definition and criteria for diagnosing IBDU does not exist, the understanding and natural history of this condition is controversial and the opinions of individual clinicians vary.

Infantile (diagnosed before 1 year of age) and very early onset (VEO)-IBD (diagnosed age less than 6 years) are more often diagnosed with IC as they present with severe growth failure, extensive colonic inflammation with lack of small bowel disease, and poor responsiveness to conventional therapies[9]. In the majority of patients diagnosed with IC the diagnosis changes over time as disease develops clearer features of an early diagnosis of either CD or UC[10]. One of the hypotheses related to IBD is that it is not a distinct disease entity, but IBDU represents a provisional descriptive term used until the true condition of the type of IBD becomes more clear, usually within a few years after diagnosis.

### 1.1.4 Phenotypical classifications

Although the clinical presentation of CD and UC is highly variable with significant diversity in phenotypes of the diseases, a comprehensive phenotype classification for characterising paediatric IBD is still not available. This diversity in adults is specified by differences in the location, the natural history and the outcome. To overcome this

question several classification methods were designed to classify IBD using clinical and epidemiological features. The latest classification for disease is the Paris classification (2011), Table 1.2 and Table 1.3, which revised the previous Rome (1991), Vienna (1998) and Montreal (2003) classifications. Since its introduction the Montreal system has brought important changes in defining the age of diagnosis, the location and the behaviour for CD and the extent and severity for UC. Although one of the merits of the Montreal classification was the introduction for the first time of a separate category for the diagnosis of paediatric patients with age less than 16 years[11], it does not capture entirely the dynamic features of paediatric IBD[9] as it was not designed or validated for paediatric patients[10]. Recent reports of the outcomes of large cohorts of children with IBD have helped clarify the differences in the disease phenotype of children presenting with IBD compared with adults, and have described the evolution of these cases over time[12]. These differences have prompted a proposed extension of the standard Montreal classification of the disease phenotype[11]. The revised Paris classification introduced modifications to age classification for children under the age of 10 years, to enable targeted studies on early onset IBD[13]. Specifically the Paris classification has recognised growth failure as a separate phenotypic description and defines early-onset IBD (EO-IBD) as diagnosis before the 18$^{th}$ year, very early-onset IBD (VEO-IBD) as diagnosis before the 6 years and infantile-onset IBD as diagnosis before one year old[13]. However, the Paris classification presents drawbacks as it is only based on endoscopic findings and does not take into account histological extent of the disease. Differences between the Paris and the Montreal classifications for UC and CD are displayed in Table 1.2 and Table 1.3.

**Table 1.2 Paris and Montreal Classification for Crohn's disease[13].**

| | Montreal Classification | Paris Classification |
|---|---|---|
| Age at diagnosis | A1: below 17 years<br>A2: 17–40 years<br>A3: above 40 years | A1a: 0–<10years<br>A1b: 10–<17 years<br>A2: 17–40 years<br>A3: >40 years |
| Location | L1: terminal ileal/ limited cecal disease<br>L2: colonic<br>L3: ileocolonic<br>L4*: isolated upper disease | L1: distal 1/3 ileum<br>6 limited cecal disease<br>L2: colonic<br>L3: ileocolonic<br>L4a: upper disease proximal to ligament of Treitz*<br>L4b: upper disease distal to ligament of Treitz and proximal to distal 1/3 ileum* |
| Behaviour | B1: nonstricturing nonpenetrating<br>B2: stricturing<br>B3: penetrating<br>p: perianal disease modifier | B1: nonstricturing nonpenetrating<br>B2: stricturing<br>B3: penetrating<br>B2B3: both penetrating and stricturing disease, either at the same or different times<br>p: perianal disease modifier |
| Growth | — | G0: no evidence of growth delay<br>G1: growth delay |

*In both the Montreal and Paris Classification systems L4 and L4a/L4b may coexist with L1, L2, and L3, respectively.

In adapting the Montreal classification for CD, the Paris Classification introduced: age at diagnosis defined as A1a (0 to <10 years), A1b (10 to <17 years), A2 (17 to 40 years), and A3 (>40 years), disease location above the distal ileum as L4a (proximal to ligament of Treitz) and L4b (ligament of Treitz to above distal ileum), the possibility of both stenosing and penetrating disease to be classified in the same patient (B2B3), the presence (G1) and absence (G0) of growth delay during disease course.

**Table 1.3 Paris and Montreal Classification for Ulcerative colitis[13].**

| | Montreal Classification | Paris Classification |
|---|---|---|
| Extent | E1: ulcerative proctitis<br>E2: left-sided UC (distal to splenic flexure)<br>E3: extensive (proximal to splenic flexure) | E1: ulcerative proctitis<br>E2: left-sided UC (distal to splenic flexure)<br>E3: extensive (hepatic flexure distally)<br>E4: pancolitis (proximal to hepatic flexure) |
| Severity | S0: clinical remission<br>S1: mild UC<br>S2: moderate UC<br>S3: severe UC | S0: never severe*<br>S1: ever severe* |

*Severe defined by Pediatric Ulcerative Colitis Activity Index (PUCAI).

In adapting the Montreal classification for UC the Paris Classification introduced: E4 to denote extent of ulcerative colitis that is proximal to the hepatic flexure and never severe (S0) and ever severe ulcerative colitis (S1) during disease course.

The Paediatric Crohn's Disease Activity Index[14] (PCDAI, Table 1.4) and the Paediatric Ulcerative Colitis Disease Activity Index[15] (PUCAI,

Table 1.5) are used to assess IBD severity. These scores are based on endoscopic and clinical evaluation of large cohorts of CD and UC patients used in order to develop a valid and reliable clinical index. A PCDAI sore of < 10 (range 0-100) indicates clinical remission, 10-30 mild disease, whereas a score > 30 indicates moderate CD activity. PUCAI activity index ranges from 0-85; a disease activity of < 10 is defined as inactive, 10-34 as mild, 35-64 as moderate, severe >65. The use of PCDAI and PUCAI are important in designing therapeutic guidelines over the disease course[16].

**Table 1.4 Paediatric Crohn's Disease Activity Index (PCDAI)**

| Item | | | | **Score** |
|---|---|---|---|---|
| | | | | |
| 1. Abdominal Pain | | | | |
| **None** | | | | 0 |
| **Mild: Brief, does not interfere with activities** | | | | 5 |
| **Moderate/severe (frequent or persistent, affecting activities)** | | | | 10 |
| | | | | |
| 2. Patient functioning, general well-being (Recall, 1 week) | | | | |
| **No limitation of activities, well** | | | | 0 |
| **Occasional difficulties in maintaining age appropriate activities, below par Frequent limitation of activities, very poor** | | | | 5 |
| **Frequent limitation of activity, very poor** | | | | 10 |
| | | | | |
| 3. Stools (per day) | | | | |
| **0-1 liquid stools, no blood** | | | | 0 |
| **2-5 liquid or up to 2 semi-formed with small blood** | | | | 5 |
| **Gross bleeding, >6 liquid stools or nocturnal diarrhoea** | | | | 10 |
| | | | | |
| 4. Laboratory | | | | |
| **HCT(%) < 10 years (male; female)** | 11-14 (male) | 11-19(female) | 15-19 (male) | |
| **>33** | > 35 | > 34 | > 37 | 0 |
| **28-33** | 30-34 | 29-33 | 32-36 | 2.5 |
| **< 28** | < 30 | < 29 | < 32 | 5 |
| | | | | |
| ESR | | | | |
| **< 20 mm/hr** | | | | 0 |
| **20-50 mm/hr** | | | | 2.5 |
| **> 50 mm/hr** | | | | 5 |
| | | | | |
| Albumin | | | | |
| **≥ 3.5 g/dL** | | | | 0 |
| **3.1-3.4 g/dL** | | | | 5 |
| **≤ 3.0 g/dL** | | | | 10 |
| | | | | |
| 5. Examination Weight | | | | |
| **Weight gain or voluntary weight stable/loss** | | | | 0 |
| **Involuntary weight stable, weight loss 1%-9%** | | | | 5 |
| **Weight loss ≥ 10%** | | | | 10 |
| | | | | |
| 6. Height at Diagnosis Score | | | | |
| **< 1 channel decrease** | | | | 0 |

| | |
|---|---|
| **≥ 1, < 2 channel decrease** | 5 |
| **> 2 channel decrease** | 10 |
| | |
| 7. Height at Follow-Up Score | |
| **Height velocity ≥ -1 SD** | 0 |
| **Height velocity < -1 SD, > -2 SD** | 5 |
| **Height velocity ≤ -2 SD** | 10 |
| | |
| 8. Abdomen | |
| **No tenderness, no mass** | 0 |
| **Tenderness or mass without tenderness** | 5 |
| **Tenderness, involuntary guarding, definite mass** | 10 |
| | |
| 9. Perirectal Disease | |
| **None, asymptomatic tags** | 0 |
| **1-2 indolent fistula, scant drainage, no tenderness** | 5 |
| **Active fistula, drainage, tenderness, or abscess** | 10 |
| | |
| 10. Extraintestinal Manifestations | |
| **None** | 0 |
| **One** | 5 |
| **Two** | 10 |

**Table 1.5 Paediatric Ulcerative Colitis Disease Activity Index (PUCAI)**

| Item | Points |
|---|---|
| | |
| 1. Abdominal pain | |
| **No pain** | 0 |
| **Pain can be ignored** | 5 |
| **Pain cannot be ignored** | 10 |
| | |
| 2. Rectal bleeding | |
| **None** | 0 |
| **Small amount only, in < 50% of stools** | 10 |
| **Small amount with most stools** | 20 |
| **Large amount (> 50% of stool content)** | 30 |
| | |
| 3. Stool consistency of most stools | |
| **Formed** | 0 |
| **Partially formed** | 5 |
| **Completely unformed** | 10 |
| | |
| 4. Number of stools per 24 hours | |
| **0-2** | 0 |
| **3-5** | 5 |
| **6-8** | 10 |
| **> 8** | 15 |
| | |
| 5. Nocturnal stools (any episode causing wakening) | |
| **No** | 0 |
| **Yes** | 10 |
| | |
| 6. Activity level | |
| **No limitation of activity** | 0 |
| **Occasional limitation of activity** | 5 |
| **Severely restricted activity** | 10 |

### 1.1.5 Epidemiology

In the last fifty years the incidence of paediatric IBD has tripled in western countries with the number of individuals affected increasing every year[17,18]. Approximately 30% of IBD diagnoses are in paediatric patients[17]. Although the peak age of onset of paediatric CD and UC is in late adolescence, 4% of paediatric IBD is diagnosed in early childhood[20]. A study from the Wessex region of Southern England, reported an increased incidence over the last decade of paediatric IBD from 6.39 per 100 000 per year to 9.37 per 100 000 per year (Figure 1.3)[19].

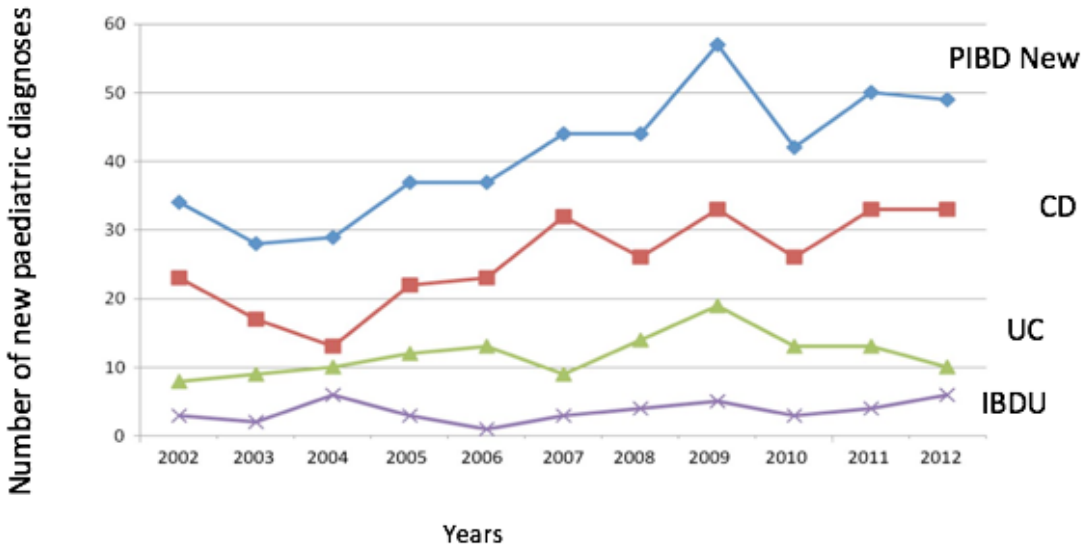



**Figure 1.3 Number of paediatric inflammatory bowel disease patients diagnosed per year in Wessex region, UK.** Blue: PIBD, Red: CD, Green: UC, Purple: IBDU. CD, Crohn's disease; IBDU, inflammatory bowel disease unclassified; UC, ulcerative colitis[19].

The highest incidence for CD has been reported in Scotland[20] and Canada[21], with an incidence of around 12 per 100 000 children per year, while for UC in Minnesota, USA, the incidence is 6.3 per 100 000 children per year[21]. The geographic trends of IBD have been reported in northern latitudes compared to the southern regions[22]. The study populations with highest IBD incidence are reported from the northern latitudes (Figure 1.4) The incidence of UC in Copenhagen, Denmark (8.1/100 000), was four times higher than in Bologna, Italy (1.9/100 000)[23]. To establish the north-south gradient in Europe a prospective study (EC-IBD) was conducted in 20 centres that focused on frequency of IBD across the continent. According to the EC-IBD study rates of UC in northern centres were 40% higher than those in the south (11.4/100 000 versus 8.9/100 000)[23]. Differences have been noticed even within countries; a Scottish

study reports a higher incidence of CD in northern Scotland compared to the southern regions of the UK[24]. The cause of the north-south gradient is not clear. The latest hypothesis suggests a link with levels of vitamin D which is lower in populations living in northern Europe. Vitamin D is known to be an inducer of *NOD2* function and the lack of vitamin D may result in the lower activity of *NOD2*, this factor may contribute to the higher incidence of CD in regions with low sun exposure[25]. Moreover, differences in diet across individuals and Countries have also been associated with IBD incident.

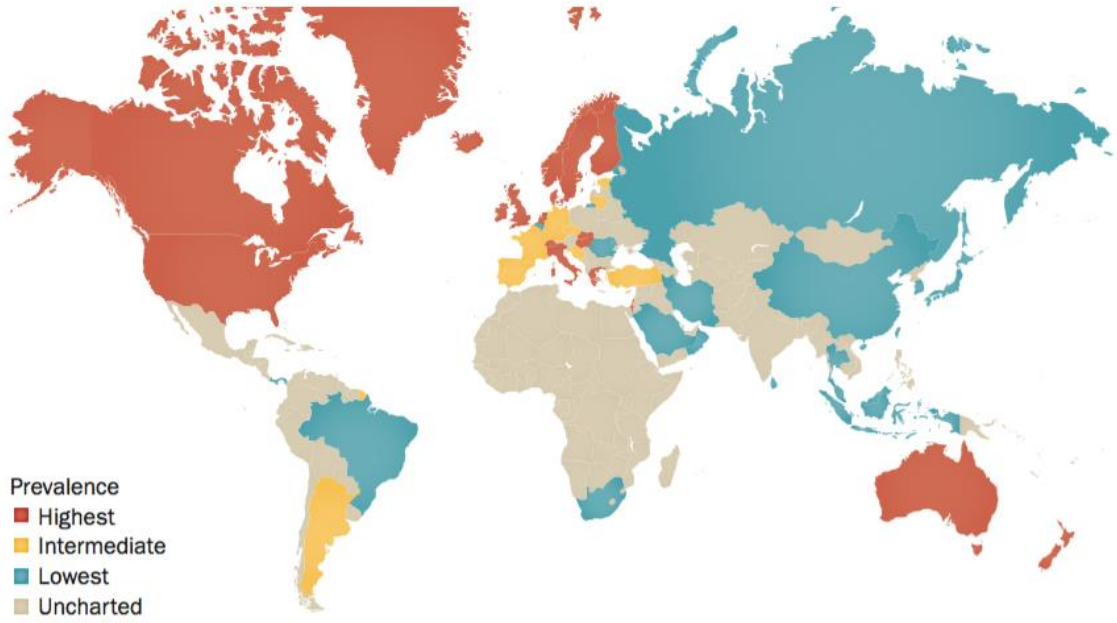



**Figure 1.4 Worldwide incidence of Inflammatory bowel disease in 2015.** Red refers to an annual incidence greater than 10/100, 000 people, orange to incidence of 5–10/100, 000 people and blue to incidence less than 4/100, 000 people. Absence of color indicates absence of data[26].

The distribution of IBD between different ethnicities is poorly documented, in particular in developing nations within Asia, Africa, and South America. Traditionally IBD is considered to be less prevalent in non-Caucasian populations. This is probably related to underrepresentation of non-Caucasians in study populations. The assessment of epidemiology in non-Caucasian populations is obstructed by factors like absence of population-based registries. A paediatric study in Wisconsin and Georgia compared the incidence and disease characteristics between African Americans and Caucasian populations, which reported a similar distribution of types of IBD in proportions of African Americans affected by IBD, suggesting that the condition is not predominantly a Caucasian disease[27]. Although environmental and genetic triggers play a prominent role in the aetiology of IBD, the cause of increasing incidence in many countries is speculated as western lifestyle or environmental factors like diet, smoking

and the hygiene hypothesis[22]. Overall, IBD represents a significant global health burden that is of growing concern.

### 1.1.6 Clinical presentation in patients with IBD

Paediatric onset IBD (pIBD) presents unique phenotypic characteristics and severity compared to adult-onset disease[28]. PIBD is more often characterized by extensive intestinal involvement, rapid early progression and a high rate of resistance to conventional therapy[29]. Moreover, early-onset IBD has a stronger familial component than adult disease[29]. These combined features indicate a stronger genetic component to pIBD compared to IBD diagnosed in adulthood. Nevertheless, both early onset and late onset IBD patients are at increased risk of emotional problems and decreased social functioning. Studies have demonstrated an impaired quality of life among patients during relapses and during remission periods[30]. The impact of IBD on everyday life has been investigated extensively using various health related quality of life measures. Such measures assess health status from the patient perspective, taking into account the physical and social aspects as well as the attitudes and behaviours of the individual[31]. A comprehensive list of issues for children and adolescents diagnosed with IBD are shown in Table 1.6. In the following section I will detail clinical and non-clinical manifestation in paediatric IBD patients.

#### *1.1.6.1 Clinical manifestation in paediatric IBD*

A study conducted by Heyman and collaborators reported that 2.7% of children present IBD symptoms before one year of age. Studies have shown that pIBD patients have higher rates of psychological disturbance, anxiety and depression compared to the control population[32]. The most common symptoms at presentation are abdominal pain, diarrhoea, weight loss, fever and blood in stools. Abdominal pain in some cases may help localize the site of inflammation: in patients suffering from ileal Crohn's disease, pain is commonly felt in the lower right quadrant, while UC patients often feel pain in the left lower quadrant[32]. One of the main challenges for children with a diagnosis of IBD is growth and skeletal development delay. An important reason for this problem is the lack of adequate nutrition, which can result from decreased caloric intake, increased needs, increased enteral losses, and altered nutrient utilization[33]. The majority of the patient's growth improves with treatment of colitis[33]. Growth delay,

more frequent in CD patients compared to UC patients, has been observed in 15-40% of IBD cases and can present before the canonical IBD symptoms.

PIBD patients have a higher risk of developing a secondary immune-mediated condition[34] (Figure 1.5 and for more details see Chapter 2). This comorbidity adds to the burden of pIBD and suggests common genetic mechanisms across immune-mediated diseases. As an example, osteopenia and osteoporosis can occur in IBD patients due to lack of vitamin D reducing intake of calcium following steroid treatment[35]. Anti-inflammatory therapy and enteral nutrition are able to help overcome growth failure and restore pubertal progression. Response to treatment is variable and adverse reaction to treatment may manifest in adolescents, even after many years on the same medications, as they progress through puberty and their metabolism changes.

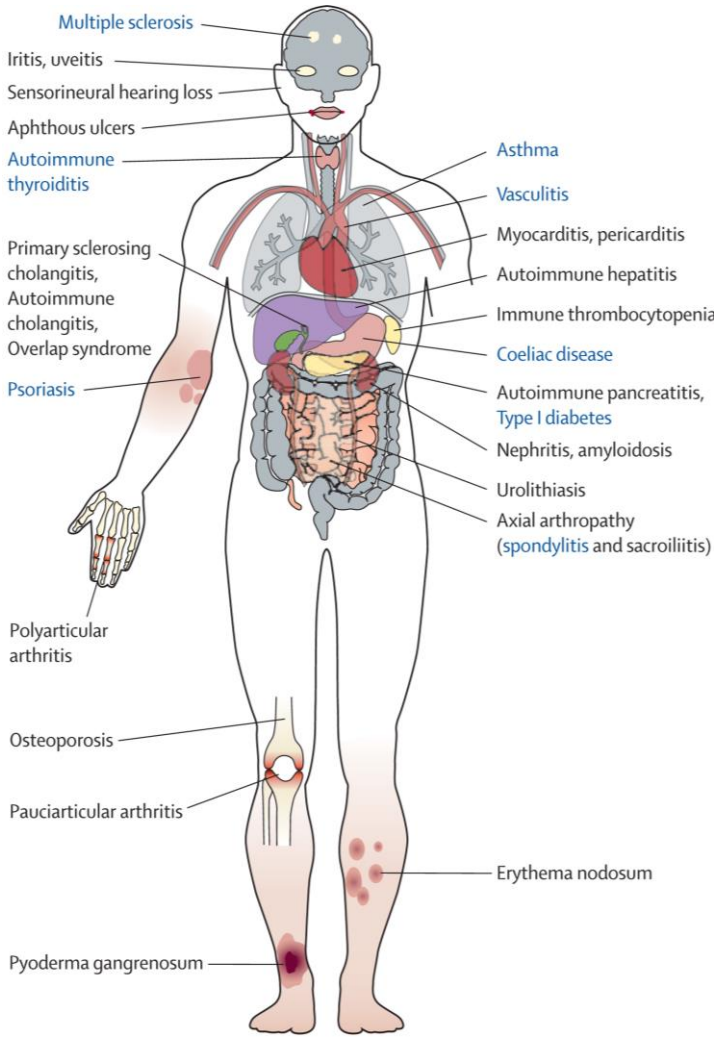



**Figure 1.5 List of comorbidities which might arise in pIBD patients[2].**

### *1.1.6.2 Non-clinical consequences in paediatric IBD*

High disease activity is the most important factor of reduced quality of life in both CD and UC. Disease activity is related to the level of fatigue and sleep difficulties and these factors are independently associated with an impaired quality of life[31]. Better quality of life was assessed in patients who had been diagnosed for a longer time, and it could be that IBD patients with a longer disease history adapt to living with the illness[36]. Newly diagnosed patients require more and different kinds of support. The impact of different treatments is poorly reported. Side effects from medications may also interfere with body image among patients. Epidemiological studies have shown that the majority of the patients who had used corticosteroids had concerns about their long-term effects[31]. Considering that disease activity is the major contributor to impaired quality of life, the use of effective treatment regimens aiming to maintain patients in remission is of great importance. Rates of poor quality of life tend to be worse in those who have had IBD surgeries compared with those who have not had surgery[36]. This increased rate could be due to the more severe nature of the symptoms of those requiring surgery or the impact of the surgery on the body. Patients in remission have a greater perception of life, lower emotional and social dysfunction compared to patients with active disease. Disease duration, age and gender do not significantly affect IBD patient's quality of life[37]. It is still questionable whether psychological support[37]should be included in the general management of IBD patients or if it should be focused on certain patients, whereas effective treatment of patient disease seems to play the greater role in improvement in quality of life[31,37]. Disease related concerns in a survey of pIBD patients is shown in Table 1.6.

**Table 1.6 Disease related concerns of pIBD patients**[38]

| Domain | Issue * |
|---|---|
| Treatment | Feeling bothered by having to take medications (1)<br>Feeling worried about needing surgery (14)<br>Feeling bothered that there don't seem to be good treatments for IBD (16)<br>Feeling bothered about treatments you have to have (21)<br>Feeling upset that you're not allowed to eat what you want (23)<br>Feeling worried about having to be admitted to hospital (24)<br>Feeling worried about X-rays or scopes (30)<br>Feeling treatment gets in the way of other things you want to do (34) |
| Body image | Being concerned about weight (4)<br>Being bothered about your height (6)<br>Being concerned or upset about the way you look because of your bowel condition or its treatment (13) |
| Emotional | Feeling worried about the possibility of having a flare-up (2)<br>Feeling upset that your bowel condition is a lifelong thing (3)<br>Feeling worried about health problems you might have in future (5)<br>Feeling that it is unfair that you have IBD (10)<br>Feeling frustrated because of your bowel condition (12)<br>Worrying about never feeling better (15)<br>Feeling worried about how your bowel condition affects your family (25)<br>Feeling stressed out (26)<br>Feeling angry that you have IBD (28)<br>Feeling your bowel condition has caused a lot of family stress/tension (32)<br>Feeling that people don't understand about your bowel condition (33)<br>Feeling guilty because of the effect of IBD on the family (37)<br>Feeling in a bad mood (39)<br>Feeling embarrassed about having to go to the bathroom a lot (44)<br>Feeling irritable (41) |
| Bowel | Feeling bothered about the stomach pain or cramps that you get (7)<br>Feeling you can't eat what you want because it gives you pain/diarrhea (17)<br>Feeling sick to your stomach or nauseated (18)<br>Feeling bloated (like your stomach is full of air) (31)<br>Worrying about having blood with a bowel movement (34)<br>Feeling bothered about bowel movements being loose or frequent (38)<br>Feeling worried or bothered about passing gas (40)<br>Worrying about having an accident without making it to the bathroom (46) |
| Functional/social | Feeling you have to give up doing things because of bowel condition (8)<br>Feeling you miss out on activities due to your bowel condition (19)<br>Feeling you will have to miss school because of your bowel condition (20)<br>Feeling unable to play sports like you could before (27)<br>Feeling worried about having problems while traveling because of your bowel condition (22)<br>Feeling worried about having to use a washroom in a public place (29)<br>Feeling like you are missing out on childhood fun (41) |
| Systemic | Feeling that you don't have the energy to do the things you want (9)<br>Feeling tired (11) |

* Numbers in parentheses indicate ranking (according to mean frequency-plus-importance score) by entire group of 117 patients completing the item reduction questionnaire.

### 1.1.7 Treatments

The chronic nature of IBD requires long-term treatment with multiple drugs. The aim of the therapy is the resolution of the symptoms and consequently the remission of the patients preventing further relapses. The treatment is decided based on several factors which can vary over the course of the disease such as anatomical localisation, behaviour of the disease, extraintestinal manifestations and is modulated based on the clinical response of the patient. According to the National Institute for Health and Care Excellence (NICE) guidelines, surgery should be reserved for managing complications (fistulae and abscesses) and treating obstruction. Three option treatment plans are widely used: nutritional, pharmacological and surgery[39,40](Figure 1.6 and Figure 1.7).

#### NUTRITIONAL THERAPY

Exclusive enteral nutrition (EEN) is an effective first line therapy in paediatric patients for small and large bowel disease, inducing remission in 60-80% of cases[6], and is used for treating Crohn's disease as an alternative therapy to corticosteroids. Initially, treatment is usually given for 3-6 weeks of exclusive liquid feeding with either elemental or polymeric formulae. Patients are not allowed to have any other dietary items except plain water and some beverages. There are two types of formulas: the elemental formula contains individual amino acids, glucose polymers, and are low fat with only about 2% to 3% of calories derived from long chain triglycerides whereas the polymeric formula, such as Modulen, contains intact proteins, complex carbohydrates and mainly long chain triglycerides. The choice of formula is dictated by clinician experience, palatability, funding, and local availability. Factors that influence enteral nutrition include patient and parent choice, compliance, palatability, lack of corticosteroid toxicity, potential benefits in terms of improved nutritional status and growth. Liquid feeding is an alternative to corticosterois and it avoids adverse effects of steroids treatment, ensuring optimal growth[41].

#### PHARMACOLOGICAL THERAPY

##### AMINOSALICYLIC ACIDS

Aminosalicylic acids (5-ASAs) are a class of anti-inflammatory drugs acting on epithelial cells by a variety of mechanisms to moderate the release of lipid mediators, inflammatory cells, cytokines and reactive oxygen species reducing the inflammatory

process and allowing damaged tissue to heal. 5-ASAs are often used long-term to maintain remission, as well as to treat mild to moderate attacks of IBD[40]. The main role for 5-ASA is maintenance of remission in UC while there is no evidence that 5-ASA is superior to placebo for the maintenance of CD[6].

### Antibiotics

Antibiotic therapies are effective in treating dysfunction of the endogenous gastrointestinal bacteria flora. There is some evidence that metronidazole and ciprofloxacin have specific uses in Crohn's disease; however there is no clear-cut evidence to support the use of these antibiotics in ulcerative colitis as disease modifying therapy [42].

### Corticosteroids

Corticosteroids are potent anti-inflammatory agents for moderate to severe relapses of both ulcerative colitis and Crohn's disease. Steroids are binding molecules of several cell types which activate a cascade of transduction signals reducing the inflammatory response. The main role of steroids is to reduce the lymphocyte and macrophage differentiation and the migration of neutrophils in the inflammation sites. Steroids are usually prescribed to treat acute attacks in both UC and CD. Although corticosteroids induce remission in 60-80% of the patients, they are not suitable for maintenance and remission due to their side effects and because patients can become corticodependent. Usually an initial high dose of the drug is given to be effective which is gradually reduced. Together with steroid treatment, patients need to take 5-ASAs to decrease the risk of relapses[43]. Steroid resistance or unresponsiveness leads to increment of treatment, or consideration of surgery.

### Azathioprine or mercaptopurines

Azathioprine (AZA) or mercaptopurines (MP) are immune suppressive drugs widely used in ulcerative colitis and Crohn's disease as adjunctive therapy. These treatments are able to induce T cell apoptosis reducing the inflammatory response. Side effects occur in 20% of the treated patients. The commonest are allergic reactions (fever, arthralgia, and rash) that typically occur after 2-3 weeks and cease rapidly when the drug is withdrawn[10]. Thiopurine S-methyl transferase (TPMT) is a key enzyme involved

in the metabolism and detoxification of azathioprine and 6-mercaptopurine. The *TPMT* gene encodes for the TPMT enzyme which confers inter-individual differences, both in terms of clinical efficacy and toxicity profiles based on the enzyme activity[44]. Current clinical guidelines recommend determining TPMT status in a given individual before commencement of thiopurine therapy in order minimise the risk of adverse effects whilst aiming for an optimal clinical response.

### BIOLOGICAL THERAPIES

Amongst current treatment regimes in the field of IBD are biological therapies. The term biological therapies refer to the treatment of IBD through the use of biological materials or molecules able to modify the biological response[45]. One of the therapies used in IBD is the blocking of the inflammatory response by using TNF-α antagonist. Tumour Necrosis Factor alpha (TNFα) is a pro-inflammatory chemokine which plays an important role in amplifying the inflammatory response in the gut of patients affected by IBD[46].

Infliximab, one of the most widely prescribed drugs in IBD, is a chimeric antibody containing 75% of human sequence and 25% of murine sequence. Infliximab is able to bind the TNFα "free" in the cytosol as well as the TNFα bound to the cellular membrane, neutralising the inflammatory effects[47]. Typically the drug is given via intravenous injections and the treatment usually last 8 weeks[48]. Due to the nature of their effects on TNF, all anti-TNF therapies share similar side effects, including increased risk of infections from intracellular pathogens, most notably other opportunistic infections, autoimmunity such as psoriasis and cutaneous lupus, infusion reactions and vascular inflammations. Drug hypersensitivity occurs in approximately 12% of the patients requiring drug suspension and switch to other medication.

### PROBIOTICS

Probiotics are a new class of IBD treatment that aim to modulate the gut microbiota directly and/or indirectly. Probiotics contain live, non-pathogenic organisms that reach the intestine in an active state improving the gut's microbiota balance towards a healthier status[49]. Probiotics include lactic acid producing bacteria and yeast. Their mechanisms of action is still not well characterised but it is hypothesized that they modulate the membrane permeability and the mucosal immune system, keeping

pathogens away from the intestinal mucosa surface[49]. In the literature there are few human studies on the efficacy of probiotics in CD and UC patients, however, these few studies have demonstrated some efficiency of probiotics in both adult and child UC patients[50]. Probiotics seem like a feasible solution for treating mild-moderate UC, however further studies are needed before recommendations can be offered on routine use of probiotics in IBD.

### OTHER MEDICATIONS

In addition to the medications for controlling gut inflammation, IBD patients might be prescribed drugs that might help relieve symptoms. Medications like anti-diarrheal medication, fibre supplements, pain relievers, iron supplements, vitamin B-12, calcium and vitamin supplements and a recommended special diet could help improve patient condition. As a lifelong disease, IBD is an important cause of psychological distress. Antidepressants are commonly prescribed in adult patients with IBD and studies have shown that young patients with IBD have a high frequency of antidepressant use, compared to individually matched population-based controls[51].

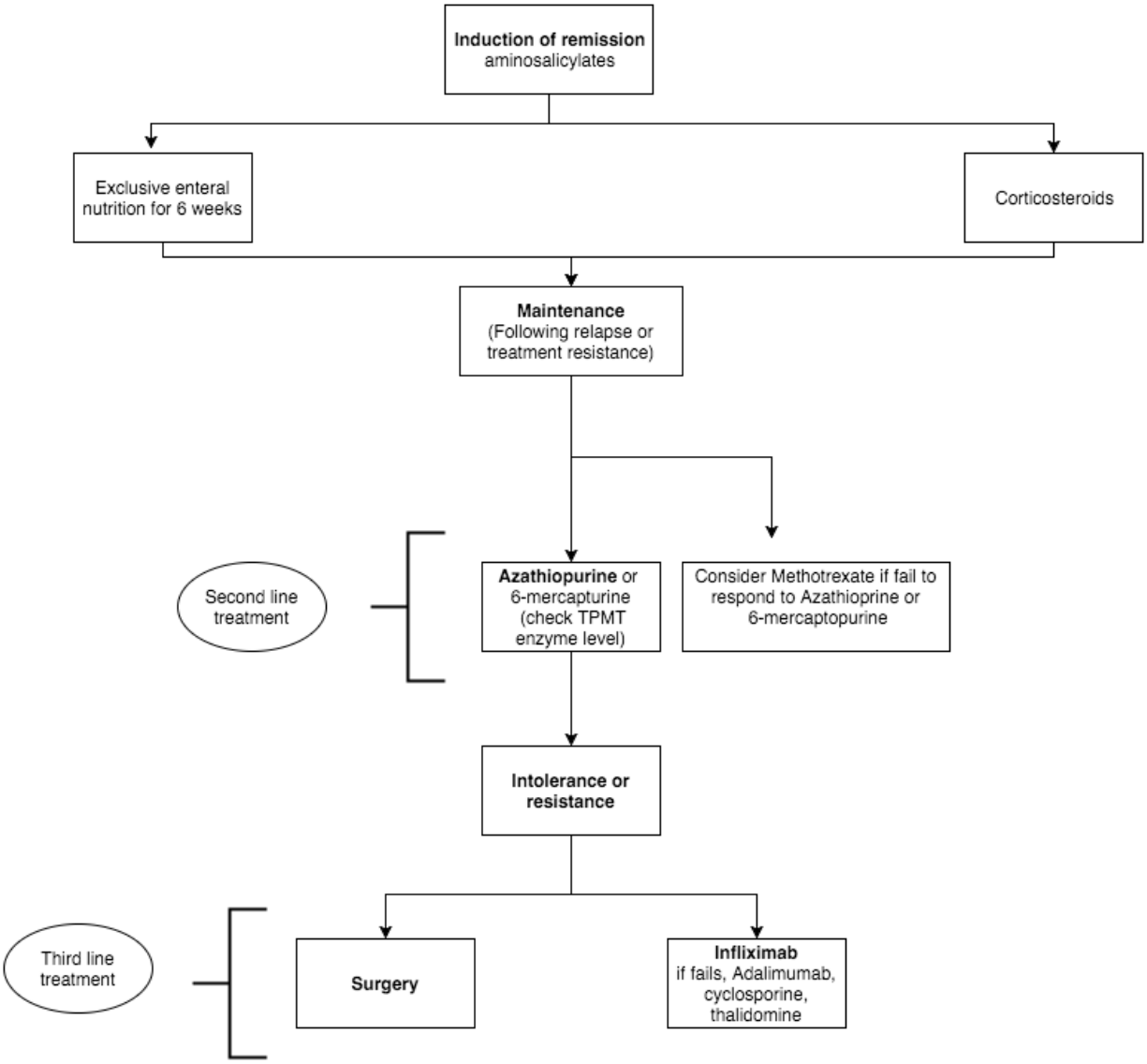



**Figure 1.6 Crohn's Disease treatment flow chart**[52]**.** In order to induce and maintain remission aminosalicylates are used in mild disease. If this is not effective, exclusive enteral nutrition and corticosteroids are the first line therapies. Thiopurine therapy may be introduced (after checking TPMT expession levels) in cases with severe disease. Surgery should be considered for patient non responsive to treatment whereas Infliximab is given to patients who are intolerant to steroids and in whom surgery is inappropriate.

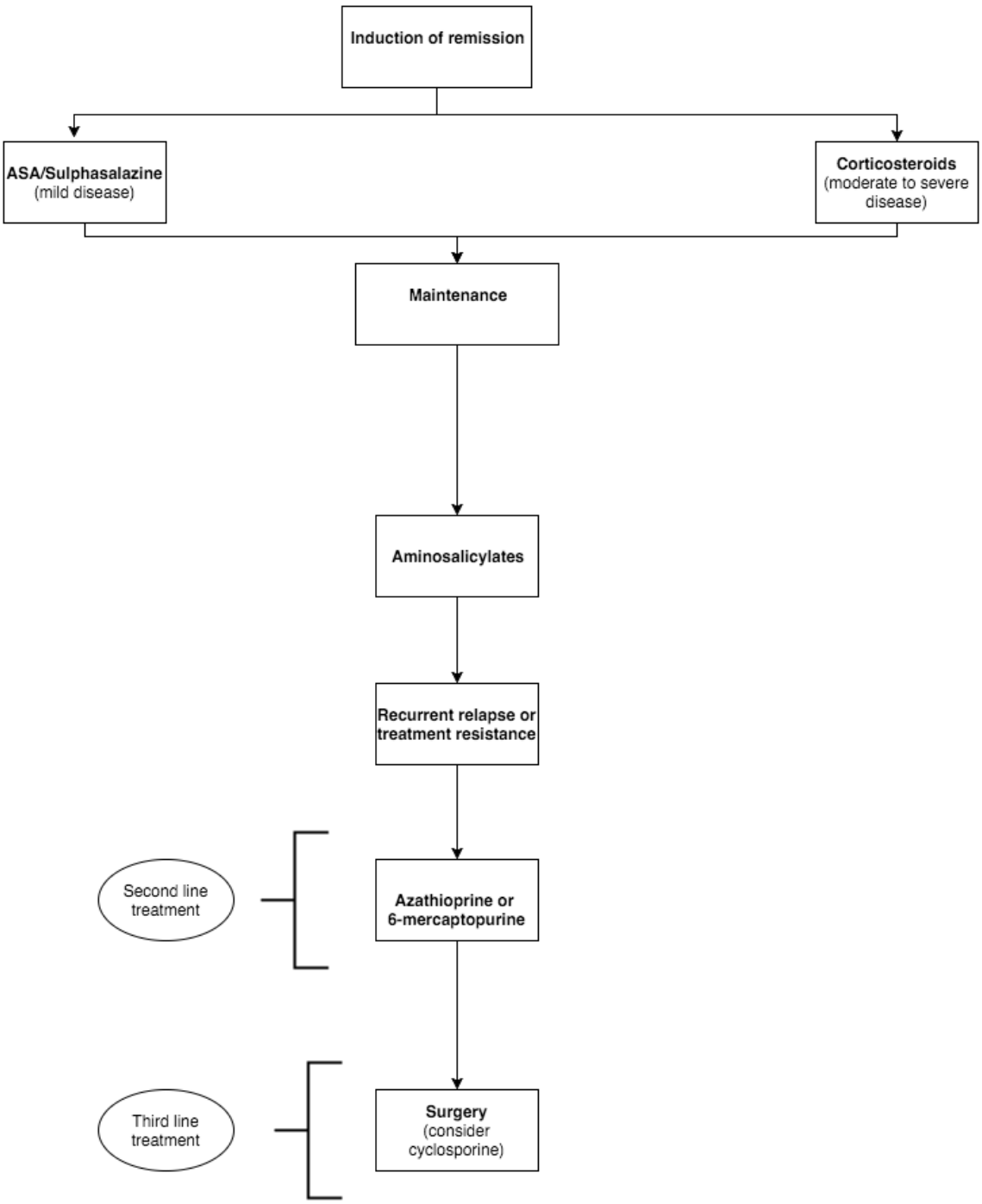



**Figure 1.7 Ulcerative colitis treatment flow chart**[52]. The therapeutic approach for ulcerative colitis consists of sulphasalazine and costicosteroids for inducing remission in mild to moderate colitis and in aminosalicylic acid for maintaining remission. Thiopurine may be introduced as maintenance therapy for patients who have failed with or cannot tolerate steroids. Surgery should be taken into account for complications of the disease whereas Infliximab should be given in patients who are intolerant to steroids and in whom surgery is inappropriate.

### SURGERY

Surgery should be considered in patients with obstructive complications and in those who have not responded to medical therapy. Proctocolectomy is the most used surgery technique in paediatric and adult patients which includes removal of the

rectum and all or part of the colon. As UC is located in the colon, the surgery completely removes the disease[6]. In CD, surgery is not curative and management is directed at minimising the impact of disease. Approximately 30% of IBD patients require surgery in the first 10 years of disease and 70–80% will have surgery in their lifetime[53].

### 1.1.8 Risk Factors

IBD is an idiopathic disease caused by a dysregulated immune response to host intestinal microflora in genetically predisposed individuals[54]. The pathogenesis of IBD is derived from the interaction of environmental and genetic risk factors (Figure 1.8).

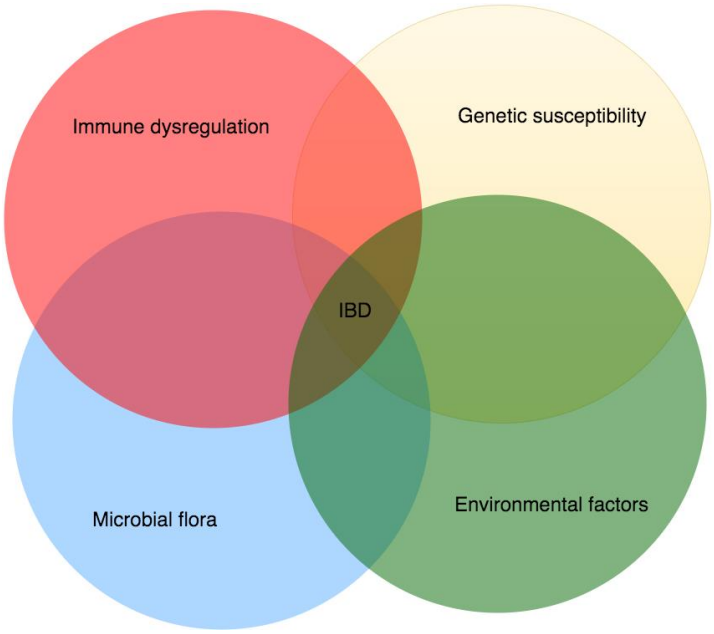



**Figure 1.8 Factors influencing the development and course of IBD.** IBD is a multifactorial disorder which arises from a combination of genetic susceptibility, host-bacteria interactions, immune dysregulation and environmental factors.

#### *1.1.8.1 Environmental risk factors*

Epidemiological changes observed in the last 50 years and discordance of IBD among monozygotic twins studies[55,56] have suggested a role of environmental factors in the pathogenesis of IBD[57]. Environmental factors increase the predisposition of the disease in genetically susceptible individuals. In the last century multiple hypotheses have been developed to explain the increasing incidence of IBD.

### GUT MICROBIOTA

The gastro intestinal (GI) microbiome of healthy individuals is composed of 45,000 phylotypes primarily dominated by four bacterial phyla: *Firmicutes*, *Bacteroidetes*, *Proteobacteria* and *Actinobacteria*[58]. Studies have observed dysbiosis in the GI microbial communities in IBD patients supporting the hypothesis that the condition arises from altered interactions between intestinal microbes and the mucosal immune system. An impaired handling of commensal microbes and pathogens is a prominent factor in disease development[59]. This is supported by several studies in both human and animal models.

### THE HYGIENE HYPOTHESIS

Another presumed contributor to the increasing incidence of IBD is improvements in hygiene. With the increase use, sometines to excess, of hygiene products, children are less exposed to microbes, and consequently the immune system is not stimulated by environmental microorganisms and antigens[57,60] . The study conducted by Koloski *et al* examined the microbial exposure of IBD paediatric patients showing that due to increased hygiene, IBD patients have a significantly lower seroprevalence of *H. pylori* compared to controls. The exposure to *H. pylori* is thought to be necessary in programming the immune system of the gut and mitigating its future inflammatory responses, perhaps even resulting in CD when the immune system is challenged[61]. *H. pylori* has been further implicated in several gastro intestinal conditions such as: gastritis, peptic ulceration, gastric cancer, and mucosa-associated lymphoid tissue lymphoma[62].

### DIET

Changes to a diet rich in sugar have led to the development of antigens that cause alterations in the gut flora with consequences in the permeability of the stomach[63] and might have a role in the development of IBD. The most consistently described dietary association with IBD has been intake of dietary fiber, fruits, or vegetables. In a paediatric[64] IBD cohort Amre *et al* demonstrated that intake of fruits and vegetables were inversely associated with the risk of CD[65]. A similar larger prospective adult study demonstrated a strong inverse association between the intake of dietary fibre and the risk of CD, with a less strong effect on UC[66]. Fibre intake from fruits and vegetables

(soluble fibre) was protective against CD, whereas insoluble fibre intake from cereals, whole grain, or bran did not reduce the risk of CD or UC[66]. This is supported by the fact that fibre is used by the microbiota of the lower GI tract as their main source of energy[67]. *Fibrolytic* bacteria degrade polysaccharides into smaller carbohydrates, which are then fermented into short-chain fatty acids to be used as a major source of energy for *colonocytes* and have immunomodulatory properties.

### SMOKING

Smoking has been shown to confer protection from ulcerative colitis but to increase risk of Crohn's disease[33]. In a French study, the adverse effect of smoking was particularly prominent among women with CD. That the incidence of IBD has traditionally been low in countries that have the highest rates of smoking supports the concept of variability in susceptibility to environmental influences. Conversely, the rate of smoking is lower than average in those countries with a high incidence of IBD, such as Sweden and Canada. Researchers have studied the systemic effects, cellular and humoral immune effects, mucosal changes, and the intestinal permeability changes with IBD and smoking. To date, none of these studies adequately explain the observed clinical patterns. It has been assumed that nicotine is the active agent in these associations, but clinical trials of nicotine chewing gum and transdermal nicotine in UC have shown limited benefit, and have been complicated by significant side-effects[33]. Gender also appears to be associated with susceptibility in relation to cigarette smoking and IBD risk[33].

### APPENDECTOMY

Similar to observations regarding smoking and IBD incidence, appendectomy also has divergent effects on CD and UC. Appendectomy appears protective in UC if it occurs prior to the age of 20 years[68]; however, appendectomy does not alter CD risk and may even be associated with an initial increase in risk, whether this represents true causality or is due to diagnostic bias remains to be definitively established.

GENDER DIFFERENCES

In paediatric, but not in adult patients, CD appears more frequent in males compared to females whereas UC affects females and males equally[33]; the use of contraceptives in women may be correlated with the development of IBD[69]. This gender difference could be caused by environmental exposure or hormonal influence[20]. In a study of 232,452 women from two prospective cohorts, Khalili and colleagues demonstrated an increased risk of CD in women who were currently using oral contraceptives with an attenuation of risk in past users compared with those who had never used them. A consistent effect was not seen for the development of UC. A meta-analysis by Cornish and colleagues identified an elevated risk for both CD and UC with oral contraceptive use[70]. This result should be interpreted carefully as meta-analysis studies usually do not take into account studies which show negative results or insignificant results as they are less likely to be published[71]. In contrast to the oral contraceptive data, postmenopausal hormone use was associated with an elevated risk of UC but not CD. These divergent results could, potentially, be due to the different intrinsic hormonal conditions that exist in premenopausal oral contraceptive users compared with users of hormonal therapy who are mostly postmenopausal.

Although environmental influences appear to be critical to the pathogenesis of CD and UC, the effect of such environmental factors on the natural history of disease, and whether interventions that modify these factors can improve patient outcomes, need further study.

#### *1.1.8.2 Genetic risk factors*

Much evidence supports the role of genetic predisposition in the pathogenesis of IBD. Since the development of technologies for analysing the human genome, it became possible to identify IBD susceptibility loci. IBD is familial in 5-14% of the diagnosed individuals and approximately 8% of IBD patients have a positive family history for the condition[72], which can contribute to an early pathogenesis of the disease[8].

Genetic studies including linkage mapping, candidate gene approaches, genome wide association studies (GWAS) and next generation sequencing studies have uncovered genetic factors underpinning this condition.

#### *1.1.8.3 Genetic basis of IBD*

IBD is defined as a complex polygenic disorder. Complex diseases, also known as multifactorial diseases, are common disorders caused by the interaction of genetic and environmental risk factors[73]. Unlike monogenic diseases, complex diseases do not follow the classical mechanism of Mendelian heredity since they are caused by the action of multiple genes. The phenotype of a monogenic disorder can be complex but depends on the action of a single gene; whereas the components of the phenotype of a complex disease may depend on interaction between multiple genes.[74] Each gene involved in a complex disease is responsible for only a part of the disease risk and for this reason they are known as "susceptibility genes". The effects of susceptibility genes can be additive, or multiplicative, with respect to the other genes involved in the same disease. In the last decade the study of multifactorial and monogenic diseases has gone from the analysis of single loci to the study of the effects of multiple loci[75]. The genetic predisposition to human disease is sometimes categorised by the frequency and effect size of the variants involved.

Causative events leading to IBD can occur any time in life making the genetic model of IBD unclear. The cumulative contribution of genetic factors, immunology and gut microbiota in the aetiology and phenotype of Crohn's disease and ulcerative colitis might differ across ages (Figure 1.9). The polygenic model suggests that a large number of genes act together to give susceptibility to the individual. The oligo-genic model suggests that one or two or a few IBD associated genes act together and alter the gut permeability[76]. As an examples, mutations in the gene encoding the X-linked inhibitor of apoptosis (*XIAP*)[77] are linked to early-onset severe colitis and the interleukin 10 (*IL10*)[78] and its associated receptor alpha and beta subunits (*IL10RA* and *IL10RB*) have been shown to cause very early onset IBD[79]. Children constitute an IBD subgroup characterised by a more severe phenotype and treatment refractoriness[80,81].

Several approaches have been used for identifying the susceptibility regions of complex disorders and the causal genes of Mendelian diseases: linkage analysis, genome wide association studies (GWAS) and more recently, whole exome and whole genome sequencing studies.

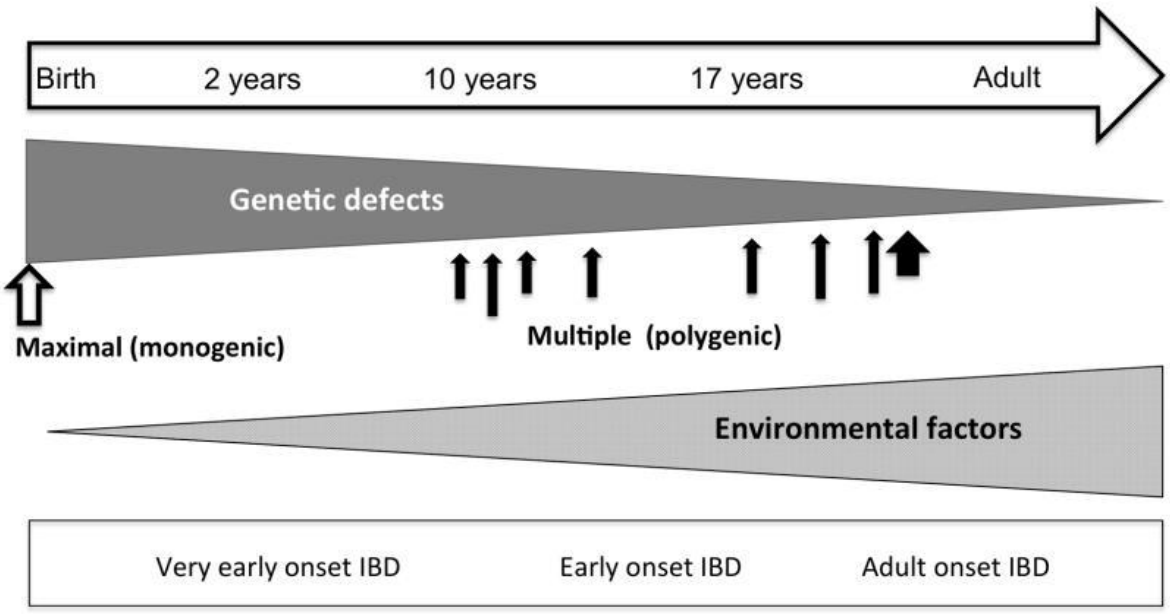



**Figure 1.9 Genetics and environmental contribution to IBD from birth to adulthood.**[12] IBD pathogenesis arises from a combination of environmental and genetic factors. However, it is now assumed that in early onset IBD rare variants with high effect have a greater role. Differently, in adult onset patients multiple common variants with low effect size and environmental factors contribute to the predisposition of disease.

#### *1.1.8.4 Immune response pathways in IBD*

Genetic and murine studies have indicated key pathogenic pathways involved in IBD pathogenesis: epithelial barrier function, microbial sensing, innate immune regulation, regulation of autophagy and regulation of adaptive immunity. Several studies have shown a dysregulation in the innate and adaptive immune response as contributors to IBD risk in both early and late onset IBD[1].

Intestinal homeostasis is crucial in determining IBD risk[1]; the weakening of the intestinal epithelium increases permeability, which might result in a loss of immunological tolerance to commensal flora caused by an excessive activation of the immune response[82–84]. Although multiple studies have described the complex relationship between host immunity and the gut flora, there are a few number of cells that are fundamental for this relationship[82,83]. When these key players are defective, there is a loss of host tolerance towards the enteric bacterial, which can result in IBD[82–84]. *In vivo* and *ex vivo* studies have shown that the key cells responsible for disease risk are T helper cells[85]. As an example, mutations within *IL10* and its receptor *IL10RA* and *IL10RB* have been associated with severe IBD in paediatric patients[86]. Dendritic cells are responsible for presenting antigens to naïve CD4+ helper T cells and maintaining the tolerance towards the commensal flora by promoting T cells regulatory (Treg)

differentiation[87]. In the case of infection, toll-like receptors (TLR) and NOD receptors recruit dendritic cells, inducing the production of pro-inflammatory cytokines (e.g. IL1 and TNF-$\alpha$ ) and the promotion of Th1 and Th2 differentiation[82]. Specifically, Th1 response has been shown to be involved in CD whereas Th2 in UC[82]. Th1 cells are characterised by increased production of IL2, IL12 and IFNγ, which are important in delayed hypersensitivity and cellular immunity, whereas Th2 lymphocytes are characterised by the production of IL5, IL10 and TGFβ which are involved in humoral immunity[88]. Studies have shown an overexpression of dendritic cells in inflamed regions of the intestine that results in loss of tolerance to the normal bacteria[82–84]. It is hypothesised that the cells and tissue damage is mediated by cytokines (such as TNF-$\alpha$) and that abnormal activation of dendritic cells might be a consequence of TLR and NOD receptors defects[88]. This is supported by genetic studies describing disease risk variation within genes coding for proteins involved in the epithelial innate immunity (e.g *NOD2* and *TLR2*)[82–84]. Other studies have supported the involvement of Th17 in IBD pathogenesis. Th17 are T cells responsible for the production of cytokine IL17 in both UC and CD patients[82–84]. Th17 differentiation is induced by IL23. Genome wide association studies have shown an association between *IL23R* and IBD, suggesting an involvement of T cells in IBD pathogenesis[1]. Cytokines involvement in IBD was firstly discovered by the study of stools sample in paediatric IBD patients[88]. TNF-$\alpha$ is a pro-inflammatory cytokine of the innate immune response. Multiple studies have shown an increased expression of TNF-$\alpha$ in IBD patients compared to healthy individuals in both adults and children[88–90].

Although CD4 T cells are the key regulators in the intestine, other cell types (e.g. B cells, natural killer cells and CD8 T cells) contribute to the maintenance of homeostasis in the gut[82–84]. Further genetic, functional and animal studies will help to identify crucial pathways involved in the mucosal immune regulation which might lead to the development of novel therapeutic drugs.

## 1.2. Methods for detecting disease genes

### 1.2.1 Genetic variation

Genetic changes that occur across the genome are broadly defined as variants because the functional alteration on the DNA is often unknown[91], the effects of the variant on the phenotype vary. Variants can occur in any part of the genome and can range in size from a single base pair (bp) to megabases. Variants can be defined as either polymorphisms or mutations, mutations occur rarely in the population (frequency less than 1%) whereas polymorphisms are more common in the population (frequency of at least 1%)[92]. Mutations are considered to be more likely disease causal changes, different to polymorphisms that are not considered to be causal by themselves but they might contribute to disease susceptibility. Simple sequence repeats are short tamdemly repetitive segments of DNA, from two to less than nine nucleotides, adjacent to each other. These repetitive regions of the DNA are predisposed sites for genetic mutations; as an example, Hungtington disease is a result of an unstable expansion the CAG repeat[93–95].

Variations can be categorised into:

- Single nucleic acid substitution – the replacement of a single nucleotide by others.[93–95]
- Insertions/deletions (Indels) – the addition of deletion of a single or hundreds of nucleotides into a DNA sequence; if the inserted DNA sequence is identical to preceding nucleotides, this is known as duplication. Insertions or deletions of tens or hundreds of bases within the DNA sequence are called copy number variants (CNV).[93–95]
- Chromosomal rearrangements – where segments of DNA are either inserted, deleted, reversed and translocated between chromosomes.[93–95]

Single nucleotide variants, are also referred to as single nucleotide polymorphisms (SNPs), and constitute the largest class of variation. Substitutions within genes are classified as synonymous, non-synonymous, stop gain/loss or start loss. The effect on the protein depends on the type of mutation and location within the protein domains and tertiary structure[93–95].

Because the genetic code is degenerate, some nucleic acid substitutions will not change the amino acid sequence and are therefore less likely to be pathogenic. These

substitutions are named synonymous. The resultant protein from a synonymous mutation is the same as the wild type; however, there will be some changes in the mRNA sequence[93–95].

Non-synonymous mutations are nucleotide substitutions in which a base substitution occurs within the coding region of a gene changing the amino acid incorporated into a protein. The effect of the change on the protein structure depends on the properties of the amino acid inserted and the importance of the original amino acid to the protein function[93–95].

Stop codons or nonsense mutations introduce a premature stop codon in the protein sequence resulting in a loss of function, or gain of function, of the protein[93–95].

Small single nucleotide insertions or deletions (indel) can be classified by their effect on protein function. Insertions or deletions of multiples of three nucleotides maintain the reading frame (non-frameshift mutations). Insertions or deletions of any other number of nucleotides cause an incorrect translation of the mRNA sequence; these are known as frameshift indels and might lead to protein truncation by encoding a premature stop codon[93–95].

Single nucleotide substitutions and indels can occur in proximity to the exon-intron boundaries within a gene and have the potential to affect the correct splicing of RNA after transcription. Specifically splicing is disturbed if mutations occur at the essential splice sites -2,-1,+1,+2 and +5[93–95]. Figure 1.10 shows the location of the most common nucleotide substitutions across a protein coding region.

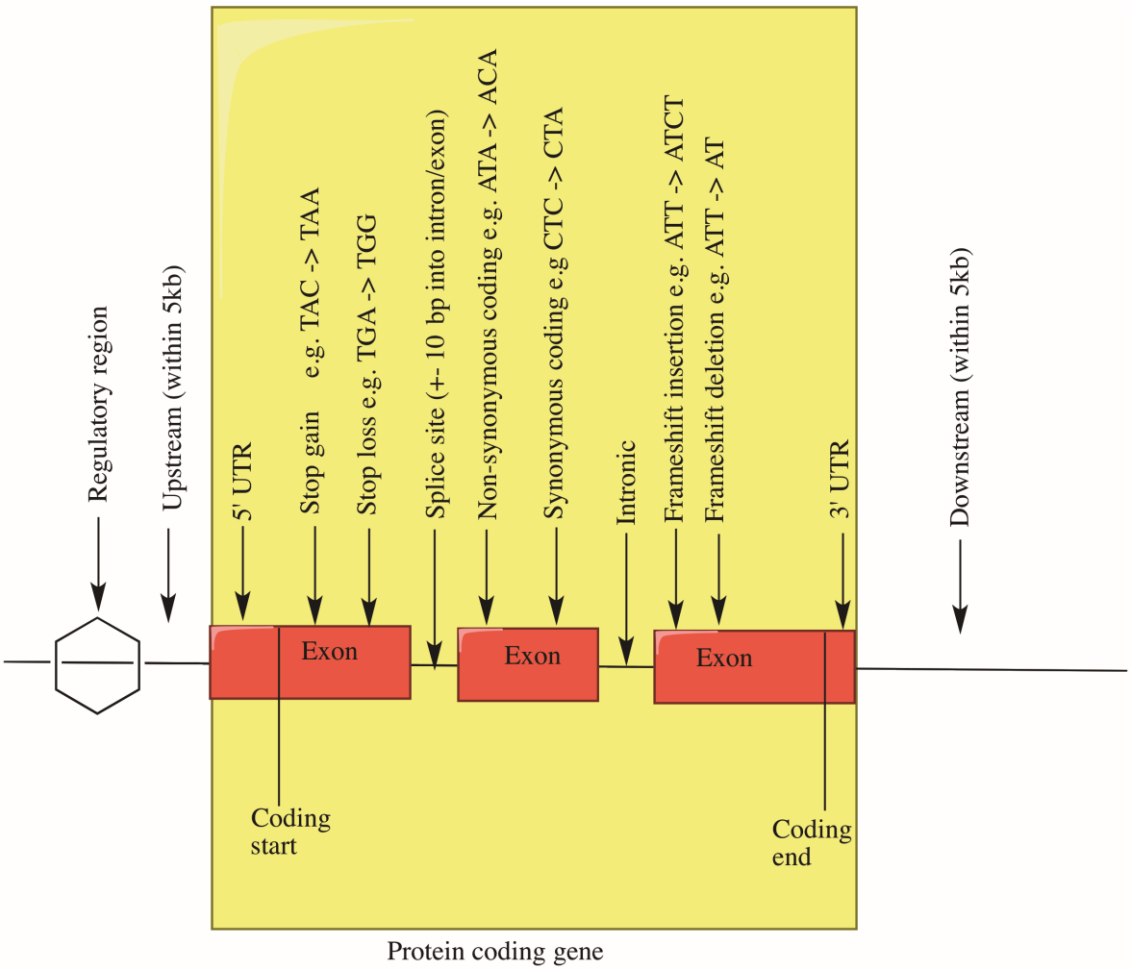



**Figure 1.10 Representation of nucleotides mutations.** Schematic representation of the location of different types of genetic changes within the protein-coding region of a gene. Stop gain mutations occur when an amino acid coding codon is replaced by a premature stop codon; stop loss mutations occur when a stop codon is replaced by an amino acid coding codon; splice site mutations occur when single nucleotide substitutions or indels occur in proximity to the exon-intron boundaries; non-synonymous mutations occur when a codon is substituted by a codon for a different amino acid; synonymous mutations occur when a codon is substituted by a different codon for the same amino acid; intronic mutations occur within an intron; frameshifts indels mutations occur when a number of nucleotides not divisible by three is inserted or deleted resulting in an erroneous translation of the genetic code[93–95].

### 1.2.2 Family and twin studies

Genetic epidemiological studies based upon families can be used to investigate familial trait aggregation and to localise disease causal genes. Family studies can be also used to characterise shared environmental risk factors and their impact on the expression of genetic predisposition. Twin studies helped to gather a better picture of disease risk and to get and estimate of disease heritability. Identical twins are genetically identical whereas non-identical twins share on average half of their polymorphic alleles. Twin studies assume that identical and non-identical twins share the same environment and therefore the difference in disease concordance rates between sets of twin pairs can

be used to determine the disease risk from genetic and environmental components. Studies have shown that the heritability estimates for CD and UC from pooled twin studies is 75% and 67% respectively[96].

IBD epidemiological studies have reported an increased rate of family history in very early onset CD and UC patients. Approximately 75-80% of family members present with the same disease, while 20% have a mixed diagnosis of CD and UC[72].

Twin studies have shown a high rate of genetic concordance for CD compared to UC[72]. A recent meta-analysis of six twin studies with a combined set of 112 monozygotic and 196 dizygotic twin pairs reported concordance rates of 30.3% and 3.6% for Crohn's disease and 15.4% and 3.9% for UC respectively[97], indicating that a large component of IBD risk is indeed genetic. Together, these family and twin studies provided the motivation for the first wave of gene-mapping studies throughout the mid-1990s aimed at identifying the regions of the genome that contribute to IBD risk.

### 1.2.3 Linkage analysis

Linkage analysis is the study of familial inheritance which assigns a disease locus to a specific region on a chromosome defined by several polymorphic markers. If a marker maps close to a causal disease gene it is expected that every affected member of the family will inherit the same allele in linkage with the allele responsible for the pathologic phenotype (Figure 1.11). Close proximity of disease and marker alleles means they will be inherited together and are unlikely to be separated by recombination during meiosis[98]. Linkage analysis is therefore reliant on an accurate genetic map of markers. The markers chosen to conduct linkage studies are usually microsatellites with a distance of 1 centimorgan apart. A centimorgan unit defines two loci with an expected recombination frequency of 1%[99]. Although a centimorgan is not a direct measure of physical distance, it roughly corresponds to one million base pairs (1cM/1Mb).

Evidence of linkage is presented in terms of the logarithm of the odds (LOD) score[100]. The score was firstly developed by J.B.S. Haldane and collaborators in 1947 and then optimised by Newton Morton in 1955. The LOD score compares the null hypothesis, which assumes that the gene associated with the disease is not linked to this genetic location, against the alternative hypothesis stating that the gene is linked to the genetic location. Genetic linkage studies are based on the maximisation of the LOD

score function[101]. A positive LOD score may suggest evidence of linkage while a negative log score may provide evidence against linkage. A typical linkage study in a Mendelian condition will report the locus with LOD scores greater than three, which corresponds to the data being 1000 times more likely to arise due to cosegregation with disease than by chance[102]. Once the genomic region of interest is identified, the region of the genome surrounding the marker is interrogated with microsatellite markes (positional cloning) to identify the likely disease causal genes.

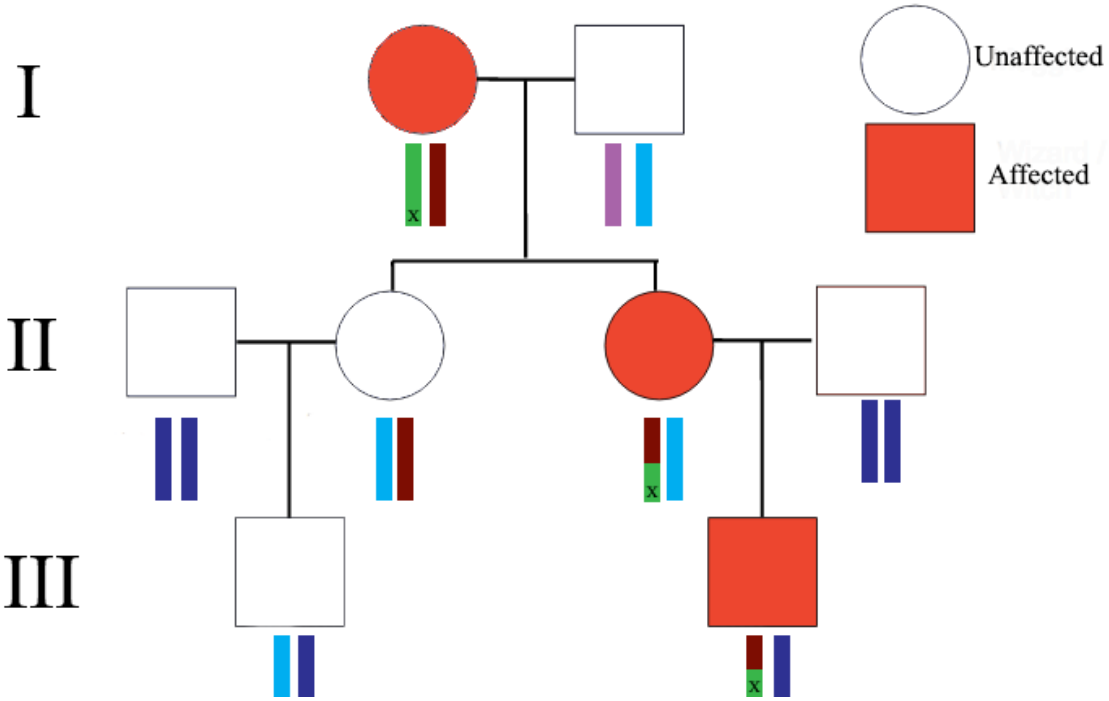



**Figure 1.11 Schematic representation of a pedigree segregating a phenotype of interest.** Square symbols represent males whereas circles females. Affected individuals are shaded in red. Alleles are indicated as bars. In this example, the affected ancestor is a known female in the top generation. The disease causal mutation at a specific locus is indicated by the X on the green haplotype. Although recombination events reduce the extent of the green haplotype transmitted through the offspring of the pedigree, there is perfect segregation of the mutation X in all affected individuals.

The first success of linkage studies was the identification of the genes underlying Duchenne muscular dystrophy and cystic fibrosis. Although successful, one of the drawbacks of linkage studies is the difficulty of assessing a risk allele when the pedigree shows reduced penetrance, in which only a small proportion of those with a given genotype develop the disease. Familial linkage studies are underpowered to determine genes causing polygenic disorders and or monogenic disorders where phenocopy prevents correct assignation of affection status[76]. The term phenocopy

describes the situation where genetically distinct individuals present with the same phenotype[93].

### 1.2.3.1 Linkage Studies in IBD

A total of 11 linkage studies have been reported for IBD,[103]. Linkage studies have identified nine susceptibility regions on chromosomes 1, 3, 4, 5, 6, 12, 14, 16 and 19[72]. (Figure 1.12). Candidate gene analysis was conducted to study potential genes of interest.

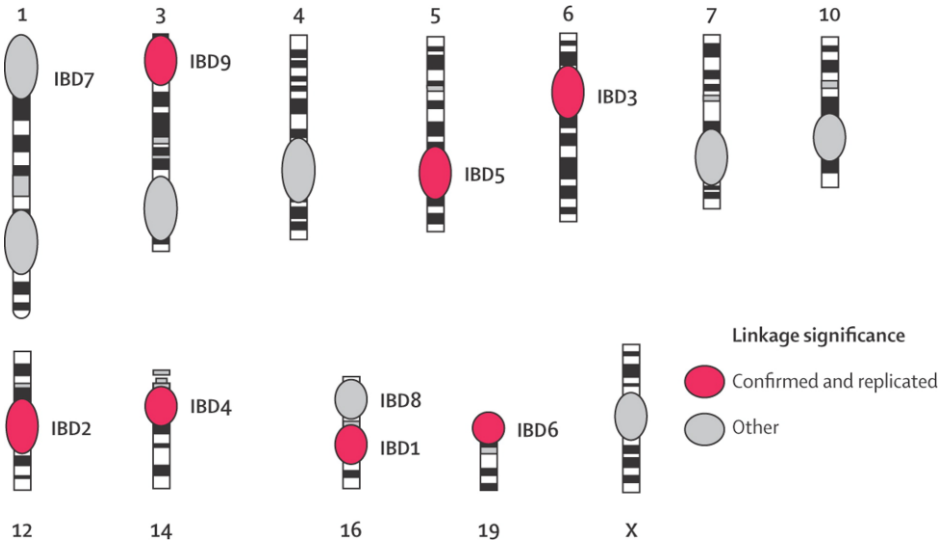



**Figure 1.12 The nine loci implicated in IBD identified by linkage studies**[104] Grey squares represent regions of suggestive, not confirmed, linkage (2 <LOD score < 3) whereas red squares represent significant linkage (LOD score >3)[105]

In 1996 Hugot and collaborators identified the IBD1 locus on chromosome 16 through a linkage study[106,107]. Within the locus, the *NOD2* gene represented the best candidate for CD causality [78,79]. The result was subsequently confirmed in association studies and the causal mutations that underlie the disease risk were identified[108,109]. NOD2 belongs to the nod-like receptor (NLR) family a group of intracellular proteins found widely in nature[108,109]. The *NOD2* gene encodes a protein involved in monocyte recognition of muramyl dipeptide-a peptidoglycan a constituent of both Gram positive and Gram negative bacteria[110]. Firstly Satsangi and collaborators in 1996, and then Duerr and collaborators in 1998 identified the association between the IBD2 locus on chromosome 12 and IBD using linkage studies. No specific genes were identified in this locus. The locus IBD3 on chromosome 6p was identified in 2001. As the IBD3 region overlaps with the *MHC* region, Dechairo *et al*. (2001) suggested that an *MHC*

autoimmune susceptibility gene may be responsible for the positive linkage results. In a study involving 1,841 ulcerative colitis cases and 1,470 controls, Fisher and collaborators found that multiple *MHC* markers showed strong association with ulcerative colitis in the first stage, around rs6927022 in a haplotype block containing the *BTNL2* gene and the *HLA loci HLA-DQA1, HLA-DRA, HLA-DRB5, and HLA-DRB1*. Clear residual association with a SNP within the *BTNL2* gene suggested a contribution of that gene or another in linkage disequilibrium with it. Two independent genome-wide scans found significant evidence and suggestive evidence for linkage on chromosome 14q11-12, also known as the IBD4 locus. The IBD4 locus contains several functional candidate genes for IBD, including the T-cell receptor genes and a number of genes involved in apoptosis. In 2000 the IBD5 locus on chromosome 5q31 was identified using linkage studies. This region contains a number of immunoregulatory cytokines which might be important in the pathophysiology of Crohn's disease: *IL4, IL5*, and *IL13*[111]. Variants within IBD5 have been shown to be associated with a more severe phenotype and with pIBD[112,113]. Nine further regions with suggestive linkage were identified, with little replication across studies[103]. Few loci were identified through linkage studies highlighting the hypothesis that complex disorders were unlikely to be driven by high penetrance loci but instead by modest effect variants[114]. Much greater statistical power to identify low-modest effect loci can be achieved by association studies in a case-control setting.

### 1.2.4 Candidate gene/locus association studies

Association studies became feasible from the late 1990s. These studies do not exploit familial inheritance patterns but are usually case-control studies based on a comparison of unrelated affected and unaffected individuals from a population. These studies compare allele frequencies between a group of healthy controls and a group of unrelated patient cases, the allele at the gene of interest is said to be associated with the trait if it occurs at a significantly different frequency compared to the control group[115]. Although association studies can be performed for any polymorphism across the genome, it is more functionally feasible to test for association within individual genes having a biological relation to the trait but with mechanism not clearly understood. Candidate genes are based on *a priori* knowledge of the gene's biological functional impact on the trait or disease of interest. This had been possible thanks to

relatively inexpensive genotyping, combined with gene mapping and variant discovery efforts. There are three causes of association between the phenotype and the locus: direct association, where the locus is causing the disease; indirect association, where the locus is not the causal risk factor but is in linkage disequilibrium with a causal locus; and spurious association derived from confounding factors such as ethnicity and gender[116]. This approach is limited by its reliance on existing knowledge about known or theoretical biology of disease. Results from candidate gene studies were underpowered to identity variants within genes that later became established IBD risk loci (e.g. *XIAP*, *LRBA* and *FOXP3*).

### 1.2.5 The human genome project

The human genome project (HGP) began in October 1990 and completed in 2003 using Sanger sequencing technology[117]. The HGP presented the architecture of the human genome to the scientific community showing that the total length of the human genome is over 3 billion base pair and the estimated number of human protein-coding genes is 20-25 000[118]. The more complete version of the genome was published in 2004 and since then the Genome Reference Consortium (GRC) continues to facilitate the curation of genome assemblies by improving the human genome reference sequence (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human). The project showed that human individuals share 99.5% sequence identity[119]. The remaining 0.5% of the DNA, including copy number variations (CNVs), shows individual variability within different populations. The sum of all these mutations confers individual variability. Single nucleotide polymorphisms (SNPs) represent the major individual variability of the human genome and they can be classified based on the predicted effect on protein structure.

SNPs are responsible not only for the phenotypic differences between individuals but also for differences in the predisposition or resistance to disease. The “common disease/common variant” hypothesis has been proposed which assumes that common SNPs are the cause of complex disorders[120]. Polymorphic mutations determine numerous pathologies and the introduction of genome wide studies have made it possible to identify new loci associated with common diseases. However, it has been shown that common genetic variations identified to date from genome-wide

association studies (GWAS) collectively explain only a small fraction of the burden of any disease in the population at large suggesting that multiple rare variations might be contributing to human common diseases[121].

### 1.2.6 The HapMap project

The International HapMap Project (http://www.hapmap.org/index.html.en) was initiated in 2002 to catalogue all the common human genetic variations. The project aimed to map polymorphisms across the genome to allow the creation of high-density polymorphism maps that, when combined with advances in genotyping technology, would facilitate association testing across all genes. The first phase of the project analysed more than one million polymorphisms in 269 individuals of four populations: Caucasian, African, Chinese and Japanese[115,122]. The first phase of the project resulted in the identification of recombination hot spots across the genome. The introduction of parent-offspring trios aided the understanding of the haplotype structure[122].

The second phase of the HapMap project was conducted using the same samples from phase I, but contained ~3.1 million markers, while the third phase of the project extended the number of genotyped non-disease individuals to 1,184 across multiple populations and a broader range of ethnic groups[123].

The dataset generated by the HapMap project provided a catalogue for conducting genome-wide association studies by providing data to discriminate recombination hotspots and being a resource for designing tag SNP sets across different populations[124]. These results, together with technological advances, made it feasible to conduct GWAS in order to identify loci associated with complex traits or disease risk.

### 1.2.7 Genome Wide Association studies

Although genome wide association studies have the same principle of candidate gene/locus association analysis, GWAS do not focus the analysis on *a priori* genomic regions of interest but they scan the entire genome for common genetic variation using hundreds of thousands of SNPs spread throughout the genome. SNPs showing significant association with disease status point to regions of the genome likely to harbour disease relevant genes. Differently from linkage studies, GWAS are not restricted to families and can be powered to detect loci with small to moderate effect sizes.

Affymetrix and Illumina, developed the most used chips for conducting GWAS[125] Each used a slightly different form of microarray, and differed in their selection of SNPs[126]. Genotypes at SNPs that were not directly assayed can be inferred through imputation algorithms based on the genotypes from a representative reference set of haplotypes, allowing for individual studies using different genotyping platforms to be effectively combined into meta-analyses.

The GWAS approach has been an important and powerful tool for the identification of genes involved in complex traits since they allow a hypothesis free simultaneous analysis of a large number of unrelated individuals. GWAS have been successfully used for the discovery of low-penetrance susceptibility genes involved in IBD (Figure 1.13).

Individual loci found with GWAS have a small effect and explain just a small part of the genetic contribution to the different diseases. As most of IBD patients carry a large number of common risk variants, the effect size of individual risk variants is low (odd ratio < 1.3)[127]. GWAS require large case-control sample sizes to identify low effect variants and to reach a sufficient statistical p-value (conventional threshold of $5x10^{-8}$ for genome-wide significance). Although the use of large-scale association studies help to increase the percentage of heritability explained, for most disorders and traits the majority of the casual genes that might have high functional consequences are still poorly understood[127] .

The phenomenon of 'missing heritability' in complex diseases remains one of the major issues in the field of genetics. The term missing heritability refers to the component of genetic variation that is not explained by variants identified to date. Scientists have suggested that missing genetic variance might be explained by gene interactions which are not captured by GWAS, structural variation including copy number variants (CNVs, insertions and deletions); copy neutral variation (inversions and translocations) which are poorly captured by the arrays used in GWAS, and rare variants present in 5% or less of the population. Rare variants are poorly represented by the genotype technologies available which were designed to capture common variation[128,129].

Today, the emerging technique for detecting and elucidating the missing variation in complex disorders is next generation sequencing (NGS) [129].

**Figure 1.13 SNP-trait associations with p-value < $5.0x10^{-8}$ (GWAS catalogue)** Each colour coded dot represents an SNP for a specific trait (http://www.ebi.ac.uk/fgpt/gwas/#diagramtab)

### *1.2.7.1 Genome Wide Association studies in IBD*

*NOD2* was the first gene found to be associated with CD by the analysis of the locus IBD1 which was previously identified by linkage studies. *NOD2* regulates the recognition of the bacteria and the activation of the inflammatory pathways: nuclear factor kappa B (NFκB) and mitogen activated protein kinase signalling pathways. Three NOD2 polymorphisms have been independently associated with CD: Arg702Trp, Gly908Arg and Leu1007fsinsC[108,130]. To date a total of 22 independent GWAS on IBD (conducted on adult and on early onset individuals) have been carried out since the first study was published, accounting for 201 IBD loci[131]. The genes identified by GWAS helped to identify important pathways and biological processes for the pathogenesis of IBD (e.g. gene involved in the autophagy pathway, genes involved in the innate and adaptive immune response), and to point out the shared genetic overlap between autoimmune diseases. A turning point in GWAS was the publication from the Wellcome Trust Case Control Consortium (WTCCC) in 2007. The WTCCC represented the largest GWAS conducted at the time including 14, 000 cases across seven diseases

and 3,000 controls. The WTCCC used techniques and methods that became the gold standard in subsequent GWAS (e.g. the genome-wide significance threshold for association as $p< 5x10^{-8}$, and the importance of performing replication in independent samples). The WTCCC study identified 24 loci, of which 14 were novel for traits like type 2 diabetes (3), rheumatoid arthritis (3), Crohn's disease (9) and type 1 diabetes (7)[132].

Only two GWAS have been conducted on early onset IBD. Kugathasan in 2008 conducted the first GWAS on a paediatric cohort. The cohort was composed of 1011 paediatric onset IBD divided into 725 CD and 261 UC. The study showed five additional novel IBD loci that met genome-wide significance: 16p11, near the cytokine gene IL27, 22q12, 10q22, 2q37 and 19q13.1[133,134]. Imielinsky and collaborators conducted the second GWAS study in 2009 on 3426 affected paediatric patients. The study replicated 29 out the 32 loci that at the time were previously associated with adult-onset CD and 13 out of the 17 loci associated with adult onset UC. Seven novel loci were associated with early-onset IBD. However, unless GWAS is performed in an exclusively well-powered paediatric onset IBD cohort, it is difficult to discard the existence of additional early-onset IBD genes. *TNFSFR6B* is a novel locus specific to the early-onset phenotype. Studies have demonstrated that the associated allele provides a protective effect. *TNFSR6B* function is involved in lymphocytes function regulating T-cell apoptosis. Patients presenting elevated levels of lymphocyte also harboured the *TNFSFR6B* variants suggesting the importance of the gene in the pathogenesis of IBD. Another locus specific to early onset IBD is 16p11. The region contains several genes including *IL27* which regulates the T-cell differentiation and the $T_h17$ pathway. *IL10* polymorphisms have shown genome-wide significance for CD in the early-onset population and not in adult disease[133]. Mutations in the genes *IL10* and *IL10RA* and *IL10RB* have been reported in immunodeficient children with severe infantile-onset IBD. *IL10* is a cytokine with anti-inflammatory properties[79]. The association with the disease has been confirmed by animal models as *IL10* deficient mice develop spontaneous colitis[79,86,135]. Although genetic differences exist in adult-onset IBD, results from early-onset GWAS have highlighted the overall similarities in these phenotypes.

With the exception of *NOD2*, the common (minor allele frequency, MAF, > 5%) polymorphisms with small effect size (OR< 1.3) discovered by GWAS explain only a

small fraction of the IBD disease heritability. Larger sample size GWAS are needed to increase the power of GWAS to detect common (MAF > 5%), low-frequency (1% <MAF<5%) variants with smaller effect size. Meta-analysis of GWAS have further identified common variants implicated in disease predisposition.

#### *1.2.7.2 Meta-analysis*

Meta-analysis combines datasets from individual GWAS to increase sample size and power. The first CD meta-analysis of three GWAS identified 21 new loci bringing the total number of IBD loci to 30. This was followed by the discovery of 71 new loci by a meta-analysis of six GWAS[107]. A study conducted by Jostin and collaborators in 2012 identified 71 new associations with IBD, increasing the number of associated loci meeting genome-wide significance thresholds to 163[59]. These results represent the most successful for any complex disease to date. Across the 163 loci, 66 loci are shared with other immune-mediated diseases and 110 are shared between UC and CD suggesting a common genetic component between the two diseases[104]. For 53% of loci the definitive casual genes have been identified but for many loci the causal allele or gene is still unknown. The functions of the identified genes are enriched for the immune response such as cytokine production, tumour necrosis factor, IL10 signalling and lymphocyte activation. The autophagy pathway was suggested to play a role in CD in early GWAS via the association of the *ATG16L1* and *IRGM*. Autophagy is the process involved in the degradation and recycling of cytosolic components and of resistance against infection and the removal of intracellular microbes. To date, an additional 38 new genomic loci have been associated with a trans-ancestry association study IBD increasing the number of known loci associated with IBD to 201[136]. Of these 38, 27 loci were associated with both Crohn's disease and ulcerative colitis, seven were specific to Crohn's disease and four were specific to ulcerative colitis. Twenty-five of the 38 newly associated loci overlapped with loci previously reported for other traits, including immune-mediated diseases, whereas 13 had not previously been associated with any disease or trait. Together, these loci explain 13.1% and 8.2% of the heritability for Crohn's disease and ulcerative colitis, respectively. However, of the 201 loci only a handful has been assigned to specific functional variants. One of the limitations of GWAS is their power of only being able to detect associations that are well covered by the SNP microarrays. Populations with a different structure from the HapMap ethnicity

group or meta-analysis combining studies with different populations can induce bias in the study[137]. Reproducibility is also a drawback of GWAS leading to reports of false positive results. It is now common knowledge and it has been proven by several studies that rare variants are likely to play an important role in the pathogenesis of complex disease due to purifying selection maintaining damaging alleles at low frequencies[129]. To be able to detect associations with these very rare variants it is necessary to utilize an enormous sample size which increases the price of GWAS arrays. Finally, the stringent correction made by multiple testing in GWAS might discard disease-associated loci. Most of the loci identified through GWAS are noncoding and are not immediately informative; to address these issues projects applying the next generation sequencing technologies and wet-lab validation techniques have been developed (e.g. the ENCODE project)[129].

### 1.2.8 Sequencing Technologies

Recent advances in next generation sequencing technology have made it possible to explore and analyse rare variations which are believed to partially contribute to the heritability not explained by GWAS[138–140].

The 1000 Genomes Project[141] described an abundance of rare and low frequency variations within the genome and how these mutations are enriched for functionally relevant disease-causal mutations. The Wellcome Trust-funded UK10K project identified 50%, 98% and 99.7% of SNP with a frequency of 0.1%, 1% and 5% respectively in 2,500 individual genomes[141]. Studies have shown that loci associated with complex traits are enriched for rare variants that cause known Mendelian disorders and it has been suggested that recessive variants confer risk related to complex diseases[142]. Rare disease causal variants are also found in genes with known common associated variants. The 1000 Genomes Project has also shown that at the most highly conserved coding sites rare and low frequency variations tend to occur: 85% of non-synonymous variants and 90% of stop-gain and splicing variants have a MAF < 0.5% in frequency compared to 65% of synonymous mutations. Rare disease causal mutations will never reach a high frequency within a population as purifying selection will act on them[141]. Although GWAS has helped discover thousands of associations, this approach alone cannot identify rare variants such as SNPs, indels, and copy number variants (CNVs) which partially contribute to the “missing

heritability". NGS has brought a major shift in the study of genetic diseases since rare variants can only be observed with deep resequencing of individuals[143]. The falling cost of sequencing has allowed the direct assaying of low-frequency variants via re-sequencing studies (Figure 1.14). The recent successes of whole-exome and whole-genome sequencing in Mendelian and complex diseases has proven to be an effective tool for the identification of rare mutations within disease-associated genes[144]. Thanks to whole genome and whole exome sequencing the state of gene discovery underlying Mendelian disorders has risen from ~166 per year between 2005 and 2009 to 236 per year between 2010 and 2014[145] (Figure 1.15).

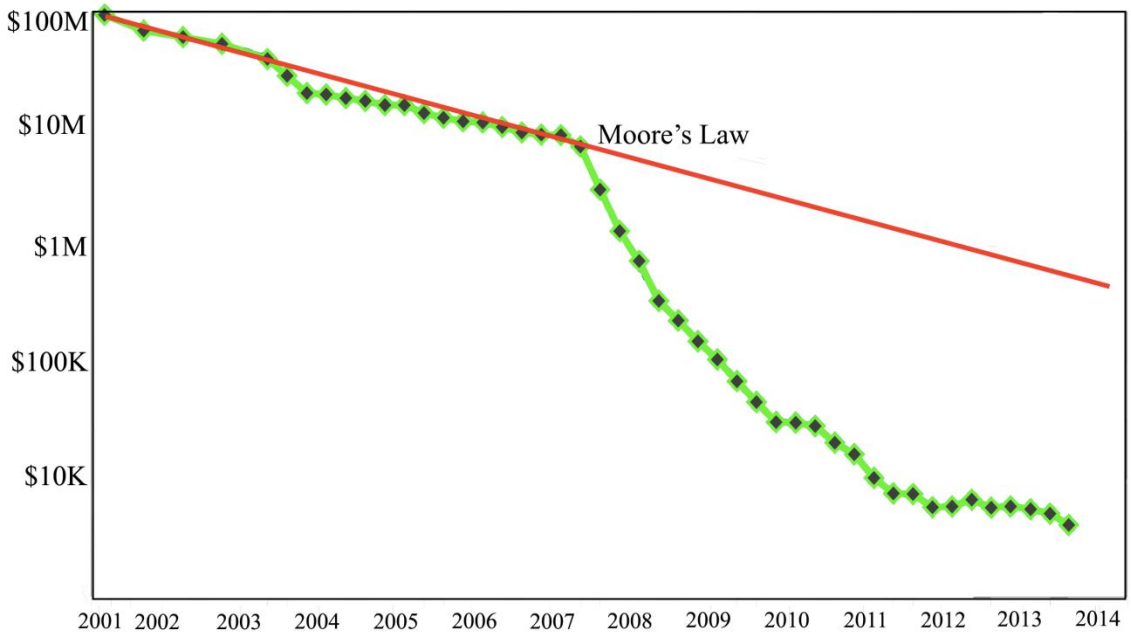



**Figure 1.14 Cost per genome sequencing over time.** **http://www.genome.gov/sequencingcosts/** the graph shows hypothetical data reflecting Moore's Law, which describes a long-term trend in the computer hardware industry that involves the doubling of 'compute power' every two years and the decreasing cost of DNA sequencing per year.

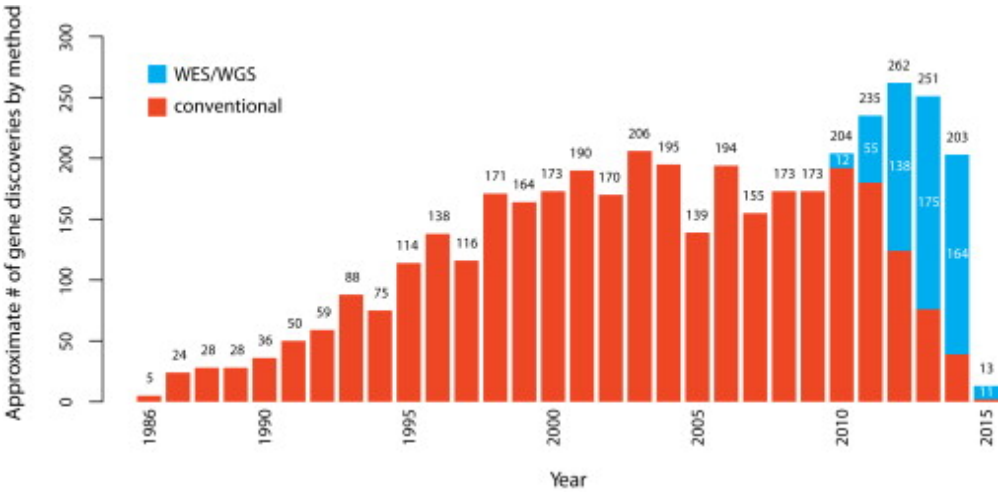



**Figure 1.15 Approximate number of gene discoveries made by WES and WGS versus conventional approaches since 2010.** From 2013, WES and WGS (blue) have discovered nearly three times as many genes as conventional approaches (red) [145].

### *1.2.8.1 Sanger sequencing*

First generation sequencing technology refers to DNA sequencing based on the Sanger method that has been widely used since its introduction in 1977. The method is defined as the chain-termination method as it uses normal and dideoxynucleotides (ddNTP's). Dideoxynucleotides are nucleotides containing a hydrogen group on the 3' carbon instead of a hydroxyl group (OH). These modified nucleotides prevent the formation of the phosphodiester bond between the dideoxynucleotide and the next incoming nucleotide resulting in the termination of the DNA chain[146,147]. Sanger sequencing (

Figure **1.16**a) requires random fragmentation of single strand DNA followed by PCR or the DNA is cloned into a high copy number plasmid which is used to transfect *E.coli*. Next the plasmid DNA undergoes several "cycle sequencing" reactions in which the modified ddNTP's nucleotides are added. The step of adding the nucleotides is conducted in four different conditions, one for each of the ddNTP's nucleotides. The incorporation of the modified nucleotide causes the termination of the DNA replication and the formation of DNA fragments of different lengths depending on the last base incorporated. The fragments are then sorted by size on a polyacrylamide gel with each of the four reactions run in a different lane and scored according to their molecular masses[148]. The read length of the Sanger protocol is 500-1000bp for a single run. Implementations of the Sanger method (DNA Sequencing by Capillary Electrophoresis) include the ability to conduct a single reaction containing all four

ddNTPs are labelled with fluorescent dyes. The extension product and then separated using electrophoresis into a single glass of capillary. DNA fragments are scored based on their masses and fluorescent labels are excited with a laser to allow interpretation of the DNA sequence[148](

Figure **1.16**b).

Dideoxynucleotides are nucleotides containing a hydrogen group on the 3' carbon instead of a hydroxyl group (OH).

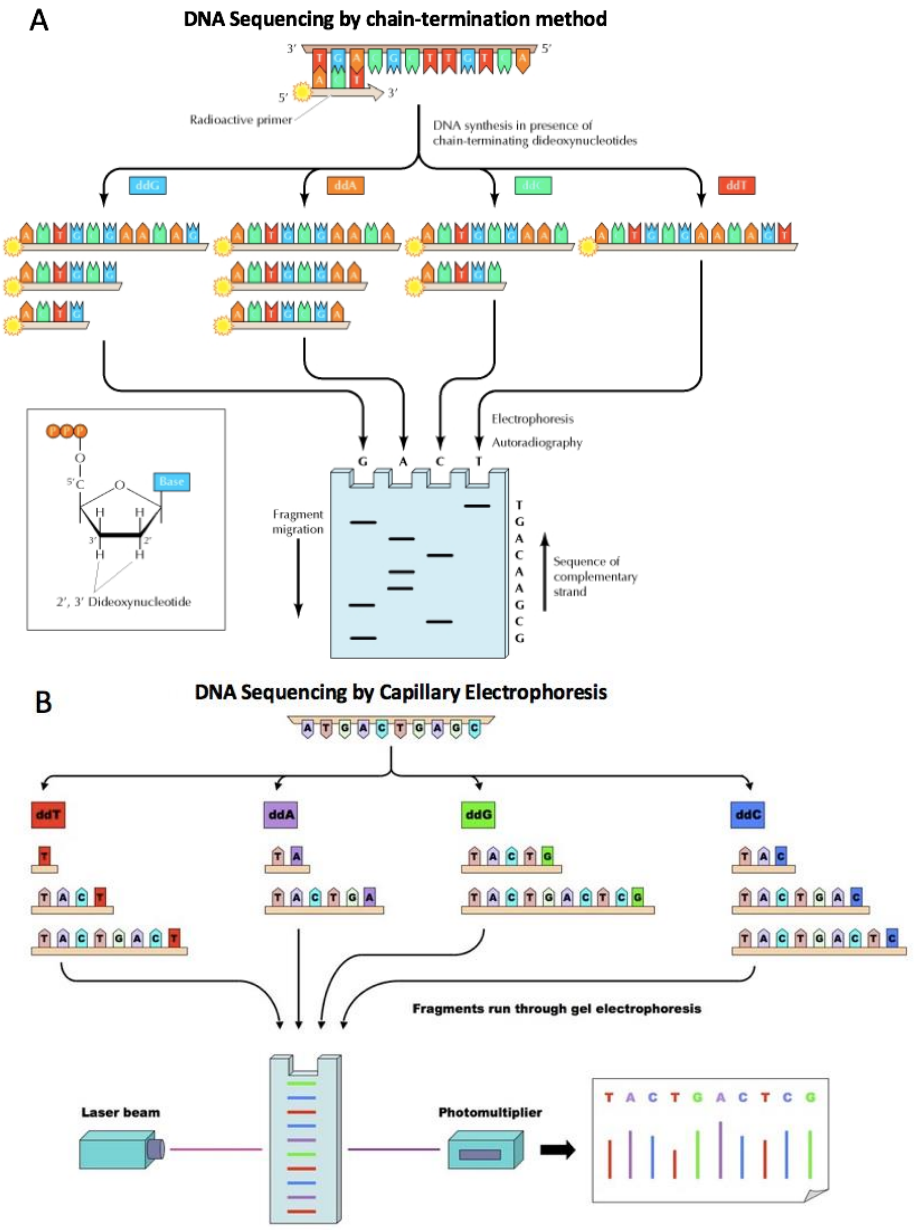



**Figure 1.16 DNA sequencing by the Sanger procedure**[146]**.** A) A single strand of DNA to be sequenced is hybridized to a 5'-end with a labelled primer. Four separate reactions are carried out, each containing one dideoxynucleotide, which lack OH groups at the 3', mixed with polymerase as well as the three other normal deoxynucleotides. The lack of the 3' OH group prevents addition of the next base, so synthesis of that DNA strand terminates. Each reaction produces a series of products extending from the radioactive primer to the base substituted by a dideoxynucleotide. Products of the four reactions are separated by electrophoresis and analysed to determine the DNA sequence. B) DNA Sequencing by Capillary Electrophoresis: all fluorescently labelled ddNTPs are added with different fluorescent dyes. The extension products are then electrophoretically separated in a single glass capillary. Fluorophores are excited by the laser at the end of the capillary. The DNA sequence is interpreted by the colour that corresponds to a particular nucleotide.

### *1.2.8.2 Next generation sequencing technology*

Next generation sequencing (NGS) has increased the number of reads sequenced up to several million bases in a single run (100-200bp per read). Although there are differences in the chemistries available for next generation technologies, such as template preparation and the sequencing process, they all share the same concept[143]. Firstly, there is the step of library preparation where DNA is stochastically fragmented into desired sizes by enzymatic reactions or sonication. Then DNA is ligated to adaptors which are short DNA sequences complementary to the oligonucleotide used in the amplification and the sequencing steps. Secondly the DNA is amplified using PCR. In paired-end protocols the two ends of each fragment are sequenced with the advantage of reducing artefacts by giving long-range positional information[149]. In the sequencing process fluorescent nucleotides are added to the chain producing a release of light which is then detected. These light signals are then translated into a nucleotide sequence, the original sequence is reconstructed by aligning the individual reads to the human genome and the alignments are examined to identify genetic variation[143].

Additional contemporary technologies have been developed over the years to study RNA, proteins and DNA methylation; however these methods are outside the scope of this thesis.

#### *1.2.8.3 Whole genome sequencing*

Whole genome sequencing (WGS) represents the sequencing of the entire genome of an organism. One of the main advantages of WGS is that it provides a complete catalogue of an individual's genetic variation[150] starting from 250ng of DNA at a concentration of 2ng/uL for conducting whole genome sequencing at 30-60x depth.

Due to the decrease in cost of sequencing (cost of $1,000 per genome at 30x coverage), projects like the 1,000 Human Genome Variation project and the International Cancer Genome Consortium were possible. More recently, the 100,000 Genome Project aims, within four years, to apply WGS as a new diagnostic tool for the NHS. The project goal is to sequence 100,000 genomes from around 70,000 people. Participants are NHS patients with rare diseases, plus their families, and patients with cancer.

However, WGS is still prohibitively expensive for many large population based studies. Additionally, the difficulties in managing and interpreting the large amount of data produced by WGS do not allow the routine application of this technology in clinical

settings. To date, data derived from the FANTOM[151] and the ENCODE[152] project provide critical information for characterizing the functional elements within the non-coding region of the genome, previously defined as "junk".

The development of sequence capture technology enabled sequencing of the whole exome, which covers ~1.5% of the human genome[153].

### *1.2.8.4 Exome sequencing*

Whole exome sequencing (WES) is currently the most powerful and cost effective approach to sequence the coding regions of the genome[154](Figure 1.17). There are approximately 120-180,000 coding exons in the human genome representing ~35Mb of the total 6.4 Gb diploid human genome[155]. The protein coding region constitutes approximately 1-2% of the genome and contains 85% of disease-causing mutations[114]. Exome sequencing has been successfully applied in Mendelian[156] and complex disorders[116] revealing additional disease causing mutations not captured by GWAS[157].

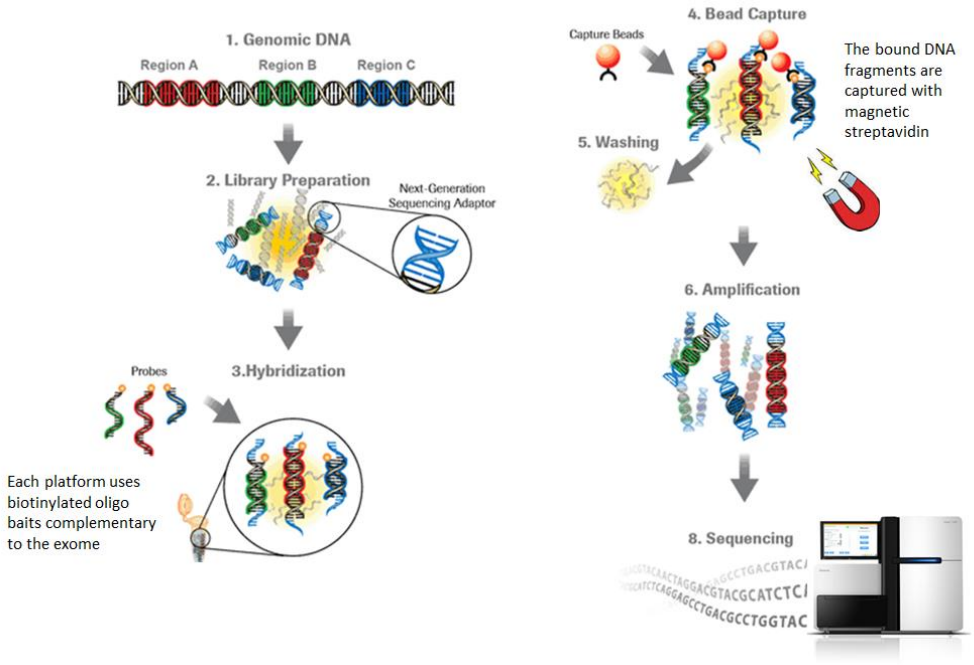



**Figure 1.17 Next generation sequencing method.** Genomic DNA is extracted and fragmented into desired sizes. DNA is then ligated to oligo baits, which are short DNA sequences complementary to the oligonucleotide used in the amplification and the sequencing steps. The bound DNA fragments are "pulled down" with magnetic streptavidin. DNA is amplified using PCR. In the sequencing process fluorescent nucleotides are added to the chain producing a release of light. The light signals are translated into nucleotide sequence[143]. (adapted from http://www.nimblegen.com/exomev3launcheq/)

The application of WES to the diagnoses of 250 patients successfully identified a disease causal mutation in 62 of them[158]. In 4% of the cases the diagnoses was revised based on WES findings and 3% of the patients obtained better management. Among the 62 patients, 33 had autosomal dominant disease, 16 had autosomal recessive disease, and nine had X-linked disease. The 25% successful molecular diagnosis was higher than other genetic tests such as karyotype analysis, Sanger sequencing of single gene and chromosomal rearrangements[158].

1.2.8.4.1 Exome sequencing in IBD

Exome sequencing is becoming more widely used in clinical research and diagnostic for patients presenting with unusual phenotypes and can lead to targeted treatment options. To date, 51 genes have been associated with monogenic disease manifesting in an early onset IBD-like phenotype.[159,160] The application of exome sequencing identified homozygous mutations in the interleukin 10 receptor (*IL10)* gene and its associated receptor alpha and beta subunits (*IL10RA* and *IL10RB*) in children presenting very-early-onset IBD (VEO, age of onset < 6 years).[79,161] The discovery of the disease causal mutations helped to personalize treatments inducing a sustained remission in the patients.[79,161] Whole-exome sequencing helped identify a mutation in the *MEFV* gene, resulting in a diagnosis of familial Mediterranean fever in a VEO patient with intractable IBD that was unresponsive to medical therapy.[162] A novel hemizygous mutation within the *FOXP3* gene was identified in three brothers from a nonconsanguineous family with atypical early-onset IBD phenotype through the application of WES study.[163] Two siblings from a consanguineous family with neonatal-onset inflammatory skin and bowel disease were identified as homozygous carriers of a four base pair deletion within the *ADAM17* gene.[164] The investigation of a child with intractable IBD using whole-exome analysis by Worthey *et al.*[77] found a hemizygous mutation in the gene X-linked inhibitor of apoptosis (*XIAP*). The same mutation was confirmed in the asymptomatic mother. Based on these findings, this patient underwent hematopoietic stem cell progenitor transplantation with a resolution of symptoms and sustained remission following this targeted treatment approach.[77]

This finding highlights the critical role of exome sequencing in carefully selected patients by providing diagnoses that can guide treatment[162,165].

#### *1.2.8.5 Targeted sequencing –gene panels*

Targeted sequencing allows the sequencing of specific regions of interest. In addition to being cheaper than whole exome and whole genome sequencing, disease-targeted testing has certain advantages and holds a firm place in diagnostic evaluation. For a many years genetic testing was only of marginal use in the diagnostic evaluation of a patient, however, with the introduction of large multi-panel testing, many more rare genes that contribute to a phenotype, and genes involved in a broader range of phenotypes can be included in testing. For example alpha-galactosidase (*GLA*) was included in a test for hypertrophic cardiomyopathy, this is a rarely considered gene for left ventricular hypertrophy, but sequencing it enabled a subset of patients to receive treatments that can slow or even cause a regression of their disease[166]. Another example is the *CFTR* gene whose sequence can suggests the likelihood of identifying the cystic fibrosis aetiology of the patient's disorder[167]. With the continued increase in the rates of positive results for disease-targeted testing, clinicians have also begun increasing the level of genetic testing in the evaluation of their patients, as a positive genetic test can save much time and cost in identifying an aetiology. Most targeted genetic tests return a result in 2–8 weeks, which is fast turnover compared to the many years some patients wait trying to understand the cause of a rare disorder.

Although whole exome and whole genome sequencing are becoming more common for identifying disease causal mutations, a targeted approach is more comprehensive in its coverage for the genes included in typical disease-targeted testing. Targeted gene sequencing can produce a higher or complete coverage of genes by integrating Sanger sequencing to fill in the missing NGS content in exome sequencing. Typical coverage of exons is approximately 90–95%[8], but when analysing small sets of genes that are implicated in a known disorder, coverage can be much lower. Poor coverage can result from various factors, including probes that are not tiled for certain genes either because the genes were not chosen for inclusion during assay development or because repetitive sequences prevented inclusion, and poorly performing probes owing to GC-richness, which interfere with the sequencing process, or low mapping quality.

#### *1.2.8.6 Next generation sequencing data analysis*

Following the creation of raw data from the sequencing process, the data analysis consists of five steps; the first step is the evaluation of the reads to confirm they meet

standardised quality criteria. This process is necessary since raw data are exposed to different biases such as base calling errors, base insertions and deletions and poor quality reads. This analysis is different for each of the NGS platforms available because of differential susceptibility of alternative chemistries and algorithms used to detect the bases[168]. Several tools like FastQC[169] are available to perform this process. The most common output format produced is the FASTQC summary report with graphs and tables of the data quality[143]. Quality values (Figure 1.18) are reported as Phred scores which are a measure of the DNA sequence quality[170,171]. The Phred score represents the probability of a base-call being incorrect. The Phred (Q) equation is: Q = -10 * log10p where p is the probability that the particular base-call is incorrect; a base-call with a Q score of 20 has a 1% probability of being incorrect; Q of 30 corresponds to 0.1% and Q of 40 to 0.01%.

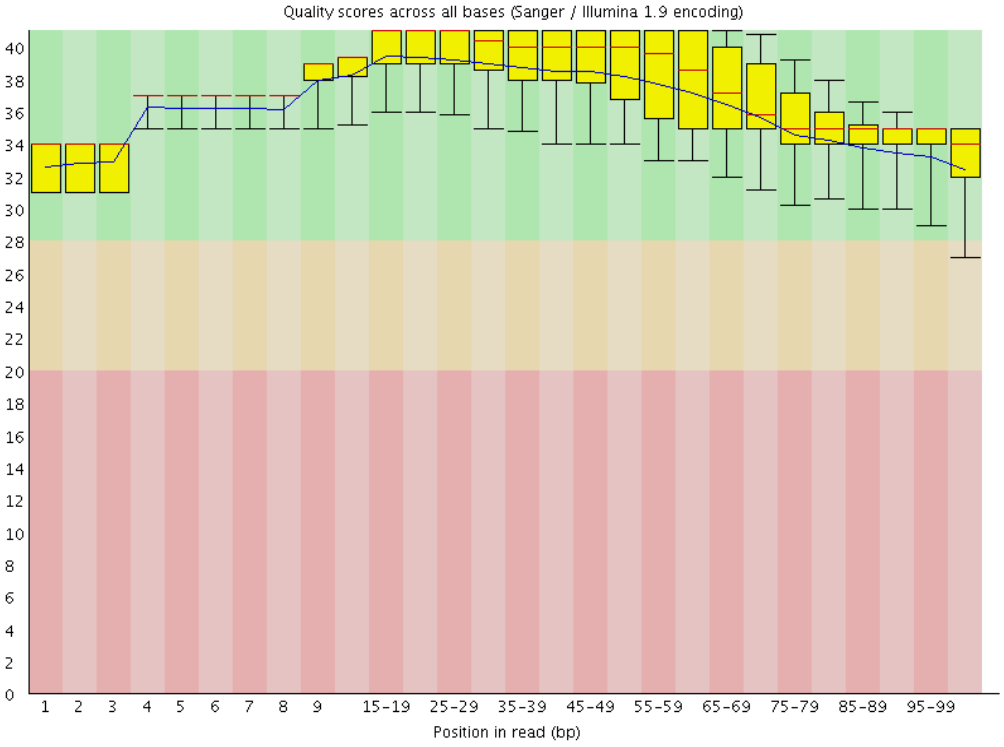



**Figure 1.18 FastQC file of the range of quality values across all bases at each position.** The central red line is the median value; yellow box represents the inter-quartile range (25-75%); upper and lower whiskers represent the 10% and 90% points and blue line represents the mean quality. The y-axis on the graph shows the quality scores. The higher the score the better the base call. The background of the graph divides the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red). The quality of calls on most platforms will degrade as the run progresses, therefore it is common that base calls fall into the orange area towards the end of a read (taken from http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/2%20Per%20Base%20Sequence%20Quality.html).

The second step aims to align the reads to an existing reference genome (Figure 1.19). Servers[152,172] such as University of Santa Cruz (UCSC)[172] hosting data for each resource

containing all historical versions of the human reference genome; the currently commonly used version is hg19/ GRCh37 although some are moving to the newer version hg38/GRCh38 of the human genome.

Alignment to the reference genome can be conducted with several alignment programs such as BWA[173], Bowtie[174], MAQ[175], Novoalign[176] and SOAP[177]. All of these software takes into account the number of expected mismatches within a read, the error sensitivity and specificity of the technology used. The alignment steps are based on the accuracy of both the base calling and of assessing the quality of the calls. The alignment algorithms are able to "index" the reads of the reference genome or the whole reference genome itself to speed up their mapping algorithm and to conduct a rapid search. Each successively aligned new character allows the software to estimate the list of positions to which the read might map[178].


**Figure 1.19 Small read alignment** NGS short reads (in black) aligned to the human genome reference. Mutations are seen as coloured vertical lines on reads (green rectangle). The depth of a specific position is the sum of the number of reads mapping to that position.

During the mapping step, the reads may overlap and cover the same region of the reference genome multiple times. Depth of a specific position is the sum of the number of reads mapping to that position. Depth of coverage can be calculated from the length of the original genome (G), the number of reads (N), and the average read length (L) as N/times L/G. For example, a hypothetical genome with 2,000 base pairs reconstructed from 8 reads with an average length of 500 nucleotides will have 2-fold redundancy.

A high read coverage in sequencing is desired because it can reduce errors in base calling, assembly and it is needed for calling heterozygotes. The result of the mapping process provides parameters for each sequenced position which include base quality, coverage, alternative allele, reference allele and the number of reads at that position. The mapping process is a critical step since the incorrect alignment of the reads can lead to errors in the identification of the variants and in inferring the genotype calls[168].

Methods for optimising the alignment is to use paired-end reads with a local realignment around the indels, to remove the duplicate reads generated during the PCR process, to reduce GC bias[179], and to exclude reads that can only be mapped to the reference genome with multiple mismatches, as they help to resolve chromosomal rearrangements such as insertions, deletions and inversions[168](Figure 1.20).

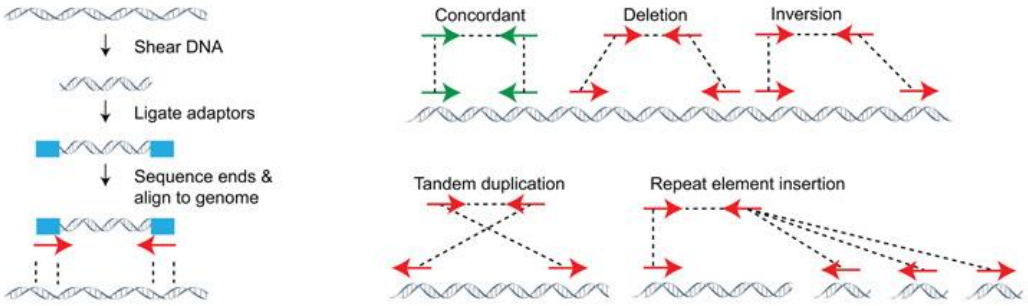



**Figure 1.20 Paired-end DNA sequencing with breakpoint detection.** Concordant read pairs map to the reference genome with the expected size and orientation (green arrows) however, read pairs that span genetic changes are discordant (red arrows)[180].

The third step consists of variant identification. SAMtools[181]and GATK[182] are two common examples of tools that detect variation from the mapping information. SAMtools deals with Sequence Alignment/Map (SAM) files and Binary Alignment/Map (BAM) files and contains a sub-command called BCFtools. The software is able to call SNPs and short indels from the alignment file[133]. The Genome Analysis Toolkit (GATK) is a software library that provides several functions for manipulating sequence data including SNPs and genotype calling[183]. The output file for this step is a VCF (Variant Call Format) file, a standardized format for storing DNA polymorphism data such as SNPs, insertions, deletions and structural variants, together with annotations (Figure 1.21). Given the large number of variant sites in the human genome and the number of individuals analysed the VCF files are usually stored in a compressed form. Compressed files can be decompressed and fast access to the file can be achieved by indexing the genomic position using tabix[181], a generic indexer for TAB-delimited files. Programs for compression, decompression and indexing are part of SAMtools.

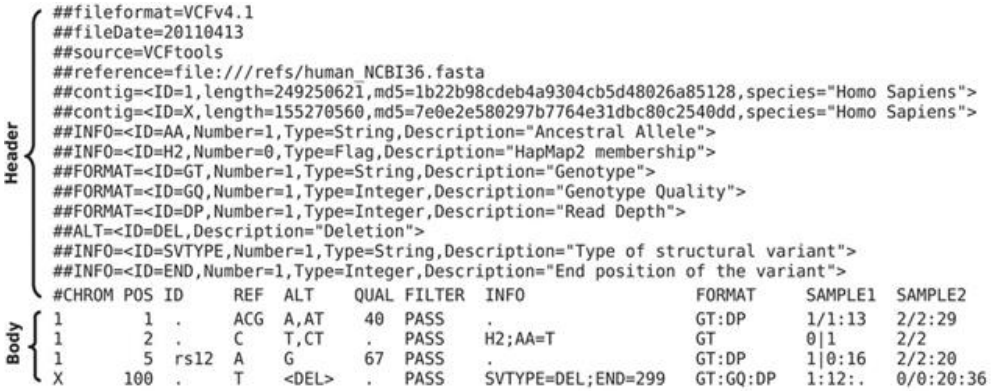



**Figure 1.21 Example of VCF file.** The header lines ##fileformat and #CHROM are mandatory lines. Each line of the body describes a variant present in SAMPLE1 and/or SAMPLE2 at a genomic position. Alternate alleles are listed in the ALT column and and reference allele in the REF column. Genotypes are indicated as 0/0, 0/1 and 1/1 for homozygous reference allele, heterozygous and homozygous for the alternative allele respectively. As an example, the second line of the body shows a SNPS (C>T) and an insertion(C>CT) in SAMPLE1 and SAMPLE2 respectively[184].

Next the vast amount of data produced is processed to predict the functional impact of the variants for determining candidate causal mutations. This step is called variant annotation. The annotation is the process of adding pertinent information about the raw DNA sequences to the genome by describing and identifying which regions can be called genes and thereby its products and functions[185]. Functional annotation represents a major bottleneck to clinical interpretation and the results can have a strong influence on the ultimate conclusions of disease studies[186]. Incorrect or incomplete annotations can cause researchers both to overlook potentially disease-relevant DNA variants and to exclude interesting variants in a pool of false positives[187]. The main challenge is to interpret the large number of apparently novel genetic variants present by chance in any single human genome, making it difficult to identify which variants are causal, even when considering only nonsynonymous variants. Recognition of functional variants is at the centre of NGS data analysis and bioinformatics. It is challenging to develop software with the ability to distinguish low-frequency alleles inherited from ancient ancestors, from *de novo* or extremely rare mutations recently introduced into the population. Computer programs are essential to this process; however, human interpretation is often required to evaluate computer-generated gene models. Although over the last years computational approaches have been developed to assess the *in silico* functional impact of mutations, there is often disagreement between approaches. The most common algorithms used for the prediction of the functional impact of the mutation on the protein are SIFT[188],

PolyPhen2[189], Gerp[190], Grantahm score[191], Phylop[192] and composite algorithms like Kggseq[193] and CADD[185].

Annotation software such as ANNOVAR[194], a command line tool for functional annotation of various genomes, are able to individually interrogate multiple of the algorithms for assessing variant causality and output the result in a single report. ANNOVAR supports single nucleotide substitution, small insertions/deletions and copy number variations (CNV) and is based on several individual databases. The tool gives a gene-based, region-based and filter-based variant annotation for the mutations of interest. The gene-based annotation identifies whether a SNPs or CNVs cause protein coding changes and the amino acids that are affected. The region-based annotations allows us to identify variants in specific genomic regions, to predicted transcription factor binding sites and GWAS hits; while the filter-based annotation reports the occurrence of the variants in dbSNP, 1000 Genome Project.

The last step of a NGS data analysis consists of the validation and visualization of results. This step is fundamental for the interpretation of the data. The Integrative Genomic Viewer (IGV) (Figure 1.22) is a tool which allows the user to load several tracks as well as the reference genome. The IGV package provides different functions as counting, sorting and indexing[195]. Another example of a visualization tool is the genome browser of the University of California Santa Cruz (UCSC) which allows users to map the experimental data together with other types of annotations such as transcriptome and the information found in other public sources[172].


**Figure 1.22 Snap-shot of genomic variants with the IGV software.** The display changes with the zoom. When zoomed in to the alignment read visibility threshold, IGV shows the reads. The colour bars in the read coverage track identify reads that differ from the reference genome. When zoomed out, IGV displays only coverage data, as shown in the figure.

### *1.2.8.7 Common variant annotation software*

Over the past decades, numerous *in silico* tools have been developed for the functional annotation of variation. Within this paragraph, I describe some of the better known annotation software that I have applied in subsequent analyses.

SORTING INTOLERANT FROM TOLERANT (SIFT)

SIFT (Sorting Intolerant from Tolerant) is a sequence homology-based tool that distinguish intolerant from tolerant amino acid substitutions and predicts whether an amino acid substitution in a protein will have a phenotypic effect. SIFT is based on the assumption that protein evolution is correlated with protein function; positions important for the protein function are conserved in an alignment of the protein family, whereas unimportant positions should appear diverse in an alignment[188]. SIFT accepts as input a protein sequence and applies multiple alignment information to predict tolerated and deleterious substitutions for every position of the submitted sequence. SIFT searches for similar sequences with a BLAST (Basic Local Alignment Search Too; Altschul e al., 1990) search. SIFT then chooses the sequences that are similar in

function and structure to the query sequence by selecting only a subset of sequences from the BLAST results. Conserved regions are then extracted from the aligned sequences and normalized probabilities for all possible substitutions from the alignment are calculated. Positions with normalized probabilities less than 0.05 are predicted to be deleterious; those greater than or equal to 0.05 are predicted to be tolerated. The SIFT score ranges from 0 to 1. Often the metric of 1-SIFT is applied so that deleterious variants achieve a score close to 1.

GRANTHAM

Grantham Scores categorize codon replacements into classes of chemical dissimilarity based on the characteristics of the amino acids[191]. The score described the difference in side chain atomic composition, polarity, and volume between any two amino acids. The score ranges from 5 to 215 and substitutions with Grantham score of 5–60 are generally considered “conservative”, 60–100 “non-conservative”, and >100 “radical” (Figure 1.23).

| Arg | Leu | Pro | Thr | Ala | Val | Gly | Ile | Phe | Tyr | Cys | His | Gln | Asn | Lys | Asp | Glu | Met | Trp | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 110 | 145 | 74 | 58 | 99 | 124 | 56 | 142 | 155 | 144 | 112 | 89 | 68 | 46 | 121 | 65 | 80 | 135 | 177 | Ser |
| | 102 | 103 | 71 | 112 | 96 | 125 | 97 | 97 | 77 | 180 | 29 | 43 | 86 | 26 | 96 | 54 | 91 | 101 | Arg |
| | | 98 | 92 | 96 | 32 | 138 | 5 | 22 | 36 | 198 | 99 | 113 | 153 | 107 | 172 | 138 | 15 | 61 | Leu |
| | | | 38 | 27 | 68 | 42 | 95 | 114 | 110 | 169 | 77 | 76 | 91 | 103 | 108 | 93 | 87 | 147 | Pro |
| | | | | 58 | 69 | 59 | 89 | 103 | 92 | 149 | 47 | 42 | 65 | 78 | 85 | 65 | 81 | 128 | Thr |
| | | | | | 64 | 60 | 94 | 113 | 112 | 195 | 86 | 91 | 111 | 106 | 126 | 107 | 84 | 148 | Ala |
| | | | | | | 109 | 29 | 50 | 55 | 192 | 84 | 96 | 133 | 97 | 152 | 121 | 21 | 88 | Val |
| | | | | | | | 135 | 153 | 147 | 159 | 98 | 87 | 80 | 127 | 94 | 98 | 127 | 184 | Gly |
| | | | | | | | | 21 | 33 | 198 | 94 | 109 | 149 | 102 | 168 | 134 | 10 | 61 | Ile |
| | | | | | | | | | 22 | 205 | 100 | 116 | 158 | 102 | 177 | 140 | 28 | 40 | Phe |
| | | | | | | | | | | 194 | 83 | 99 | 143 | 85 | 160 | 122 | 36 | 37 | Tyr |
| | | | | | | | | | | | 174 | 154 | 139 | 202 | 154 | 170 | 196 | 215 | Cys |
| | | | | | | | | | | | | 24 | 68 | 32 | 81 | 40 | 87 | 115 | His |
| | | | | | | | | | | | | | 46 | 53 | 61 | 29 | 101 | 130 | Gln |
| | | | | | | | | | | | | | | 94 | 23 | 42 | 142 | 174 | Asn |
| | | | | | | | | | | | | | | | 101 | 56 | 95 | 110 | Lys |
| | | | | | | | | | | | | | | | | 45 | 160 | 181 | Asp |
| | | | | | | | | | | | | | | | | | 126 | 152 | Glu |
| | | | | | | | | | | | | | | | | | | 67 | Met |

**Figure 1.23 Grantham table.** Numbers represent the pairwise score for substituting one amino acid with another. The greater is the score, the larger the change in the chemical nature of the amino acid[191].

PHYLOGENETIC P-VALUES (PHYLOP)

PhyloP (phylogenetic P-values) is freely available as part of the PHAST package (http://compgen.bscb.cornell.edu/phast)[192]. The software, developed by Pollard and collaborators in 2010, allows the detection of sites under negative or positive

selection[192]. The PhyloP score can measure acceleration (faster evolution than expected under neutral drift) as well as conservation (slower than expected evolution). The score goes from -11.958 to 2.941; the larger the score the more conserved is the site.

POLYMORPHISM PHENOTYPING (POLYPHEN2 HDIV AND POLYPHEN2 HVAR)

PolyPhen2[189] (Polymorphism Phenotyping v2) is a software which predicts possible impact of an amino acid substitution on the structure and function of a human protein using physical and comparative considerations. PolyPhen-2 is able to predict the functional significance of an allele replacement by using a trained supervised machine-learning model. Two pairs of datasets were used to train and test PolyPhen-2 prediction models. The first pair, PolyPhen-2 HumDiv, was compiled from all damaging alleles with known effects on the molecular function causing human Mendelian diseases, present in the UniProtKB database, together with differences between human proteins and their closely related mammalian homologs, assumed to be non-damaging. The second pair, PolyPhen-2 HumVar, consisted of all human disease-causing mutations from UniProtKB, together with common human nsSNPs (MAF>1%) without annotated involvement in disease, which were treated as non-damaging. The user can choose between HumDiv- and HumVar-trained PolyPhen-2 models. The diagnostics of Mendelian diseases requires distinguishing mutations with drastic effects from all the remaining human variation, including abundant mildly deleterious alleles, therefore HumVar-trained is more adequate to use for this task. In contrast, HumDiv-trained is better for evaluating rare alleles at loci potentially involved in complex phenotypes, dense mapping of regions identified by genome-wide association studies, and analysis of natural selection from sequence data, where even mildly deleterious alleles must be treated as damaging. Mutations are classified as damaging (posterior probability >=0.957), probably damaging (0.453<= posterior probability <=0.956) and benign (posterior probability <=0.452).

GENOMIC EVOLUTIONARY RATE PROFILING (GERP NR AND GERP RS)

Genomic Evolutionary Rate Profiling (GERP)[192], implemented in the Gerp++ software (http://mendel.stanford.edu/SidowLab/downloads/gerp/), identifies evolutionary constrained elements in multiple alignments by quantifying substitution deficits. GERP

uses mammalian genome alignments to determine and asses a rejected substitution score for each variant which indicates the difference between observed and expected evolution rate. Positive scores represent that the location might be under selection constrains whereas negative scores indicates that the location is evolving under neutral selection[192].

COMBINED ANNOTATION DEPENDENT DEPLETION (CADD)

CADD[185] is a tool for scoring the deleteriousness of single nucleotide variants as well as insertion/deletions variants in the human genome. Differently from many variant annotation and scoring tools which tend to exploit a single information type (e.g. conservation) and/or are restricted in scope (e.g. to missense changes), combined Annotation Dependent Depletion (CADD) integrates multiple annotations into one metric by contrasting variants that survived natural selection with simulated mutations. CADD can quantitatively prioritize functional, deleterious, and disease causal variants across a wide range of functional categories, effect sizes and genetic architectures and can be used prioritize causal variation in both research and clinical settings.

KGGSEQ: LOGIT MODEL

KGGSeq[193] is a free available command line platform for the analysis of monogenic and complex diseases using sequencing data. To assess the functional impact of a list of mutation of interest, the platform uses a logit model which combines multiple prediction methods and compute an unbiased probability of a rare variant being pathogenic. Specifically, the model uses 13 available functional impact scores (CADD score, FATHMM score, GERP++ NR/RS, LRT, MutationAssessor, MutationTaster, PhyloP, Polyphen2 HDIV/HVAR, SIFT, SiPhy and SLR). A subset of original prediction tools (out of the 13 tools) are used for the combined prediction by the logistic regression model which have the largest posterior probability among all possible combinatorial subsets is reported in the output. By default, the model tries all subsets of the scores (at least two) for a combinatorial prediction and finally presents a prediction that can pass the threshold to classify the SNV as a potentially disease causal (KGGSeq output “Y”) or neutral (KGGSeq output “N”). The cutoff used to determine the variants leads to the maximal Matthews correlation coefficient (MCC):

the corresponding true positive and true negative value at the maximal MCC. The MCC is used in machine learning as a measure of the quality of binary (two-class) classifications. It takes into account true and false positives and negatives and is generally regarded as a balanced measure that can be used even if the classes are of very different sizes.

ExoVar is the training set used in the machine learning method by the logistic regression. The set is composed of 5,340 alleles with known effects on the molecular function causing human Mendelian diseases from the UniProt database, which are treated as positive control variants, and 4,752 rare (alternative allele frequency <1%) nonsynonymous variants with at least one homozygous genotype for the alternative allele in the 1000 Genomes Project, which are treated as negative control variants.

#### *1.2.8.8 Filtering and prioritisation strategy in exome sequencing*

To identify the genes underlying Mendelian and complex phenotypes, it is critical to apply an objective filtering criteria (Figure 1.24). The American College of Medical Genetics and Genomics (ACMG)[196] has developed guidance for the interpretation of sequence variants to separate disease causing or disease-associated genetic variants from neutral variants present in all human genomes that are rare and potentially functional, but not involved in the pathogenic of disease under study. According to the ACMG guidelines variants should be classified as *pathogenic* if they contribute mechanistically to disease, but are not necessarily fully penetrant (i.e., may not be sufficient in isolation to cause disease); *implicated* if there is evidence supporting the pathogenic role, with a defined level of confidence; *associated* if found significantly enriched in disease cases compared to matched controls; *damaging* if they alter the normal levels or biochemical function of a gene or gene product; *deleterious* if they reduce the reproductive fitness of carriers, and would be removed by purifying natural selection. This terminology is mainly recommended to describe variants identified in genes causing Mendelian disorders. The classification of variants in these five categories is based on criteria using different types of evidence (e.g., population data, computational data, functional data, and segregation data). The Working Group also specified a set of disorders, the relevant associated genes and certain categories of variants that should be reported regardless of the reason why clinical sequencing is performed.

Filtering strategies are required in exome sequencing studies in order to refine the large number of variants called to focus on disease causing mutations. Exomes contain not only disease variations but also neutral variants with no effect on the phenotype. In order to reduce the number of calls that are likely to be disease associated, variants are first filtered based on quality criteria, such as the total number of independent reads showing the variant and the percentage of reads showing the variant. Moreover, mutations outside the coding regions can be filtered out, as well as synonymous coding variants, assuming that that these mutations will have minimal effect on the protein function. Further filtering consists of excluding common variants as identified by publically available databases or in-house databases.

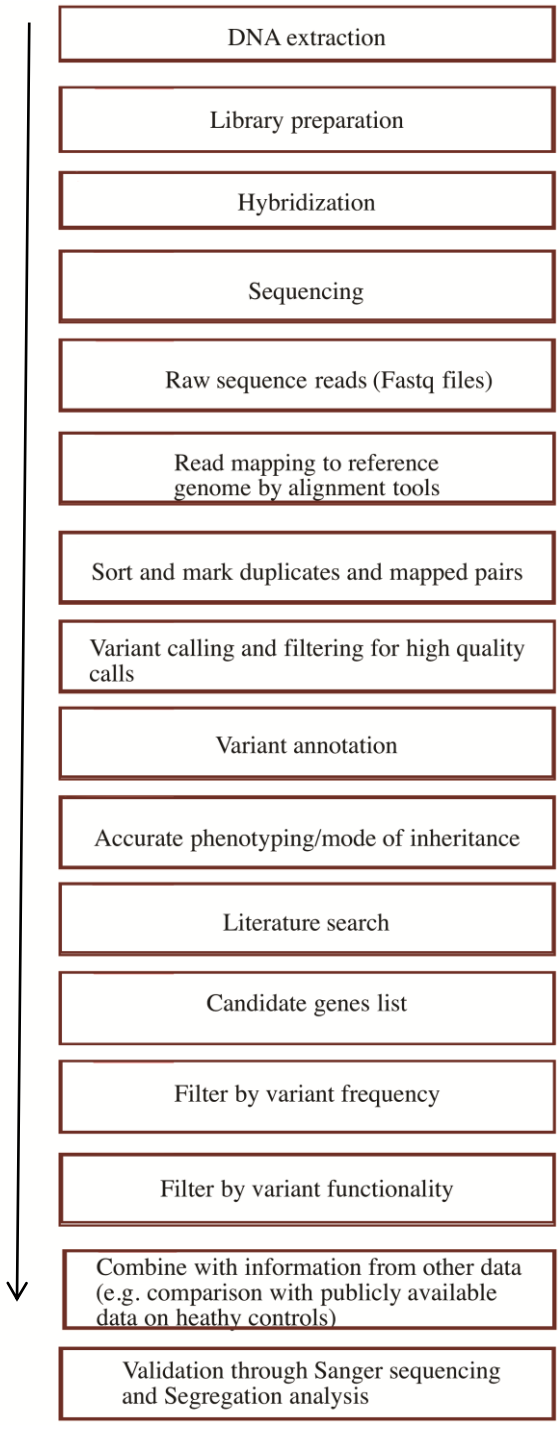



**Figure 1.24 Proposed pipeline for data processing and data filtering for whole exome sequencing data.** After the sequenced data have been generated and good quality called genotypes have been selected, variant filtering and prioritisation is required to determine a list of high priority variants from the vast number obtained (circa 25,000 per individual). After compiling a list of candidate genes, where applicable, allele frequency filter is used as a first step to prioritise variants. Variants are further filtered based on their functionality. Indels and stop loss/stop gain mutations are widely accepted as evidence for being disease-causal. The addition of other data can be helpful in prioritising candidate variants. Segregation analysis is conducted to validate variant causality.

### Candidate gene filtering

A candidate gene filtering strategy limits the interrogation of exome data to a set of genes already known or suspected to be implicated in the disease of interest. Candidate genes can be selected because they resemble genes associated with similar diseases or because the predicted protein function seems relevant to the physiology of the disease.

### Pathway analysis

Pathway analysis is an alternative way to identify deleterious genes within pathways known to be involved in the disease pathogenesis. By focusing on key pathways previously described by genome wide association studies, for example, the application of whole exome and further whole genome sequence will help to identify associations with novel common and rare variants within known genes and genes not previously detected. The combination of technologies may help to further interpret and understand the overall variant contribution to disease.

### Familial sequencing

For families with several family members affected by the same disorder, multiple individuals can be sequenced to identify shared variation. Unaffected relatives can be sequenced to exclude benign variation. As affected siblings are expected to share 50% of their genetic material, by selecting the most distantly related affected individulas the amount of shared mutation can be minimised. Ng and collaborators were the first to apply this filtering strategy to reduce genetic variations and identify the disease causing mutation of two brothers affected by Miller syndrome[156]. Exome data of three siblings affected by hyperphosphatasia mental retardation syndrome were used in the study conducted by Krawitz to identify causal mutations by selecting those variants identical by descent[197].

### Overlap strategy

Another filtering strategy is to pull out shared mutations of unrelated patients with similar phenotype. The number of genes with mutations in multiple affected patients will decrease by combining data from increasing numbers of patients, resulting in less

candidate genes for follow-up. This filtering strategy can be efficient in analysing dominant and monogenic disorders[198].

### Novel genes and mutations

Spontaneous mutations are frequent causes of single cases of sporadic diseases[199]. The *de novo* mutation rate in humans is high, corresponding to 1 in 100 million positions in the haploid genome[200,201], therefore novel mutations are generated at a high frequency in a population. Novel mutations in genes leading to severe diseases, such as lethality in embryonic stages or reduced reproductive fitness, are unlikely to be transmitted to multiple family members, and therefore will not be detected by linkage gene mapping or association studies. The identification of *de novo* mutation using WES is advantageous when both healthy parents and the affected proband are sequenced to identify and therefore exclude the inherited variants. The Deciphering Developmental Disorders (DDD) study was the first study within the UK to perform genome-wide microarray and whole exome sequencing on children with undiagnosed developmental disorders and their parents to better characterise the role of de novo mutations within known developmental disorder genes[202]. According to the ACMG, *de novo* variants are considered probably deleterious or pathogenic if the variant is confirmed to be *de novo* through Sanger sequencing, if the patient has a family history of disease that is consistent with *de novo* inheritance; if the phenotype in the patient matches the gene's disease association with reasonable specificity and if the variant is validated through functional studies.

### Variants of unknown significance

Variants of unknown significance (VUS) are defined as variations whose association with disease risk is unknown. The ACMG recommendation for variants with unknown clinical significance is to conduct investigation to demonstrate evidence for a role of a candidate gene and one or more variants disrupting it. To address this important clinical problem, various types of evidence may help to classify such variants as deleterious or neutral, with respect to the disease of interest (Table 1.7). Experimental approaches to investigating the impact of a sequence variant on gene function, or cell or organism phenotype, can also have a role in demonstrating that a variant is damaging to gene function and in identifying the molecular mechanisms underlying a

variant's effect on disease risk. When a gene has already been confidently implicated in disease, and it is known what class of variant is causal (for instance, loss or gain of function as represented by a specific assay), then an experiment that places a variant of unknown significance into such a functional class can be particularly informative[186].

**Table 1.7 Types of evidences for classifying variants of unknown significance[196].**

| Line of Evidence | Advantage(s) | Disadvantage(s) |
|---|---|---|
| Frequency in cases and controls | Provides a direct estimate of disease risk | Variants are rare, so such studies would need to be prohibitively large (10,000+) |
| Co-occurrence (in *trans*) with deleterious mutations | If homozygotes and compound heterozygotes are assumed to be embryonically lethal (or vanishingly rare), it is possible to classify a variant as neutral on the basis of a single observation | Much less power to show causality; quantification is dependent on the assumed fitness of the homozygous genotype, which is not known with precision |
| Cosegregation with disease in pedigrees | Easily quantifiable and directly related to disease risk; not susceptible to uncertainties in mutation frequencies or population stratification | Requires sampling of additional individuals in the pedigrees (particularly additional cases), which may be difficult to achieve |
| Family history | Usually available for most variants without additional data or sample collection; potentially very powerful | Dependent on family ascertainment scheme; could be biased in stratified populations with heterogeneous ascertainment, so not as robust as cosegregation; power may be low for infrequent variants |
| Species conservation and amino acid–change severity | Can be applied to every possible missense change in genes; does not require extensive family history; complete conservation is predictive if enough evolutionary time sequence is available | Only indirectly related to disease risk; the magnitude of odds ratios is not sufficient to classify variants without additional information |
| Functional studies | Can evaluate biologically the variant's effect on the protein's ability to perform some key cellular functions | May only be relevant for variants in certain domains of the protein; the function tested may not be related to cancer causation |
| Loss of heterozygosity | Straightforward to quantify as an adjunct to cosegregation data; robust | Requires tumor material |
| Pathological classification | Potentially powerful for traits in which the pathological characteristics are quite distinct and quantifiable | Prediction is weak when routine pathology data are used; systematic evaluation requires biological material and it could be weakly predictive |

### INCIDENTAL FINDINGS

Incidental (or secondary) findings are results that are not related to the indication for ordering the sequencing but that may nonetheless be of medical value or utility to the ordering clinicians and the patient.

ACMG guidelines regarding incidental findings recommends to return the incidental findings to the doctor and those doctors should manage the information with the patients for the specified Mendelian disease specified in Table 1.8. In most cases, the ACMG Working Group recommends only pathogenic variants with a higher likelihood of causing disease should be reported as incidental findings.

**Table 1.8 Disorders and genes recommended to be checked in sequencing studies according to the ACMG guidelines[196].**

| Phenotype | Age of Onset | Gene | Inheritance | Variants to Report |
|---|---|---|---|---|
| Hereditary Breast and Ovarian Cancer | Adult | *BRCA1*<br>*BRCA2* | AD | KP & EP |
| Li-Fraumeni Syndrome | Child/adult | *TP53* | AD | KP & EP |
| Peutz-Jeghers Syndrome | Child/adult | *STK11* | AD | KP & EP |
| Lynch Syndrome | Adult | *MLH1 MSH2*<br>*MSH6 PMS2* | AD | KP & EP |
| Familial adenomatous polyposis | Child | *APC* | AD | KP & EP |
| MYH-Associated Polyposis; Adenomas, multiple colorectal, FAP type 2; Colorectal adenomatous polyposis, autosomal recessive, with pilomatricomas | Adult | *MUTYH* | AR | KP & EP |
| Von Hippel Lindau syndrome | Child/adult | *VHL* | AD | KP & EP |
| Multiple Endocrine Neoplasia Type 1 | Child/adult | *MEN1* | AD | KP & EP |
| Multiple Endocrine Neoplasia Type 2 | Child/adult | *RET* | AD | KP |
| Familial Medullary Thyroid Cancer (FMTC) | Child/adult | *RET*<br>*NTRK1* | AD | KP |
| PTEN Hamartoma Tumor Syndrome | Child | *PTEN* | AD | KP & EP |
| Retinoblastoma | Child | *RB1* | AD | KP & EP |
| Hereditary Paraganglioma-Pheochromocytoma Syndrome | Child/adult | *SDHD SDHAF2*<br>*SDHC SDHB* | AD | KP & EP |
| Tuberous Sclerosis Complex | Child | *TSC1*<br>*TSC2* | AD | KP & EP |
| WT1-related Wilms tumor | Child | *WT1* | AD | KP & EP |
| Neurofibromatosis type 2 | Child/adult | *NF2* | AD | KP & EP |
| EDS - vascular type | Child/adult | *COL3A1* | AD | KP & EP |
| Marfan Syndrome, Loeys-Dietz Syndromes, and Familial Thoracic Aortic Aneurysms and Dissections | Child/adult | *FBN1 TGFBR1*<br>*TGFBR2 SMAD3*<br>*ACTA2*<br>*MYLK MYH11* | AD | KP & EP |
| Hypertrophic cardiomyopathy, Dilated cardiomyopathy | Child/adult | *MYBPC3 MYH7*<br>*TNNT2 TNNI3*<br>*TPM1 MYL3*<br>*ACTC1 PRKAG2*<br>*MYL2 LMNA GLA* | AD<br>XL | KP & EP |
| Catecholaminergic polymorphic ventricular tachycardia | | *RYR2* | AD | KP |
| Arrhythmogenic right ventricular cardiomyopathy | Child/adult | *PKP2 DSP*<br>*DSC2 TMEM43*<br>*DSG2* | AD | KP & EP |
| Romano-Ward Long QT Syndromes Types 1, 2, and 3, Brugada Syndrome | Child/adult | *KCNQ1 KCNH2*<br>*SCN5A PCSK9* | AD | KP & EP |
| Familial hypercholesterolemia | Child | *LDLR APOB* | SD<br>SD | KP & EP<br>KP |
| Malignant hyperthermia susceptibility | Child/adult | *RYR1 CACNA1S* | AD | KP |

AD: autosomal dominant, AR: autosomal recessive; XL: X linked

KP = known pathogenic, sequence variation is previously reported and is a recognized cause of the disorder; EP = expected pathogenic, sequence variation is previously unreported and is of the type which is expected to cause the disorder. The recommendation to not report expected pathogenic variants for some genes is due to the recognition that truncating variants, the primary type of expected pathogenic variants, are not an established cause of some diseases on the list.

### *1.2.8.9 Challenges and success of whole exome sequencing*

Although exome sequencing is an extremely efficient and robust technology for the analysis of part of the functional region of the human genome, it has limitations. Drawbacks of WES are the assumption that disorders are probably caused by coding variants; there is incomplete coverage of all exons across the genome and lack of complete targeting of 5' and 3' untranslated regions by the capture kits due to the intricate nature of the genome (e.g. exon size and guanine-cytosine (GC) context)[203]. Another limitation of WES is the inability to detect non-exonic causal variants in the 98% of the genome not sequenced. The evaluation and detection of rare heterozygous variants is a complex process as they are more affected by technical errors and require read depths of at least 10.[204] As previously discussed, the identification of the disease causal mutation requires efficient filtering techniques. Patients with rare recessive diseases[205] might harbour disease variants in compound heterozygosity. Damaging variants in compound heterozygosity could be often overlooked if the second risk variant is common (MAF > 5%) and therefore it will be discarded by standard filtering approaches which focus on rare (MAF < 5%) mutations. Moreover, another challenge in filtering sequence variants for compound heterozygotes is to determine whether the two heterozygous variants affect different copies of the chromosome or the same copy (phase). The final step of genomic data processing involves literature review for each of the prioritised genes making it difficult to interpret potential causative variants.

## 1.3 Summary

IBD is an idiopathic, chronic condition with multifactorial determinants: the intestinal inflammation arises from abnormal host-microbe interactions in genetically predisposed individuals which can also result in the development of comorbid autoimmune diseases in the same patients. The precise aetiology of this complex disease remains unclear. IBD location, progression, and response to therapy have age-dependent characteristics[28,206,207]. Early onset patients present with more severe manifestations and clinical studies have shown that paediatric IBD patients are more often treated with steroids and immunosuppressive drugs and have a more severe course compared to adults[29].

Familial linkage studies, genome-wide association and more recently next generation sequencing studies have advanced the understanding of its aetiology. Whole exome sequencing is an affordable and efficient method to target the exome which contain 85% of monogenic disease causal variations[203]. However, this technology is not without limitations.

Next generation sequencing technologies are increasingly used in the study of inherited disease with major successes in the identification of genes involved in Mendelian traits. However, application of these cutting edge technologies in the diagnosis of complex diseases is not yet routinely applied.

In this thesis I apply whole exome sequencing to a cohort of paediatric IBD patients (pIBD) to study the effect of rare variants underlying the condition. It is now widely accepted that genetics may play a particularly important role in paediatric cases, as children typically lack the environmental exposures associated with IBD in adults and have had less time to be influenced by such immune triggers[160].

The systematic evaluation of all type of variations in these patients might help us to understand the genetics underlying this complex pathology.

## Thesis aims and specific contribution

In the introduction chapter I have outlined a summary of the characteristics of inflammatory bowel disease and provided a brief background to the history of gene and locus discovery experiments with examples in Mendelian and complex traits. This thesis describes analyses to better understand the genetic background of IBD using contemporary sequencing technology.

**Aim 1 – determine the incidence of comorbidities in paediatric IBD**

Chapter 2 fully investigate the incidence of comorbidities within the Southampton pIBD cohort and the application of whole exome sequencing for the identification of disease risk variants. This analysis aims to provide insight into the relationship between paediatric IBD and concurrent autoimmune diseases in children.

The work conducted in this chapter was predominately conducted by myself in collaboration with Dr James Ashton and under the supervision of Prof Sarah Ennis. As first author of the paper I was responsible in facilitating recruitment process, in the curation of the research database, data processing, data quality control, development of a pipeline for the application rare variants statistical analysis, data interpretation and manuscript preparation.

**Aim 2 – application of burden of mutation test within genes of the NOD receptor pathway**

In chapter 3 I apply the SKAT-O test for detecting association between genes involved in the NOD signalling pathway and disease. *NOD2* was the first gene found to be associated with CD and 11 genes in the same pathway (out of 41) have been already implicated in IBD by GWAS. For this reason, we decided to test the application of this software on this established pathway and determine evidence for previously unknown genes in IBD susceptibility. By limiting our analysis of sequencing data to a key pathway implicated by GWAS, we massively reduce the genomic search space for causal variation to concentrate on regions with high prior probability of containing pathogenic mutation while maximising sensitivity to rare and private mutations.

As first author of this work my contribution was the design of the pipeline for the execution of the gene based statistical analysis within the validation and replication cohort, data quality control, data interpretation, manuscript write-up and submission. Prof Sarah Ennis provided significant input.

**Aim 3 – application of burden of mutation test to investigate the role of heat shock proteins in IBD**

Chapter 4 focuses on the analysis conducted in collaboration with the University of Stanford in California, USA. The data presented is based on the genomic and functional studies conducted on the *HSPA1L* gene from a core family (Stanford data) and our Caucasian cohort of 136 paediatric IBD patients and 106 controls. Our findings provide insights into the pathogenesis of IBD, as well as expanding our understanding of the roles of heat shock proteins in human disease.

As a joint first author, together with Dr Takahashi (University of Stanford), of this work my contribution was to comprehensively investigate the role of *HSPA1L* within our pIBD cohort, design of the pipeline and application of appropriate gene based statistical analysis, data quality control, primer design for segregation analysis, extraction of variation within known IBD genes from the Stanford and Southampton exome data, data interpretation, manuscript write-up. Prof Sarah Ennis and Prof Mike Snyder (Stanford University) provided significant input.

**Aim 4 – determine the incidence of disease causal variations within genes causing monogenic forms of IBD**

Chapter 5 focuses on the analysis of variants within the 51 genes known to cause monogenic forms of IBD. Observed mutations were filtered according to the ACMG guidelines. The aim of this analysis was to apply whole exome sequencing to identify known and *de novo* potentially causative mutations within genes associated with monogenic IBD. Whilst monogenic IBD is a rare condition, it is vital to have early identification of causative mutations in order to provide a prognosis and improve treatment.

The work presented in this chapter was predominately conducted by myself in collaboration with Dr James Ashton. As joint first author of the paper I was responsible in facilitating recruitment process, research database curation, data processing, data quality control, data interpretation, design of primers for segregation analysis and manuscript preparation. The work was conducted under the supervision of Prof Sarah Ennis.

The focus of chapter 6 is a discussion of the major points drawn from the previous chapters and conclusions of the thesis.

# Chapter 2 **Common genes within complex autoimmune diseases**

## 2.0 Summary

The work presented in this chapter investigates the incidence of comorbidities within the Southampton pIBD cohort and the application of whole exome sequencing as a first step into a personalised approach to common diseases. This analysis aims to provide insight into the relationship between paediatric IBD and other autoimmune diseases (clinician diagnosed). For a subset of patients with pIBD and concurrent asthma, exome data was interrogated to ascertain the burden of pathogenic variants within genes implicated in asthma. Association testing was conducted between cases and population controls of non-IBD individuals wiusing the SKAT-O test. By interrogating the exome data for a subset of children we were able to show that for a group of patients the relationship between concurrent pIBD and the second autoimmune condition could be caused by a systemic immune dysregulation rather than organ specific immune dysfunctions.

As a joint first author of this analysis my contribution was to facilitate recruitment process, curate and update the research database, data processing, quality check, development of a pipeline for the application rare variants statistical analysis, data interpretation and manuscript preparation.

## 2.1 Background

Patients with inflammatory bowel disease can be affected by comorbid autoimmune diseases. The term comorbidity refers to either: any secondary health problem that affects a person arising as a result of a primary condition or a clinical manifestation that is not pathologically linked to the primary condition[208]. In IBD the second condition can occur prior to the gastro-intestinal disease and persist after remission of gut symptoms. Comorbidity conditions can significantly change the clinical manifestations of IBD, its activity and its prognosis[208]. Although the second disease presents differences in clinical presentation, genomic analysis of autoimmune and inflammatory disorders suggests shared genetic components and common cellular and molecular immune pathways for these conditions[209]. This leads to the hypothesis that, unlike Mendelian disorders, common autoimmune and inflammatory conditions may

arise from the combined effects of common non-disease specific and disease specific loci, and environmental triggers[120]. In some cases the shared genes have opposite function in the two pathologies; for example the mutation R620W in gene *PTPN22* is protective in IBD but a risk factor in systemic lupus erythematosus (SLE) and rheumatoid arthritis (RA) [209]. With regards to IBD, various studies have looked at the incidence of other autoimmune-mediated disorders in IBD patients and at genetic risk loci shared across autoimmune disorders. [210,211]. A study conducted by Lees *et al* shows 51 IBD genes overlapping with coeliac disease, psoriasis, SLE, RA, mycobacterial infection such as leprosy, and non-immune conditions such as obesity[209]. The number of loci shared by IBD and other autoimmune disease involved innate immunity (i.e. *IRF5*), T-cell activation (i.e. *PTPN22*, *IL2*, *IL2RA* and *IL21*) or the activation of the unfolded protein response (i.e. *ORMDL3*)[212]. These shared loci offer an insight into the biology of the conditions but does not resolve the identification of the true causal variants.

In paediatric patients studies into the co-existence of pIBD and other autoimmune diseases have reported strong association between pIBD and RA, SLE and hypothyroidism with a trend towards increased prevalence of other autoimmune conditions including asthma and eczema in both CD and UC[213]. Primary sclerosing cholangitis (PSC) is a chronic progressive disorder of unknown aetiology characterised by chronic inflammation and stricture formation of the biliary tree. The disease is rare in the general population but is strongly associated with IBD affecting up to 5% of patients with UC, with a slightly lower prevalence (up to 3.6%) in CD[214]. Immunogenetic studies have identified a number of HLA haplotypes associated with PSC, HLA-B8/DR3 haplotype being particularly common in patients with PSC and UC and infrequent in patients with PSC alone[214]. This haplotype is also associated with other autoimmune disease including coeliac disease, thyrotoxicosis, lupoid autoimmune hepatitis, myasthenia gravis, and type 1 diabetes mellitus. Possible explanations for the PSC association with IBD include the development of autoantibodies to an unknown antigen in an immunogenetically susceptible host which cross reacts with biliary and colonic epithelium and is capable of inducing complement activation[215]. Alternatively, the initiation of the immune response may be the ingress of bacteria or other toxic metabolites through the diseased bowel wall.

The delineation of genes and pathways that relate more specifically to certain autoimmune conditions than to others provides valuable information that can be used to target autoimmune phenotypes with interventions that are relevant to those pathways. The highlighted biologic pathways then provide a focus for more fundamental research, aimed to identify the underlying disease mechanisms in autoimmunity, and they could inform the development of novel therapies. An example is anti-TNF targeted therapies which are successfully applied to a diverse group of autoimmune disorders, including RA, IBD, psoriasis and others[216]. The identification of autoimmune associated genes have enabled researchers to elucidate the extent of genetic overlap across this broad group of disorders. Loci that are shared between various autoimmune disorders and involved in a wide range of immune pathways might help explain common pathogenic features and inform the development of novel therapies. Further, the lack of overlap for other loci and pathways also suggests distinct pathogenic mechanisms that could explain, at least in part, the phenotypic diversity across the spectrum of autoimmune disease.

Predisposition to IBD and other autoimmune disease has a strong genetic component, and analyses of whole exome data of these patients may yield variations associated with both groups of disease. The work presented in this chapter examines the autoimmune disease burden in patients diagnosed with pIBD and interrogates exome data in a subset of patients.

## 2.2 Methods

### 2.2.1 The Southampton paediatric cohort

Since 2010 clinicians and research nurses in the Southampton IBD paediatric research study have been recruiting children diagnosed with IBD through tertiary referral clinics at University Hospital Southampton. This hospital is the regional centre for paediatric gastroenterology, providing a tertiary paediatric gastroenterology and endoscopy service for the Wessex region, and draws upon a patient population of 3.5 million. The service has a live rolling database of over 200 paediatric IBD cases and approximately 50–70 patients are diagnosed each year. All children aged less than 18 at the point of diagnosis are eligible for recruitment to the Southampton Genetics of pIBD study. Up to December 2015 the cohort stands at 814 participants, 290 of which are children (CD n = 176; UC n = 86; IBDU n = 31). All parents and any first, second or third degree relatives diagnosed with any form of IBD are routinely recruited (n = 344). Of the 290 pedigrees recruited to date, 100 probands have at least one additional family member with a clinical diagnosis of IBD and half of these families (50 probands) present a mixed diagnosis of both CD and UC reflecting the genetic overlap in subtype aetiology. Breakdown of the cohort is shown in Figure 2.1.

Diagnosis is established using the Porto criteria[185]. The Porto criteria are a collection of criteria for the diagnosis of CD, UC and IC based on clinical signs and symptoms, endoscopy and histology and radiology. According to the criteria, every IBD suspected child should undergo a complete diagnostic program consisting of colonoscopy with ileal intubation, upper gastrointestinal endoscopy and radiologic contrast imaging of the small bowel[217]. Multiple biopsies from all segments of the gastrointestinal tract are needed for a complete histologic evaluation; the diagnosis of indeterminate colitis cannot be made unless a full diagnostic program has been performed[217]. Informed written consent is provided by an attending parent or legal guardian for all paediatric recruits as part of normal outpatient clinics and followed through their treatment, at which point clinical data and biological samples are collected. Collection of biological samples entails in blood or saliva, biopsy specimen and faeces. Biopsy specimen are taken during the routine clinical endoscopy at various levels such as the oesophagus, stomach, duodenum, ileum, colon and the rectum. At each level, the endoscopist may take anywhere between 0-10 biopsies depending on the disease expression. The

collection of biological samples in treatment naïve patients will enable a truer/more powerful assessment of disease aetiology. Clinical data recorded for each patient include date of birth, sex, age at diagnosis, symptoms, disease extent and severity using the Paris classification of paediatric IBD[13]. Details of clinical investigations, treatment, previous illnesses, familial history of IBD or any other autoimmune disease and life style information (i.e. smoking, diet) are also collected. Parents and relatives diagnosed with IBD were routinely recruited as well. During the initial recruitment clinical information and blood samples or saliva where appropriate, are taken for DNA extraction in preparation for genetic analysis. Patients have been prioritised for exome analysis on the basis of presenting with a severe phenotype is chosen to undergo exome analysis. To date, 147 paediatric IBD patients underwent whole exome sequencing. By conducting our analysis in a local cohort with ethical approval for additional biological sample collection, we have biological samples, facilities and staff to confirm immunological functional relevance of precise gene variants by selecting the exact patients in whom these variants occur. This concomitantly stratifies the patient subset most likely to benefit from treatment with new therapeutics.

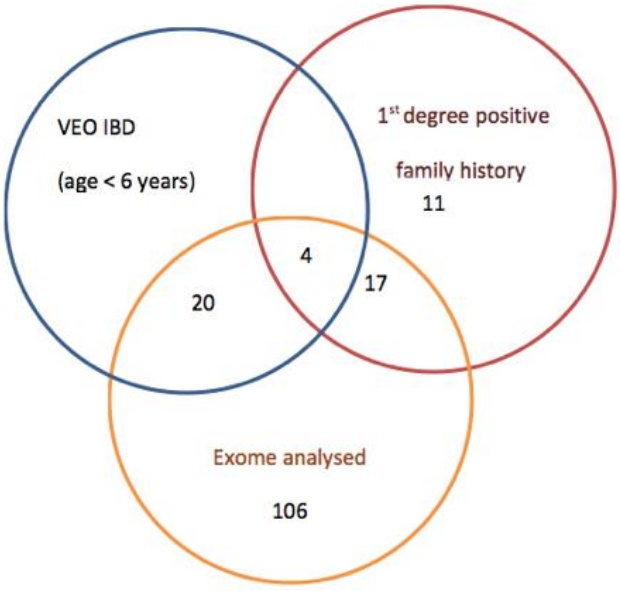



**Figure 2.1 Breakdown of the Southampton paediatric IBD cohort**. The cohort consists in 814 recruits of which 290 are children (119 females and 171 males). Up do December 2015 we have comprehensive genetic data for 147 children.

#### *2.2.1.1 Study population*

Anonymised patient data was interrogated to identify paediatric IBD patients with comorbidity of other medically diagnosed autoimmune diseases and positive family history of autoimmune diseases other than IBD. Medical notes were consulted for ambiguous diagnoses and review allowed for exclusion of any unconfirmed diagnoses.

At the time of the analysis, data from 173 patients prospectively collected at recruitment was interrogated to identify pIBD patients with: 1) Comorbidity of other autoimmune diseases (clinician diagnosed) and; 2) positive family history of autoimmune diseases other than IBD.

### 2.2.2 DNA extraction for whole exome sequencing

Genomic DNA is routinely extracted from patient bloods by the laboratory technician. DNA was extracted from Ethylenediamine tetraacetic acid (EDTA) anticoagulated peripheral venous blood samples using the salting out method [218]. DNA concentration was estimated using the Qubit 2.0 Fluorometer and α260:280 ratio calculated using a nanodrop spectrophometer. The average DNA yield obtained was 150µg/ml and approximately 20ug of each patient DNA was extracted for next generation sequencing.

### 2.2.3 The Southampton whole exome analysis pipeline

Exome sequencing was carried out by targeted exome capture using Agilent SureSelect Human All Exon 51Mb V4 followed by sequencing on the Illumina HiSeq system (outsourced to collaborators at Wellcome Trust Centre for Human Genetics, Oxford). The Agilent SureSelect Human All Exon version 4 capture kit was used for sequencing the protein-coding fraction of the genome.

An overview of the analysis workflow for whole-exome sequencing data is shown in Figure 2.2. Paired end sequencing data were aligned against the human genome reference sequence (hg19/GRCh37) using Novoalign (novoalignMPI V3.00.01). Picard was used to mark duplicate reads resulting from PCR clonality or optical duplicates. Sequencing depth and breadth statistics were calculated with custom scripts and the BedTools package[219] (v2.13.2). Variants were excluded if they had a PHRED quality score of <20 and/or a depth of <4. Variants were annotated with respect to genes and RefSeq transcripts and cross-referenced using the Annovar software tool (v2013

Feb21) with: i) databases of known variation downloaded from the Annovar website (April 2013), ii) 1000 Genomes Project (2012 April release), iii) dbSNP137, iv) 4500 European American samples from The National Heart Lung and Blood Institute Exome Sequencing Project, Exome Variant Server (http://evs.gs.washington.edu/EVS/), (ESP6500 release), data from Complete Genomics containing 46 unrelated human subjects. Genes were cross referenced with the list of possible false positive genes from published[220] and in-house exome data repositories. This produced a list of novel and rare variants for each individual. *In silico* functional annotation using evolving bioinformatic resources (e.g. SIFT, Polyphen, GERP++ and Keggseq model) were applied. Resultant variants files for each subject were subjected to further in-house quality control tests to detect DNA sample contamination and ensure sex concordance by assessing autosomal and X chromosome heterozygosity. Variant sharing between all pairs of individuals was assessed to confirm sample relationships. Sample provenance was confirmed by independent genotyping of a validated SNP panel, developed specifically for exome data [221]. All analyses were conducted using the University of Southampton supercomputing machine (Iridis).

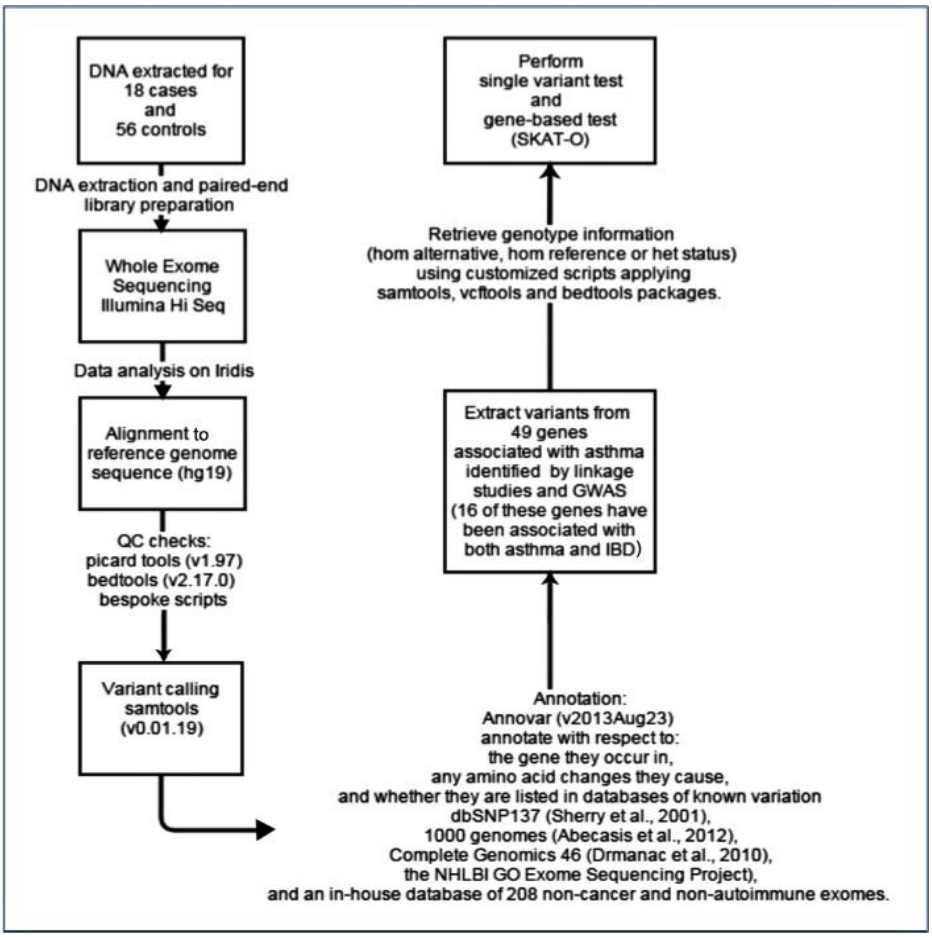



**Figure 2.2 Pipeline used for whole exome data processing and analysis** Briefly**,** Novoalign was used for the alignment step, Samtools for the variant calling and Annovar for annotation.

### 2.2.4 Quality control tests

To check the accuracy and the quality of the sequencing data, five different quality control analyses were conducted. Quality controls are necessary to ensure quality of the DNA sample as samples are often processed in batches. Sequencing protocols require multiple steps of manual sample handling and manipulation, for this reason it is possible that DNA from more than one individual may be placed in the same well or sample switched.

The first test involved assessing the quality of reads after mapping, target coverage at selected sequencing depths and the mean sequence coverage (minimum mean coverage 50). Secondly a similarity matrix for identifying the percentage of shared called variants between any two samples on the same run was constructed. The term similarity matrix refers to a pairwise matrix of the percentage of identical variants for all samples sequenced in the same batch. The matrix enables identification of 1$^{st}$ and

2nd degree relatives which share approximately 60% and 50% of genetic variations respectively (Table 2.1).

**Table 2.1 Similarity matrix: percentage of shared variant across samples sequenced on the same plate.**

| | PR0158 | PR0159 | PR0160 |
|---|---|---|---|
| PR0158 | 100 | 44.27 | 44.02 |
| PR0159 | 44.03 | 100 | 43.96 |
| PR0160 | 43.92 | 44.1 | 100 |

In each square is the percentage of share variants between samples sequenced on the same plate. Samples are coloured based on the percentage. Usually two unrelated individuals share approximately 43% of genetic component.

The third quality control test performed consisted in counting the number of heterozygous variants on the X chromosome to predict the sex of the samples and conduct a blind cross check with the medical records. Males present a percentage of called variants on the X chromosome < 50% whereas females > 50% (Figure 2.3).

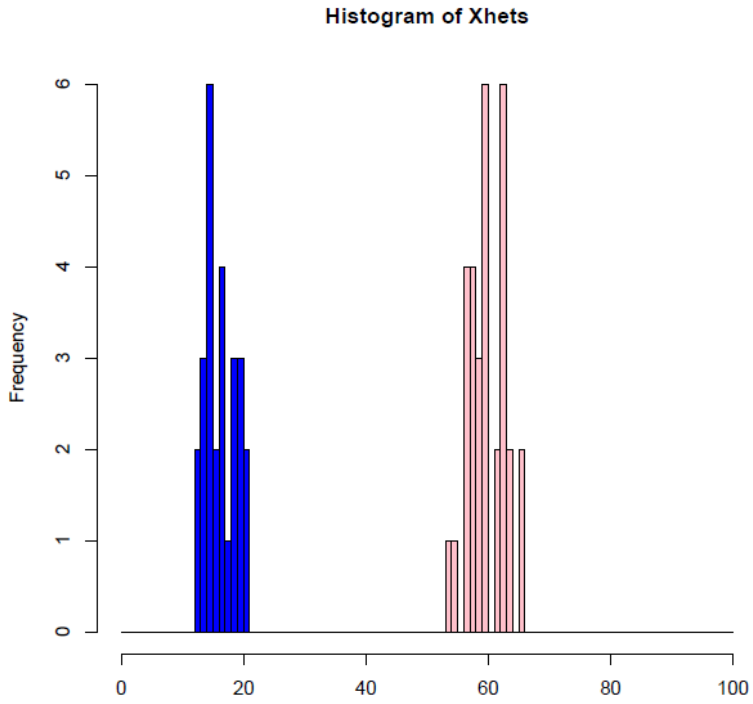



**Figure 2.3 Gender check.** Frequency is plotted on y-axis and the percentage of all heterozygous variants called on the X chromosome for 48 samples is plotted on the x-axis. Males and females are clearly distinguishable in blue and pink respectively.

Following the gender check, the fourth analysis was performed to assess DNA contamination within the samples of the plate. We assess sample contamination by measuring for each sample the deviation of the autosome chromosomes heterozygosity (heterozygous genotypes to alternative allele homozygous genotypes

ratio) from the expected allele ratio of 1.55 (Figure 2.4). If contamination is present due to mixing of two samples, the number of reference alleles present in variants called homozygous for the alternative allele is higher for the two contaminated samples. This test does not take into account sample ethnicity and inbreeding which are biological processes that can alter the heterozygous genotypes to alternative allele homozygous genotypes ratio.

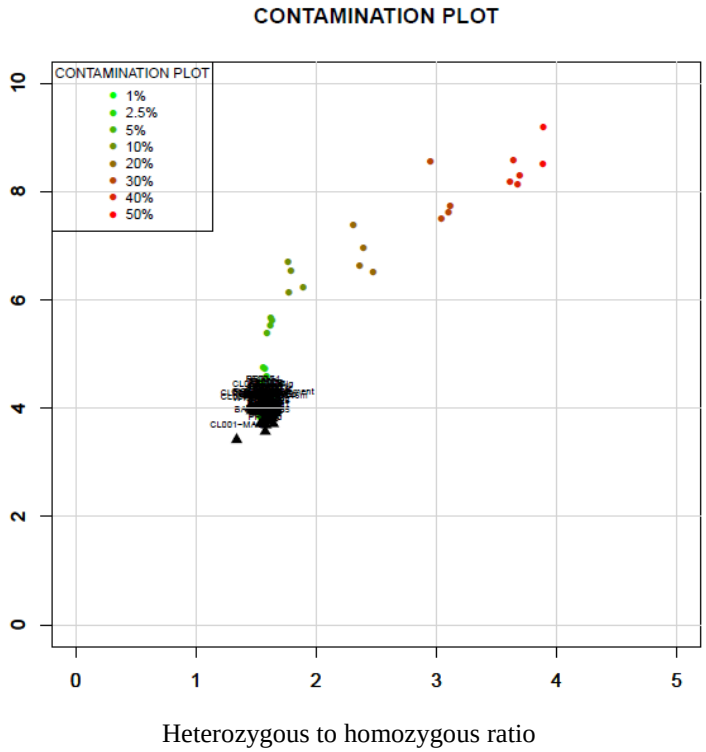



**Figure 2.4 Contamination plot.** The Deviation metric is plotted on the y-axis and the heterozygous to homozygous ratio is on the x-axis. The sequenced samples are represented with triangular shapes while simulated contaminated samples are in a circle shaped.

As a final quality control check we genotyped a copy of the original DNA plate sent to be sequenced using a panel of 24 SNPs designed by our research group including myself to compare the exome data and unambiguously discriminate the samples in the exome sequencing plate. The panel is currently being used by companies such as LGC and other research groups. The list of SNPs was compiled by establishing a list of SNP which overlap regions between the major whole-exome enrichment kits: Agilent SureSelect Human All Exon V4, Illumina TruSeq Exome Enrichment and Nimblegen SeqCap EZ Human Exome Library V3.0 kits. The common regions within capture kits were then compared against common SNPs taken from dbSNP 137 for extracting regions of overlap. From the resulting list of SNPs, SNPs were excluded if they

represented the complementary bases transversions, A↔T and G↔C; if presented in large-scale genomic repeats or in homopolymeric tracts of ≤5 bp; if they were not genotyped in phase 3 HapMap; and if they alter the primary sequence of the protein they encode or they have been reported in OMIM, Online Mendelian Inheritance in Man. Next, SNPs were further excluded if: they were located within 10 bp from exon boundaries; if they were situated in regions with a high sequence similarity to non-target regions (BLAT score >100); and if they were in linkage disequilibrium with any other selected SNPs[221].

### 2.2.5 Gene selection

The latest genome-wide meta-analysis of IBD reported 193 genes across 163 loci with statistically independent signals of association at genome-wide significance (P<5x10-8)[59]. These genes were cross-referenced with 49 genes associated with asthma identified by linkage studies and GWAS[222]. Sixteen of these genes have been associated with both asthma and IBD (Figure 2.5). Gene names were cross-referenced with the HUGO webserver to confirm the approved gene symbol.

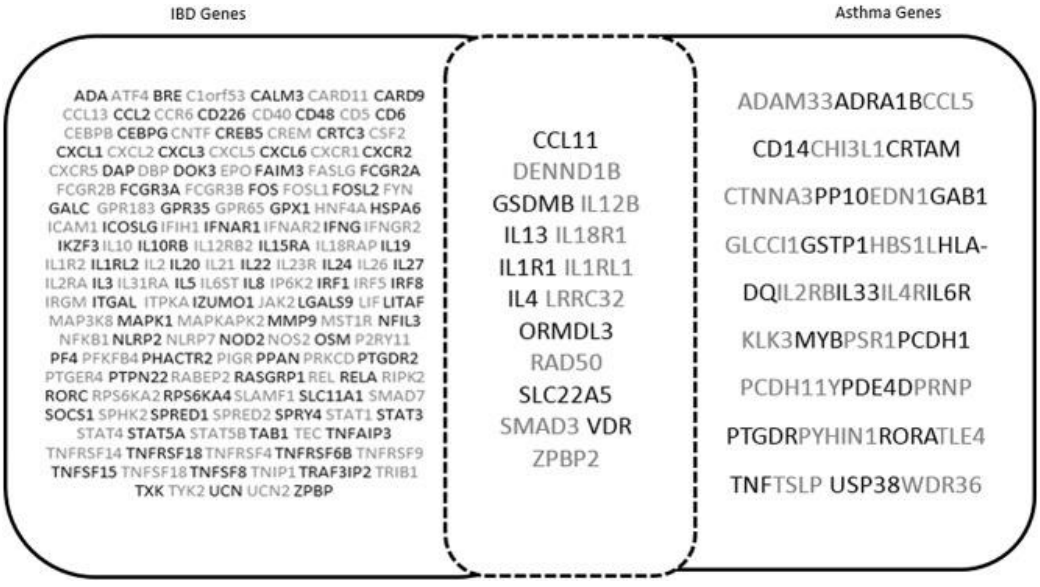



**Figure 2.5 Overlap of GWAS significant gene loci in IBD (left) and asthma (right).** 177 IBD only genes, 33 asthma only genes and 16 genes common to both disorders.

### 2.2.6 Variant association testing

Our findings and those of others[223,224] indicate asthma is the most common concurrent autoimmune disease in IBD patients. For this reason we wanted to further investigate if a subset of patients with a concurrent diagnosis of both IBD and asthma present with

a significant burden of mutation within known genes associated with asthma. Although no test was performed to determine the power of the test, our modest sample size was underpowered to extend this analysis to all 193 IBD genes.

To detect association between the genetic component and disease status first a single variant test and then a gene-based test (SKAT-O) were performed. In order to run these tests, genotype information (homozygous alternative, homozygous reference or heterozygous status) were retrieved using customized scripts applying samtools[181], vcftools[184] and bedtools[219] packages. All variant sites across 49 genes (comprising 33 genes specific to asthma and 16 genes common to both diseases) were used to generate the variant call file (VCF) for each of the 18 exome analyzed patients and 56 unrelated, germline, non-IBD controls without a primary immunodeficiency diagnoses. Our Genomic Informatics group has a rolling database of non-IBD clinical exomes. Controls without any clinical diagnosis of autoimmune disease were selected from this in-house database.

Variations were further excluded based on the Hardy-Weinberg equilibrium status ($p<0.001$) in the control group, by using vcftools[184]. VCF files containing genotype information for all cases and controls were merged together and annotated. Both single and joint analysis were carried out using the EPACTS (Efficient and Parallelizable Association Container Toolbox) software[225].

### 2.2.7 Single variant association testing

The single variant logistic score test[225] was performed in order to detect differences in variant frequency between cases and control group. The test was not performed on mutations occurring in only one individual in either case or control group.

### 2.2.8 Rare variant profile filtering

The burden of rare and novel damaging variation was described for each of the 18 patients across 49 asthma genes. Synonymous variations were excluded from the analysis on the assumption of their low impact on protein function. All novel to individual, novel to Southampton pIBD cohort and clinical variants as well as frameshift insertion, frameshift deletion, stop gain and stop loss mutations were retained for further analysis. Novel to individual denotes variants not previously reported in dbSNP137 database, 1000 Genomes Project, Exome Variants Server (EVS) of European

Americans in the NHLI-ESP project with 6500 exomes [http://evs.gs.washington.edu/EVS/], in 46 unrelated human subjects sequenced by Complete Genomics[226], in other individuals of the Soton IBD cohort, or in the Southampton reference exome database. Novel to Soton cohort denotes variants not previously reported in dbSNP137 database[227], 1000 Genomes Project[119], Exome Variant Server (EVS) of European Americans in the NHLI-ESP project with 6500 exomes [http://evs.gs.washington.edu/EVS/], in 46 unrelated human subjects sequenced by Complete Genomics[226] but has been seen in other individuals of the Soton IBD cohort.

To refine this list to variations most likely to have a biological impact, common variants occurring in ≥ 5% of individuals from 1000 genomes project[119,141] were excluded and variants less likely to impact on protein function as expressed by the logit categorical score[189]were excluded (logit = N). Pathways were determined using DAVID (Database for Annotation, Visualization and Integrated Discovery;)[228] and KEGG pathway[229].

### 2.2.9 Joint variant association testing

The sequence kernel association testing optimal unified test (SKAT-O)[230] is a gene-based test for assessing the contribution of rare and common variations within a genomic locus with trait[230]. Specifically, SKAT-O encompasses both burden test and SKAT[230] test to offer a powerful way of conducting association analysis on combined rare and common variation as single variant tests are often underpowered due to the large sample size needed to detect a significant association. SKAT-O test returns the most significant result between burden test and SKAT testing.

SKAT is a gene based test developed for testing the associations of rare and common variants with a dichotomous, case vs. control, or quantitative trait. SKAT uses a multiple regression model to regress the phenotype on genotype and on covariates such as age and ethnic group. Differently from burden test, SKAT testing is powerful when a region contains both risk, neutral or protective variants. Burden tests aggregate all variants within a region into a single variable and regress the phenotype on this variable. Moreover, burden tests are based in the assumptions that all rare variants within a region are disease causal. SKAT do not make these assumptions and collapse individual variant with weights. SKAT uses weights to up-weight causal variants and down-weight neutral variants. Weights are determined using the beta

distribution density function which uses information derived from the minor allele frequency of each variant within all cases and controls of the cohort under study.

SKAT test can be less powerful than burden tests if a large proportion of the rare variants in a region are truly causal and influence the phenotype in the same direction. Therefore, SKAT-O uses burden and SKAT testing to maximize test power for the particular scenario under investigation[10].

In this analysis we executed the SKAT-O test with the small sample adjustment and by applying a MAF threshold of 0.05 to define rare variations and using default weights. To conduct the test, a group file with mutations of interests (missense, nonsense, splice-site variants and coding indels) was created for each of the 49 genes.

## 2.3 Results

### 2.3.1 Demographic data and prevalence of autoimmune comorbidity

At the time of the analysis the Southampton pIBD study cohort comprised 173 children (142 CD, 69 UC and 20 IBDU). Demographics for all patients is shown in Table 2.2; a marginal excess of males was observed, with the majority of patients being of white British ethnicity.

**Table 2.2 Demographic data representing the Southampton paediatric IBD cohort.**

| | CD | UC | IBDU | Total IBD |
|---|---|---|---|---|
| **Number of patients**<br>**VEO-IBD** | 98<br>7 | 55<br>9 | 20<br>2 | 173<br>18 |
| **Median Age**<br>**($25^{th}$/$75^{th}$ centile)** | 12.28<br>(9.47/14.29) | 11.73<br>(9.67/13.59) | 12.15<br>(8.72/14.10) | 12.28<br>(9.56/14.17) |
| **Female (n)** | 52 | 30 | 12 | 94 |
| **Mean age of Onset** | 11.74 | 11.19 | 11.43 | 11.55 |

CD: Crohn's disease, UC: ulcerative colitis, IBD: inflammatory bowel disease unclassified. The cohort is composed of 231 recruits (142 CD, 69 UC and 20 IBDU). Median age, number of female and mean age of onset is presented. The mean age for the entire cohort is 12 years with a mean age of onset of almost 12 years. In the cohort is composed by a higher component of male patients (n=137) compared to female patients (n=94).

The interrogation of patient data revealed concurrent diagnosis of pIBD with six distinct autoimmune mediated conditions (Table 2.3). Asthma and atopic dermatitis were the most common autoimmune comorbidities affecting 28 and 24 children respectively. Across the cohort, less frequent cases of sclerosing cholangitis (n=4), coeliac disease (n=2) and vitiligo (n=2) were present. Forty-nine children (28.3%) presented with a second autoimmune condition. Across the cohort there was a family history of asthma, atopic dermatitis, coeliac disease, sclerosing cholangitis and vitiligo although no probands had diagnosis of these conditions at the time of analysis. Of the 28 IBD patients who had a concurrent diagnosis of asthma we selected the 18 youngest of these for exome analysis. All 18 were of white British ancestry.

**Table 2.3 Prevalence of autoimmune disease in the pIBD cohort (173 patients)**

| Autoimmune disease | Crohn's Disease (n = 98) | Ulcerative Colitis (n = 55) | IBDU (n= 20) | Overall pIBD cohort prevalence (n=173) | Overall Population paediatric prevalence (%) |
|---|---|---|---|---|---|
| **Asthma** | 19 (19.40%) | 9 (16.40%) | 0 | 28 (16.18%) | 15.00[224,231,232] |
| **Atopic dermatitis** | 18 (18.40%) | 6 (8.11%) | 0 | 24 (13.87%) | 16.50[233] |
| **Coeliac disease** | 1 (1.020%) | 1 (1.35%) | 0 | 2 (1.15%) | 0.99[234] |
| **Sclerosing cholangitis** | 1 (1.02%) | 2 (3.64%) | 1 (50%) | 4 (2.31%) | 0.01*[235] |
| **Vitiligo** | 1 (1.02%) | 0 | 1 (50%) | 2 (1.15%) | 1.00[236] |

*Data for general population prevalence only and not specific to general IBD.

### 2.3.2 Quality control checks on exome data

No unexpected relationship was identified from the similarity matrix. Tables in appendix I and II represent quality of reads and similarity matrix showing the percentage of variants shared between each pair of samples. Table 2.4 reports the results of the gender and DNA contamination checks. Excess heterozygosity was ascertained if the percentage of autosome heterozygosity ratio was greater than to 2 standard deviations above the mean. The measure of X-chromosome heterozygosity followed by blind gender cross check between the sex predicted by the pipeline and the clinical research database revealed a discrepancy between sample PR0150 and PR0151. The identification of the switch was further supported by the application of the 24 SNP panel suggesting a reciprocal transposition, Table 2.5. The arrangement of the samples on the plate sent for exome sequencing showed that the two samples, PR0150 and PR0151, were adjacent to each other suggesting the sample transposition during one of the multiple wet lab handling steps of exome capture and sequence analysis. The identification of the switch and resolution of identity permitted downstream analyses.

**Table 2.4 Results from the contamination and gender checks for each of the 18 samples**

| Sample ID | % Autosome heterozygosity | Exome sex | Clinical database sex |
|---|---|---|---|
| PR0007 | 59.85 | M | M |
| PR0011 | 60.99 | M | M |
| PR0031 | 59.89 | M | M |
| PR0032 | 59.64 | M | M |
| PR0036 | 60.86 | F | F |
| PR0039 | 60.22 | M | M |
| PR0085 | 60.02 | M | M |
| PR0110 | 61.17 | F | F |
| PR0158 | 60.58 | F | F |
| PR0160 | 61.26 | M | M |
| PR0167 | 60.67 | M | M |
| PR0188 | 60.91 | M | M |
| PR0068 | 61.73 | F | F |
| PR0083 | 61.96 | F | F |
| PR0107 | 60.51 | M | M |
| PR0146 | 59.99 | M | M |
| PR0148 | 60.77 | M | M |
| ***PR0151*** | ***61.03*** | ***M*** | ***F*** |

X heterozygosity represents the percentage heterozygosity on theX chromsome; male if <50 otherwise female; % autosome heterozygosity represents the heterozygotes:homozygous ratio over all autosome chromosomes (contamination is suspected if the ratio exceed 62%). Gender discordance for sample PR0151 is in bold-italics.

**Table 2.5 Exome and genotype data for PR0150 and PR0151 in adjacent rows**

| Sample | rs2228611 | rs497692 | rs1410592 | rs2229546 | rs10203363 | rs2819561 | rs4688963 | rs309557 | rs2942 | rs17548783 | rs4735258 | rs1381532 | rs10883099 | rs4617548 | rs7300444 | rs9532292 | rs2297995 | rs4577050 | rs2070203 | rs1037256 | rs2298628 | rs10373 | rs4148973 | rs4675 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PR0150 | C | A | C | C | A | T | T | G | T | C | A | G | G | C | A | G | G | G | G | N | C | G | G | T |
| exome | A | A | C | T | G | T | C | A | C | C | A | G | G | T | A | G | A | A | G | N | C | G | G | T |
| PR0150 | C | G | T | C | G | T | T | A | C | C | G | A | A | C | A | A | G | A | G | T | T | A | G | C |
| genotype | A | A | C | T | G | T | C | A | C | C | G | G | G | C | A | A | A | A | A | C | C | G | G | T |
| PR0151 | C | G | T | C | G | T | T | A | C | C | G | A | A | C | A | A | G | A | G | N | T | A | G | C |
| exome | A | A | C | T | G | T | C | A | C | C | G | G | G | C | A | A | A | A | A | N | C | G | G | T |
| PR0151 | C | A | C | C | A | T | T | G | T | C | A | G | G | C | A | G | G | G | G | T | C | G | G | T |
| genotype | A | A | C | T | G | T | C | A | C | C | A | G | G | T | A | G | A | A | G | C | C | G | G | T |

Top row lists the SNPs of the panel. Markers for the resolution of the switch are in yellow. N indicates missing data.

### 2.3.3 Single variant association test for variants in asthma and dual susceptibility genes

Among the 28 patients affected by asthma, the 18 youngest patients were selected for exome sequencing (9 CD and 9 UC). Characteristics for each of the patients that underwent exome sequencing are presented in Table 2.6. Thirty-six of the 49 genes either specific to asthma and common to both asthma and IBD were analysed, as no coding variants were called in *ADRA1B, CCL5, CD14, HLA-DQ, IL12B, IL13, IL4, ORMDL3, PCDH1, RAD50, TNF, TSLP* and *SLC22A5* across cases and controls and these genes were excluded from single variant and joint testing.

**Table 2.6 Clinical profile of the 18 patients with concurrent IBD and asthma selected for exome analysis**

| Study ID | Sex | Age at Diagnosis of IBD | Diagnosis | Paris classification of IBD | Other autoimmune disease status other than asthma |
|---|---|---|---|---|---|
| **PR0007** | M | 11.41 | CD | A1L34B1 | |
| **PR0011** | M | 15.51 | CD | A1L1B2 | |
| **PR0031** | M | 11.43 | CD | A1L3B1p | |
| **PR0032** | M | 7.29 | CD | A1L24B1p | Atopic dermatitis |
| **PR0036** | F | 9.67 | CD | A1L3B1 | |
| **PR0039** | M | 10.30 | UC | E3- | |
| **PR0068** | F | 11.23 | UC | E2- | |
| **PR0083** | F | 9.68 | UC | E3S3 | |
| **PR0085** | M | 13.12 | UC | E3S3 | |
| **PR0107** | M | 9.22 | CD | A1L1- | |
| **PR0110** | F | 2.98 | UC | E3- | |
| **PR0146** | M | 14.52 | UC | A1L3- | |
| **PR0148** | M | 9.13 | CD | A1L34B3p | Atopic dermatitis |
| **PR0151** | F | 13.30 | CD | A1L24B1 | |
| **PR0158** | F | 15.25 | UC | E3S3 | Atopic dermatitis |
| **PR0160** | M | 12.55 | UC | E3S1 | Atopic dermatitis |
| **PR0167** | M | 13.30 | UC | E2S2 | Atopic dermatitis |
| **PR0188** | M | 11.03 | CD | A1L1- | |

- Denotes missing classification data.

A total of 175 different variants were identified across 36 genes in the case and control exomes. 73 occurred only in one individual (case or control) and these were not analyzed in the single variant test. The single variant test was applied to 102 variants across 33 genes. Three of these variants showed significant association with disease status (association p <0.05; *PYHIN1, ZPBP2* and *LRRC32*). However, none of these variants withstood multiple testing corrections (Table 2.7). *ZPBP2* and *LRRC32* are known to be involved in both asthma and IBD. Within these genes, *ZPBP2*

nonsynonymous variant at position 38027030 bp and *LRRC32* synonymous variant at position 76372052 bp (p= 0.011 and p=0.043 respectively) were found at higher frequency in cases compared to controls. *PYHIN1* is known to be involved in asthma pathogenesis. In this gene the nonsynonymous variant at position 158943483bp (p= 0.008) was observed at higher frequency in cases compared to control group. The frequency of these mutations suggests their possible deleterious effects in increasing disease risk in genetically susceptible individuals.

**Table 2.7 Single variant test results for the 36 known asthma genes in which variation was found across the cohort.** Only variants with a p<0.1 are shown.

| Gene set | Gene | Chr | Bp position (hg19) | Var | MAF in the cohort | Allele frequency in cases | Allele frequency in controls | P value unadjusted |
|---|---|---|---|---|---|---|---|---|
| **Asthma/IBD** | ZPBP2 | 17 | 38027030 | ns | 0.014 | 0.056 | 0 | 0.011 |
| **Asthma** | PYHIN1 | 1 | 158943483 | ns | 0.027 | 0.083 | 0.009 | 0.015 |
| **Asthma/IBD** | LRRC32 | 11 | 76372052 | sn | 0.425 | 0.722 | 0.527 | 0.044 |
| **Asthma/IBD** | SMAD3 | 15 | 67457698 | ns | 0.034 | 0.083 | 0.018 | 0.054 |
| **Asthma** | NPSR1 | 7 | 34917768 | ns | 0.034 | 0.083 | 0.018 | 0.056 |
| **Asthma** | IL6R | 1 | 154426970 | ns | 0.419 | 0.556 | 0.375 | 0.07 |
| **Asthma/IBD** | SMAD3 | 15 | 67457335 | ns | 0.135 | 0.944 | 0.839 | 0.08 |
| **Asthma** | PYHIN1 | 1 | 158908886 | ns | 0.02 | 0.056 | 0.009 | 0.081 |
| **Asthma** | NPSR1 | 7 | 34917702 | ns | 0.108 | 0.028 | 0.134 | 0.097 |

*Found 175 variant in 36 genes. 76 variants were removed because occurring in 1 individual in both cases and controls. No variants in *ADRA1B, CCL5, CD14, HLA-DQ, IL12B, IL13, IL4, ORMDL3, PCDH1, RAD50, TNF, TSLP and SLC22A5*. ns, nonsynonymous and sn, synonymous

### 2.3.4 Individual profiles of rare and deleterious variants

Individual burden of variation revealed 24 variants (Table 2.8). Several dual susceptibility genes (for pIBD and Asthma) were identified as harbouring one or more variants. Mutations fall within 3 pathways consistently reported in KEGG and DAVID. *ZPBP2* (zona pellucida binding protein 2[237]) and *SMAD3* (involved in the adherens junction pathway[238]) have variants observed across both CD and UC but these variants also occur in 1% and 2% of the 1000 Genomes reference population .

Patient PR0085 carries the nonsynonymous *ZPBP2* mutation, but also carries a novel frameshift deletion in *DENND1B* gene expressed by natural killer cells and dendritic cells[107,239]. The same patient harbours a nonsynonymous mutation at position 67353579 bp within *GSTP1* which is reported to be involved in asthma pathogenesis

only. Also of interest amongst the genes previously implicated in both diseases are two distinct and very rare mutations in the *RAD50* gene located within the IBD5 cytokine cluster on chromosome 5q31[111]. This gene contains the locus control region required for the Th2 cytokine gene expression[240]. In asthma specific genes, variations were found in 8 genes. A more common variant (rs3918396) is seen to recur within the *ADAM33* gene, a second variant in the same gene has been identified in PR0158 and other patients within the Southampton pIBD cohort.

PR0110 is a patient diagnosed aged 2 years with severe UC. She is seen to harbour a mutation that could impact splicing at position 8009439 bp in *GLCCI1* and a novel frameshift insertion at position 69407255 within *CTNNA3*. This gene encodes the α –T-catenin protein; a key component of the adherens junctional complex in epithelial cells necessary for cellular adherence[241].

**Table 2.8 Deleterious variant profiles across asthma genes for all 18 patients with either CD or UC**

| | Chr | Position hg19 | Gene | Var | Nucleotide change | Protein change | NtI* | Phylop | Gerp | Logit Model[35] | Grantham | rsID | 1000 genome project MAF< 5% | EVS MAF< 5% | Crohn's disease | | | | | | | | | Ulcerative colitis | | | | | | | | | Pathways |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | | | | | PR0007 | PR0011 | PR0031 | PR0032 | PR0036 | PR0107 | PR0148 | PR0151 | PR0188 | PR0039 | PR0068 | PR0083 | PR0085 | PR0110 | PR0146 | PR0158 | PR0160 | PR0167 | |
| Asthma Genes | 20 | 3651765 | ADAM33 | ns | G2050A | V684I | | 0.982259 | 3.71 | Y | 29_Con | rs3918396 | 0.05 | 0.088085 | . | . | . | X | . | X | . | . | . | . | . | . | . | . | . | . | . | . | |
| | 20 | 3652641 | ADAM33 | sp | 1598-9C>A | . | NtSC | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | X | . | . | |
| | 5 | 159344587 | ADRA1B | sn | T675C | R225R | NtSC | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | X | |
| | 10 | 69407255 | CTNNA3 | fi | 17_18insA | P6fs | NtI | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | X | . | . | . | . | ① | |
| | 4 | 144359490 | GAB1 | ns | C932T | P311L | . | 0.999098 | 5.45 | Y | 98_ModCon | rs28925904 | 0.01 | 0.020349 | . | . | . | X | . | . | . | . | . | . | . | . | . | . | . | . | . | . | |
| | 7 | 8009439 | GLCCI1 | sp | 457+1G>A | . | NtI | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | X | . | . | . | . | |
| | 11 | 67352183 | GSTP1 | ns | G172A | D58N | . | 0.999701 | 5.42 | Y | 23_Con | rs45543438 | . | 0.00048 | . | . | . | . | X | . | . | . | . | . | . | . | . | . | . | . | . | . | |
| | 11 | 67353579 | GSTP1 | ns | C341T | A114V | . | 0.793 | -2.42 | Y | 64_ModCon | rs1138272 | 0.04 | 0.082182 | . | . | . | . | . | X | . | . | . | . | . | . | X | . | . | X | . | . | |
| | 22 | 37531375 | IL2RB | ns | G811A | G271R | . | 0.11428 | -4.47 | Y | 125_ModRad | rs116250343 | 0.01 | 0.000698 | X | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | ② ③ |
| | 16 | 27357927 | IL4R | ns | C26T | T9M | . | | | . | 81_ModCon | rs2234895 | 0.05 | 0.088488 | . | . | . | . | . | X | . | X | . | . | . | . | . | . | . | . | . | X | ② ③ |
| | 6 | 135522781 | MYB | ns | A1907C | E636A | . | 0.998806 | 5.32 | Y | 107_ModRad | . | . | 0.000116 | X | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | |
| | 5 | 141244811 | PCDH1 | ns | G1085A | R362Q | . | 0.998974 | 4.8 | Y | 43_Con | rs200086291 | 0.0014 | 0.000581 | . | . | X | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | |
| | 5 | 141248151 | PCDH1 | ns | G886A | G296S | . | 0.998159 | 4.26 | Y | 56_ModCon | rs148156227 | . | 0.000116 | . | . | . | . | . | X | . | . | . | . | . | . | . | . | . | . | . | . | |
| | 14 | 52734972 | PTGDR | ns | G440A | R147H | NtI | 0.99145 | 3.25 | Y | 29_Con | . | . | . | . | X | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | |
| | 15 | 60884594 | RORA | ns | C23A | A8E | . | | | Y | 107_ModRad | rs150952015 | 0.05 | 0.043895 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | X | . | |
| | 15 | 60806942 | RORA | sn | A396G | R132R | NtSC | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | X | . | . | . | . | . | . | . | . | |
| Asthma/IBD genes | 1 | 197479900 | DENND1B | fd | 2015_2018del | 672_673del | NtI | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | X | . | . | . | . | . | |
| | 1 | 197480986 | DENND1B | ns | G1687A | G563S | . | . | . | Y | 56_ModCon | rs10494755 | 0.05 | 0.097093 | . | . | . | . | . | . | . | . | X | . | . | . | . | . | . | . | . | . | |
| | 5 | 158743788 | IL12B | ns | G892T | V298F | . | 0.99929 | 5.35 | Y | 50_Con | rs3213119 | 0.02 | 0.030349 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | X | . | ② | |
| | 2 | 102781649 | IL1R1 | ns | C371G | A124G | . | 0.86635 | 1.3 | Y | 60_ModCon | rs2228139 | 0.05 | 0.064186 | . | . | . | . | . | . | X | X | . | . | . | X | . | . | . | . | . | . | |
| | 5 | 131915673 | RAD50 | ns | G671A | R224H | . | 0.993424 | 3.85 | Y | 29_Con | rs28903088 | . | 0.00093 | . | . | . | . | . | . | . | . | . | . | . | X | . | . | . | . | . | . | |
| | 5 | 131953758 | RAD50 | sp | 3165-4A>T | . | CS | . | . | . | . | rs104895050 | 0.001 | . | . | . | . | . | X | . | . | . | . | . | . | . | . | . | . | . | . | . | |
| | 15 | 67457698 | SMAD3 | ns | A193G | I65V | . | 0.997447 | 4.6 | Y | 29_Con | rs35874463 | 0.02 | 0.054664 | . | . | . | . | . | . | . | . | . | . | X | . | . | . | . | X | . | . | ① |
| | 17 | 38027030 | ZPBP2 | ns | C136G | P46A | . | 0.998789 | 4.66 | Y | 27_Con | rs35591738 | 0.01 | 0.010385 | . | . | . | . | . | X | . | . | . | . | . | . | X | . | . | . | . | X | |
| Variant/s per individual | | | | | | | | | | | | | | | 2 | 1 | 1 | 2 | 2 | 4 | 2 | 2 | 1 | 1 | 1 | 2 | 3 | 2 | 0 | 3 | 2 | 3 | |

Cross (X) indicates presence of a variant

ns, nonsynonymous; sn, synonymous; fi, frameshift insertion, fd, frameshift deletion; sp, splicing; nfi, non-frameshift insertion; nfd, non-frameshift deletion.

Con, conservative; ModCon, moderative conservatie; ModRad, moderative radical.

*NtI (novel to individual) denotes variants not previously reported in dbSNP137 database, 1000 Genomes Project, Exome Variants Server (EVS) of European Americans in the NHLI-ESP project with 6500 exomes[http://evs.gs.washington.edu/EVS/], in 46 unrelated human subjects sequenced by Complete Genomics[226], in other individuals of the Soton IBD cohort, or in the Southampton reference exome database.

NtSC (novel to Soton cohort) denotes variants not previously reported in dbSNP137 database[227], 1000 Genomes Project[119], Exome Variants Server (EVS) of European Americans in the NHLI-ESP project with 6500 exomes[http://evs.gs.washington.edu/EVS/], in 46 unrelated human subjects sequenced by Complete Genomics[226] but has been seen in other individuals of the Soton IBD cohort.

CS (Clinical significance) assigned dbSNP value of probable-pathogenic or pathogenic[227].

Pathways: ① Adherens junction[229], ②Jak-STAT signaling pathway[229] and ③Cytokine-cytokine receptor interaction[229]

### 2.3.5 Joint rare variant association test for variants in IBD and dual susceptibility genes

The joint test for assessing the contribution of private, rare and common mutation between disease status and genes highlighted 6 genes with a p-value < 0.05 prior to Bonferroni correction (Table 2.9).

Of these 6 genes, 3 are known susceptibility genes for both IBD and asthma (*ZPBP2* p= 0.009, *IL1R1* p= 0.036 and *IL18R1* p= 0.038); the remaining genes were asthma specific (*PYHIN* p= 0.025; *IL2RB* p=0.036; *GSTP1* p= 0.040). These genes are all key determinants of the immune response and have variants observed across both CD and UC.

**Table 2.9 Joint variant test (SKAT-O) result for the 36 known asthma genes in which variations was found across the entire cohort. Only genes with a p<0.1 are shown.**

| Gene set | Gene | Chr | Bp position (hg19) | Total number of samples 18 cases; 56 controls) | Fraction of individuals who carry rare variants under the MAF thresholds (MAF < 0.05)* | Number of all variants defined in the group file | Number of variant defined as rare (MAF < 0.05)* | P value unadjusted | Weighted (W) or Unweighted (UW) p value |
|---|---|---|---|---|---|---|---|---|---|
| **Asthma/IBD** | ZPBP2 | 17 | 38024626-38032996 | 74 | 0.027 | 4 | 1 | 0.009 | W |
| **Asthma** | PYHIN1 | 1 | 158906777-158943483 | 74 | 0.108 | 5 | 5 | 0.025 | UW |
| **Asthma/IBD** | IL1R1 | 2 | 102781629-158943483 | 74 | 0.014 | 5 | 2 | 0.037 | W |
| **Asthma** | IL2RB | 22 | 37524329-37539651 | 74 | 0.014 | 4 | 1 | 0.037 | UW |
| **Asthma/IBD** | IL18R1 | 2 | 102984279-103001402 | 74 | 0.014 | 3 | 1 | 0.039 | UW |
| **Asthma** | GSTP1 | 11 | 67352183-67353970 | 74 | 0.014 | 4 | 1 | 0.041 | W |
| **Asthma** | TLE4 | 9 | 82187750-82336794 | 74 | 0.108 | 4 | 4 | 0.056 | UW |
| **Asthma** | NPSR1 | 7 | 34698177-34917768 | 74 | 0.135 | 12 | 4 | 0.066 | UW |
| **Asthma** | CTNNA3 | 10 | 67680203-69407255 | 74 | 0.162 | 6 | 3 | 0.081 | UW |

* These variants received different weights in the SKAT-O joint test.

No variants were found in *ADRA1B, CCL5, CD14, HLA-DQ, IL12B, IL13, IL4, ORMDL3, PCDH1, RAD50, TNF, TSLP* and *SLC22A5* across cases and controls.

## 2.4 Discussion

Inflammatory bowel disease is associated with several extraintestinal manifestations that may produce greater morbidity than the underlying intestinal disease and may even be the initial presenting symptoms of the IBD. Recent studies have shown that 36% of patients with IBD have at least one extraintestinal manifestation. Some are more commonly related to active colitis (joint, skin, ocular, and oral manifestations). Others are especially seen with small bowel dysfunction (cholelithiasis, nephrolithiasis, and obstructive uropathy), and some are nonspecific disorders (osteoporosis, hepatobiliary disease, and amyloidosis)[6]. Extraintestinal IBD-related immune disease can be classified into two major groups: the first includes reactive manifestations often associated with intestinal inflammatory activity and therefore they represent a pathogenic mechanism common with intestinal disease (arthritis, erythema nodosum, pyoderma gangrenosum, aphthous stomatitis, uveitis); the second includes many autoimmune diseases independent of the bowel disease that reflect only a major susceptibility to autoimmunity. They are not considered (apart for primary sclerosing cholangitis) as specific IBD features but only as autoimmune associated diseases such as ankylosing spondilitis, primary biliary cirrhosis, alopecia areata, and thyroid autoimmune disease and others. Many studies in genetically susceptible animal models suggest the important role of enteric flora in activating the immune system against bacterial antigens. The sharing of these colonic antigens by extraintestinal organs, associated with a genetic susceptibility, would lead to an immune attack to these organs. Studies show that children with pIBD are more likely to have other autoimmune mediated conditions suggesting shared genetic components play a major role in the predisposition of the individual to both groups of disease[80].

Our cohort of 173 children with pIBD revealed forty-nine children (28.3%) with a concurrent diagnosis of an autoimmune disease. Asthma and atopic dermatitis occurred with the highest frequency; the prevalence of clinically diagnosed asthma was 19.4% in children with CD and 16.4% in patients with UC, exceeding UK disease estimates (15.3%[242,243]).

Although this study is not designed to demonstrate a statistically significant increase in autoimmune disease burden in children with pIBD, our observations indicate prevalence estimates approaching the upper limit recorded in the literature[80]. Our findings are consistent with literature indicating children with pIBD are more likely to

have other autoimmune conditions, and that a common genetic components etiology may predispose individuals to multiple autoimmune manifestations[80,223].

Even in a very modest cohort, SKAT-O association analysis revealed six genes with significant burden of mutation. While significance levels would not withstand a Bonferroni correction for 36 genes tested, the strong prior hypothesis to the analysis of these genes might suggest such a multiple testing correction would be inappropriate.

*ZPBP2* is located on the chr17q12-q21 region which has been associated with early-onset asthma, and variants in the same linkage disequilibrium block have been associated with Crohn's disease, type 1 diabetes and primary biliary cirrhosis[244].

*IL1R1* encodes for a cytokine receptor that belongs to the interleukin-1 receptor family. The gene was found to be associated with asthma in a GWAS on 933 European ancestry individuals with severe asthma based on Global Initiative for Asthma (GINA) criteria[245]. At the same genomic region of *IL1R1, IL18R1* was also identified as associated with asthma. Specifically the gene was evaluated in a GWAS conducted on Mexican pediatrc patients[246]. The association was further replicated in a family-based study on Denmark, United Kingdom and Norway families[247].

*GSTP1* is involved in the detoxification of a wide variety of exogenous and endogenous compounds, including reactive oxygen species. This gene was associated by a GWAS early onset asthma[248]. *IL2RB* is involved in lymphoid cell differentiation and it was firstly associated by GWAS conducted by the GABRIEL consortium in 2012[249]. *PYHIN1* (Pyrin And HIN Domain Family, Member 1) encodes a protein that belongs to the HIN-200 family of interferon inducible proteins, important in controlling cell cycle, differentiation and apoptosis[250]. It has been noted to be an asthma susceptibility locus, specifically in those of African descent[251]. *PYHIN1* was identified as associated with asthma in 2011 through a meta-analysis conducted on 5,416 European American, African American or African Caribbean, and Latino ancestry individuals with asthma. The *PYHIN1* association was specific to the African descent groups[251].

*PYHIN1* and *ZPBP2* were significantly associated with asthma in both single variant testing and following SKAT-O testing. Variants within these genes were found with higher frequency in cases compared to controls suggesting a deleterious role of the mutations in the pathogenesis of disease. Susceptibility genes for both IBD and asthma are most commonly involved with immune regulation raising the possibility of an

overall immune dysregulation underlying both diseases. These genes may be implicated in the same pathways as found in other probands but may not yet have been associated with IBD/asthma or did not hold enough significance association to be included in GWAS meta-analyses.

This study demonstrates robust data collection, all pIBD diagnoses are made using strict criteria[217] and autoimmune comorbidity was validated through integration of the medical notes (paper and electronic). This study looked only at genes identified through GWAS (of asthma and IBD) this increased the probability of finding causal rare and private mutation within known implicated genes. By design, GWAS are powered only to implicate genes in which common variant alleles are overrepresented in the disease population. It is highly likely that pathogenic coding changes that are either very rare or even private to individuals in other genes have gone undetected by these methods.

Exome sequencing allows capture of extremely large and useful amounts of data. Limitations of this sequencing technique still exist and can have an impact on research data; inefficiencies in the exon targeting process can lead to uneven capture and result in exons with low sequence coverage and off-target hybridizations. Alongside this unknown or yet-to-be-annotated exons, evolutionary conserved non-coding regions and regulatory sequences (such as enhancers or promoters) involved in IBD and asthma will not be captured. Exome sequencing is not designed to capture information regarding the methylation state of DNA and therefore epigenetic factors in disease are not investigated. Necessary filtering of vast datasets intrinsic to NGS may lead to the exclusion of valid variants.

## 2.5 Conclusions

In this study, we identified the prevalence of concurrent autoimmune diagnoses in a cohort of children with childhood onset IBD. We observe a frequency of asthma and atopic dermatitis at the highest end of the normal range. As shown in Table 2.8, we demonstrated that in children with asthma the spectrum of known, rare and novel variation in established disease-related genes is extensive and varied, even when restricted to mutations predicted to be pathogenic. NGS may be set to become a key routine diagnostic tool of the future, and it is important we begin to elucidate the role of key genes and pathways already known to us. Improved assessment of true functional significance of mutations will require substantial improvements to *in silico* annotation informed by rigorous and extensive functional validation of rare variants using patient cell lines, but such extensive studies of functional relevance are beyond the scope of this this analysis. However, perfect annotation of single variants in isolation cannot predict outcome in patients who harbour a profile of variants and across genes and pathways. This bottleneck to the interpretation of genomic data may be aided by assessment of highly selected patient groups[252]. Our study uncovers the patient specific burden of pathogenic mutations in known disease genes. We find evidence to support causality of key genes such as *ZPBP2* and *PYHIN1* and further postulate that for a subset of patients, the relationship between concurrent pIBD and autoimmune disease lies in systemic immune dysregulation rather than organ specific immune dysfunction.

Exome sequencing may be set to become a key routine diagnostic tool of the future, and it is important we begin to elucidate the role of key genes and pathways already known to us.

# Chapter 3 **Exome Analysis of Rare and Common Variants within the NOD Signaling Pathway**

## 3.0 Summary

Next generation sequencing has enabled the analysis of rare and private variations. Rare variants are believed having a stronger impact on complex diseases compared to common variants[253]. For rare mutations, single association testing has limited power because (often inaccessibly) large sample sizes are required to actually detect association with a single variant. In order to overcome this problem, several statistical tests in which rare and common variants are jointly analysed within a gene or a region have been developed. In this chapter we aim to apply the SKAT-O test, a joint test for rare and common variants that can be adjusted for covariates, for detecting association between genes involved in the NOD signalling pathway and disease. *NOD2* was the first gene found to be associated with CD and 11 genes of the pathway (out of 41) have been already implicated in IBD by GWAS. For this reason, we decided to test the application of this software to this established pathway and determine evidence for previously unknown genes in IBD susceptibility.

My contribution to this analysis was the design of the pipeline for the execution of gene based statistical analysis, collection of data and analysis of the validation and replication cohort, data interpretation, manuscript write-up and submission.

## 3.1 Background

Since the discovery of *NOD2* in 2001 as the first susceptibility gene for IBD[108], over 200 loci have been associated with IBD risk in humans through GWAS.[136] GWAS have provided substantial insight into the understanding of the biology of complex diseases by providing robust and replicated evidence for autophagy[254], immune response[254] and bacterial recognition[254] patterns. However, an intrinsic limitation of these studies is their focus on common variation, typically those with a MAF ≥ 5% in the general population. The combined contribution of these common mutations to IBD heritability only account for 13.6% of CD and 8.2% of UC respectively.[140] It is hypothesized that low frequency (MAF of 0.05 -5%) and rare (MAF ≤ 0.05%) variation may contribute significantly towards some fraction of the missing heritability of IBD.[138–140]

Recent technological advances in DNA sequencing have made it possible to sequence large tracts of the genome in a cost-effective manner. This has enabled large-scale studies of the impact of rare variants on complex diseases.[156] WES and WGS have improved the understanding of genetic cause of diseases by revealing variants not captured by GWAS.[157] It is estimated that ~85% of disease-causing mutations reside within the coding regions of the genome.[255] Therefore, targeting these expressed regions of the genome represents the most cost-effective means to uncover causal disease genes.[256] GWAS are powered to assess common variation in large patient cohorts that are often necessarily composed of adults in order to amass sizeable patient groups. Large cohorts of patients with disease onset in childhood are less easily ascertained and also likely enriched for rare or private variation of large effect.[162] However, several properties of rare variants make their genetic effects difficult to detect with traditional statistical approaches. Rare variants by definition have a low allele frequency that makes population-based methods difficult to implement, and a combination of multiple rare variants might contribute to population risk. The general idea underlying rare variants statistical tests is to assess a set of rare variants in a defined region or regions, by collapsing or aggregating rare variants[242]. To improve the statistical power several strategies for weighting rare variants and/or adding the informative covariates in the model have been published[257].

The maintenance of the equilibrium in the intestinal mucosa is an important process in order to preserve the normal mucosal physiology and prevent triggers which might contribute to the development of many gastrointestinal disorders including IBD[254]. The intestinal barrier presents endogenous defensive mechanism, such as antimicrobial peptides (AMPs), to modulate immune response e maintain intestinal homeostasis. The release of AMPs is regulated by Toll-like receptor (TLR) and NOD2 signals triggered by gut flora[254]. Since the discovery of *NOD2*, there has been a great interest in understanding the role and function of NOD2 in IBD risk with computational and functional studies[258].

NOD2 belongs to the nod-like receptor (NLR) family a group of phylogenetically conserved intracellular proteins. NLR proteins share a domain structure made up of three components: a carboxy-terminal leucine rich domain, involved in the ligand recognition, a central NOD domain, and an amino-terminal domain composed of CARDs or pyrin domains that allows molecular interactions[259]. Proteins containing

other CARD or pyrin domains perform different molecular functions but both types of molecules trigger the nuclear factor κB (NF-κB), and mutations in the CARD or pyrin domain have been associated with inflammatory disorders[260]. The NOD2 protein is exclusively expressed in the gut by several types of cell such as monocytes, macrophages, dendritic cells, epithelial cell, as well as terminal ileum Paneth cells[261]. The NOD2 protein is composed of two CARD N-terminal domains (residues 28-220), a central nucleotide--binding domain for oligomerisation, NBD, (residues 273-577) and a C-terminal leucine-rich domain (residues 744-1020). CARD and NBD domains are responsible for protein-protein interactions and the activation of the NF-κB pathway whereas the leucine rich domain, as for the other member of the NLR family, interacts with the bacterial peptidoglycans[262] (Figure 3.1).

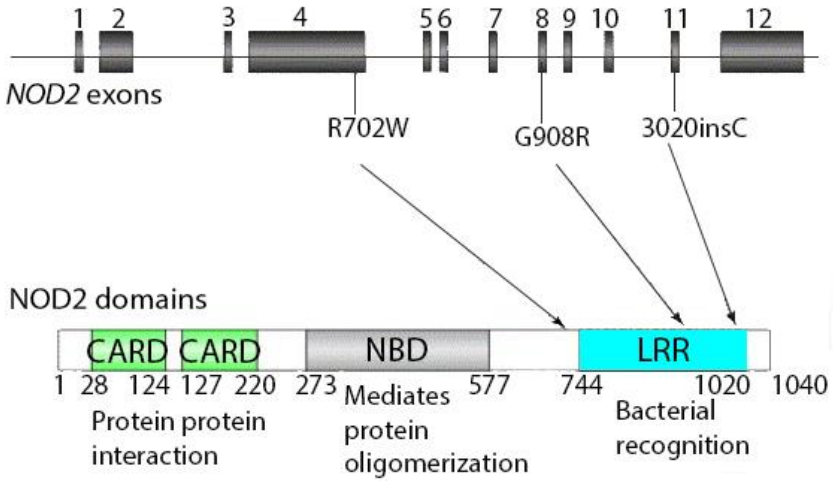



**Figure 3.1: Structure of the NOD2 gene and protein.** The figure shows predicted functional domains and locations of variants associated with CD.

NOD2 and NOD1 proteins are highly conserved cytoplasmatic receptors that sense microbial effectors. These proteins are an active molecule of the immune system as they are an intracellular sensor of the bacterial cell wall protein peptidoglycan, which is presented in both Gram-positive and Gram-negative bacteria[263]. NOD2 specifically recognises muramyl dipeptide, a product of bacteria cell wall degradation, which is hypothesised to enter the cell via phagocytosis or through other uptake mechanisms[213]. NOD2 has multiple roles in response to the pathogen recognition: it activates the NF-κB and MAPK (mitogen-activated protein kinase) pathways, and it regulates the expression of bactericidal peptides such as cryptdins and defensins in the gut flora[259]. Once NOD2 binds to the ligand and oligomerises (Figure 3.2), the conformational change in the protein enables association with RICK/RIP2 proteins

through its CARD domains. RIP2, a member of the receptor-interacting protein (RIP) family of kinases, mediates the ubiquitination of the protein NEMO/IKKγ leading to the activation of NF-κB and the production of inflammatory cytokines. The transcription factor NF-κB binds to inhibitory proteins called inhibitors of κB (IκBs) in the cytoplasm. The phosphorylation of the inhibitor molecule mediated by NEMO/IKKγ leads to the translocation of NF-κB into the nucleus where gene transcription is induced[264]. Moreover, the protein kinase RIP2 recruits the mitogen-activated protein kinase 7 MAP3K7/TAK1 which leads to the activation of the MAPK pathway[265].

Although more than 30 genetic variations have been detected within *NOD2*, only three have confirmed associations with CD and are present in approximately 40% of affected patients[266]: two missense mutations R702W and G908R and a frameshift insertion Leu1007fiC[266]. All three mutations are found within or adjacent to the LRR domain, and corrupt the protein function with consequent reductions in NF-κB activation and chemokine production[262]. The frameshift mutation Leu1007fiC (3020InsC) causes a truncation of 33 amino acids in the LRR domain leading to a loss of the entire function of the whole protein[109]. As frameshift mutations affect all amino acids downstream of the mutation, they are likely to be highly deleterious to the gene's function. The other two single nucleotide mutations are 2104 T>C, which leads to substitution of the arginine at position 702 with a tryptophan (R702W), and 2722 G>C, where the glycine at position 908 is changed to an arginine (G908R)[108,266]. These two SNPs alter the protein conformation and therefore interfere with the binding of the muramyl dipeptide[259]. A meta-analysis conducted by Economou and collaborators assessed the impact of these three variations across several populations[266]. The analysis included forty-two studies and showed that the strongest association is found with the Leu1007finC mutation. Approximately 8-17% of CD patients have one of these three mutations in both their copies of *NOD2*: heterozygous carriers for one of the *NOD2* mutations have a 2-4 fold increased risk of CD, while homozygous carriers increase the risk by 20-40 fold[267,268]. In addition to CD, *NOD2* mutations are risk factors for other syndromes, such as Blau Syndrome, an autosomal disorder characterised by familial granulomatous arthritis, uveitis, and skin granulomas[269]. Unlike in CD, the *NOD2* mutations that confer susceptibility to Blau Syndrome are found in the NOD encoding region[270]. *NOD2* is also associated with several types of cancer, including carcinoma of the colon[271], early-onset breast cancer[272] and leukaemia[273].

In this study, we hypothesis that rare and private genetic variation across genes involved in the NOD signaling pathway may contribute to childhood onset IBD. We interrogate WES data to extract all genetic variation across the frequency spectrum in a pIBD cohort and evaluate the joint effect of rare and common variants with a gene-based statistical test (SKAT-O[274]). We further validate our findings in an independent cohort.

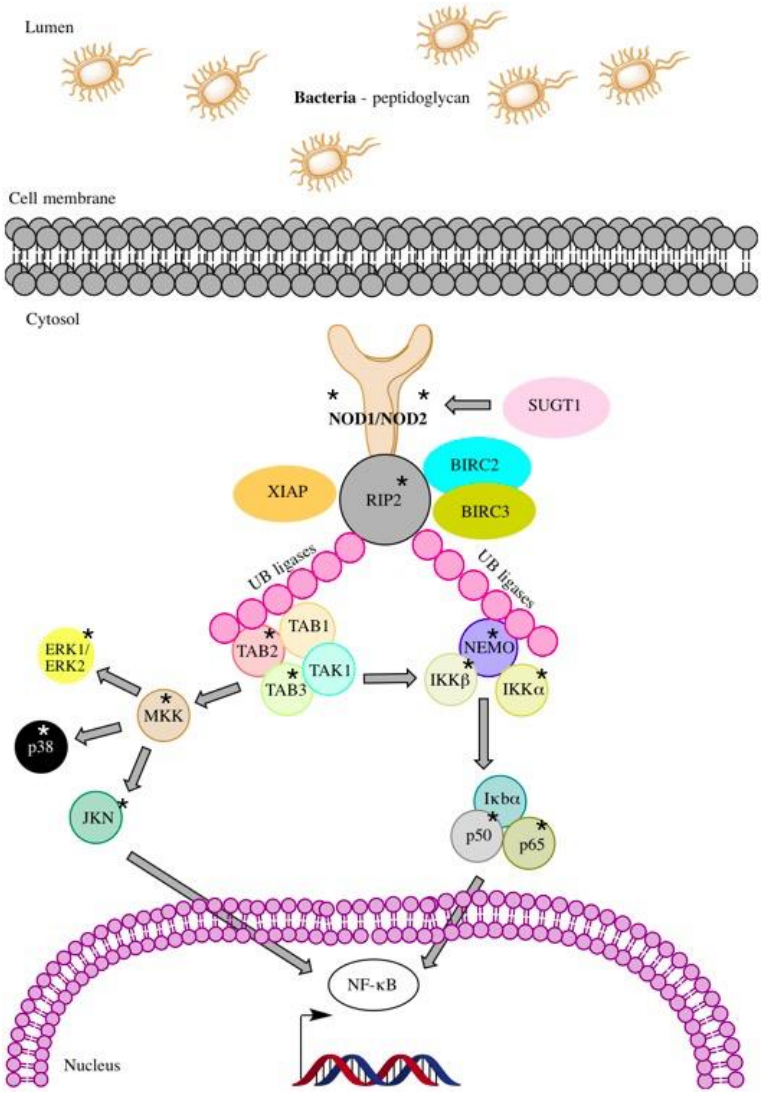



**Figure 3.2 Proteins acting within the NOD signaling pathway**. The recognition of NOD1 and NOD2 of the bacterial peptidoglycan (PGN) promotes the formation of the multi-protein complex the inflammasome. The complex recruits the kinase receptor interacting protein 2 (RIP2), which is ubiquitinated by the ubiquitin ligases XIAP, BIRC2 and BIRC3 proteins. The polyubiquitinated RIP2 recruits the TAK binding protein 1 (TAB1), TAB2 and TAB3 and the kinase TAK1 which leads to the activation of the MAPK kinases, p38 and c-Jun N-terminal kinase (JNK) through the activation of mitogen-activated protein kinase kinase (MKK). RIP2 polyubiquitinated also interacts with the IκB kinase (IKK) complex (IKKα, IKKβ and NEMO). The IKK complex mediates the phosphorylation of the IKKβ subunit of IKK by TAK1 and results in the phosphorylation and degradation of the NF-κB inhibitor (IκBα) which results in the cytoplasmic release and translocation of NF-κB dimers p65 and p50 in the nucleus to activate of the expression of the NF-κB proinflammatory genes. Proteins with asterisks indicate coding genes previously identified by GWAS

## 3.2. Methods

### 3.2.1 Cases and samples

For the discovery cohort, patients were recruited through paediatric gastroenterology clinics at University Hospital Southampton (UHS), as described in paragraph 2.2.1. Whole-exome sequencing data were available for 146 independent children diagnosed with IBD within the discovery cohort. Demographic data for the IBD cohort are shown in Table 3.1. We accessed control samples through our local database of germline exome sequence data for 126 unrelated patients with no inflammatory-related disease.

**Table 3.1 Patient demographics for the cohort of 146 paediatric IBD patients that underwent whole-exome sequencing**

| | CD | UC | IBDU |
|---|---|---|---|
| n | 90 | 32 | 24 |
| Male (%) | 57 (63.3) | 18 (56.25) | 8 (33.3) |
| Mean age in years (range) | 11.25(2-17) | 9.97(1-15) | 11.17(2-16) |

CD, Crohn's disease; UC, ulcerative colitis; IBDU, inflammatory bowel disease unclassified

We used an independent replication cohort derived from the Children's Mercy Kansas City IBD cohort and the Critical Assessment of Genome Interpretation (CAGI, 2013)[275] dataset to validate significant results from the discovery cohort. CAGI (https://genomeinterpretation.org/) is a community of blinded predicted experiments aiming to predict phenotypes from human variation. For every challenge, the community provides confidential genetic data (e.g. experimental data, WES and WGS data) and phenotypes information.

The Children's Mercy Kansas City cohort consists of 13 independent IBD patients and 1 control while the CAGI dataset is composed of 20 unrelated adult CD patients and 8 healthy controls. We merged 102 additional control samples of British ethnicity from the 1KG phase3 dataset resulting in the retention of 33 unrelated cases and 111 independent controls for subsequent analysis in the validation cohort.

### 3.2.2 Discovery cohort DNA extraction

Genomic DNA for each of the Southampton patients undergoing exome sequencing was extracted as described in paragraph 2.2.2.

### 3.2.3 Whole-exome sequencing data generation and analysis

For the discovery and Children's Mercy Kansas City cohort, whole-exome capture was performed using Agilent SureSelect Human all Exon 51 Mb (versions 4 and 5) capture kits and TruSeq Expanded Exome and Nextera Expanded Exome capture kits. Capture technology is characterized by rapid progress, including new content and improved probe design, and we applied the optimal capture chemistry available at the time of sample sequencing. All samples were sequenced on the Illumina HiSeq 2000 and HiSeq 2500 platforms. As previously described[276](paragraph 2.2.3), fastQ raw data generated from Illumina paired-end sequencing protocol were aligned against the human genome reference 19 using Novoalign (2.08.02). SAMtools mpileup tool (samtools/0.1.19) was used to call SNPs and short indels from the alignment file. Variants were excluded if they had a PHRED quality score of <20 and/or a depth of <4. ANNOVAR (annovar/2013Feb21) was applied for variant annotation against a database of RefSeq transcripts. A bespoke script was used to assign individual variants as: "novel" if they were not previously reported in the dbSNP137 databases, 1000 Genomes Project (1KG) and the Exome Variant Server (EVS) of European Americans of the NHLI-ESP project with 6500 exomes, or in the Southampton database of reference exomes. Resultant variant call files for each individual were subjected to further in-house quality control tests to detect DNA sample contamination and ensure sex concordance by assessing autosomal and X chromosome heterozygosity. Variant sharing between all pairs of individuals was assessed to confirm that subjects were not related. Sample provenance was confirmed by application of a validated SNP tracking panel developed specifically for exome data[221].

For the CAGI subgroup of the replication cohort, whole-exome sequencing was performed using the TruSeq capture kit and sequenced on Illumina platforms. Alignment against the human genome (hg19) was conducted with BWA. PICARD was used to remove duplicate reads and GATK for genotype calling.[275]

### 3.2.4 Gene selection

Genes involved in the NOD receptor pathway were extracted by interrogating the KEGG Pathway database.[229] This pathway (KEGG ID: hsa04621) is composed of 56 genes, of which 41 are intrinsic to NOD signaling. Across these 41 genes, 25 and 39

genes had coverage greater than 50% within the Agilent V4 and V5 capture kit respectively. Therefore, the poor exome capture kit coverage has influenced our ability to detect variation within these genes Eleven of the 41 genes have been previously identified through GWA studies (Table 3.2). Gene names were cross-referenced with the HUGO webserver to confirm the approved gene symbol. (Table 3.2). All good quality (Depth ≥ 4 and Phred ≥ 20) variants within these genes were extracted using local scripts and retained for analyses. SKAT-O statistical test was performed on the 41 genes directly involved in the *NOD1* and *NOD2* signaling cascade.

**Table 3.2 Percentage of gene coverage for each of the 40 genes involved in the NOD2 pathway according to the Agilent SureSelect V4 and Agilent SureSelect V5 all Human exome capture kits.**

| Gene | Protein | Agilent v4 % gene coverage | Agilent V5 % gene coverage |
|---|---|---|---|
| *BIRC2* | BIRC2 (cIAP1) | 67.61 | 86.71 |
| *BIRC3* | BIRC3 (cIAP2) | 38.88 | 98.45 |
| *CARD6* | CARD6 | 82.43 | 99.85 |
| *CARD9* | CARD9 | 93.58 | 93.58 |
| *CASP8* | CASP8 | 82.56 | 84.88 |
| *CCL2** | CCL2 | 75.64 | 100.00 |
| *CCL5* | CCL5 | 45.00 | 80.39 |
| *CHUK* | IKKα | 72.19 | 100.00 |
| *CXCL1** | CXCL1 | 40.58 | 97.04 |
| *CXCL2** | CXCL2 | 38.90 | 100.00 |
| *ERBB2IP* | LAP2 | 64.45 | 95.90 |
| *IKBKB* | IKKβ | 76.52 | 88.15 |
| *IKBKG* | NEMO | 15.16 | 77.39 |
| *IL6* | IL6 | 83.11 | 100.00 |
| *IL8** | IL8 | 43.64 | 100.00 |
| *MAP3K7* | M3K7 | 43.22 | 98.78 |
| *MAPK1** | MK01 | 26.15 | 26.15 |
| *MAPK10* | MK10 | 30.51 | 86.73 |
| *MAPK11* | MK11 | 57.24 | 100.00 |
| *MAPK12* | MK12 | 85.12 | 100.00 |
| *MAPK13* | MK13 | 23.00 | 30.28 |
| *MAPK14* | MK14 | 39.65 | 100.00 |
| *MAPK3* | MK03 | 73.08 | 73.08 |
| *MAPK8* | MK08 | 100.00 | 100.00 |
| *MAPK9* | MK09 | 25.89 | 97.46 |
| *NFKB1** | NFKB1/p50 | 75.48 | 88.84 |
| *NFKBIA* | IKBα | 85.19 | 100.00 |
| *NFKBIB* | IKBB | 68.85 | 71.69 |

| | | | |
|---|---|---|---|
| *NOD1* | NOD1 | 71.09 | 91.06 |
| *NOD2** | NOD2 | 77.79 | 100.00 |
| *RELA** | TF65 | 61.86 | 100.00 |
| *RIPK2** | RIP2 | 73.81 | 100.00 |
| *SUGT1* | SUGT1 | 73.74 | 98.32 |
| *TAB1** | TAB1 | 54.17 | 95.55 |
| *TAB2* | TAB2 | 53.90 | 92.45 |
| *TAB3* | TAB3 | 43.80 | 91.56 |
| *TNF* | TNFα | 9.72 | 12.84 |
| *TNFAIP3** | TNAP3 | 47.21 | 81.53 |
| *TRAF6* | TRAF6 | 22.58 | 55.42 |
| *TRIP6* | TRIP6 | 100.00 | 100.00 |
| *XIAP* | XIAP | 15.50 | 92.70 |

Genes with asterisks indicate genes previously identified by GWAS.

### 3.2.5 Principle component analysis

Population stratification can represent a source of bias in association studies as genotype differences between case and control group could be due to different ancestry rather than the effect on disease susceptibility. The oversampling of individuals of one population for cases in an association studies might lead to spurious associations. In order to minimize bias for association analysis, we conducted a principle component analysis (PCA) using the SNPRelate[277] package in order to discriminate ethnic clusters. The PCA was conducted on the whole discovery dataset merged with the 1,092 subjects from the 1KG phase 1 dataset (20101123). PCA was applied to 1363 samples with 305,950 biallelic SNPs.

The same PCA procedure was conducted on the CAGI and 1KG data (209,029 biallelic SNPs across 1158 samples) and on the Kansas and 1KG data (224, 786 biallelic SNP across 1134 samples) to discriminate ethnic clusters.

### 3.2.6 Variant calling and quality control

Next generation sequencing pipelines typically identify genomic locations at which any given sample *differs* from the human genome reference sequence on a case-by-case basis. After compiling the list of all variants identified in all cases and controls it was necessary to positively re-call the genotypic state (for the full set of all variants from all samples) in order to distinguish allelic genotypic status from missing data for each individual. As this information is not given by the standard in-house pipeline for exome

data, customised scripts using samtools[181], vcftools[184] and bedtools[219] packages were written and used to retrieve the genotypes. Figure 3.3 summarises the steps involved in the analysis.

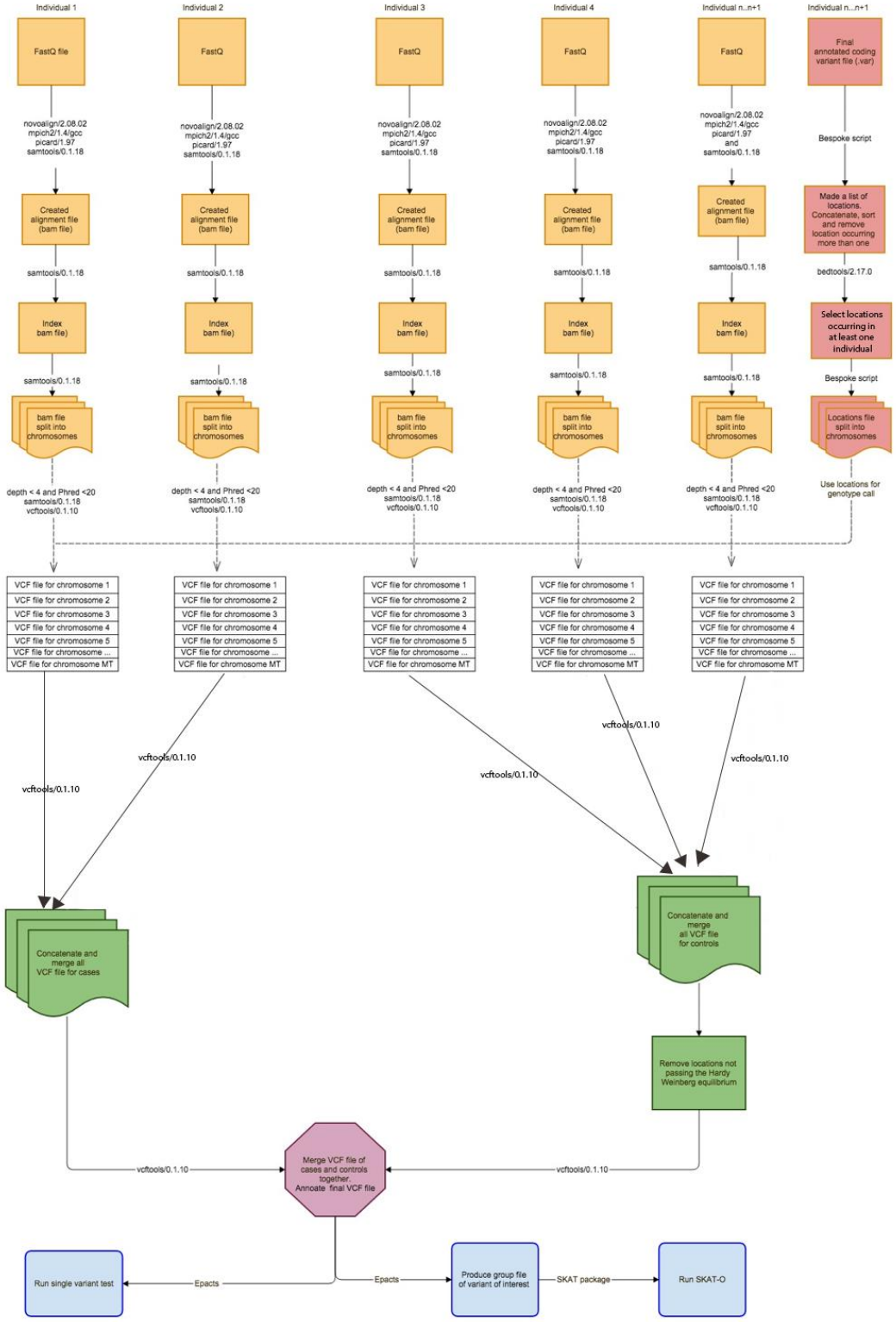



**Figure 3.3 Steps in running rare variants association tests.** From the fastQ file the alignment (bam) files were created for each sample using novoalign/2.08.02, mpich2/1.4/gcc, picard/1.97 and samtools/0.1.18. Bam files were then indexed and split into chromosomes using samtools/0.1.18. Red squares on the right indicate steps to create a unique list of non-redundant variants from the entire cohort. The variant list was used to create VCF files for each individual for each chromosome using

samtools/0.1.18. Variants in the VCF file were filtered out if presenting depth < 4 and phred <20. VCF files were concatenated in two files for cases and controls respectively using vcftools/0.1.10 (green squares). Variants not passing the Hardy-Weinberg equilibrium test on control samples only were removed using vcftools/0.1.10. Case and control VCF files were merged together with vcftools/0.1.10 (pink octagon); to conduct the single variant and SKAT-O test (blue boxes).

The first step consisted of recreating the alignment files for each of cases and controls of the discovery and validation cohort. The alignment files for cases and controls were firstly indexed and then divided into chromosomes to execute in parallel in a more efficient way. The second step was to make a unique list of all the genetic locations in which at least one individual harbour a variant. The third step consisted of creating the variant-calling (VCF) file with all the genotype information using the list of variant locations for each individual. The exclusion of markers deviating from the HWE is an important step in association studies as this can help prevent the inclusion of calling error and of loci associated with disease observed in controls[278]. Therefore, variations were further excluded if they deviated significantly from the Hardy-Weinberg equilibrium status in the control group, by using vcftools[184] (genotypes out of Hardy-Weinberg equilibrium, default value of $p<0.001$, were excluded). In the fourth step, VCF files for cases and controls were merged together and annotated. The fifth step was the construction of a group file where gene and variations of interest were specified.

To detect association between genetic variant and disease status, the sequence kernel association optimal unified test[274] (SKAT-O) was performed using the EPACTS software package[225] in the discovery cohort. SKAT-O test was further conducted on the replication cohort to validate significant results from the discovery cohort.

### 3.2.7 Burden of mutation association testing in the discovery cohort

SKAT-O test was applied to further investigate the joint effect of rare and low frequency variants. Specifically, SKAT-O encompasses both a burden test and a SKAT test to offer a powerful means of conducting association analyses on combined rare and common variation as single variant tests are often underpowered due to the large sample size needed to detect a significant association, as described in paragraph 2.2.9.

To conduct the test, a group file containing all mutations of interest (synonymous, non-synonymous, splicing, frameshifts and non-frameshifts, stop gain and stop loss) was created for each of the 41 genes. SKAT-O was executed using the small sample

adjustment, by applying a MAF threshold of 0.05 to define rare variations within the sample size and using default weights.[274]

### 3.2.8 Burden of mutation testing in the validation cohort

As the validation cohort comprises whole-exome and whole-genome subjects, only variants falling within the consensus target region were considered. By limiting variants assessed to only those found in the genomic regions captured by both technologies, we limited the potential for bias when using data from two different capture technologies. A bespoke script using Bedtools was used to select only locations covered by both sequencing technologies. Variant sites across the four genes requiring replication were used to generate a subset of the VCF file for each dataset. Ultimately, VCF files for all individuals were merged and annotated. SKAT-O testing was conducted using the same settings applied in the discovery cohort.

SKAT-O testing was further conducted using the same approach on the combined discovery and replication cohorts ($n_{cases}$ = 169 and $n_{controls}$ = 217).

## 3.3 Results

PCA removed 10 cases and 20 controls reducing the final number of cases to 136 and controls to 106 within the discovery cohort (Figure 3.4). No individuals were removed from the replication cohort (Appendix III and IV).

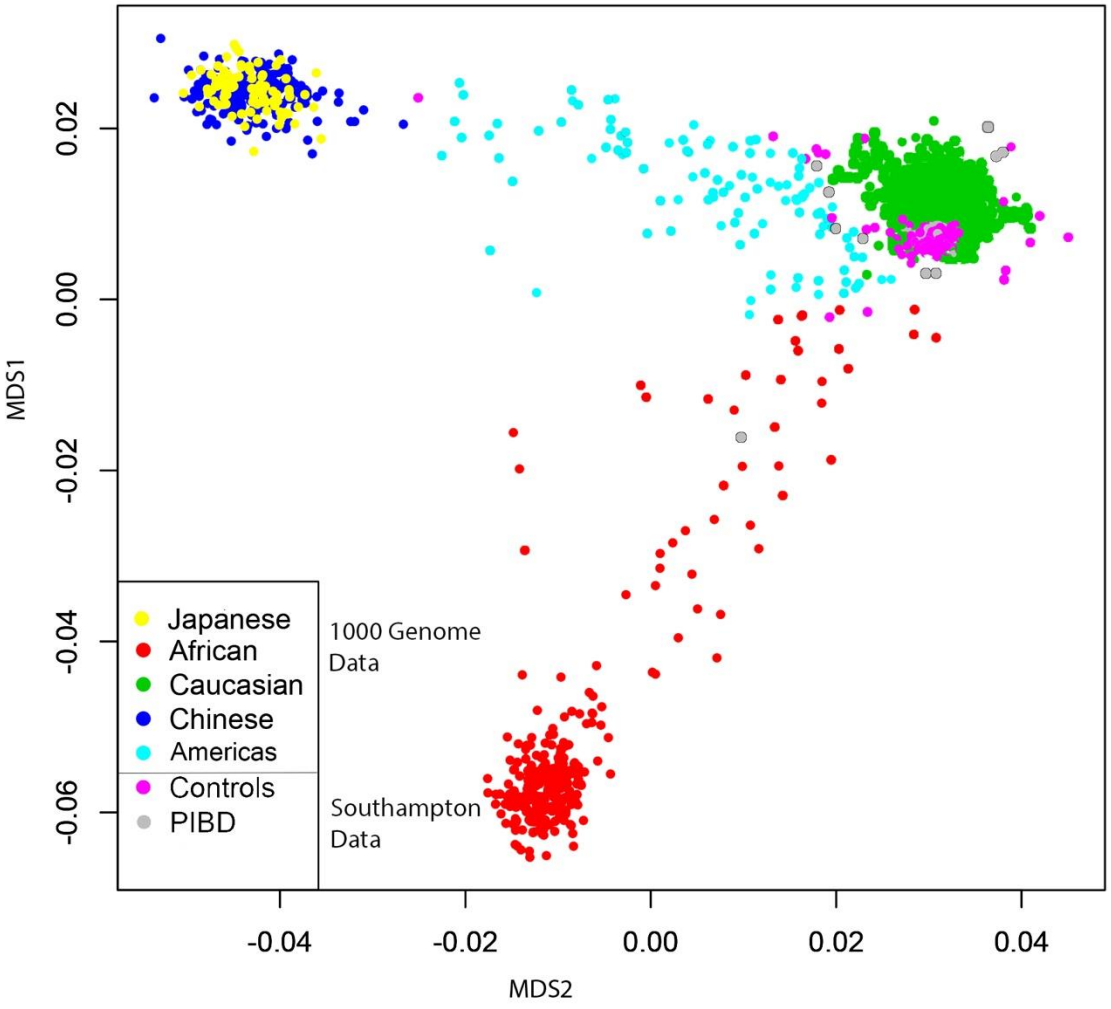



**Figure 3.4 Principle component analysis (PCA) across five ethnic groups from 1000 genome project and the discovery cohort (146 paediatric IBD cases and 126 non-IBD controls).** The five ethnic groups from 1000 Genome are colored as indicated. The Southampton IBD cohort and controls are in pink and light blue respectively. Southampton IBD samples excluded from the SKAT-O test because of ethnic status are represented with a black outline.

Mutations were identified in either cases and/or controls in all but one gene (*CCL5*) from the NOD signaling pathway in the discovery cohort ($n_{cases}$ = 136 and $n_{controls}$ = 106). A total of 250 variants (Appendix V) that occurred in at least one individual (either case or control) across 41 genes were called in order to extract and create the VCF file for

all 242 individuals. We observed 67 novel variants, 94 rare variants with a $MAF_{1KG}$ < 1%, 41 low frequency mutations (1% ≤ $MAF_{1KG}$ ≥ 5%) and 48 common mutations ($MAF_{1KG}$ > 5%). A total of 146 (of 250) mutations occurred with higher frequency in cases compared to controls suggesting a possible deleterious effect in increasing disease risk in susceptible individuals. Although within the same genes presented the same frequency in case and control group, we observed variations in the same gene with high allelic frequency in cases or in controls suggesting both protective and risk effect.

### 3.3.1 Variants within the *NOD2* gene

Across 126 pIBD cases and 85 controls of the discovery cohort, we observed 31 mutations within the 12 exons of *NOD2.* Of these, 26 had a MAF < 0.05 within the sample size (Table 3.3). These mutations were identified in seven different exons of the gene (Figure 3.5). In addition to the known IBD biomarkers, Arg702Trp, Gly908Arg and Leu1007fsinsC[108,109], we observed two novel variants, 20 rare ($MAF_{1KG}$ < 0.01), two low frequency (0.01≤ $MAF_{1KG}$ ≤ 0.05) and four common mutations ($MAF_{1KG}$ > 0.05) (Table 4). Ten of the 26 mutations were annotated as deleterious by SIFT and 13 of them are described in HGMD as pathogenic.[279] Twenty six (out of 31) mutations observed would not have been assessed in any GWAS due to their rarity.

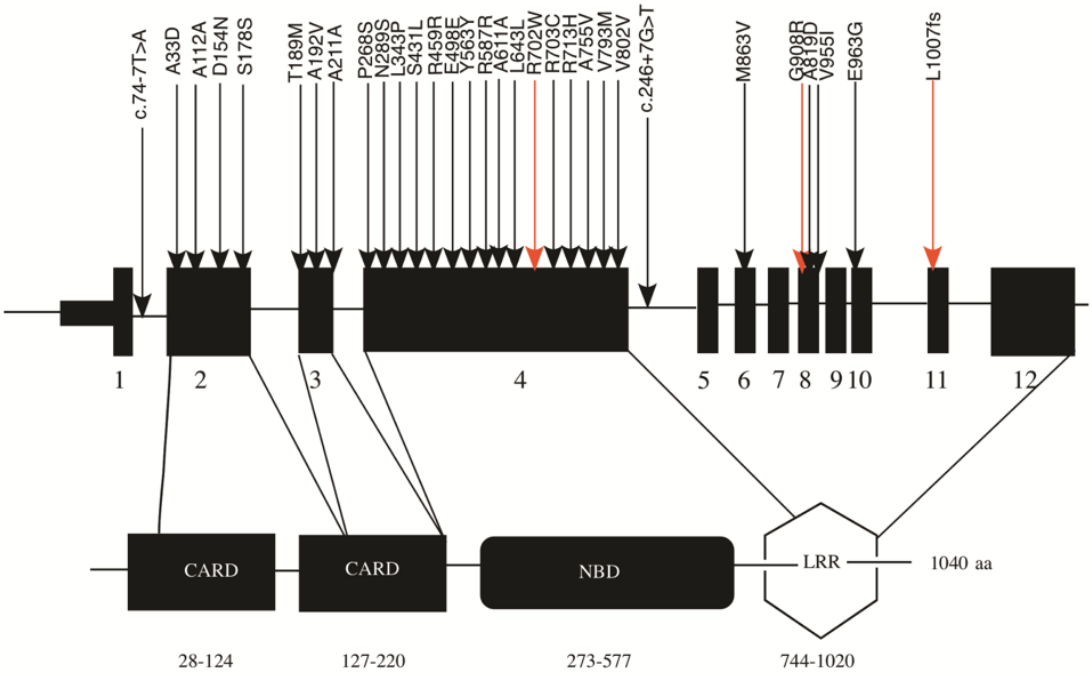



**Figure 3.5 *NOD2* gene and protein structures** *NOD2* is a gene composed of 12 exons (black rectangles). The NOD2 protein consists of two N-terminal CARD (caspase activation recruitment) domains, a central NBD (nucleotide-binding oligomerization) domain and a terminal sequence rich in leucine. The CARD domains interact with RIP2 protein to activate the immune response in the gut and the leucine-rich domain recognizes the bacterial peptidoglycan. Mutations within the NBD have been shown to increase the inflammatory cascade[263]. The 31 mutations observed by interrogating exome data from 136 pIBD and 106 controls are indicated with arrows. Known IBD biomarkers are in red.

**Table 3.3. List of 31 *NOD2* variants observed across the discovery cohort**

| Bp position (hg19) | Variant | Coding change | Protein change | 1-SIFT | Gerp | Maxent score | dbSNP | Frequency in 1KG | Frequency in NHLBI ESP | HGMD | Frequency in cases (n=136) | | | Frequency in controls (n=106) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | Homozygous Reference | Heterozygous | Homozygous Alternative | Homozygous Reference | Heterozygous | Homozygous Alternative |
| 50733392 | sp | c.74-7T>A | . | . | . | 1.83 | rs104895421 | 0.0014 | 0.001861 | listed | 1 | 0 | 0 | 0.99 | 0.01 | 0 |
| 50733423 | ns | c.98C>A | p.A33D | T | 1.13 | . | . | 0.000008 | . | not listed | 0.99 | 0.01 | 0 | 1 | 0 | 0 |
| 50733661 | sn | c.336C>T | p.A112A | . | . | . | . | 0.00002 | . | not listed | 1 | 0 | 0 | 0.99 | 0.01 | 0 |
| 50733785 | ns | c.460G>A | p.D154N | T | 0.958 | . | rs146054564 | . | 0.002093 | not listed | 1 | 0 | 0 | 0.99 | 0.01 | 0 |
| 50733859 | sn | c.534C>G | p.S178S | . | . | . | rs2067085 | 0.26 | 0.409302 | not listed | 0.39 | 0.54 | 0.07 | 0.47 | 0.38 | 0.15 |
| 50741791 | ns | c.566C>T | p.T189M | T | 3.48 | . | rs61755182 | 0.0014 | 0.004419 | listed | 0.99 | 0.01 | 0 | 1 | 0 | 0 |
| 50741800 | ns | c.575C>T | p.A192V | D | 0.916 | . | rs149071116 | 0.00004 | . | not listed | 1 | 0 | 0 | 0.99 | 0.01 | 0 |
| 50741858 | sn | c.633C>T | p.A211A | . | . | . | rs5743269 | 0.0009 | 0.001744 | not listed | 1 | 0 | 0 | 0.99 | 0.01 | 0 |
| 50744624 | ns | c.802C>T | p.P268S | T | -9.98 | . | rs2066842 | 0.12 | 0.26907 | not listed | 0.43 | 0.47 | 0.1 | 0.56 | 0.35 | 0.09 |
| 50744688 | ns | c.866A>G | p.N289S | D | 4.56 | . | rs5743271 | 0.01 | 0.006279 | listed | 0.99 | 0.01 | 0 | 0.98 | 0.02 | 0 |
| 50744850 | ns | c.1028T>C | p.L343P | D | 5.4 | . | . | . | 0.000116 | not listed | 1 | 0 | 0 | 0.99 | 0.01 | 0 |
| 50745114 | ns | c.1292C>T | p.S431L | D | 3.64 | . | rs104895431 | 0.0005 | 0.001395 | listed | 0.99 | 0.01 | 0 | 1 | 0 | 0 |
| 50745199 | sn | c.1377C>T | p.R459R | . | . | . | rs2066843 | 0.13 | 0.270993 | not listed | 0.42 | 0.48 | 0.1 | 0.5 | 0.4 | 0.1 |
| 50745316 | sn | c.1494A>G | p.E498E | . | . | . | . | . | . | not listed | 1 | 0 | 0 | 0.99 | 0.01 | 0 |
| 50745511 | sn | c.1689C>T | p.Y563Y | . | . | . | rs111608429 | 0.0005 | . | not listed | 0.99 | 0.01 | 0 | 1 | 0 | 0 |
| 50745583 | sn | c.1761T>G | p.R587R | . | . | . | rs1861759 | 0.25 | 0.402558 | not listed | 0.4 | 0.54 | 0.07 | 0.47 | 0.39 | 0.14 |
| 50745655 | sn | c.1833C>T | p.A611A | . | . | . | rs61736932 | 0.0046 | 0.010698 | not listed | 0.99 | 0.01 | 0 | 1 | 0 | 0 |
| 50745751 | sn | c.1929C>T | p.L643L | . | . | . | . | 0.000008 | . | not listed | 1 | 0 | 0 | 0.99 | 0.01 | 0 |
| 50745926 | ns | c.2104C>T | p.R702W | D | 2.42 | . | rs2066844 | 0.02 | 0.043488 | listed | 0.88 | 0.1 | 0.01 | 0.87 | 0.13 | 0 |
| 50745929 | ns | c.2107C>T | p.R703C | D | 2.89 | . | rs5743277 | 0.0023 | 0.006977 | listed | 0.98 | 0.02 | 0 | 1 | 0 | 0 |
| 50745960 | ns | c.2138G>A | p.R713H | T | 4.13 | . | rs104895483 | . | 0.000233 | listed | 0.99 | 0.01 | 0 | 1 | 0 | 0 |
| 50746086 | ns | c.2264C>T | p.A755V | D | 5.12 | . | rs61747625 | 0.0005 | 0.004651 | listed | 0.99 | 0.01 | 0 | 1 | 0 | 0 |
| 50746199 | ns | c.2377G>A | p.V793M | D | 3.51 | . | rs104895444 | 0.0005 | 0.001628 | listed | 0.99 | 0.01 | 0 | 0.98 | 0.02 | 0 |
| 50746228 | sn | c.2406G>T | p.V802V | . | . | . | rs104895495 | . | 0.00186 | not listed | 0.99 | 0.01 | 0 | 0.99 | 0.01 | 0 |
| 50746291 | sp | c.2462+7G>T | . | . | . | 0.83 | rs202111813 | 0.0005 | 0.000581 | not listed | 0.99 | 0.01 | 0 | 1 | 0 | 0 |
| 50750842 | ns | c.2587A>G | p.M863V | T | -9.48 | . | rs104895447 | . | 0.00186 | listed | 0.99 | 0.01 | 0 | 1 | 0 | 0 |
| 50756540 | ns | c.2722G>C | p.G908R | D | 5.56 | . | rs2066845 | 0.01 | 0.014535 | listed | 0.96 | 0.04 | 0 | 0.96 | 0.04 | 0 |
| 50756571 | ns | c.2753C>A | p.A918D | D | 5.56 | . | rs104895452 | 0.0009 | 0.000814 | listed | 1 | 0 | 0 | 0.99 | 0.01 | 0 |
| 50757276 | ns | c.2863G>A | p.V955I | T | -9.14 | . | rs5743291 | 0.05 | 0.096047 | listed | 0.83 | 0.17 | 0 | 0.81 | 0.19 | 0 |
| 50759405 | ns | c.2888A>G | p.E963G | T | 5.29 | . | . | . | . | not listed | 0.99 | 0.01 | 0 | 1 | 0 | 0 |
| 50763778 | fr | c.3019dupC | p.L1007fs | . | . | . | rs2066847 | 0.006 | . | listed | 0.9 | 0.09 | 0.01 | 0.99 | 0.01 | 0 |

ns, non-synonymous; sn, synonymous; fi, frameshift insertion, fd, frameshift deletion; sp, splicing; nfi, non-frameshift insertion; nfd, non-frameshift deletion, sp, splicing

B, benign; C, Conservative; D, deleterious; MC, moderately Conservative; MR, moderately Radical; NR, not reported; P, possibly damaging; R

### 3.3.2 Gene based burden of mutation testing in the discovery cohort

The gene-based test for assessing the combined association of novel, rare and common mutation with disease status showed significant evidence for association with four genes across the discovery cohort (*BIRC2*, *NFKB1*, *NOD2,* and *SUGT1* see Table 3.4). *NFKB1* (p=0.005) and *NOD2* (p=0.029) are known IBD genes with multiple previous publications implicating their role in IBD. *BIRC2* (p=0.004) and *SUGT1* (p=0.047) represent previously unreported genes.

**Table 3.4 Joint variant test (SKAT-O) result for the 41 genes within the NOD signaling pathway in which variations was found across the entire discovery cohort.**

| Gene | Chr | bp position (hg19) | Total number of samples (136 cases; 106 controls) | Fraction of individuals who carry rare variants under the MAF thresholds (MAF < 0.05)* | Number of all variants defined in the group file | Number of variant defined as rare (MAF < 0.05)* | P-value unadjusted | Weighted (W) or Unweighted (UW) p value |
|---|---|---|---|---|---|---|---|---|
| *BIRC2* | 11 | 102220918-102248410 | 242 | 0.07851 | 6 | 6 | 0.004 | W |
| *NFKB1* | 4 | 103488139-103537672 | 242 | 0.11983 | 10 | 9 | 0.005 | UW |
| *NOD2* | 16 | 50733392-50763778 | 242 | 0.21488 | 31 | 25 | 0.029 | W |
| *SUGT1* | 13 | 53231709-53261936 | 242 | 0.33058 | 6 | 5 | 0.047 | UW |
| *MAPK11* | 22 | 50703796-50706381 | 242 | 0.07024 | 7 | 5 | 0.061 | UW |
| *CARD6* | 5 | 40841561-40853404 | 242 | 0.0909 | 10 | 8 | 0.074 | UW |
| *MAPK8* | 10 | 49609720-49642974 | 242 | 0.01652 | 3 | 3 | 0.075 | W |
| *BIRC3* | 11 | 102195774-102201850 | 242 | 0.05371 | 6 | 6 | 0.075 | UW |
| *IL8* | 4 | 74606393-74607328 | 242 | 0.01239 | 2 | 2 | 0.091 | W |
| *MAPK3* | 16 | 30128224-30134507 | 242 | 0.02066 | 4 | 4 | 0.111 | W |
| *TAB2* | 6 | 149699333-149730846 | 242 | 0.07024 | 6 | 6 | 0.117 | W |
| *MAPK14* | 6 | 36063793-36075286 | 242 | 0.04132 | 4 | 4 | 0.129 | UW |
| *NFKBIB* | 19 | 39395836-39398201 | 242 | 0.04545 | 4 | 4 | 0.238 | UW |
| *IKBKB* | 8 | 42128942-42188489 | 242 | 0.06198 | 6 | 6 | 0.249 | W |
| *ERBB2IP* | 5 | 65307924-65372200 | 242 | 0.17355 | 13 | 9 | 0.292 | W |
| *TNF* | 6 | 31544562-31544562 | 242 | 0.00826 | 1 | 1 | 0.293 | UW |
| *IL6* | 7 | 22767137-22771156 | 242 | 0.08677 | 3 | 3 | 0.313 | UW |
| *MAPK1* | 22 | 22123519-22162126 | 242 | 0.01239 | 4 | 3 | 0.314 | W |
| *CASP8* | 2 | 202122956-202149864 | 242 | 0.0909 | 8 | 6 | 0.319 | W |
| *MAPK12* | 22 | 50691870-50699668 | 242 | 0.19835 | 16 | 13 | 0.35 | W |
| *TAB3* | X | 30849697-30877801 | 242 | 0.02066 | 5 | 4 | 0.362 | W |
| *CHUK* | 10 | 101964267-101980355 | 242 | 0.02892 | 6 | 4 | 0.38 | W |
| *MAPK9* | 5 | 179665354-179676062 | 242 | 0.00826 | 2 | 2 | 0.593 | W |
| *TNFAIP3* | 6 | 138196066-138202378 | 242 | 0.06198 | 7 | 7 | 0.653 | W |
| *NOD1* | 7 | 30487954-30496518 | 242 | 0.08264 | 17 | 13 | 0.657 | UW |
| *MAPK10* | 4 | 86952589-86952590 | 242 | 0.00413 | 2 | 1 | 0.781 | UW |
| *XIAP* | X | 123034511-123040945 | 242 | 0.004132 | 2 | 1 | 0.781 | W |
| *TRIP6* | 7 | 100465128-100469223 | 242 | 0.07438 | 11 | 9 | 0.783 | UW |
| *TRAF6* | 11 | 36514122-36518769 | 242 | 0.00413 | 1 | 1 | 0.79 | UW |
| *TAB1* | 22 | 39795831-39832516 | 242 | 0.02479 | 7 | 6 | 0.795 | W |
| *IKBKG* | X | 153780386-153780386 | 242 | 0.00413 | 1 | 1 | 0.798 | UW |
| *MAPK13* | 6 | 36098410-36107131 | 242 | 0.02892 | 9 | 8 | 0.799 | UW |
| *RIPK2* | 8 | 90770315-90802611 | 242 | 0.07438 | 5 | 4 | 0.803 | W |
| *RELA* | 11 | 65422007-65427183 | 242 | 0.02479 | 4 | 4 | 0.813 | UW |
| *NFKBIA* | 14 | 35872068-35873770 | 242 | 0.01239 | 4 | 2 | 0.841 | UW |
| *CARD9* | 9 | 139258615-139266519 | 242 | 0.16942 | 12 | 10 | 0.866 | UW |
| *CXCL2* | 4 | 74964625-74964830 | 242 | 0.03719 | 2 | 2 | 0.88 | UW |
| *CXCL1* | 4 | 74735244-74736235 | 242 | 0.00826 | 2 | 1 | 1 | UW |
| *MAP3K7* | 6 | 91256978-91266350 | 242 | 0.00413 | 2 | 1 | 1 | UW |
| *CCL2* | 17 | 32583269-32583269 | 242 | 0.00826 | 1 | 0 | 1 | UW |
| *CCL5* | - | | - | - | - | - | Not tested | #N/D |

chr: chromosome; * These variants received different weights in the SKAT-O joint test. Genes are ordered by p-value

### 3.3.3 Replication of the gene based burden of mutation test in the validation cohort

We conducted a replication analysis of *NFKB1*, *BIRC2*, *NOD2* and *SUGT1* in the replication cohort ($n_{cases}$ = 33; $n_{controls}$ = 111). A total of 13 variants were identified across the regions sequenced in all individuals in the replication cohort. No variant was observed in *SUGT1* in the validation cohort and therefore SKAT-O test was not conducted on this gene. Although no power analyses were conducted, SKAT-O test showed statistical association for *BIRC2* (p=0.041) and *NOD2* (p=0.045) and was not powered to detect significant association for *NFKB1* (p=0.223, Table 3.5). Gene based test on the combined discovery and replication cohort ($n_{cases}$ = 169 and $n_{controls}$ = 217) confirmed statistical association for *NOD2* (p=0.011), *NFKB1* (p=0.017) and *BIRC2* (p= 0.030), Table 3.5.

**Table 3.5 SKAT-O test result for the four significant genes within the NOD signaling pathway in which variations was found across the replication cohort only and across the combined discovery and replication cohort.**

| Gene | Dataset | Chromosome | bp position (hg19) | Total number of samples | Fraction of individuals who carry rare variants under the MAF thresholds (MAF < 0.05)* | Number of all variants defined in the group file | Number of variant defined as rare (MAF < 0.05)* | P-value unadjusted |
|---|---|---|---|---|---|---|---|---|
| ***BIRC2*** | Replication cohort (33 cases; 111 controls) | 11 | 102219940-102249151 | 144 | 0.11806 | 3 | 2 | 0.041 |
| | Combined replication and validation cohort (169 cases; 217 controls) | 11 | 11:102248377-102248377 | 386 | 0.04663 | 1 | 1 | 0. 030 |
| ***NOD2*** | Replication cohort (33 cases; 111 controls) | 16 | 50733859-50763778 | 144 | 0.11111 | 8 | 3 | 0.045 |
| | Combined replication and validation cohort (169 cases; 217 controls) | 16 | 50733859-50763778 | 386 | 0.041451 | 4 | 2 | 0.011 |
| ***NFKB1*** | Replication cohort (33 cases; 111 controls) | 4 | 103505961-103514658 | 144 | 0.02777 | 2 | 1 | 0.223 |
| | Combined replication and validation cohort (169 cases; 217 controls) | 4 | 103505961-103514658 | 386 | 0.05699 | 2 | 1 | 0.017 |

No rare variants found in *SUGT1*. * These variants received different weights in the SKAT-O joint test. Genes are ordered by p-value

## 3.4 Discussion

Since 2005 NGS has proven to be an effective technology for the study of rare and low frequency mutations within disease-associated genes.[204] More than 100 types of Mendelian disorders have been studied using WES with a diagnostic rate of success of 25%–30%.[145] This success represents a substantially higher rate than that afforded by classical clinical genetic testing such as karyotyping (<5%) or array comparative genomic hybridization (~15%–20%).[145] The combination of traditional genetic testing and WES/WGS technology has rapidly accelerated the discovery of new disease-associated genes underlying Mendelian traits: from an average of 166 per year between 2005 and 2009 to 236 per year between 2010 and 2014, with the numbers increasing every year. WES/WGS has made gene discovery for all phenotypes feasible and cost effective.[145] The rapid growth and success of the next generation sequencing technologies in Mendelian traits has brought a great interest in their application to complex traits. WES and WGS have enable diagnosis and alternative treatment in patients with monogenic IBD.[160]

In our study we applied WES and the SKAT-O statistical test on a discovery cohort of 242 individuals. We conducted the analysis with no assumption with regard to IBD diagnosis (CD, UC or IBDU) because in half of the families recruited in the study we observed mixed diagnoses reflecting the substantial genetic overlap between IBD subtypes. We targeted our analysis to all genes across the entire NOD signaling pathway. Despite a modest cohort size, we detected significant association in four genes (*NFKB1*, *NOD2*, *SUGT1* and *BIRC2*).

*NFKB1* and *NOD2* are known IBD genes (*NFKB1* with ulcerative colitis and *NOD2* with Crohn's disease). The identification of these known genes provided support of biologically meaningful results within SKAT-O significant genes.

*BIRC2* and *SUGT1* have not been previously associated with this chronic autoimmune condition. The significance of *BIRC2* and *NOD2* was further replicated in an independent cohort of IBD patients.

*NFKB1* encodes a transcription factor that regulates the transcription of genes involved in immune and inflammatory responses, cell growth, and apoptosis.[280] It is activated by a variety of triggers including cytokines and bacterial products through the activation of the toll-like and NOD receptors.[280] The gene lies within the known IBD2 locus[281] and has been repeated identified through association studies as causal gene in

IBD.[107,140,282] Although *NFKB1* does not replicate in our validation cohort as it is too small to detect association between the variation observed in this gene and the phenotype; the combined dataset of validation and discovery cohort showed significant association.

*NOD2* is the earliest gene implicated in IBD pathogenesis and the most strongly associated in association studies with IBD.[130] Polymorphisms within *NOD2* are known to increase the risk of developing CD.[283] *NOD2* patient carriers of one of the three allelic biomarker variants have an increased risk of developing CD: heterozygous carriers have a 2–4-fold increased risk of CD, while homozygous or compound heterozygous carriers have a 20–40-fold increased risk.[130]

Little information is available on the function of *SUGT1*, a ubiquitin ligase-associated protein, which positively regulate NOD receptors activation in order to assembly the "inflammasone" complex.[284] This gene is marginally significant within the discovery cohort and does not replicate in the validation cohort. Further testing are needed in order to fully investigate the role of *SUGT1* in IBD pathogenesis.

*BIRC2* (Figure 3.6) belongs to a gene family encoding three conserved proteins characterized by the presence of 1-3 baculovirus IAP repeat (BIR) motifs.[285] These are *XIAP*, *BIRC2* (also known as cIAP1) and *BIRC3* (also known as cIAP2). *XIAP* is located on the X chromosome while *BIRC2* and *BIRC3* are both localised on chromosome 11.

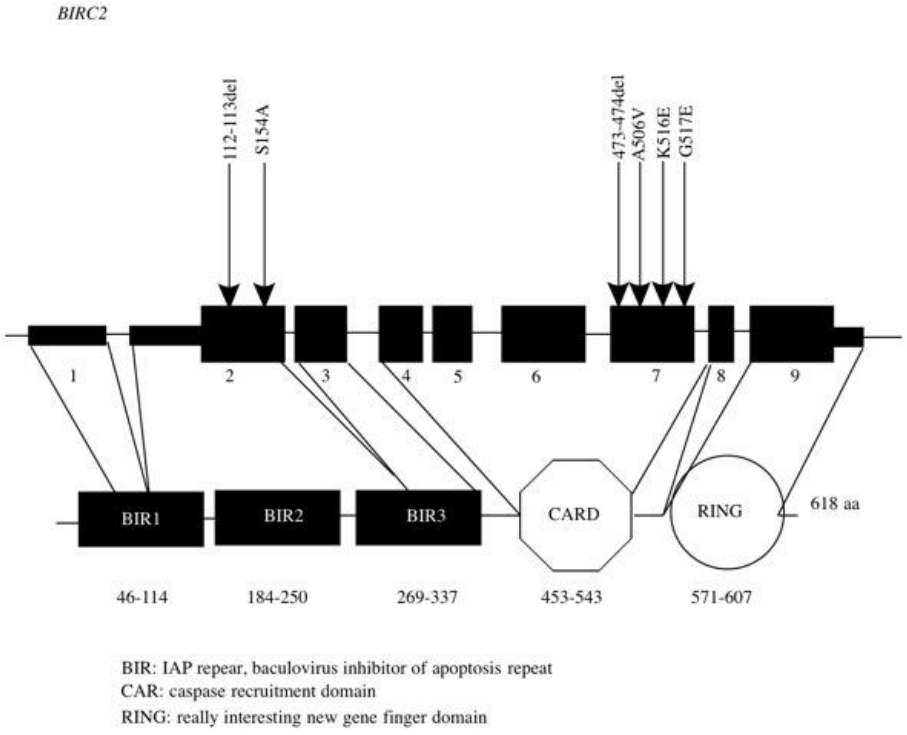



**Figure 3.6 *BIRC2* gene and protein structures.** *BIRC2* is a gene composed of 9 exons (black rectangles). *BIRC2* encodes an inhibitor of apoptosis protein, which contributes to innate immune responses by

acting as inhibitor of cell death downstream of tumour necrosis factor receptor-associated factors TRAF1 and TRAF2. The inhibitor of apoptosis genes share three tandem specific motifs named BIR, belonging to the zinc-finger domain, mediating protein-protein interaction, a CARD domain, involved in protein-protein interaction though CARD-CARD mediated interaction and a C-terminal RING domain, conferring an E3-ubiquitin ligase activity. The RING domain of BIRC2, BIRC3, and XIAP is required for the ubiquitin activity of the IAPs. Studies have reported that the CARD domain in BIRC2 and BIRC3 act as an inhibitor of the ubiquitin ligase activity. Mutations within the BIR1 domain in *BIRC2* alters molecular interaction with TNF receptor associated factor 1 (TRAF1) and TRAF2[286,287]. The six mutations found by interrogating exome data from 136 pIBD and 106 controls are represented by arrows.

Several studies[286,287] have demonstrated the importance of these genes in regulating the expression of proinflammatory cytokines, such as TNF$\alpha$, through NF-$\kappa$B and MAPK pathways primarily through their ubiquitin-ligase activity. *XIAP*, *BIRC2* and *BIRC3* are key players in regulating the NOD1 and NOD2 signaling pathway by directly promoting RIPK2 ubiquitylation and they facilitate activation of NF-$\kappa$B pathway to promote cell survival.[288] Cellular studies on *BIRC2*, *BIRC3* and *XIAP* deficient macrophages were defective for MAPKs and NF-$\kappa$B activation[286,287] This defect in the NOD signaling was also further observed *in vivo* in *BIRC2*, *BIRC3* and *XIAP* knockout murine IBD models.[287] *BIRC2* and *BIRC3* are inhibitors of the Fas signaling cascade in human intestinal cell line.[286] The expression profile of *BIRC3* was further investigated in 14 UC patients indicating an overexpression in colonic specimens during disease flares.[289] Additional studies on the interleukin (IL)-11 expression suggested a possible protective role of IAP, indicating that an over-expression of the IAP proteins could promote healing of the gut.[290] It is therefore feasible that mutations within these genes might impact gut healing and contribute to flares in IBD. Six variants within *BIRC2* were observed in the discovery cohort across 15 cases and 4 controls. Three of these were novel (p.112_113del, p.S154A and p.G517E), two were rare (p.K516E and p.S318S) and one was low frequency (p.A506V,). Across the 15 cases (four with CD, four with IBDU and seven with UC), four were diagnosed aged < 6 years, seven had a positive family history for IBD and nine were diagnosed with a second autoimmune condition other than IBD. While our observed enrichment of variation within *BIRC2* directly implicates this gene in paediatric IBD, further functional analyses are necessary for a comprehensive understanding of the role of individual variants in this protein and their wider impact on the signaling pathway. While mutations in *XIAP* are known to cause up to 4% of male early onset IBD, it is has been postulated that *BIRC2* and *BIRC3* might contribute to IBD pathogenesis by regulating the inflammatory cascade through their

ubiquitin-ligase activity. Although *BIRC2* is not located on the X chromosome and does not appear in a X-linked mode of inheritance, it does not indicate that this gene does not have the same effect on disease as *XIAP*. Our findings are the first to directly implicate this genes in pIBD.[291]

Novel drugs that mimic the natural endogenous inhibitor of the IAP (the mitochondria-derived activator of caspases, SMAC) have been proposed to suppress the pro-inflammatory immune response in the gastro-intestinal tract for patient with moderate to severe disease activity.[292] It is possible that increased application of genetic finding might benefit the patients harbouring these mutations with these novel treatment therapies.

## 3.5 Conclusions

A gene based burden of mutation test for association using sequencing data on a small cohort have supported the involvement of *NFKB1* and *NOD2* in the pathogenesis of IBD and confirmed a role for *BIRC2* in the pathogenesis of disease. This is the first study highlighting the role of *BIRC2* in IBD through targeted exome sequencing.

It is possible that the 15 patients harbouring *BIRC2* mutations may benefit from new treatments targeting the IAP expression and function. Further studies are required to assess the role of targeted therapy in the clinical management of these patients.

# Chapter 4 **De Novo and Rare HSPA1L Mutations for Inflammatory Bowel Disease Revealed by Whole Exome Sequencing**

## 4.0 Summary

The work presented in this chapter deals with the analysis conducted in collaboration with the University of Stanford in California, USA. This analysis aimed to thoroughly investigate the role of the heat shock proteins HSPA1L in IBD pathogenesis**.** Whole exome sequencing was conducted on a proband with IBD and the unaffected first degree relatives from an core family (Stanford) with no prior family history of IBD to investigate the possible genetic contribution. A rare de novo mutation in *HSPA1L* was identified in the affected child that resulted in a non-synonymous amino acid change in a highly conserved nucleotide binding domain of the HSPA1L protein. Subsequently, we comprehensively screened for rare and de novo mutations in *HSPA1L* in our paediatric IBD cohort and found additional rare non-synonymous mutations in highly conserved residues. Since HSPA1L is involved in protein refolding, *in vitro* assays to characterized these mutations were performed. The assays indicated that the found mutations caused varying degrees of altered protein function with some functioning as a dominant negative. Our findings provide evidence of support that variants in *HSPA1L* that impair protein function might contribute to IBD pathogenesis.

My contribution to this work was to comprehensively investigate the role of *HSPA1L* and the two most common form of HSP70, *HSPA1A* and *HSPA1B*, within our pIBD cohort, design of the pipeline and application of appropriate gene based statistical analysis, primer design for segregation analysis, extraction of variation within known IBD genes from the Stanford and Southampton exome data, data interpretation and manuscript write-up.

## 4.1 Background

The human leucocyte antigen (HLA) complex is a highly polymorphic region comprising more than 130 genes on chromosome 6. Linkage and genome-wide studies have consistently shown evidence of association between IBD and the IBD3 locus on chromosome 6, which enclose the HLA complex[293,294]. The HLA region includes four sub

regions: class-I, class-III, class-II and extended class-II regions, which encode for proteins with immunoregulatory function. The class-III region contains multiple genes which encode for cytokine, transcription regulation and protein–protein interactions, signalling and chaperone function (e.g. *HSPA1A*, *HSPA1B*, and *HSPA1L*) [293,294], figure 4.1.

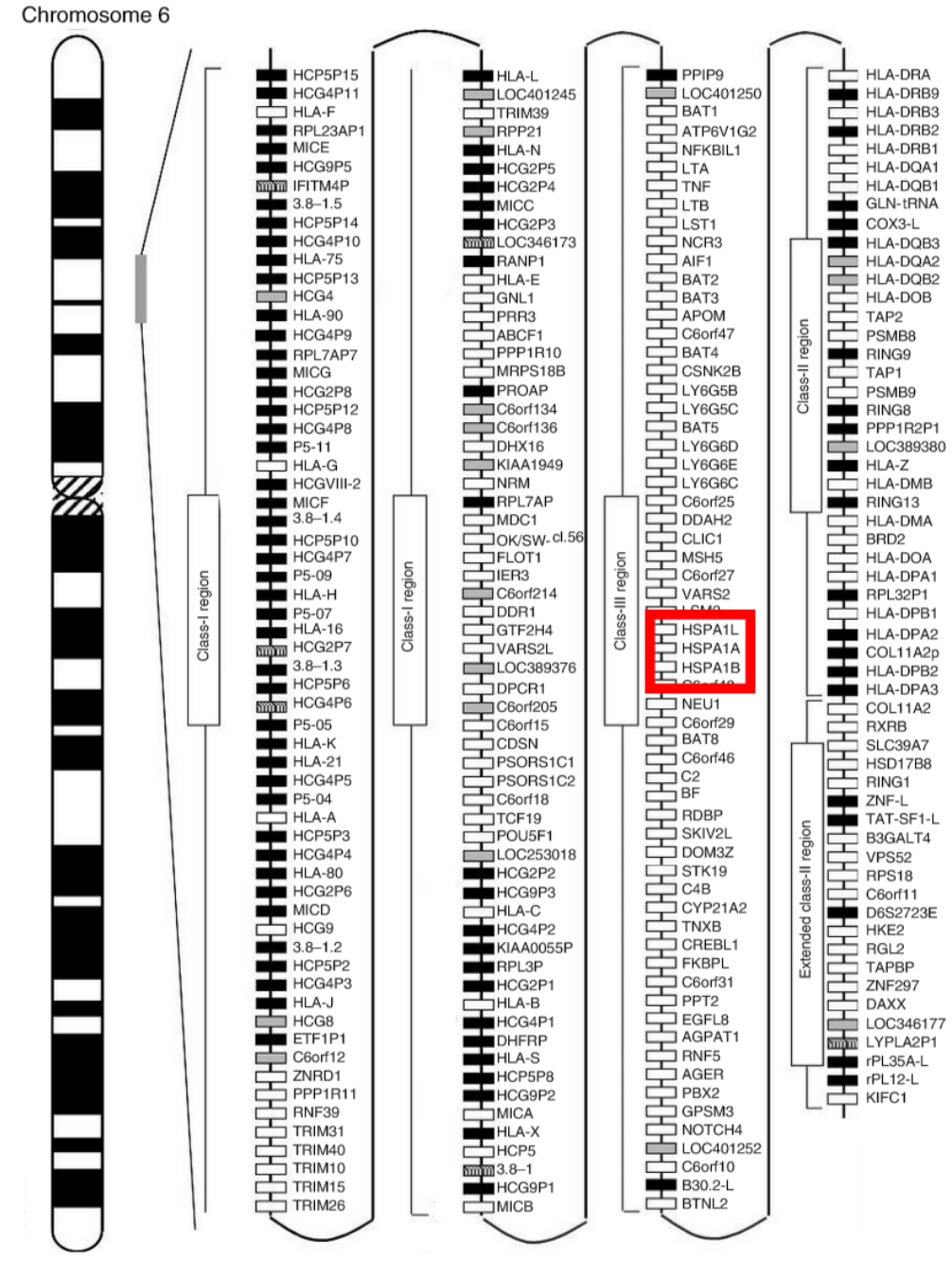



**Figure 4.1 Gene map of the HLA region**. Genes within the HLA sub-regions: classical class-I, class-III, classical class-II and extended class-II regions. White, grey, striped, black boxes show expressed genes, gene candidates, non-coding genes and pseudogenes respectively. Red box indicates location of *HSPA1L*, *HSPA1A* and *HSPA1B*. Modified from Shiina et al[295].

*HSPA1L*, *HSPA1A* and *HSPA1B* are member of the 70-kD heat shock protein family (HSP70) [296]. Heat shock proteins (HSPs) are highly conserved molecular chaperone proteins firstly discovered as a response to an increase in temperature. Chaperonins are a class of specific proteins that use adenosine triphosphate (ATP) energy to assist a polypeptide as it folds into the proper tertiary structure or to degrade unfolded proteins. HSPs are classified into families based on their molecular weights, e.g. the 70 kDa protein family is indicated as HSP70. HSP70 proteins play multiple roles in protein quality control of the cell, including refolding denatured proteins, preventing aggregation, and intracellular protein transport[297].

The transcription factor heat shock factor 1 (HSF1) induces the expression of HSP proteins genes. In stress situations HSF1 activates HSP70 genes and studies have shown that in mice defective for HSF1 there is a lack of HSP70 expression[297]. HSP70s have been shown to be up-regulated in response to injury stimuli, to modulate inflammatory response[296] and to have anti-apoptotic functions[298] by inhibiting apoptosis regulating proteins. All these functions are related to IBD pathogenesis. However, a distinct role for each HSP70 family member is still not well understood and their potential role in IBD has not been fully established yet. In a normal intestinal mucosa HSP70 genes are expressed in the surface of the epithelium of the colon and their function is to act as a resistance to bacterial toxins[299].

It has been demonstrated that HSP70 expression is higher in UC patients and decreases to a level in line with the expression found in healthy controls in UC patient in remission[300].

HSPs inhibit cytokine production mediated by inflammatory genes such as NF-κB and MAPK resulting in a decrease of the inflammatory process[301,302]. This HSPs function makes them a possible therapeutic target[302].

However, the interaction through which HSP proteins suppress inflammatory cytokine is not understood. More research needs to be done to elucidate the actual mechanism of HSPs leading to IBD pathogenesis.

The work presented in this chapter was initiated by the findings of a *de novo* mutation in *HSPA1L* in a core family of Eastern European and Middle-Eastern origins by our collaborators at Stanford. The family was composed of the affected UC proband, both unaffected parents and an unaffected sibling. This finding initiated a comprehensive investigation of HSP70 family in our pIBD cohort to further elucidate role of *HSPA1L*,

*HSPA1A* and *HSPA1B* genes in IBD. This study represents the first report on the association between IBD and *de novo* and rare non-synonymous mutations in the heat shock protein *HSPA1L*, thereby demonstrating a functional role for this protein in IBD and expanding our knowledge of the role of these proteins in human disease.

## 4.2 Methods

### 4.2.1 Cases and samples

Written informed consent was provided by an attending parent or legal guardian for all paediatric recruits. For the Family A, the proband was recruited at the Stanford Hospital, California (USA). She was diagnosed with UC at the age of 16 and there was no family history of UC. For the Soton pIBD cohort, patients were recruited through paediatric gastroenterology clinics at University Hospital Southampton (UHS), as described in paragraph 2.2.1.

### 4.2.2 Whole exome sequencing and data analysis

For Family A, whole exome sequencing and data analysis was performed as previously described[303] by the Stanford group. In brief, whole exome enrichment was performed with the Agilent SureSelect Human All Exon V5+UTRs kit (Agilent Technologies, Santa Clara, CA) and sequenced with the Illumina HiSeq 2000 sequencer (Illumina, San Diego, CA). Paired-end, 101-b short reads generated from each library were mapped to the reference genome hg19 with Burrows-Wheeler Aligner (version 0.7.7), and variants were called with the Genome Analysis Toolkit[304] (GATK; version 1.6). Called SNV and Indel variants were further annotated with ANNOVAR[194] (version 2013Aug23). Potentially damaging Indels were predicted with the SIFT[305] (< 0.05) and PolyPhen-2[306] (> 0.85) algorithms.

For the Soton pIBD cohort, whole-exome capture was performed using Agilent SureSelect Human All Exon 51 Mb (versions 4 and 5) capture kit as previously described in paragraph 2.2.3[34,307]. Summary statistics for each individual are listed in Appendix VI. PICARD (picard/1.97) was used to remove duplicate reads and SAMtools[181] mpileup (samtools/0.1.18) was used to call SNPs and short INDELs from the alignment file. Variants were excluded if they had a PHRED quality score of <20 and/or a depth of <4. ANNOVAR (annovar/2013Feb21)[194] was applied for variant annotation. Following our first process of high quality variant detection, fastQ raw data for the pIBD cohort were further analysed to investigate the contribution of non-uniquely mapped reads. These reads are considered poor quality and usually discarded. However, it is possible that the analysis of these reads might impact

identification of SNPs and INDELs in highly homologous genes such as *HSPA1L*, *HSPA1A* and *HSPA1B*. The raw data generated from paired-end sequencing protocol were re-aligned against hg19 using Novoalign[176] with the option to report all alignment types. PICARD was not used to remove duplicate reads and any SNP and INDEL was retained in the downstream analysis regardless of depth or phred score.

### 4.2.3 Variants in *HSPA1L*, *HSPA1A* and *HSPA1B* across pIBD cohort

Information for all variants called in *HSPA1L*, *HSPA1A* and *HSPA1B* genes was extracted across 136 pIBD patients and 106 controls. Rare ($MAF_{1KG}$ < 0.01), non-synonymous *HSPA1L* mutations were selected and verified by Sanger sequencing in the proband and relatives where applicable (Fig. S1a-g in the Appendix VIII). Primers (Table 4.1) were designed using Primer-BLAST with default settings. Primer-BLAST default settings allow only two mismatches per primer pair and that any primer with at least six mismatches or more should be ignored. Primer melting temperature default setting is set to minimum of 57 °C, a maximum of 63°C with an optimal temperature set at 60 °C. As *HSPA1A* and *HSPA1B* share 98% of sequence similarity, specificity to discriminate *HSPA1A* and *HSPA1B* was ensured using a nested PCR. Specific primers were used to distinguish *HSPA1A* from *HSPA1B* and then a second set of primers was used on each PCR product to amplify the site of the mutation of interest (Figure 4.2). For each sample, PCR was conducted and enriched DNA samples and the samples sent to be sequenced at Source BioScience.

**Table 4.1 Characteristics of primer pairs used for Sanger sequencing**

| | Sequence (5'->3') | Length | Tm | GC% | Product length |
|---|---|---|---|---|---|
| ***HSPA1L* c.A1674T p.E558D** | | | | | |
| Forward primer | ACTGCCCTGATAAAGCGCAA | 20 | 60.32 | 50 | 894 |
| Reverse primer | GGGGCCTAGTTTTCCTGAGTC | 21 | 60.07 | 57.14 | |
| ***HSPA1L* c.515_517del p.172_173del,** | | | | | |
| Forward primer | GCTAAACGTCTGATCGGCAG | 20 | 59.08 | 55 | 575 |
| Reverse primer | CTCACGGCTCGCTTGTTCT | 19 | 60.37 | 57.89 | |
| ***HSPA1L* c.G802A p.A268T & HSPA1L c.C800T p.T267I** | | | | | |
| Forward primer | TTGACAACAGGCTTGTGAGC | 20 | 58.98 | 50 | 209 |
| Reverse primer | AAATCGAGCTCTGGTGATGG | 20 | 57.68 | 50 | |
| ***HSPA1L* c.G229A p.G77S** | | | | | |
| Forward primer | CTACGTGGCCTTCACAGACA | 20 | 59.68 | 55 | 196 |
| Reverse primer | CACAAGGACTTTGGGCTTGC | 20 | 59.97 | 55 | |
| ***HSPA1A* gene** | | | | | |
| Forward primer | TCTCGCGGATCCAGTGTTC | 19 | 60 | 57.9 | 2030 |
| Reverse primer | TCCAAAACAAAAACAGCAATCTTGG | 25 | 47.7 | 36 | |
| ***HSPA1B* gene** | | | | | |
| Forward primer | TTGTCGCGGATCCCGTCCG | 19 | 64 | 68.4 | 2106 |
| Reverse primer | GAAGTGAAGCAGCAAAGAGCTGAAGC | 26 | 54.4 | 50 | |
| ***HSPA1A*_HSPA1B c.G1652T p.S551I** | | | | | |
| Forward primer | GCAAGGCCAACAAGATCACC | 20 | 62 | 55 | 267 |
| Reverse primer | GGGTTACACACCTGCTCCAG | 20 | 64 | 60 | |

For each primer sequence, length, melting temperature (Tm), percentage of CG (%CG) and the final product length is given.

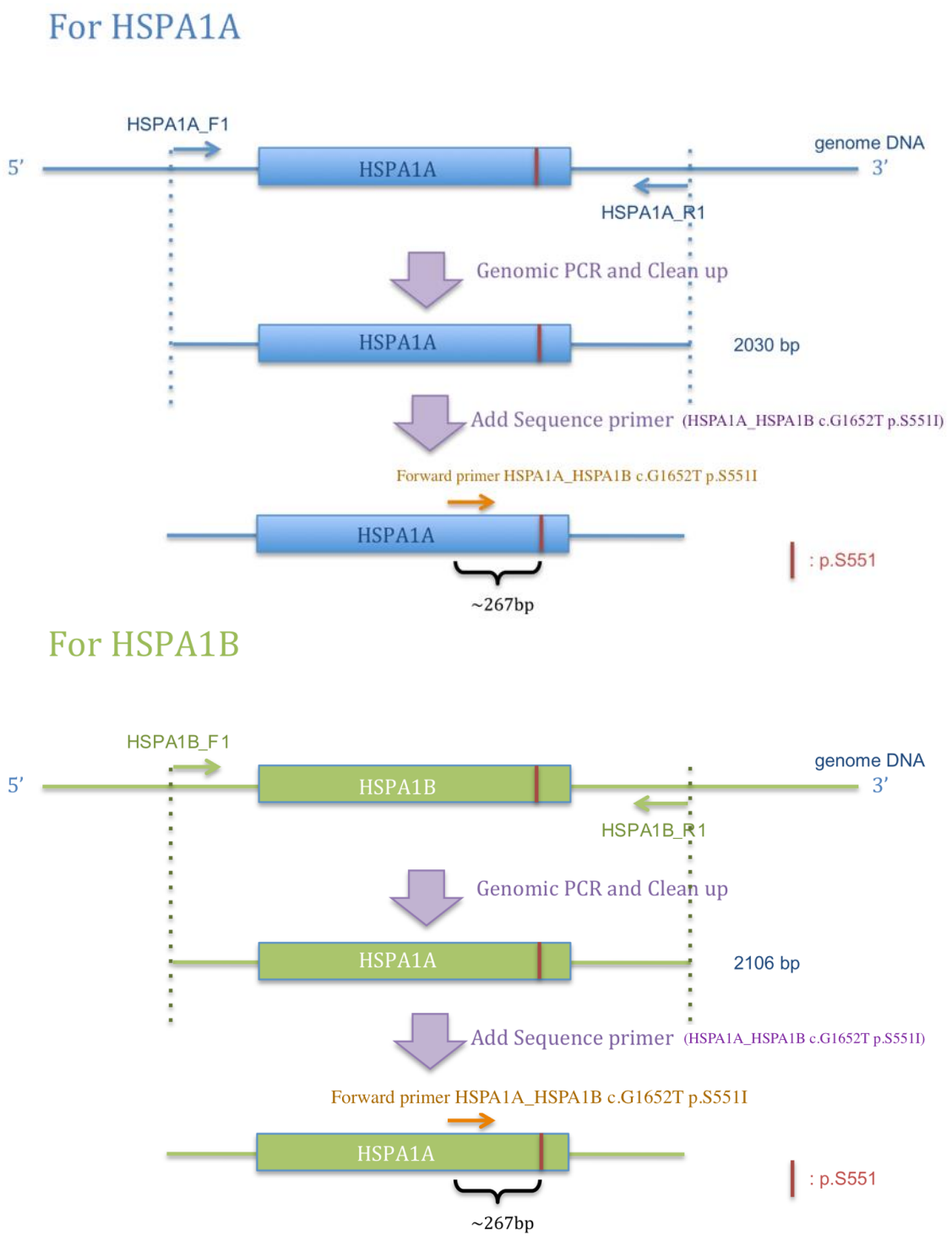



**Figure 4.2 Nested PCR method for discriminating *HSPA1A* and *HSPA1B*.** Specific primers were used to amplify *HSPA1A* and *HSPA1B* respectively. Each resultant PCR product (2030bp for *HSPA1A* and 2106bp for *HSPA1B*) were subsequently cleaned and used for a second PCR using the specific set of primers for the mutation of interest (p.S55I). The product of the second PCR for each gene (267 bp) was sent to Source BioScience for sequencing.

### 4.2.4 Rare variant profiling across known IBD genes

We generated a list of 336 known IBD genes. This list comprises 78 genes involved in early onset and monogenic form of IBD[133,134,160,308,309] and 274 genes falling within 163 loci implicated in IBD[140] (16 genes were in common to both disease groups). Variation within these 336 genes was extracted from the VCF generated for each of six patients harbouring one of the four *HSPA1L* mutations of interest.

Synonymous variants, those with a $MAF_{1KG} > 0.05$, those falling within a homopolymer or a repeat region, those representing alignment artefacts or those flagged as likely false-positive[220] were excluded. All remaining novel, non-synonymous, frameshift and non-frameshift indels, splicing, stop gain and stop loss mutations were considered.

### 4.2.5 Burden of mutation testing across heat shock protein genes

Whole exome sequencing data was available on 146 children diagnosed with IBD. The Southampton genomic informatic group also has access to germline exome sequence data for 126 unrelated patients with no inflammatory-related disease. As described in paragraph 3.2.5, in order to minimize bias for association analysis, we conducted a principle component analysis using the SNPRelate package on combined set of patients and controls and excluded non-Caucasian samples. PCA was applied to 1,363 samples with 305,950 biallelic SNPs. This procedure removed 10 cases and 20 controls reducing the final number of cases to 136 and controls to 106. All variants identified in any individual for *HSPA1L*, *HSPA1A* and *HSPA1B* genes were positively called to distinguish homozygous reference from zero coverage regions in all samples across the pIBD patients and controls. These genotypes were selected for further analysis. To detect association between genetic variant and disease status, the SKAT-O gene-based test was performed as described in paragraph 2.2.9. To conduct the test, a group file of non-synonymous and non-frameshift only variants was created for each of the three genes[310–312]. SKAT-O was conducted excluding synonymous variants as these are less likely to impact the protein function, as previously described in Auer *et al*[310–312]. SKAT-O was executed with the small sample adjustment, by applying MAF threshold of 0.01 to define rare variations within the whole cohort, and using default weights. The EPACTS software package[225] was use to perform this test.

## 4.3 Results

### 4.3 .1 Family-based whole exome sequencing analysis revealed a de novo mutation in *HSPA1L*

The Stanford group analysed the exomes of Family A, comprising the affected proband (12s) diagnosed with UC, both unaffected parents and an unaffected sibling. After excluding implausible genes such as those encoding olfactory receptors and mucins, and applying the *in silico* predictions (SIFT < 0.05 and PolyPhen-2 > 0.85), they found a *de novo* heterozygous mutation c.830C>T (encoding p.Ser277Leu) affecting the gene *HSPA1L* only in 12s but not in other family members (Table 4.2). All genotypes were confirmed by Sanger sequencing (Figure 4.3). The mutation observed within Family A resides at a nucleotide-binding site, and is highly conserved between species and within paralogous members of the human HSP70 family. This result has initiated a comprehensive investigation of HSP70 family in a larger cohort to further elucidate role of *HSPA1L*, *HSPA1A* and *HSPA1B* genes in IBD.

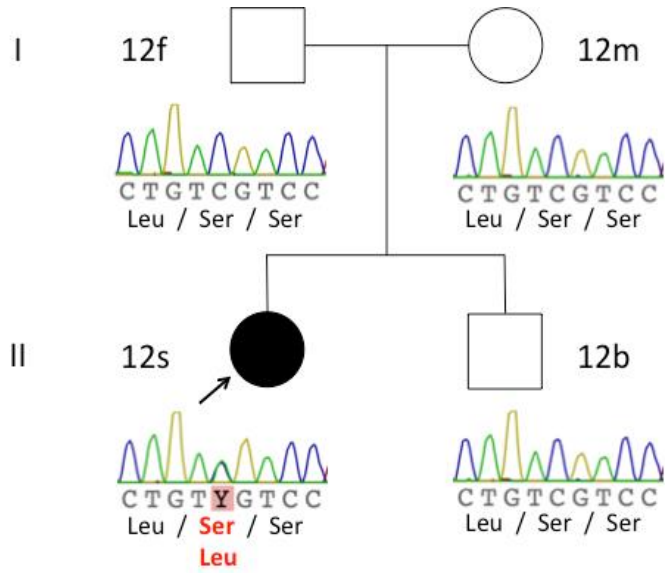



**Figure 4.3 Pedigree and Sanger traces of Family A.** The ulcerative colitis patient (filled symbol) has a *de novo* heterozygous mutation of c.830C>T (encoding p.Ser277Leu).

**Table 4.2 Homozygous and heterozygous mutations unique to the index patient with ulcerative colitis (12s)**

| Band | 12f | 12m | 12b | 12s | Gene | AAChange | esp6500_all | 1000g2012apr_all | snp138 | SIFT_score | PolyPhen2_HDIV_score | note |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3q29 | 0/1 | 0/2 | 0/1 | 1/2 | *MUC4* | p.T181I | NA | NA | rs729593 | 0.03 | 0.311 | Mucin 4, Cell Surface ssociated |
| 8p23.1 | 0/1 | 0/1 | 0/1 | 1/1 | *RP1L1* | p.A1319G | NA | NA | rs4840501 | 1 | 0 | Retinitis Pigmentosa 1-Like 1 |
| 11p11.2 | 0/2 | 0/1 | 0/2 | 1/2 | *OR4B1* | p.T274M | NA | NA | rs7130086 | 0 | 1 | Olfactor Receptor |
| 11p15.4 | 0/1 | 0/1 | 0/1 | 1/1 | *OR51I2* | p.R263C | 0.020006 | 0.01 | rs75620804 | 0.01 | 1 | Olfactory Receptor |
| 11q25 | 0/1 | 0/1 | 0/1 | 1/1 | *NCAPD3* | p.R622Q | 0.04209 | 0.03 | rs12292394 | 0.59 | 0 | Condensin-2 complexsubunit D3 |
| 14q31.1 | 0/1 | 0/1 | 0/1 | 1/1 | *TSHR* | p.P52T | 0.046594 | 0.03 | rs2234919 | 0.5 | 0.007 | Thyroid Stimulating Hormone Receptor |
| 22q13.2 | 0/1 | 0/1 | 0/0 | 1/1 | *EFCAB6* | p.T1030P | 0.041827 | 0.03 | rs34955597 | 0.27 | 0.168 | EF-hand calcium binding domain 6 |
| 1p36.23 | 0/0 | 0/0 | 0/0 | 0/1 | *SLC45A1* | p.A565V | NA | NA | NA | 0.18 | 0.607 | Solute Carrier Family 45, Member 1 |
| 2p14 | 0/0 | 0/0 | 0/0 | 0/1 | *SLC1A4* | p.P22L | 0.000786 | NA | rs201175768 | 0.29 | 0.1 | Transporter for Ala, Ser, Cys, and Thr |
| 5q14.1 | 0/0 | 0/0 | 0/0 | 0/1 | *MSH3* | p.P63A | NA | NA | rs2405876 | . | 0.235 | Post-replicative DNA mismatch repair system |
| 5q14.1 | 0/0 | 0/0 | 0/0 | 0/1 | *MSH3* | p.P64A | NA | NA | rs2405877 | . | 0.043 | Post-replicative DNA mismatch repair system |
| 6p21.33 | 0/0 | 0/0 | 0/0 | 0/1 | *MUC21* | p.E304G | NA | NA | rs201896109 | . | 0.011 | 6p21.33, Mucin 21, Cell Surface Associated |
| 6p21.33 | 0/0 | 0/0 | 0/0 | 0/1 | *HSPA1L* | p.S277L | NA | NA | NA | 0 | 1 | 6p21.33, HSP70-Hom |
| 7q22.1 | 0/0 | 0/0 | 0/0 | 0/1 | *MUC17* | p.P2716A | NA | NA | rs34924040 | 0.92 | 0.103 | Mucin 17, Cell Surface Associated |
| 8q12.1 | 0/0 | 0/0 | 0/0 | 0/1 | *RPS20* | p.T23P | NA | NA | NA | 0.03 | 0.289 | Ribosomal Protein S20 |
| 11p15.5 | 0/0 | 0/0 | 0/0 | 0/1 | *MUC6* | p.S1842P | NA | NA | rs111373859 | 0.36 | 0.04 | Gastric Mucin-6 |
| 11p15.5 | 0/0 | 0/0 | 0/0 | 0/1 | *MUC6* | p.N1686S | NA | NA | rs200243990 | 1 | 0 | Gastric Mucin-6 |
| 11q12.1 | 0/0 | 0/0 | 0/0 | 0/1 | *OR8U1 OR8U8* | p.G242S | NA | NA | rs77614949 | 0.45 | 0.025 | Olfactory Receptor |

0/0 indicates reference homozygote; 0/1 indicates heterozygote (reference/alternative); 1/1 indicates alternative homozygote; 0/2 indicates heterozygote (reference/2nd alternative); 1/2 indicates heterozygote (alternative/2nd alternative). esp6500_all, alternative allele frequency in all subjects in the NHLBI-ESP project with 6500 exomes; 1000g2012apr_all, alternative allele frequency data in 1000 Genomes Project.

### 4.3.2 Investigation of rare mutations in *HSPA1L* in a larger cohort of IBD patients

To determine the prevalence of rare *HSPA1L* mutations in IBD patients, we interrogated the exomes of 136 IBD patients and 106 non-IBD control subjects of the Southampton cohort. Fourteen *HSPA1L* variants across the exomes of children diagnosed with IBD and controls were observed. Five were synonymous and thus unlikely to have a functional impact on the encoded protein. Two non-synonymous variants were common ($MAF_{1KG} > 0.05$). Of the remaining seven variants, two were low frequency ($MAF_{1KG}$ 0.01–0.05) non-synonymous mutations, four were rare ($MAF_{1KG} <$ 0.01) non-synonymous mutations and one was a novel non-frameshift 3 base pair (bp) deletion found in an IBD patient. One low frequency mutation was unique to controls, but was synonymous. Thus, the rare non-synonymous and de novo variants were observed in cases only (Table 4.3).

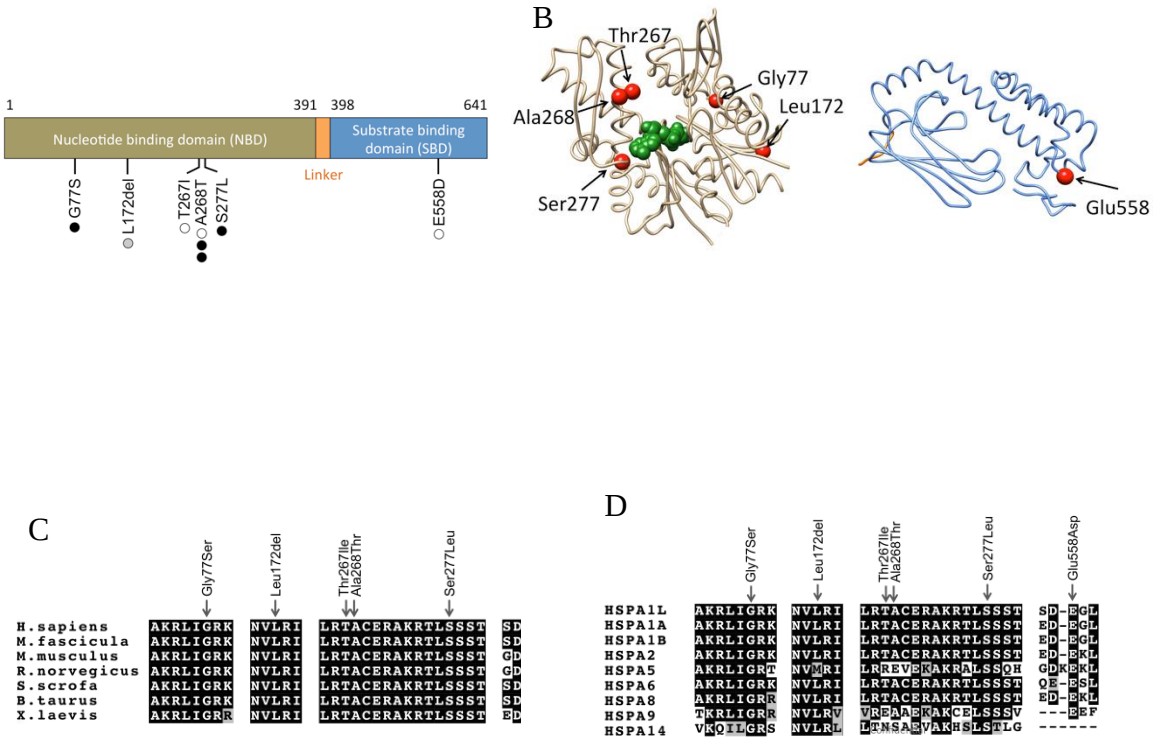
Of interest, the four rare non-synonymous mutations (p.Gly77Ser, p.Thr267Ile, p.Ala268Thr and p.Glu558Asp) and the novel frameshift mutation (p.Leu172del) reside at highly conserved residues throughout speciation and human paralogs (Figure 4.4A, B, C and D). The p.Gly77Ser and Leu172del variants are at or adjacent to the nucleotide-binding site; p.Thr267Ile and p.Ala268Thr are located at a nucleotide exchange factor binding domain, and pGlu558Asp resides in a substrate binding domain. These five variants were deemed to be of highest functional interest and were verified by Sanger sequencing in the probands and all relatives for whom DNA was available (Appendix VIII). Together with the index family case, these results indicate that five of six *HSPA1L* IBD mutations may affect nucleotide binding or exchange.

**Table 4.3 Variants found in IBD patients and controls in *HSPA1L* (no filtering applied)**

| Base pair location in hg19 | Varinat type | Nucleotide change | Protein change | phylop | 1-sift | PolyPhen2 | Grantham | dbSNP137 | Frequency in 1000 Genome Project | Cases* genotypes (homozygous reference allele, heterozygous, homozygous alternative allele) | Controls+ genotypes (homozygous reference allele, heterozygous, homozygous alternative allele) | MAF within combined cases and controls cohort |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **31779233** | **nfd** | **c.515_517del** | **p.L172del** | **.** | **.** | **.** | **.** | **.** | ***De novo*** | 135,1,0 | 106,0,0 | 0.0020• |
| **31778076** | **ns** | **c.1674A>T** | **p.E558D** | **0.108385** | **T** | **B** | **C** | **.** | **0.0000089** ✚ | 135,1,0 | 106,0,0 | 0.0020• |
| **31778948** | **ns** | **c.802G>A** | **p.A268T** | **0.997482** | **D** | **D** | **MC** | **rs34620296** | **0.0014000** | 134,2,0 | 106,0,0 | 0.0041• |
| **31778950** | **ns** | **c.800C>T** | **p.T267I** | **0.998993** | **D** | **D** | **MC** | **rs139868987** | **0.0014000** | 135,1,0 | 106,0,0 | 0.0020• |
| **31779521** | **ns** | **c.229G>A** | **p.G77S** | **0.936178** | **D** | **D** | **MC** | **rs368138379** | 0.0000770 ◊ | 135,1,0 | 106,0,0 | 0.0020• |
| 31779728 | ns | c.22G>C | p.A8P | 0.995889 | D | D | C | rs9469057 | 0.0130000 | 136,0,0 | 103,3,0 | 0.0061• |
| 31778077 | ns | c.1673A>C | p.E558A | 0.995982 | T | P | MR | rs2227955 | 0.0480000 | 129,7,0 | 98,8,0 | 0.0309• |
| 31777946 | ns | c.1804G>A | p.E602K | 0.997651 | D | B | MC | rs2075800 | 0.2900000 | 57,57,22 | 48,49,9 | 0.3471 |
| 31778272 | ns | c.1478C>T | p.T493M | 0.008994 | T | B | MC | rs2227956 | 0.8800000 | 6,33,97 | 2,23,81 | 0.1487 |
| 31778697 | sn | c.1053G>C | p.L351L | . | . | . | . | rs199780750 | 0.0000400 ✚ | 135,1,0 | 106,0,0 | 0.0020 |
| 31779003 | sn | c.747G>A | p.R249R | . | . | . | . | rs116768554 | 0.0027000 | 135,1,0 | 106,0,0 | 0.0020 |
| 31778322 | sn | c.1428C>T | p.I476I | . | . | . | . | rs35347921 | 0.0040000 | 135,1,0 | 106,0,0 | 0.0020 |
| 31778831 | sn | c.919T>C | p.L307L | . | . | . | . | rs35326839 | 0.0200000 | 133,3,0 | 102,4,0 | 0.0144 |
| 31778529 | sn | c.1221G>A | p.T407T | . | . | . | . | rs2075799 | 0.1400000 | 123,13,0 | 90,14,2 | 0.0640 |

14 variants ordered by variant type and within type ordered by Frequency in 1000 Genome Project. • Variants used in the SKAT-O test * Soton PIBD exomes, n=136 ; + Soton controls, n=106 ; ◊ Frequency in NHLBI ESP; ✚Frequency in ExAC Browser; Novel variants and rare (MAF < 0.01) non-synonymous variants of interest are shown in bold; Dots denote missing data. Ns = non-synonymous; sn = synonymous; nfd = non-frameshift deletion. B, Benign; C, Conservative; D, deleterious; MC, Moderately Conservative; MR, Moderately Radical; P, possibly damaging; T, tolerated.

**Figure 4.4 Figure A, B C and D of *de novo* and rare variants in HSPA1L**




A) Schematic representation of the *HSPA1L* gene describing *de novo* and rare variants identified within Family A and the 136 IBD cohort. Black, white, and grey circles represent ulcerative colitis, Crohn's disease, and IBD unclassified, respectively. B) The identified rare variants (left) on the structure of nucleotide binding domain (NBD) of HSPA1L (PDB entry codes: 3GDQ[313]) and (right) on homology-based model of substrate binding domain (SBD) of HSPA1L created by using Phyre2[314]. The variant sites are shown in red, and adenosine diphosphate and phosphate ($PO_4$) are depicted as a space-filling representation in green. C) Amino acid conservation of HSPA1L among species. D) Amino acid conservation among paralogs of HSPA1L in human. Amino acid sequences were aligned using Clustal Omega and annotated using BOXSHADE.

### 4.3.3 Mutations in *HSPA1A* and *HSPA1B*

We also examined the highly homologous *HSPA1A* and *HSPA1B* genes in the Southampton cohort. Although *HSPA1L* is expressed at a low level in the intestine, *HSPA1A* and *HSPA1B* are abundantly expressed in this tissue[315]. Two common synonymous variants in *HSPA1A* and five synonymous variants in *HSPA1B* were found, of which three were low frequency (MAF 0.01–0.05) in 1000 Genome Project (Table 4.4). We also performed variant calling only on the reads that are non-uniquely mapped to the *HSPA1A* and *HSPA1B*, and identified one additional novel non-synonymous mutation in *HSPA1A* (p.S551I). However, Sanger sequencing of this variant in the proband and all available pedigree members was negative validating the quality metrics applied in our sequencing pipeline. Interestingly, we did not find non-synonymous mutations in either of the *HSPA1A* and *HSPA1B* homologs (Figure 1f in Appendix VIII).

**Table 4.4 . *HSPA1A* and *HSPA1B* variants identified in IBD patients and controls (no filtering applied)**

| Gene | chromosome | Base pair location in hg19 | Variant type | Nucleotide change | Protein change | dbSNP137 | Frequency in 1000 Genome Project | Cases* genotypes (homozygous reference allele, heterozygous, homozygous alternative allele)* | Controls+ genotypes (homozygous reference allele, heterozygous, homozygous alternative allele) | MAF within cases and controls |
|---|---|---|---|---|---|---|---|---|---|---|
| HSPA1A | 6 | 31783755 | sn | c.222T>C | p.I74I | rs1043620 | 0.95 | 125,0,11 | 95,1,10 | 0.08884 |
| HSPA1A | 6 | 31785228 | sn | c.1695G>C | p.A565A | rs33998554 | 0.83 | 27,32,77 | 10,30,66 | 0.28099 |
| HSPA1B | 6 | 31795745 | sn | c.18G>A | p.A6A | rs34004874 | 0.01 | 134,2,0 | 104,1,1 | 0.01033 |
| HSPA1B | 6 | 31795949 | sn | c.222T>C | p.I74I | rs140434649 | 1 | 121,0,15 | 93,0,13 | 0.11570 |
| HSPA1B | 6 | 31797272 | sn | c.1545C>A | p.I515I | rs17854926 | 0.04 | 135,1,0 | 106,0,0 | 0.00206 |
| HSPA1B | 6 | 31797422 | sn | c.1695G>C | p.A565A | rs33998554 | 0.01 | 134,2,0 | 103,3,0 | 0.01033 |
| HSPA1B | 6 | 31797587 | sn | c.1860C>G | p.G620G | rs539689 | 0.56 | 34,62,40 | 36,47,23 | 0.48554 |

*****Occurrence in Soton PIBD exomes (n = 136), **+**Occurrence in Soton control exomes (n = 106), sn, synonymous

### 4.3.4 Joint rare variant association test

We conducted a gene-based test for assessing the combined association of coding novel, rare and common mutations between affected and unaffected individuals within the whole cohort. This analysis was limited to variants most likely to impact protein function and discounted synonymous changes. Since we did not observe any

non-synonymous variants in *HSPA1A* and *HSPA1B*, we did not conduct the SKAT-O test for these genes. Therefore, for *HSPA1L*, SKAT-O testing was conducted on the four rare non-synonymous mutations (p.G77S, p.T267I, p.A268T and p.E558D) and one novel non-frameshift (p.L172del). The test showed a significant association between *HSPA1L* variants and the IBD phenotype (p=0.024, Table 4.4). When the SKAT-O test was repeated to include the two low frequency non-synonymous mutations (p.A8P and p.E558A) in addition to the five rare mutations, the association remained significant (p=0.034, Table 4.5). Overall, these analyses suggest the rare mutations in *HSPA1L* are associated with IBD. The fact that the majority of mutations reside in specific domains (i.e. at nucleotide binding or exchange) further suggests these variants are not randomly associated with IBD and likely to be causative mutations.

**Table 4.5 Results of SKAT-O within *HSPA1L***

| Gene | bp position (hg19) | Total number of samples (136 cases; 106 controls) | Frequency of individuals with rare (MAF < 0.01)* variants | Number of all variants defined in the group file | Number of variants defined as rare (MAF < 0.01)* | P value unadjusted | Weighted (W) or Unweighted (UW) p value |
|---|---|---|---|---|---|---|---|
| Using non-synonymous and non-frameshift variants, excluding low frequency (0.01 < MAF < 0.05) and common (MAF > 0.05) variants. | | | | | | | |
| *HSPA1L* | 6:31778076-31779521 | 242 | 0.024793 | 5 | 5 | 0.02428 | W |
| Using non-synonymous and non-frameshift variants, excluding common variants (MAF > 0.05). | | | | | | | |
| *HSPA1L* | 6:31778076-31779728 | 242 | 0.033058 | 7 | 6 | 0.034846 | W |

* These variants received different weights in the SKAT-O joint test. The minor allele frequency is calculated within the full sample size.

### 4.3.5 Characterization of mutations in genes known to be associated with IBD

As IBD is complex, polygenic disorder, in order to understand the mutational background that might contribute to this disorder, I have further analysed the seven patients carrying one of the six *HSPA1L* variants of interest. We called a total of 2288 variants in 279 genes within the 336 known IBD genes. A total of 100 mutations (Table 4.6) remained following filtering criteria described in 4.2.4. Eighty-eight of these were non-synonymous and were annotated with predicted likelihood to be deleterious (1-SIFT, Grantham and PolyPhen-2) including *NOD2* p.R703C, CARD11 p.R794C and *HSPA6* p.V521G and these results are presented graphically in Figure 4.5. Nine variants represented as triangles in the top right quadrant of this graph represent those where all three measures of deleteriousness concur in assigning functional importance.

Nevertheless, we found no previously known IBD gene in common between the seven patients.

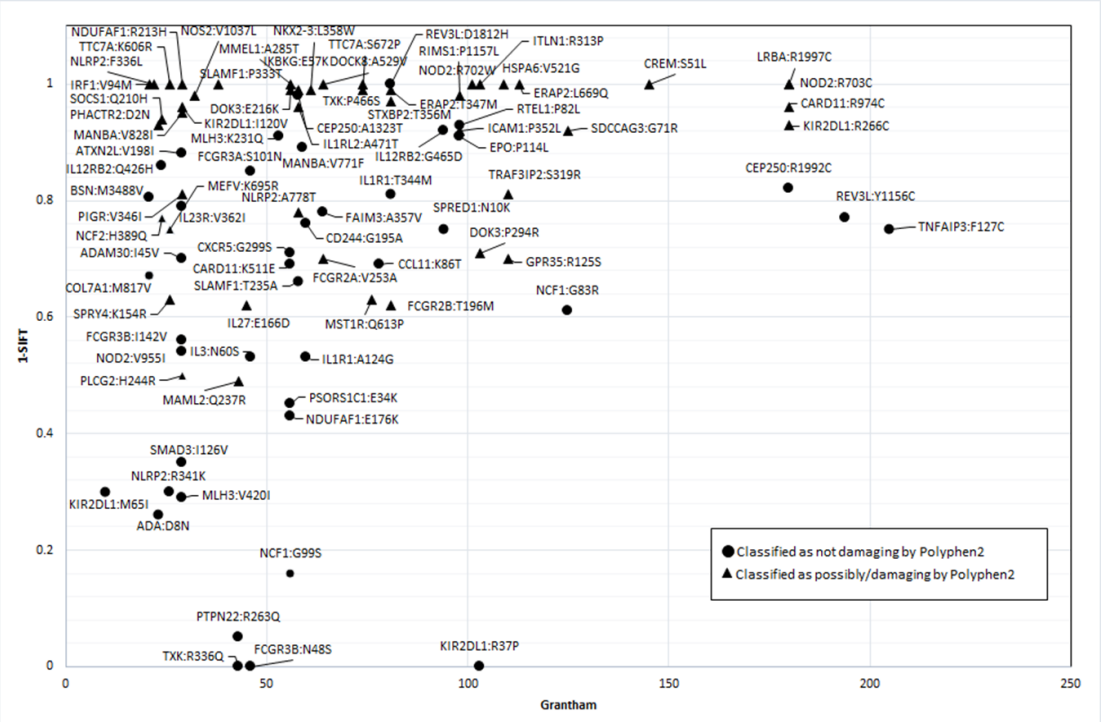



**Figure 4.5 Correlation between 1-SIFT and Grantham scores for the variants found within the seven patients harbouring *HSPA1L* mutations of interest.** *In silico* functional predictions of the 88 non-synonymous variants across the 279 associated IBD genes[59]. On the plot are represented three functional annotation tools: SIFT on the Y axes, Grantham on the X axes and Polyphen2 in circles or triangular shapes. SIFT distinguishes intolerant from tolerant amino acid substitutions. The SIFT score ranges from 0 to 1, with one indicating deleterious substitutions. The Grantham scores categorize codon replacements into classes of increasing chemical dissimilarity based on the chemical characteristics of the amino acids; the higher the score the more deleterious is the change. PolyPhen2 assesses the impact of an amino acid substitution on the structure and function of a protein; damaging mutations are represented as triangle.

**Table 4.6 The hundred variants across the 336 known IBD gene for the seven patients carrying HSPA1L variants of interest.**

| Gene | Base pair location in hg19 | Chromosome | Variant type | Nucleotide change | Protein change | 1-SIFT | PolyPhen2 | Grantham | MaxEnt score | rs ID number in dbSNP138 | Frequency in 46 CG | Frequency in 1KG Project | Frequency in NHLBI ESP | PR0034 | PR0142 | PR0151 | PR0156 | PR0161 | PR0244 | 12s |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **FRAMESHIFT INSERTION** | | | | | | | | | | | | | | | | | | | | |
| CXCL6 | 74702810 | 4 | fi | 239_240insT | V80fs | . | . | . | . | . | . | . | . | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| PSORS1C1 | 31106501 | 6 | fi | 112dupC | P38fs | . | . | . | . | . | 0.029 | . | . | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| **NON-FRAMESHIFT DELETION and NON-FRAMASHIFT INSERTION** | | | | | | | | | | | | | | | | | | | | |
| MAML2 | 95825372 | 11 | nfd | 1821_1823del | 607_608del | . | . | . | . | . | . | . | . | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| MAML2 | 95825375 | 11 | nfd | 1818_1820del | 606_607del | . | . | . | . | . | . | . | . | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| INPP5E | 139327607 | 9 | nfi | 1159_1160ins | E387delins | . | . | . | . | rs71384081 | 0.022 | . | . | 0 | 1 | 0 | 1 | 1 | 0 | 0 |
| **NON-SYNONYMOUS** | | | | | | | | | | | | | | | | | | | | |
| ADA | 43280227 | 20 | ns | G22A | D8N | T | B | C | . | rs73598374 | 0.022 | 0.04 | 0.0419 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| ADAM30 | 120438827 | 1 | ns | A133G | I45V | T | B | C | . | rs41276636 | . | 0.0046 | 0.0101 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| ATXN2L | 28837687 | 16 | ns | G592A | V198I | T | B | C | . | rs117987062 | 0.011 | 0.0037 | 0.0119 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| BSN | 49699740 | 3 | ns | A10462G | M3488V | T | T | C | . | . | . | . | . | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| CARD11 | 2953020 | 7 | ns | C2920T | R974C | D | D | R | . | rs201847585 | . | 0.0005 | 0.0002 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| CARD11 | 2972208 | 7 | ns | A1531G | K511E | T | B | MC | . | . | . | . | . | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| CCL11 | 32614672 | 17 | ns | A257C | K86T | T | B | MC | . | rs34262946 | 0.022 | 0.01 | 0.0002 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| CD244 | 160806019 | 1 | ns | G584C | G195A | T | B | MC | . | rs35224927 | . | 0.02 | 0.025 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| CEP250 | 34092171 | 20 | ns | C5974T | R1992C | T | B | R | . | rs41290926 | . | 0.0013 | 0.0037 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| CEP250 | 34089740 | 20 | ns | G3967A | A1323T | D | B | MC | . | rs114063154 | . | 0.01 | 0.0144 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| COL7A1 | 48626213 | 3 | ns | A2449G | M817V | T | B | C | . | rs147017402 | . | 0.0005 | 0.0009 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| CREM | 35495901 | 10 | ns | C152T | S51L | D | D | MR | . | rs52806860 | . | 0.0018 | 0.0043 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| CXCR5 | 118765283 | 11 | ns | G895A | G299S | T | B | MC | . | rs665648 | 0.022 | 0.01 | 0.0192 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| DOCK8 | 368128 | 9 | ns | C1586T | A529V | D | D | MC | . | rs17673268 | 0.022 | 0.04 | 0.106 | 1 | 0 | 0 | 1 | 1 | 0 | 0 |
| DOK3 | 176931911 | 5 | ns | G646A | E216K | D | D | MC | . | rs139699386 | . | . | 0.0001 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| DOK3 | 176931594 | 5 | ns | C881G | P294R | T | D | MR | . | rs61749657 | 0.065 | 0.04 | 0.0487 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| EPO | 100320381 | 7 | ns | C341T | P114L | T | B | MC | . | rs11976235 | . | 0.0027 | 0.0071 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| ERAP2 | 96228072 | 5 | ns | C1040T | T347M | D | D | MC | . | rs75263594 | 0.022 | 0.02 | 0.0312 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| ERAP2 | 96239258 | 5 | ns | T2006A | L669Q | D | D | MR | . | rs17408150 | 0.033 | 0.02 | 0.0569 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| FAIM3 | 207078467 | 1 | ns | C1070T | A357V | T | B | MC | . | rs41304091 | . | 0.01 | 0.0149 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| FCGR2A | 161483703 | 1 | ns | T758C | V253A | T | P | MC | . | . | . | . | . | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| FCGR2B | 161642981 | 1 | ns | C587T | T196M | T | D | MC | . | rs137950262 | . | . | 0.0043 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| FCGR3A | 161518336 | 1 | ns | G302A | S101N | T | B | C | . | rs448740 | . | . | . | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| FCGR3B | 161599693 | 1 | ns | A143G | N48S | T | B | C | . | rs448740 | 0.478 | . | . | 2 | 0 | 0 | 2 | 0 | 2 | 1 |
| FCGR3B | 161599571 | 1 | ns | A424G | I142V | T | B | C | . | rs199890941 | . | . | . | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| GPR35 | 241569742 | 2 | ns | C373A | R125S | T | D | MR | . | rs34778053 | 0.011 | 0.02 | 0.0232 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| HSPA6 | 161496010 | 1 | ns | T1562G | V521G | D | D | MR | . | rs199677197 | . | . | . | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| ICAM1 | 10395208 | 19 | ns | C1055T | P352L | T | P | MC | . | rs1801714 | 0.011 | 0.01 | 0.0292 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| IL12RB2 | 67833643 | 1 | ns | G1394A | G465D | T | B | MC | . | rs2307153 | . | 0.01 | 0.019 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| IL12RB2 | 67833527 | 1 | ns | G1278C | Q426H | T | B | C | . | rs2307145 | 0.065 | 0.04 | 0.0485 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| IL1R1 | 102781649 | 2 | ns | C371G | A124G | T | B | MC | . | rs2228139 | 0.043 | 0.05 | 0.0642 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| IL1R1 | 102791086 | 2 | ns | C1031T | T344M | T | B | MC | . | rs28362304 | 0.043 | 0.01 | 0.005 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| IL1RL2 | 102851470 | 2 | ns | G1411A | A471T | D | D | MC | . | rs75091099 | 0.022 | 0.01 | 0.0044 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| IL23R | 67705900 | 1 | ns | G1084A | V362I | T | B | C | . | rs41313262 | . | 0.01 | 0.0149 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| IL27 | 28511206 | 16 | ns | G498C | E166D | T | P | C | . | rs147413292 | 0.007 | 0.03 | 0.0588 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| IL3 | 131396676 | 5 | ns | A179G | N60S | D | B | C | . | rs35482671 | . | 0.0027 | 0.0034 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| IRF1 | 131822730 | 5 | ns | G280A | V94M | D | D | C | . | . | . | . | . | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| ITLN1 | 160846458 | 1 | ns | G938C | R313P | D | D | MR | . | rs8144 | . | 0.0032 | 0.0057 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| KIR2DL1 | 55294454 | 19 | ns | C796T | R266C | T | P | R | . | rs151328241 | 0.054 | . | 0.1308 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| KIR2DL1 | 55285072 | 19 | ns | A358G | I120V | D | P | C | . | rs138345877 | . | 0.01 | 0.0128 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| KIR2DL1 | 55284824 | 19 | ns | G110C | R37P | T | B | MR | . | rs139078925 | 0.087 | . | . | 0 | 0 | 2 | 0 | 0 | 0 | 0 |
| KIR2DL1 | 55284909 | 19 | ns | G195A | M65I | T | B | C | . | . | . | . | . | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| LRBA | 151520216 | 4 | ns | C5989T | R1997C | D | D | R | . | rs35879351 | 0.011 | 0.02 | 0.0323 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| MAML2 | 95826485 | 11 | ns | A710G | Q237R | T | D | C | . | rs61749253 | . | 0.01 | 0.02 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| MANBA | 103553372 | 4 | ns | G2482A | V828I | T | D | C | . | rs75826658 | 0.011 | 0.01 | 0.0176 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| MANBA | 103556049 | 4 | ns | G2311T | V771F | T | B | MC | . | rs201779762 | . | . | 0.0002 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| MEFV | 3293403 | 16 | ns | A2084G | K695R | T | D | C | . | . | . | 0.0032 | 0.0038 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| MLH3 | 75515101 | 14 | ns | G1258A | V420I | T | B | C | . | rs28756982 | 0.011 | 0.01 | 0.0147 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| MLH3 | 75515668 | 14 | ns | A691C | K231Q | T | B | MC | . | rs28756981 | 0.011 | 0.01 | 0.0192 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| MMEL1 | 2535684 | 1 | ns | G853A | A285T | D | D | MC | . | . | . | . | . | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| MST1R | 49935526 | 3 | ns | A1838C | Q613P | T | D | MC | . | rs35986685 | 0.011 | 0.0041 | 0.0091 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| NCF1 | 74193668 | 7 | ns | G295A | G99S | T | B | MC | . | rs10614 | . | . | . | 1 | 0 | 0 | 2 | 0 | 1 | 0 |
| NCF1 | 74193620 | 7 | ns | G247A | G83R | T | D | MR | . | rs139225348 | . | 0.009059 | 0.0118 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| NCF2 | 183532580 | 1 | ns | C1167A | H389Q | T | D | C | . | rs17849502 | 0.011 | 0.03 | 0.0502 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NDUFAF1 | 41688732 | 15 | ns | G526A | E176K | T | B | MC | . | rs35227875 | 0.011 | 0.02 | 0.0436 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| NDUFAF1 | 41687178 | 15 | ns | G638A | R213H | D | D | C | . | rs144437724 | . | . | 0.0003 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| NKX2-3 | 101295456 | 10 | ns | T1073G | L358W | D | D | MC | . | rs151053941 | . | 0.01 | 0.0011 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| NLRP2 | 55494141 | 19 | ns | T1006C | F336L | D | P | C | . | rs62124644 | . | 0.01 | 0.0112 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| NLRP2 | 55501424 | 19 | ns | G2332A | A778T | T | P | MC | . | rs117066658 | . | 0.01 | 0.0149 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| NLRP2 | 55494157 | 19 | ns | G1022A | R341K | T | B | C | . | rs41514352 | 0.065 | 0.05 | 0.0374 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| NOD2 | 50745926 | 16 | ns | C2104T | R702W | D | D | MR | . | rs2066844 | 0.022 | 0.02 | 0.0435 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| NOD2 | 50745929 | 16 | ns | C2107T | R703C | D | P | R | . | rs5743277 | . | 0.0023 | 0.007 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| NOD2 | 50757276 | 16 | ns | G2863A | V955I | T | B | C | . | rs5743291 | 0.036 | 0.05 | 0.096 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| NOS2 | 26087106 | 17 | ns | G3109C | V1037L | D | P | C | . | rs145383683 | . | 0.0014 | 0.002 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| PHACTR2 | 143929450 | 6 | ns | G4A | D2N | T | P | C | . | rs41285023 | 0.011 | 0.01 | 0.0258 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| PIGR | 207110449 | 1 | ns | G1036A | V346I | T | P | C | . | rs12748810 | . | 0.0041 | 0.0056 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| PLCG2 | 81916912 | 16 | ns | A731G | H244R | T | P | C | . | rs11548656 | . | 0.02 | 0.0349 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| PSORS1C1 | 31106489 | 6 | ns | G100A | E34K | T | B | MC | . | rs1265096 | 0.011 | 0.03 | 0.0666 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| PTPN22 | 114394689 | 1 | ns | G788A | R263Q | T | B | C | . | rs33996649 | 0.011 | 0.01 | 0.021 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| REV3L | 111694124 | 6 | ns | G5434C | D1812H | D | B | MC | . | rs3218599 | 0.011 | 0.01 | 0.0217 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| REV3L | 111696091 | 6 | ns | A3467G | Y1156C | T | B | R | . | rs458017 | 0.007 | 0.04 | 0.0673 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| RIMS1 | 72984123 | 6 | ns | C3470T | P1157L | T | D | MC | . | rs41265501 | . | 0.01 | 0.0334 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| RTEL1 | 62292793 | 20 | ns | C245T | P82L | T | B | MC | . | rs143461704 | . | 0.0002592 | 0.0009 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| SDCCAG3 | 139304551 | 9 | ns | G211A | G71R | T | D | MR | . | rs192537312 | . | 0.0009 | 0.0008 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| SLAMF1 | 160580549 | 1 | ns | C997A | P333T | D | D | C | . | rs3796504 | 0.043 | 0.04 | 0.0888 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| SLAMF1 | 160593973 | 1 | ns | A703G | T235A | T | B | MC | . | . | . | . | . | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| SMAD3 | 67457698 | 15 | ns | A376G | I126V | T | B | C | . | rs35874463 | 0.011 | 0.02 | 0.0547 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| SOCS1 | 11348706 | 16 | ns | G630C | Q210H | T | D | C | . | rs11549428 | 0.072 | 0.0013 | 0.0085 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| SPRED1 | 38545416 | 15 | ns | C30A | N10K | T | B | MC | . | rs201692618 | . | . | 0.0002 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| SPRY4 | 141694213 | 5 | ns | A461G | K154R | T | P | C | . | rs78310959 | . | . | 0.0033 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| STXBP2 | 7708058 | 19 | ns | C1067T | T356M | D | D | MC | . | rs117761837 | 0.011 | 0.01 | 0.0162 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| TNFAIP3 | 138196066 | 6 | ns | T380G | F127C | T | B | R | . | rs2230926 | 0.138 | 0.02 | 0.0321 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| TRAF3IP2 | 111901465 | 6 | ns | C957A | S319R | T | P | MR | . | rs146226365 | . | . | 0.0016 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| TTC7A | 47273468 | 2 | ns | A1817G | K606R | D | D | C | . | rs139010200 | . | 0.0018 | 0.0047 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| TTC7A | 47277182 | 2 | ns | T2014C | S672P | D | D | MC | . | rs149602485 | . | 0.0018 | 0.0045 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| TXK | 48082095 | 4 | ns | G1007A | R336Q | T | B | C | . | rs11724347 | 0.054 | 0.03 | 0.0735 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| TXK | 48073653 | 4 | ns | C1396T | P466S | D | D | MC | . | . | . | . | . | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| IKBKG | 153780386 | X | ns | G169G>A | E57K | D | D | MC | | rs148695964 | . | 0.00119 | 0.0018 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| **SPLICING** | | | | | | | | | | | | | | | | | | | | |
| CD5 | 60893322 | 11 | sp | 1490+9G>C | . | . | . | . | 0.23 | rs574843 | . | 0.04 | 0.0891 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| EIF3C | 28734759 | 16 | sp | 937-10G>A | . | . | . | . | 0.63 | rs186557331 | 0.011 | . | . | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| IL12B | 158743829 | 5 | sp | 856-5T>C | . | . | . | . | 0.87 | . | . | . | . | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| PLCG2 | 81904447 | 16 | sp | 565-10A>G | . | | . | . | 0.38 | rs62046684 | . | 0.02 | 0.0438 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| SP140 | 231176163 | 2 | sp | 2362-4C>A | . | . | . | . | 0.94 | rs199837567 | . | . | . | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| STXBP2 | 7707311 | 19 | sp | 828-4C>T | . | . | . | . | 1.26 | rs151257815 | . | 0.01 | 0.018 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| TTC7A | 47205921 | 2 | sp | 649-10C>T | . | . | . | . | 0.05 | rs149360779 | 0.011 | 0.0037 | 0.009 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |

Dots denote missing data;
Novel variants are shown in grey.
Zero (0) denotes reference homozygote; one (1) indicates heterozygous genotype; two (2) indicates homozygous genotype.
ns, non-synonymous; fi, frameshift insertion, sp, splicing; nfi, non-frameshift insertion; nfd, non-frameshift deletion.
B, benign; C, Conservative; D, deleterious; MC, moderately Conservative; MR, moderately Radical; P, possibly damaging; R, Radical; T, tolerated.
◊ indicates variants that despite not being reported in dbSNP137 or 1000 Genomes Project are reported in dbSNP137 or seen in our in-house reference exomes and are therefore not characterised as novel.

### 4.3.6 Patient profile

Summary of each patient phenotype and characteristics are shown in Table 4.7.

**Table 4.7 Summary of patient phenotypes and characteristics with HSPA1L mutation of interest**

| Sample ID | HSPA1L mutation | Age at diagnosis | Sex | Disease | Phenotype description | Ethnicity | Surgery | Family history |
|---|---|---|---|---|---|---|---|---|
| **12s** | p.S277L (c.830C>T) | 16 | F | UC | Initially left-sided colitis and proctitis, subsequently, pancolitis | Northern, Eastern European and Middle-Eastern mixed ancestry | - | - |
| **PR0034** | p.E558D (c.1674A>T) | 13 | M | CD | Nonstricturing ileocolonic | White British | + | - |
| **PR0142** | p.G77S (c.229G>A) | 13 | M | UC | Extensive mild to moderate pancolitis / Maternal grandmother has UC | Polish | - | + |
| **PR0151** | p.A268T (c.802G>A) | 13 | F | CD | Panenteric colitis | White British | - | - |
| **PR0161** | p.A268T (c.802G>A) | 10 | F | UC | Extensive mild to moderate pancolitis and autoimmune sclerosing cholangitis/ Sister has UC (Dx age 13 years) | White British | - | + |
| **PR0156** | p.T267I (c.800C>T) | 15 | M | CD | Terminal ileitis | White British | - | - |
| **PR0244** | p.L172del (c.515_517del) | 13 | F | IBDU | Mild chronic inactive gastritis | White British | - | - |

Patients are ordered according to *HSPA1L* mutations. UC, ulcerative colitis; CD, Crohn's disease; IBDU, inflammatory bowel disease unclassified; F, female; M, male; -, negative; +, positive

12s, from the Stanford Family A, was identified with the p.S277L *de novo* mutation in *HSPA1L*. She was diagnosed with UC at age 16. She was initially treated with oral and rectal 5-aminosalicylic acid and subsequently has been on 5-aminosalicylic acid and azathioprine. There is no family history of IBD. She carries a rare (MAF = 0.0016) mutation (p.L358W) in the *NKX2-3* gene, a member of the *NKX* family of homeodomain-containing transcription factor, which is known as a IBD susceptibility gene by GWAS studies. She also harbours a novel non-synonymous (p.P466S) mutation in the *TXK* gene that is also known as a susceptibility gene for IBD and Behcet's disease

and a non-synonymous mutation within *IKBKG* (p.E57K) which is a gene involved in the NOD2 pathway.

PR0034, harbouring the *HSPA1L* p.E558D, was diagnosed with Crohn's disease at age 13 years. He presented with diarrhea, abdominal pain and a significant weight loss. Initial endoscopy and radiological investigations showed ileo-colonic Crohn's disease. He was initially started on enteral nutrition, but responded poorly necessitating treatment with intravenous steroids. Due to poor nutritional improvement, he needed a prolonged support with naso-gastric tube feeds. He was subsequently started on thiopurines following which he remained stable for a few months, but then developed significant relapses necessitating treatment with biologics. Further endoscopic and radiological investigations revealed structuring disease for which, he underwent a de-functioning ileostomy and right hemicolectomy due to failure of medical management at age 13. He is currently stable and in remission on infliximab and azathioprine.

PR0142, a male patient diagnosed with ulcerative colitis at 13 years of age. He presented with bloody diarrhea for over a year. Initial endoscopy showed moderate degree of pan-colitis with backwash ileitis. He was initially treated with steroids and then subsequently maintained in remission on 6-mercapturine. There was a family history of colitis in the maternal grand-mother. He carries the *HSPA1L* p.G77S mutation and two damaging heterozygous non-synonymous mutations in *DOK3* - a gene involved in the B-cell receptor signaling pathway. However, discontinuous exome sequencing data is incapable of resolving phase and we cannot confirm compound heterozygosity without additional sequencing. This patient also carries a potential damaging variant in *CARD11* (p.R974C) and in *LRBA* (p.R1997C). *CARD11* is a UC associated gene encoding for cytoplasmic proteins involved in the apoptotic signaling cascade and activation of NF-kB[138], while *LRBA* is involved in regulating endosomal trafficking[316].

PR0151, a female patient having the p.A268T mutation, presented at 13 years of age, with a two year history of abdominal pain and diarrhea following an infection with cryptosporidium. Endoscopy at diagnosis showed moderately severe Crohn's disease with granulomatous gastritis, ileitis and colitis. She also had significant weight loss and a reduced height velocity at diagnosis. Following initial treatment with Modulen, then 5-azathioprine, she was subsequently started on infliximab due to recurrent disease. She has maintained remission for the last three years on a combination of azathioprine

and infliximab. She harbours potential compound heterozygous (p.T347M and p.L669Q) variants in the *ERAP2* gene that are likely to impact the antigen presentation function of the protein. Both variants occur at a 2% allele frequency in 1000 Genomes Project. This patient also harbours a rare possibly damaging mutation in *NDUFAF1* (p.R213H) that encodes a protein involved in the mitochondrial respiratory chain, and a low frequency variant in *NLRP2* (p.A801T), an inhibitor of the NF-κB signaling pathway[317].

PR0156 was diagnosed with ileal-caecal CD aged 15 years. He presented with long-standing history of abdominal pain and weight loss over several months. There was also a background of asthma and hay fever. Initial therapy with enteral nutrition and then subsequently steroids did not show significant improvement. He was started on infliximab within the first year of diagnosis due to a protracted disease course. He has maintained remission on infliximab for the last three years. He carries the *HSPA1L* p.A267T and the *NOD2* p.R702W variants in heterozygous state, the latter representing one of the three biomarkers for IBD that is associated with a twofold increase in odds ratio of CD[318]. In addition he also has a novel mutation within *FCGR2A* (p.V253A) reported as possibly damaging by PolyPhen2 but as tolerated and moderate conserved by 1-SIFT and Grantham respectively. He carries two heterozygous mutations in *SLAMF1*, regulator of the microbicidal mechanisms in macrophages[319], of which variant (p.P333T) is annotated as deleterious but the second novel (p.T235A) variant is annotated as benign, and two damaging heterozygous non-synonymous mutations in *TTC7A* - a gene involved in intestinal development[303].

PR0161 is an early onset UC patient diagnosed at age 10 years with concurrent autoimmune sclerosing cholangitis. She presented with abdominal pain, rectal bleeding and deranged liver function tests. Her disease course has remained relatively stable on a combination of azathioprine, ursodeoxycholic acid and a low dose prednisolone (for the liver disease). She carries a rare *NOD2* variation (p.R703C, MAF = 0.002) located in the leucine rich domain that is implicated in intracellular receptor function for components of microbial pathogens[320]. The patient also harbours a rare frameshift insertion in *CXCL6*. This variant is not reported in public repositories (1KG, dbSNP, EVS[321], 46 CG[226]) but it has been observed in one other Southampton IBD pediatric proband as well as four controls (patients diagnosed with non-autoimmune conditions) within the Southampton control group of reference exomes. The *HSPA1L*

mutation carried by this patient (p.A268T) was also confirmed in her affected sister who is also diagnosed with UC.

PR0244 was diagnosed with inflammatory disease unclassified (IBDU) aged 13 years. She was started on 5-azathioprine, budesonide and prednisolone. Non-resolving symptoms have necessitated continuous steroidal therapy since January 2015. She carries the *HSPA1L* p.L172del and a novel mutation (p.V94M) in the *IRF1* gene shown to regulate apoptosis, DNA damage and tumor suppression[322]. She also harbours a rare (MAF = 0.0032) non-synonymous (p.R313P) damaging variant in the *ITLN1* gene implicated in pathogen recognition[323].

These seven patients had no variants in known IBD-associated genes in common besides the rare mutations in the *HSPA1L* (Table 4.5).

## 4.4 Discussion

This exome sequencing study was initiated by the finding of a *de novo* mutation in *HSAP1L* in an index family by our colleagues. The first phase was a family-based analysis, which revealed a *de novo*, novel HSPA1L mutation as a candidate potential causative mutation in the index IBD patient with no family history of IBD. The second phase was the comprehensive investigation for a role of HSPs in IBD with the analysis of rare *HSPA1L* mutations using an independent exomes cohort. In the analysis, in 136 cases five different novel or rare non-synonymous variants were identified in six pIBD patients, whereas no rare variants were found in 106 controls. Although the minor allele frequency for each of these five variants is too low to assess its association to the disease individually, a gene-based SKAT-O test revealed significant burden of mutation (p=0.024) when assessing non-silent rare variants observed in *HSPA1L*. The association maintained significance when reassessed to include all rare and low frequency variants (p=0.034).

The *HSPA1L* gene is located in the MHC class III region, which is within the IBD3 locus, a genetic linkage region for both UC and CD[324]. Likewise, in our study, rare *HSPA1L* mutations were observed both in UC, CD, and IBDU patients. These data suggest that *HSPA1L* might play a common pathogenic role in IBD. HSPA1L is constitutively expressed but its expression is at a lower level compared to other members of the HSP70 family, such as HSPA1A and HSPA1B[325]. Although the distinctive role of each isoform and substrate protein specificity for each HSP70 family member have not been well studied, it is reported that each HSP70 has binding preferences to purified peptides[326]. Also in a recent study, Hasson *et al.* demonstrated that HSPA1L and not HSPA1A, promotes translocation of Parkin to damaged mitochondria[327], which is required for mitophagy, suggesting that HSPA1L has specific protein substrates and functions. Further analysis of IBD phenotype, response to therapy, and histopathological data of patients with or without HSPA1L mutation will lead to a better understanding of disease mechanisms.

Considering that HSPA1A and HSPA1B have identical amino acid sequence and other HSP70s have high similarity in sequence to HSPA1A or HSPA1L, the low but constitutively expressed dominant negative variants may disturb the whole HSP70 chaperone system. The Stanford group conducted an independent biochemical assay to evaluate the effect of the rare non-synonymous *HSPA1L* mutations on its protein

function. Effects of the rare non-synonymous variants on chaperone function was determined by measuring the refolding of heat-inactivated luciferase substrate using recombinant HSPA1L proteins. The HSPA1L mutations were heterozygous in each patient. To evaluate whether the mutant alleles have dominant negative effects on wild type (WT) HSPA1L protein, the activity of a 1:1 mixture (molar) of WT and mutant protein was compared against the activity of HSPA1L WT alone. This was further compared against the activity of the previously-known loss-of-function mutation p.Lys73Ser (equivalent to Lys71Ser in HSPA1A[328]) as a control for dominant negative effect (the protocols can be found in Appendix VII). Among the identified variants, p.Gly77Ser, p.Leu172del, and p.Ser277Leu were more deleterious in that they showed almost complete loss of function and significant dominant negative effects in *in vitro* assays (Figure 4.6 and Figure 4.7). This severity is consistent with their low allele frequency (i.e. *de novo* or novel) in the human population. We hypothesized that these deleterious variants might be associated with more severe clinical observations, such as very early onset of IBD or severe symptoms; however, no such correlation was evident in our modest group of subjects. For example, the American patient 12s with severe p.Ser277Leu mutation had relatively late onset at age 16 and Southampton patient PR0161, having a less-harmful p.Ala268Thr mutation, diagnosed aged 10 years. We speculate that other genetic and/or environmental factors are likely to contribute to disease severity.

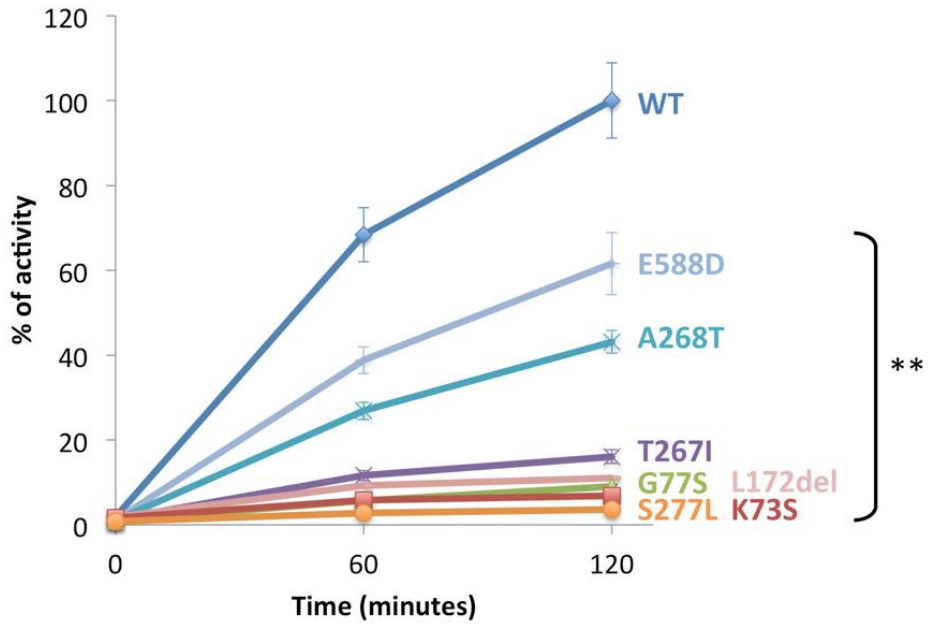



**Figure 4.6 In vitro chaperone activity assays.** Reactivation of heat-denatured luciferase in the presence of each HSPA1L variant (4 μM). Luciferase activity in the presence of HSPA1L WT at 120 minutes was set as 100%. ** indicates P<0.01 for the comparison between HSPA1L WT and each variant by Dunnett's

test (n = 3-6). The activity of the previously-known mutation p.Lys73Ser was measured as a positive control for loss-of-function and dominant negative mutant. The bars represent the standard deviation. Data are representative of two independent experiments.

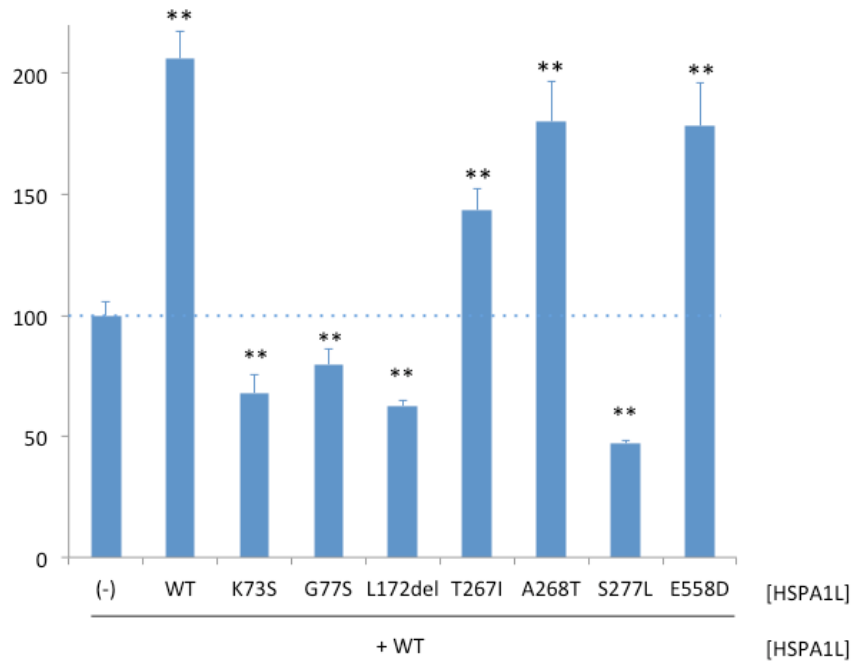



**Figure 4.7 Effect of the HSPA1L variant on HSP70/HSP40 mediated-refolding heat denatured luciferase.** Dominant negative effects of Gly77Ser, Leu172del, and Ser277Leu in refolding activity of each HSPA1L variant (2 μM) in the presence of HSPA1L WT (2 μM). Refolding activity of HSPA1L WT (2 μM) only was set as 100%. ** indicates P<0.01 for the comparison between HSPA1L WT only and each variant by Dunnett's test (n = 3-6). The activity of the previously-known mutation p.Lys73Ser was measured as a positive control for loss-of-function and dominant negative mutant. The bars represent the standard deviation. Data are representative of two independent experiments.

It is of interest that unlike *HSPA1L*, non-synonymous variants were found in neither the *HSPA1A* nor the *HSPA1B* genes given that HSP70 family proteins are highly homologous in sequence. While it is possible that the HSPA1L protein may have a specialized function, or that the *HSPA1A* and *HSPA1B* genes cannot be readily mutated without causing lethality, we are unable to completely exclude technical limitations as a factor due to the following reasons: a) Poor exome capture kit coverage of *HSPA1A* and *HSPA1B* substantially impacted our power to detect variation within these genes (Table 4.8); b) An excess of unmapped reads suggest limitations in mapping sequences on highly homologous genes. However, since few non-synonymous mutations in the nucleotide binding domain of *HSPA1A* and *HSPA1B* have been reported in public SNV repositories, such as dbSNP, it is plausible that mutations in these genes are not well tolerated or that this is a result of technical limitations of the capture and sequencing technologies.

**Table 4.8 Gene percentage coverage for *HSPA1L*, *HSPA1A* and *HSPA1B* in the Agilent SureSelect V4 and V5 kits**

| Gene Name | Gene size (bp) | Coding size (bp) | Agilent V4 % gene coverage (87 IBD cases and 56 controls) | Agilent V5 % gene coverage (59 IBD cases and 70 controls) |
|---|---|---|---|---|
| HSPA1L | 6,042 | 1,923 | 16.7 | 18.7 |
| HSPA1A | 2,433 | 1,923 | 5.9 | 20.0 |
| HSPA1B | 2,520 | 1,923 | 3.2 | 20.0 |

To understand the background mutational profile in the patients with the rare *HSPA1L* mutation, we further identified low frequency, rare and novel variations within known susceptibility genes for IBD across the exome data of the seven patients harbouring rare *HSPA1L* mutations. We identified 59 genes encoding 82 variants of which 22 mutations (27%) were assessed as deleterious by 1-SIFT; 23 (28%) as damaging by PolyPhen2 and six (7.3%) as radical by Grantham score. Functional annotation represents a major bottleneck to the clinical interpretation and the results can have a strong influence on the ultimate conclusions of disease studies[186]. For fifteen variants, including *NOD2* p.R703C, *CARD11* p.R794C and *HSPA6* p.V521G, three major measures of *in silico* annotation concur in assigning functional importance, and therefore these mutations are likely to contribute to the phenotypic heterogeneity in individuals with *HSPA1L* mutations. Evidence of functional protein-protein interactions between HSPA1L and NFkB1, a known IBD protein involved in the NOD signalling pathway and TNF-α signalling pathway, have been shown[329,330]. Protein interactions have also been reported between HSPA1L and RELA, the complexing partner of NF-kB1[229,330]. On the other hand, the rare *HSPA1L* mutations observed in the Southampton cohort were inherited from unaffected parents, which might indicate that potential cumulative effect from other genetic defects may act either independently or together with HSPA1L to influence disease susceptibility. Nevertheless, it is important to note that these patients had no other shared known IBD-associated genes in common that harbour potentially damaging variants other than the rare mutations in *HSPA1L*, suggesting that HSPA1L might be an important contributor to IBD although its phenotypes may be influenced by other genetic or environmental factors.

Through whole exome sequencing analysis of Family A and 136 Caucasian IBD patients, 1 *de novo* and 5 rare damaging mutations in the *HSPA1L* gene that are potentially

associated with the aetiology of this disease were observed. These variants caused loss of function of HSPA1L protein to varying degrees, and three of them also exhibited dominant negative effects on the wild type protein, which may in turn contribute to the disease phenotype as well as other potential biological dysfunctions. Of interest, three of the identified mutants showed dominant negative effects on another HSP70 protein (i.e. HSPA1A). Previous papers reported that dominant negative mutants blunted HSP70's protective effects in *Drosophila* models of neurodegenerative diseases, such as polyglutamine disease[331] or Parkinson's Disease[332]. We believe these findings would provide insights into the pathogenesis and treatment of IBD and as well as the general role of HSP70 proteins in human biology and disease.

## 4.5 Conclusions

The results from this international collaboration indicate that *de novo* and rare mutations in *HSPA1L* are associated with IBD. These findings provide insights into the pathogenesis and treatment of IBD, as well as expand our understanding of the roles of HSP70s in human disease.

# Chapter 5 **Identification of variants in genes associated with monogenic inflammatory bowel disease by whole exome sequencing**

## 5.0 Summary

Although most cases of IBD are caused by complex host-environment interaction, there are a number of conditions associated with a single gene mutation. Typically, monogenic IBD is very early onset (aged < 6 years), presents with a unique form of disease, may have atypical features and is genetically distinct from late-onset IBD. Within this chapter we aimed to interrogate whole-exome data for 147 paediatric IBD patients (of which 22 were very early onset) for a panel of 51 genes known to be associated with monogenic IBD. Observed variation was categorised according to the American College of Medical Genetics (ACMG) guidelines to identify rare, *novel* and known variants that might contribute to IBD. Variants were categorised into 'pathogenic' and 'likely pathogenic' variants. Six pathogenic variants were identified and segregation analysis was conducted. Although we have rigorously applied the recently updated ACMG guidelines for variant classification, assigning variants to these categories is fraught with difficulty. It is currently not clear how strongly these rare variants influence the genetic susceptibility to IBD, particularly in patients with nonconventional forms of IBD, the identification of variants of unknown significance can lead to the therapeutic dilemma of whether to wait for the disease to progress or start early treatment. Because some of the disease-specific treatment options have potentially severe adverse effects, careful evaluation of genetic variants is required not only to validate sequence data and statistical association but to provide functional evidence that those variants cause disease.

For this work I was responsible in the curation of the research database, data processing, execution of the quality check, data interpretation, design of primers for segregation analysis and manuscript preparation.

## 5.1 Background

In a small subgroup of pIBD patients, typically with early onset (age < 10 years) and very early onset (age < 6 years), there could be a single gene cause for the disease[309].

Whether or not IBD should be the primary diagnosis or the condition be labelled as IBD-like phenotype of an underlying immune defect is uncertain.

Although the overall incidence of pIBD is increasing[19], the frequency of IBD caused by a single genetic variant (monogenic IBD) is very low[309]. Individuals who have monogenic IBD are important to recognise as they have increased risk of developing significant concurrent problems, such as immunodeficiency, and this will impact on treatment options and prognosis next[160]. In addition, monogenic IBD may also be associated with specific features not typically associated with IBD such as nail and hair abnormalities[164], epidermolysis bullosa and autoimmune haemolytic anaemia[309]. Identification of atypical signs such as these should trigger further testing in these patients. To date, 51 genes have been identified linked to monogenic IBD, with the majority also associated with another condition; particularly a functional immune disorder (such as chronic granulomatous disease or severe combined immunodeficiency syndrome)[159]. There is a potential for misdirected treatment of patients with monogenic disease; receiving escalated treatment regimens with extreme forms of surgery and medical therapies rather than treating the underlying immune or other defect.

The accessibility of next generation sequencing technology has allowed identification of rare and novel pathogenic variants in pIBD[276]. Furthermore, variants in genes associated with primary immunodeficiency have been identified in patients with very early onset IBD[333], alongside specific mutations in genes associated with monogenic IBD[334]. Previously whole-exome sequencing has helped identify an association between and children presenting very-early-onset IBD and homozygous mutations in the interleukin 10 receptor (*IL10*) gene, *IL10* associated receptor alpha and beta subunits (*IL10RA* and *IL10RB*), homozygous mutations in *ADAM17* and hemizygous mutations within *FOXP3*[79].

This study utilises whole-exome sequencing data from a cohort of children diagnosed with IBD to extract all variants across 51 genes associated with monogenic IBD and identify potentially pathogenic mutations.

## 5.2 Materials and Methods

### 5.2.1 Recruitment

Children are recruited following diagnosis by the paediatric gastroenterology service at University Hospital Southampton (UHS). All children aged under 18 years are eligible for inclusion and all are diagnosed in line with the Porto criteria as described in paragraph 2.2.1.

### 5.2.3 DNA extraction

DNA was extracted as previously described in 2.2.2.

### 5.2.4 Whole-exome sequencing and data processing

Whole-exome capture was performed using Agilent SureSelect Human All Exon 51 Mb (versions 4 and 5) capture kit. Raw data generated from paired-end sequencing protocol were processed as described in 2.2.3. Variants were excluded if they had a PHRED quality score of <20 and/or a depth of <4. Copy number variations (CNVs) were assessed using the software ExomeDepth.

### 5.2.5 Gene selection and filtering strategy

A list of 50 genes taken from Uhlig et al[309] and updated with a single gene from Li *et al* [159] gave a total of 51 genes for interrogation after comprehensive literature review (Table 5.1). Any variation within these 51 genes was extracted from the variant call files generated for each of the pIBD patients.

**Table 5.1 Genes associated with Monogenic IBD**

| Gene | Associated condition | Inheritance | Agilent V5 % gene coverage (51 patients) | Agilent V4 % gene coverage (96 patients) |
|---|---|---|---|---|
| *ADA* | Severe combined immunodeficiency | AR | 100.00 | 97.85 |
| *ADAM17* | ADAM17 deficiency | AR | 100.00 | 79.74 |
| *AICDA* | Hyper IgM syndrome | AR | 93.27 | 30.92 |
| *BTK* | Agammaglobulinaemia | X | 93.76 | 87.21 |
| *CD3y* | Severe combined immunodeficiency | AR | 49.35 | 49.35 |
| *CD40LG* | Hyper IgM syndrome | X | 49.35 | 49.35 |
| *COL7A1* | Dystrophic bullosa | AR | 100.00 | 59.52 |
| *CYBA* | Chronic granulomatous disease | AR | 100.00 | 99.17 |
| *CYBB* | Chronic granulomatous disease | X | 100.00 | 96.68 |
| *DCLRE1C* | Omenn syndrome | AR | 100.00 | 46.34 |
| *DKC1* | Hoyeraal-Hreidarsson syndrome | X | 65.01 | 55.30 |
| *DOCK8* | Hyper IgE syndrome | AR | 100.00 | 78.77 |
| *FERMT1* | Kindler syndrome | AR | 99.29 | 88.98 |
| *FOXP3* | IPEX | X | 85.03 | 54.46 |
| *G6PC3* | Congenital neutropenia | AR | 93.03 | 63.06 |
| *GUCY2C* | Familial diarrhoea | AD | 68.80 | 71.77 |
| *HPS1* | Hermansky-Pudlak 1 | AR | 100.00 | 89.35 |
| *HPS4* | Hermansky-Pudlak 4 | AR | 91.68 | 60.22 |
| *HPS6* | Hermansky-Pudlak 6 | AR | 99.30 | 71.34 |
| *ICOS* | CVID 1 | AR | 100.00 | 98.80 |
| *IKBKG* | X-linked ectodermal immunodeficiency | X | 100.00 | 35.11 |
| *IL10* | IL-10 signalling defects | AR | 92.88 | 51.38 |
| *IL10RA* | IL-10 signalling defects | AR | 100.00 | 55.56 |
| *IL10RB* | IL-10 signalling defects | AR | 97.04 | 66.65 |
| *IL21* | IL-21 deficiency | AR | 100.00 | 100.00 |
| *IL2RA* | IPEX-like | AR | 100.00 | 38.77 |
| *IL2RG* | Severe combined immunodeficiency | X | 100.00 | 90.55 |
| *ITGB2* | Leukocyte adhesion deficiency type 1 | AR | 86.80 | 81.29 |
| *LIG4* | Severe combined immunodeficiency | AR | 86.99 | 67.37 |
| *LRBA* | CVID8 | AR | 99.10 | 90.65 |
| *MASP2* | MASP deficiency | AR | 100.00 | 93.76 |
| *MEFV* | Familial Mediterranean fever | AR | 89.56 | 75.59 |
| *MVK* | Mevalonate kinase deficiency | AR | 100.00 | 62.99 |
| *NCF1* | Chronic granulomatous disease | AR | 23.61 | 0.00 |

| | | | | |
|---|---|---|---|---|
| *NCF2* | Chronic granulomatous disease | AR | 96.96 | 83.77 |
| *NCF4* | Chronic granulomatous disease | AR | 100.00 | 100.00 |
| *PIK3R1* | Agammaglobulinaemia | AR | 96.19 | 39.86 |
| *PLCG2* | Phospholipase C-γ2 defects | AD | 93.33 | 93.68 |
| *RAG2* | Severe combined immunodeficiency | AR | 74.73 | 61.29 |
| *RTEL1* | Hoyeraal-Hreidarsson syndrome | AR | 87.14 | 83.26 |
| *SH2D1A* | X-linked lymphoproliferative syndrome type 1 | X | 100.00 | 39.13 |
| *SKIV2L* | Trichohepatoenteric syndrome | AR | 16.66 | 16.16 |
| *SLC37A4* | Glycogen storage disease type 1b | AR | 53.53 | 54.86 |
| *STAT1* | IPEX-like | AD | 91.69 | 61.99 |
| *STXBP2* | Familial haemophagtocytic lymphohistiocytosis type 5 | AR | 100.00 | 100.00 |
| *TTC37* | Trichohepatoenteric syndrome | AR | 95.85 | 85.94 |
| *TTC7A* | TTC7A deficiency | AR | 100.00 | 57.68 |
| *WAS* | WAS | X | 100.00 | 95.85 |
| *XIAP* | X-linked lymphoproliferative syndrome type 2 | X | 92.72 | 15.51 |
| *ZAP70* | Severe combined immunodeficiency | AR | 93.20 | 88.49 |

AR: autosomal recessive; X: X-linked

The following list of filters was applied in order to exclude variation unlikely to have clinical impact: all synonymous variants; variants common within the general population ($MAF_{1KG}$> 0.05); variants within intron-exon splice boundaries considered unlikely to impact splicing (MaxEnt score < 3); poorly conserved variants (Phylop < 0.95); variants within homopolymer tracts or repeat regions; those representing alignment artifacts or flagged as likely false-positive. All remaining novel, non-synonymous, frameshift and non-frameshift insertion/deletions, splicing, stop gain and stop loss mutations were considered and grouped based on the American College of Medical Genetics (ACMG) guidelines into the categories 'Pathogenic', 'Likely Pathogenic' and 'Benign'.

The ACMG guidelines on classification of variants specify that the functional impact of mutations must have been assessed to classify the variant as pathogenic; all pathogenic variants have previous functional work and are listed in the human gene mutation database (HGMD) (34). Likely pathogenic variants have functional impact inferred from similar mutations and demonstrate compelling clinical correlation.

### 5.2.6 Sanger sequencing and segregation analysis

Variants within the `Pathogenic` group occurring in the correct zygosity to be causal and assessed as deleterious by in silico annotation tools were verified by Sanger sequencing in the probands and all relatives for whom DNA was available (Appendix. IX). Primers were designed using primerBLAST and sequencing was outsourced at Source Bioscience, Nottingham, UK.

## 5.3 Results

### 5.3.1 Southampton pIBD cohort

At the time of analysis 147 individual patient exomes have been sequenced from the Southampton PIBD cohort. Demographic data for patients is shown in Table 5.2.

**Table 5.2 Southampton PIBD cohort demographics**

| | Crohn's Disease | Ulcerative colitis | IBDU | Total IBD |
|---|---|---|---|---|
| **Number of patients** | 87 | 37 | 23 | 147 |
| **Median Age** | 12.24 | 10.04 | 12.30 | 12.24 |
| **Female no. (%)** | 34 (39.09) | 16 (43.24) | 14 (60.87) | 64 (43.54%) |
| **Mean age of Onset (SD)** | 11.30 (3.53) | 9.38 (3.98) | 11.17 (3.54) | 11.04 (3.80) |

### 5.3.2 Characterization of mutations within genes associated with monogenic form of IBD

Within the cohort, 574 variants were called across all 51 genes specifically associated with a monogenic form of IBD. A total of 67 mutations remained following standard filtering criteria (Figure 5.1). Following the ACMG guidelines, four of these were determined as 'Pathogenic' category, two as 'Likely Pathogenic' category and 61 as 'Benign' category. No CNVs were identified in genes with 'Pathogenic' or 'Likely Pathogenic' mutations.

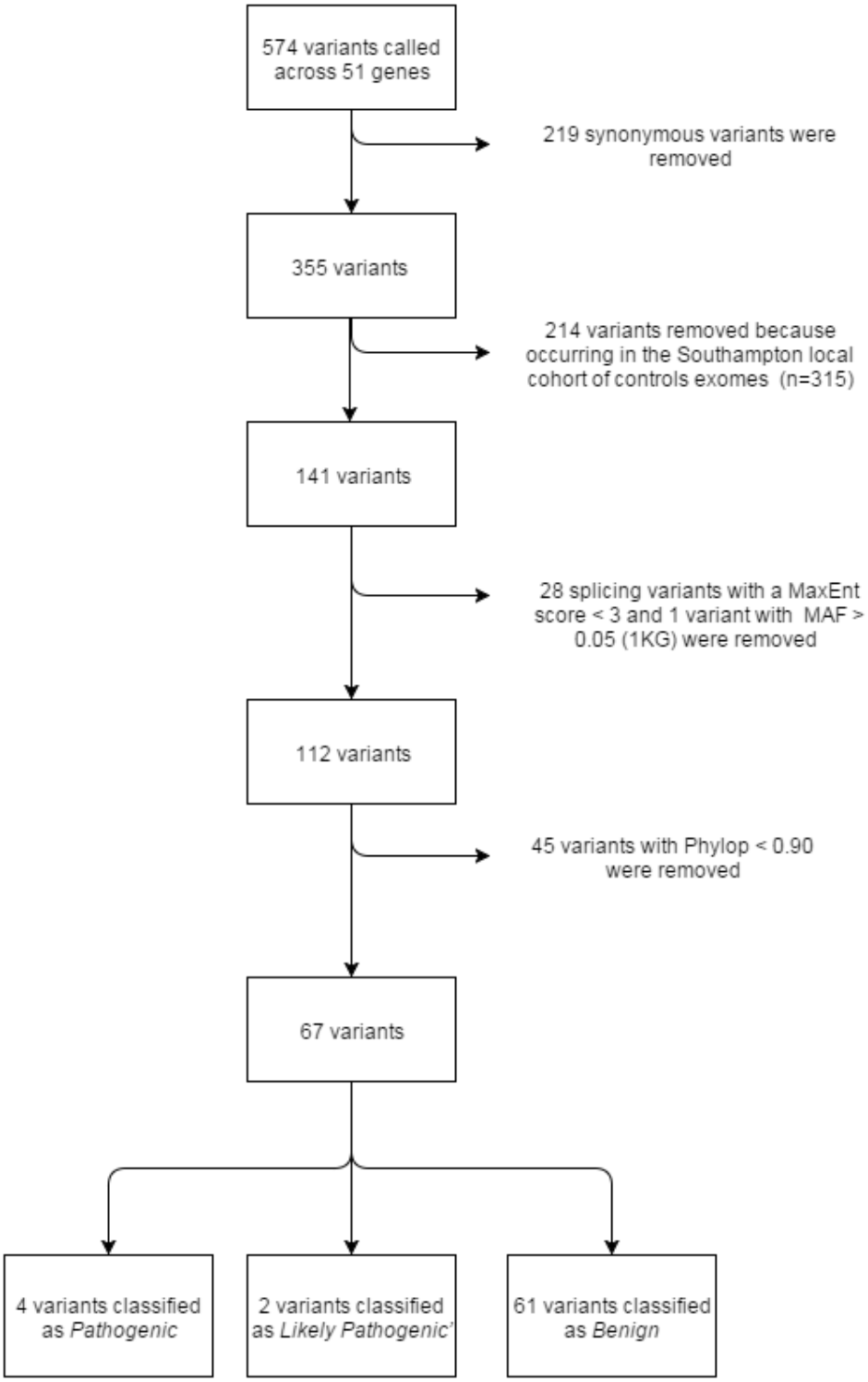



**Figure 5.1 Variant filter steps.** A total of 574 variants in all 51 known monogenic IBD genes were extracted, of which 219 synonymous variants were discarded due to their low likelihood to impact protein function. Of the remaining 355 variants, 214 variants were removed as these were observed in in any zygosity state within the local control cohort of exomes (n=315), 1 variant was discounted due to a MAF > 0.05 (1000 Genome Project) and 28 splicing variants with a MaxEnt score < 3 were removed. Of the 112 mutations remaining, 45 variants were removed due to their low conservation across species (Phylop < 0.90). A total of 67 mutations remained of which four of these were allocated into the 'Pathogenic' category, two into the 'Likely Pathogenic' category and 61 into the 'Benign' category of the ACMG guidelines.

#### *5.3.2.1 'Pathogenic' and 'Likely Pathogenic' mutations*

Four pathogenic, *CYBA* (c.287+2T>C), *COL7A1* (c.6501+1G>C), *LIG4* (p.R814X), and *XIAP* (p.T470S), and 2 likely pathogenic, *FERMT1* (p.R271Q) and *SKIV2L* (c.354+5G>A), variants were identified in six independent probands. Genes *CYBA*, *COL7A*, *LIG4*, *FERMT1* and *SKIV2L* are known to cause disease in an autosomal recessive mode whereas *XIAP* occurs in an X-linked recessive mode of inheritance (Table 5.4).

As *CYBA*, *COL7A*, *LI4G*, *FERMT1* and *SKIV2L* are known to cause disease in homozygous state exome data for the patients harbouring heterozygous variants within these genes were further interrogated in order to identify a second, pathogenic, common variant which might contribute to the phenotype (Table 5.3). We observed common variant in all genes but one (*LIG4*): *CYBA* (2), *COL7A* (1), *FERMT1* (5) and *SKIV2L* (2). Patient specific mutations, characteristics and associated disease types can be seen in table 5.3.

***COL7A1* (c.6501+1G>C)**: Patient 1 is a female patient diagnosed with with ileo-colonic Crohn's disease at 12 years of age with concurrent autoimmune hypothyroidism. She carries a rare splicing mutation in the *COL7A1* gene (c.6501+1G>C) in heterozygous state. This mutation was previously associated with Hallopeau-Siemens recessive dystrophic epidermolysis bullosa, a condition causing blistering of the skin and digestive tract. This patient did not have any dermatological features suggestive of epidermolysis bullosa. The histological findings were consistent with a diagnosis of Crohn's disease. This patient also carries a common ($MAF_{1KG}$ = 0.67) heterozygous synonymous (p.P939P) variant in *COL7A1* which the authors do not believe to contribute to disease phenotype.

***SKIV2L* (c.354+5G>A):** Patient 2 was diagnosed with ulcerative colitis at the age of 5 years. She presented an extremely severe course in the initial years needing prolonged periods of steroid dependency. She carries a novel splicing mutation c.354+5G>A within the *SKIV2L* gene, which is associated with trichohepatoenteric syndrome. This is an autosomal recessive condition presenting with intractable diarrhoea, woolly hair, liver derangements and facial abnormalities. It presents in infancy needing intensive nutritional interventions. Our patient did not have clinical features seen in this condition. She is currently well with a prolonged period of remission on 6-mercaptopurine. Patient 2 harbours a second common, not conserved and benign nonsynonymous variant (p.M214L) in homozygous state and a common synonymous

(p.Y1067Y) variant in homozygous state within *SKIV2L*. There was no significant difference in depth of coverage for exome 8 and 26, in which mutations p.M214L and p.Y1067Y occur, for this patient and two other samples indicating a low chance of multiple exon deletion in 1 allele.

***LIG4* (p.R814X):** Patient 3 is a female patient diagnosed aged 12 with IBDU when she was 6 years of age. She was diagnosed with vitiligo, which is chronic, autoimmune skin condition in which patches of the skin undergo depigmentation. She harbours a known stop gain (p.R814X) variant within the LIG4 gene, associated with so called *LIG4* syndrome. This syndrome is characterised by immunodeficiency, skin abnormalities (including photosensitivity) and IBD presenting with protracted diarrhoea. The course of her disease is currently stable on immunosuppressive medications. It is possible that some features of her clinical presentation particularly the dermatological phenotype are contributed by this genetic variation. The patient carries the variant in a heterozygous state and she does not harbour any other common or rare mutation within *LIG4* identified using WES data analysis.

***CYBA* (c.287+2T>C)**: Patient 4, a female patient diagnosed at the age of 16 with an extremely severe stricturing Crohn's disease requiring urgent resection of her ileum. She carries a heterozygous splicing mutation within the *CYBA* gene (c.287+2T>C) known to be associated with chronic granulomatous disease (CGD), representing a heterogeneous group of immune deficiencies. Granuloma is a histological hallmark of CD characterised by a collection of non-degradable inflammatory cells. However, granuloma formation can be seen in other conditions such as CGD. Patients with CGD have usually present with recurrent bouts of infections such abscess formation, pneumonias and osteomyelitis. Our patient had a stable course following ileal resection, with no evidence of recurrent infections. This patient was last reviewed in the clinic 4 years ago and is currently followed up out of region. A widely used diagnostic test for CGD is nitroblue tetrazolium test (NBT). Our patient has not been formally assessed for CGD. This patient also harbours two common ($MAF_{1KG}$ > 5%) variants within *CYBA*: a non-synonymous (p.V174A*)* and assessed by *in silico* tools to be benign (Gerp= -7.6, Phylop=0.000535) variant and a synonymous (p.E12E) mutation.

***FERMT1* (p.R271Q):** Patient 5 is an early onset UC patient diagnosed at age 9 years. He was subsequently diagnosed with severe oral pemphigus at age 12. He carries a rare (MAF = 0. 000116) non-synonymous mutation within the *FERMT1* (c.G812A, p.R271Q)

gene on chromosome 20 in heterozygous state. This identified variant causes a mutation in the same codon of the known *FERMT1* stop gain mutation (c.C811T, p.R271X) which is known to cause Kindler's syndrome, a blistering skin disease that may present with ulcerative colitis. Within the same gene the patient also carries a second common, conserved nonsynonymous variant (p.R526K) in homozygous state, four common synonymous variants (p.F565F, p.K525K, in homozygous state, and p.L385L and p.H38H in heterozygous state) and two splicing variants with a MaxEnt score < 3 (c.532+8T>C and c.152-4G>A with a score of 1.2 and 1.72 respectively). There was no significant difference in depth of coverage for exome 12 and 13, in which p.R526K and p.F565F occur, between patient 5 and two other samples sequenced at the same time. It is possible that part of the clinical presentations observed within patient 5 could be explained by the variation observed in *FERMT1*. Variants p.R271Q and p.R526K were confirmed by Sanger sequencing in the proband and relatives where applicable (Figure 5. 2 and Figure 1a-b in Appendix IX).

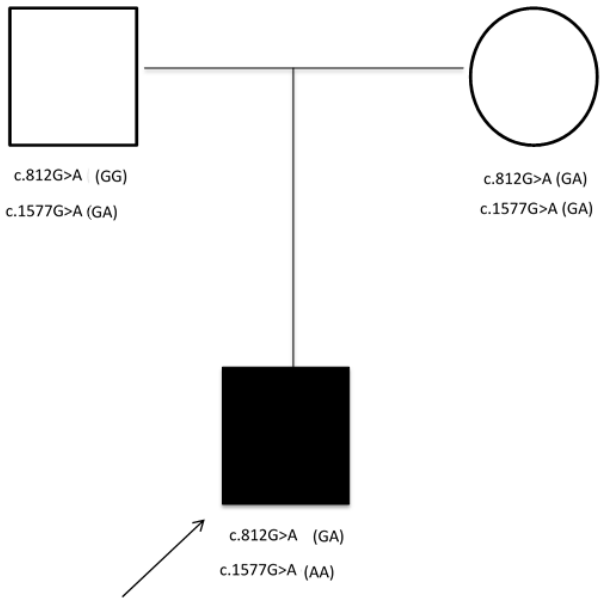



**Figure 5.2 Segregation analysis for *FERMT1* variant c.1577G>A and c.812G>A.** c.1577G>A is in heterozygous status in proband and mother while c.812G>A is in homozygous status in proband and heterozygous status in mother and father. We think that the cumulative effect of these *FERMT1* variants in the proband might contribute to disease risk.

***XIAP* (p.T470S)**: We identified a known non-synonymous variant within the *XIAP* (c.A1408T, p.T470S) gene on the X chromosome, known to be associated with X-linked lymphoproliferative disease type 2 (XLP2) in patient 6. XLP2 is an extremely rare condition, characterised by an inability to mount an immune response to Epstein-Barr virus. EBV infection in these patients can be fatal presenting with bone marrow failure,

hepatitis and malignant lymphoma. Another feature seen in this condition is the presence of dysgammaglobulinemia, characterised by a deficiency of gamma globulins. This condition can also present in some patients with an IBD-like phenotype with severe indolent and fistulating perianal disease. Our patient presented at 4 years of age with a severe fistulating Crohn's disease; he had an extremely severe disease course with recurrent perianal abscesses and fistulae. An isolated IgA deficiency was also detected in this patient, in keeping with features of dysgammaglobinemia. He is currently stable on azathioprine, although he suffers from periodic bouts of recurrent stomatitis. Segregation analysis in the proband and available family member confirmed hemizygous variant in the proband. The variant is absent in the unaffected father and present in the unaffected mother in heterozygous state (Figure 5.3 and Figure 1c in Appendix IX). Of note patient 6 has a half-sister (same father) with Crohn's disease. The sister had an history of lip swelling, underwent surgery in 2012 and she is currently stable on infliximab. Sanger sequencing in the half-sister for the *XIAP* variant was normal.

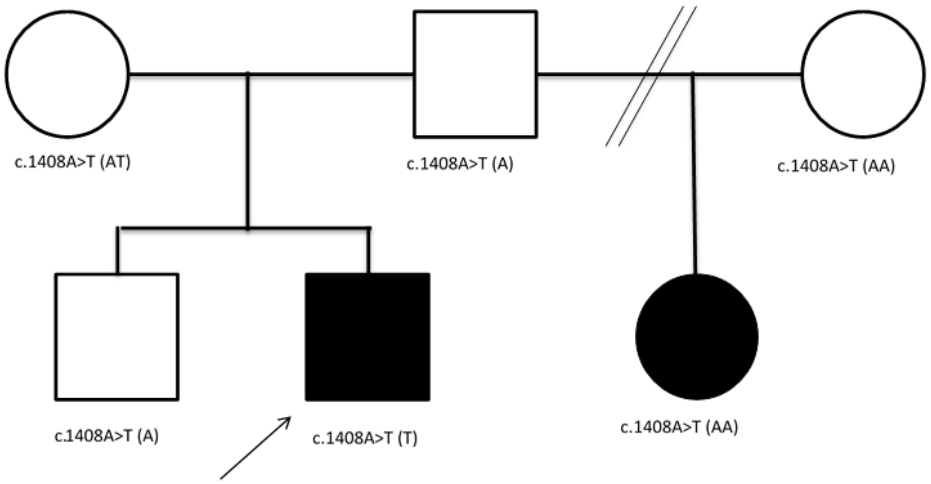



**Figure 5.3 Patient 6 family pedigree**. Segregation analysis for *XIAP* variant c.1408A>T. The variant is present in heterozygous and hemizygous status in the mother and in the proband respectively.

**Table 5.3 Clinical details of patients with 'Pathogenic' and 'Likely Pathogenic' variants**

| Patient ID | Mutation | Novel or Known | ACMG guidelines- Pathogenic or Likely Pathogenic | Previously identified causative variant and functional impact | Phenotype associated with previous gene variant | Disease | Age at diagnosis (years) | Gender | Paris classification at diagnosis | Other clinical features | Clinical course since diagnosis | Family history |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | **COL7A1** (c.6501+1G>C) | Known | P | c.6501+1G>C (45) | **Hallopeau-Siemens recessive dystrophic epidermolysis bullosa** when homozygous- severe skin and digestive tract blistering (31, 45) | CD | 12 | F | L3 | Autoimmune hypothyroidism diagnosed age 7 (Anti-thyroid peroxidase antibodies 2864iu/ml (<75iu/ml)<br>Mouth ulcers | Mild course over 2 years follow-up | - |
| **2** | **SKIV2L (**c.6501+1G>C) | Novel | LP | c.355-2A>C (33) | **Trichohepatoenteric syndrome**- intractable diarrhoea, hair/facial abnormalities presenting in infancy (42) | UC | 5 | F | E4 | No additional features at diagnosis | Turbulent with frequent relapses and prolonged steroid dependency over 11 year follow-up | Paternal ulcerative colitis |
| **3** | **LIG4 (**c.C2440T:p.R814X) | Known | P | c.C2440T<br>p.R814X (32) | **Lig4 syndrome**- immunodeficiency, skin abnormalities (photosensitivity, psoriasis), protracted diarrhoea (32) | IBDU | 12 | F | NA | Vitiligo diagnosed age 6, presented with diarrhoea | Mild course over 5 year follow-up | Maternal great grandfather suffered from ulcerative colitis |
| **4** | **CYBA (**c.287+2T>C) | Known | P | c.287+2T>C (46) | **Chronic granulomatous disease**- recurrent infections, Crohn's-like colitis, perianal disease, granuloma are not always present on biopsy (47) | CD | 16 | F | L3 + structuring disease (B2) | Extremely severe stricturing disease requiring surgery on presentation to remove terminal ileal stricture.<br>Granuloma on histology and extensive granulation tissue on resected specimen | Subsequent right hemicolectomy 1 year after diagnosis<br>Progressed to anti-TNF therapy quickly and now dependant after 4 years follow-up. | - |
| **5** | **FERMT1** (c.G812A:p.R271Q) | Novel | LP | c.811C>T p.R271X (34) | **Kindler's syndrome**- blistering skin disease (34) | UC | 9 | M | E2 | Severe oral pemphigus (blistering skin disease) diagnosed age 12 by immunofluorescence | Disease controlled with azathioprine after initial frequent relapses. 9 year follow-up | - |
| **6** | **XIAP (**c.A1408T:p.T470S) | Known | P | c.A1408T<br>p.T470S (36) | **X-linked lymphoproliferative disease type 2**- dysgammaglobulinemia (can be low) and lymphoma. IBD-type presents with perianal disease (36) | CD | 4 | M | L1 + L4 | Severe perianal disease presenting with abscess and fistula age 3.<br>Mouth ulcers<br>IgA deficiency (<0.07) but other immune work-up normal including neutrophil burst and ANCA | Ongoing perianal disease with subsequent fissures and recurrent fistulae.<br>Turbulent course after 3 years follow-up | Brother has mouth ulcers and has been investigated for IBD aged 7 years. No diagnosis at time of writing.<br>Half-sister (paternal) has severe Crohn's disease |

P: pathogenic; LP: likely pathogenic; UC, ulcerative colitis; CD, Crohn's disease; IBDU, inflammatory bowel disease unclassified.

**Table 5.4 Pathogenic, Likely Pathogenic and 'second hit' variants identified in genes known to cause monogenic IBD**

| Patient | sex | Zygosity* | Gene | Mode on inheritance | Chromosme | Location hg19 | Variant type | Variant info | phylop | 1-sift | ,polyphen2, | mutationtaster, | gerp++ | MaxEnt | dbSNP135 | Frequency in 1KG | Frequency in EVS | Frequency in ExAc | ACHG annotation | Occurrence in controls (n=315) | Occurrence in cases (n=147) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | F | 1 | COL7A1 | AR | 3 | 48611694 | sp | COL7A1:NM_000094:exon80:c.6501+1G>C | . | . | . | . | . | 8.27 | . | Not seen | Not seen | 0.000008245 | pathogenic | Not seen | 1 |
| 1 | F | 1 | COL7A1 | AR | 3 | 48625266 | sn | COL7A1:NM_000094:exon21:c.A2817G:p.P939P | . | . | . | . | . | . | rs1264194 | 0.66 | 0.70825 | 0.6799 | benign | 237 | 138 |
| 2 | F | 1 | SKIV2L | AR | 6 | 31928119 | sp | SKIV2L:NM_006929:exon4:c.354+5G>A | . | . | . | . | . | 3.69 | - | Not seen | Not seen | Not seen | pathogenic | Not seen | 1 |
| 2 | F | 2 | SKIV2L | AR | 6 | 31929014 | ns | SKIV2L:NM_006929:exon8:c.A640C:p.M214L | 0.222828 | . | . | . | 1.8 | . | rs437179 | 0.78 | 0.71116 | 0.7699 | benign | 269 | 135 |
| 2 | F | 2 | SKIV2L | AR | 6 | 31936668 | sn | SKIV2L:NM_006929:exon26:c.T3201C:p.Y1067Y | . | . | . | . | . | . | rs410851 | 0.81 | 0.71030 | 0.7826 | benign | 257 | 135 |
| 3 | F | 1 | LIG4 | AR | 13 | 108861177 | sg | LIG4:NM_002312:exon2:c.C2440T:p.R814X,LIG4:NM_001098268:exon2:c.C2440T:p.R814X,LIG4:NM_206937:exon3:c.C2440T:p.R814X | 0.991331 | 0.903176 | 0.734134 | 1 | 4.25 | . | rs104894419 | 0.0005 | 0.00023 | 0.00008237 | pathogenic | Not seen | 1 |
| 4 | F | 1 | CYBA | AR | 16 | 88713161 | sp | CYBA:NM_000101:exon5:c.287+2T>C | . | . | . | . | . | 7.75 | rs747774702 | Not seen | Not seen | 0.00004987 | pathogenic | Not seen | 1 |
| 4 | F | 2 | CYBA | AR | 16 | 88709828 | ns | CYBA:NM_000101:exon6:c.T521C:p.V174A | 0.000535 | . | . | 0.000001 | -7.6 | . | rs1049254 | 0.71 | 0.62831 | 0.6892 | benign | 135 | 122 |
| 4 | F | 2 | CYBA | AR | 16 | 88717386 | sn | CYBA:NM_000101:exon1:c.A36G:p.E12E | . | . | . | . | . | . | rs8053867 | 1 | 1 | 0.9991 | benign | 229 | 146 |
| 5 | M | 1 | FERMT1 | AR | 20 | 6088216 | ns | FERMT1:NM_017671:exon6:c.G812A:p.R271Q, | 0.998967 | 0.99 | 1 | 0.999994 | 5.34 | . | rs144791466 | Not seen | 0.00011 | 0.00004207 | pathogenic | Not seen | 1 |
| 5 | M | 2 | FERMT1 | AR | 20 | 6064710 | sn | FERMT1:NM_017671:exon13:c.T1695C:p.F565F | . | . | . | . | . | . | rs753927 | 0.44 | 0.36244 | 0.3683 | benign | 187 | 92 |
| 5 | M | 2 | FERMT1 | AR | 20 | 6065729 | ns | FERMT1:NM_017671:exon12:c.G1577A:p.R526K | 0.963443 | . | . | 0.000272 | 5.01 | . | rs2232074 | 0.48 | 0.05162 | 0.3977 | benign | 194 | 93 |
| 5 | M | 2 | FERMT1 | AR | 20 | 6065731 | sn | FERMT1:NM_017671:exon12:c.A1575G:p.K525K | . | . | . | . | . | . | rs2232073 | 0.46 | 0.05395 | 0.3978 | benign | 194 | 93 |
| 5 | M | 1 | FERMT1 | AR | 20 | 6069723 | sn | FERMT1:NM_017671:exon10:c.C1153T:p.L385L | . | . | . | . | . | . | rs35413391 | 0.04 | 0.07651 | 0.05306 | benign | 35 | 24 |
| 5 | M | 1 | FERMT1 | AR | 20 | 6093116 | sp | FERMT1:NM_017671:exon5:c.532+8T>C | . | . | . | . | . | 1.20 | rs41308641 | 0.07 | 0.12930 | 0.0956 | benign | 69 | 33 |
| 5 | M | 1 | FERMT1 | AR | 20 | 6096695 | sp | FERMT1:NM_017671:exon4:c.152-4G>A | . | . | . | . | . | 1.72 | rs2295435 | 0.4 | 0.43 | 0.4489 | benign | 194 | 98 |
| 5 | M | 1 | FERMT1 | AR | 20 | 6100088 | sn | FERMT1:NM_017671:exon2:c.T114C:p.H38H | . | . | . | . | . | . | rs10373 | 0.51 | 0.53209 | 0.5243 | benign | 238 | 114 |
| 6 | M | hemizygous | XIAP | X | X | 123040945 | ns | XIAP:NM_001167:exon7:c.A1408T:p.T470S,XIAP:NM_001204401:exon7:c.A1408T:p.T470S | 0.998574 | 0.94 | 0.004 | 0.274905 | 4.11 | . | rs143165174 | Not seen | 0.00059 | 0.0004818 | pathogenic | Not seen | 1 |

*One (1) denotes heterozygous state; two (2) denotes homozygous state; AR: autosomal recessive; X: X-linked; ns: nonsynonymous, sn, synonymous; sp: splicing; benign and second hit variants are shaded in grey; pathogenic hit variants are shown in white background

## 5.4 Discussion

This study applies whole-exome sequencing to 147 paediatric IBD patients and interrogates a panel of 51 genes identified as being associated with monogenic causes of IBD. We have rigorously applied the recently updated ACMG guidelines for variant classification to our results and identified four known and two novel variants that fit the 'pathogenic' or 'likely pathogenic' categories respectively[335]. These guidelines were created to provide standard filtering criteria for clinical laboratories. Importantly according to ACMG guidelines variants are classified based on the specific variant, with the zygosity of that variant ignored for classification. The mutations identified in this study may contribute to the disease seen in individuals, however this is a hypothesis and further validation is required to confirm or refute this.

Patient 6, harbouring a known causative mutation in the *XIAP* gene for X-linked lymphoproliferative disease type 2 has a phenotype which closely resembles that described for the condition; early onset severe Crohn's like symptoms, dysgammaglobinaemia and perianal disease. This condition is extremely rare, previous estimates of disease associated with *XIAP* mutations, including the more common X-linked lymphoproliferative disease type 1, have put the prevalence of *XIAP* mutations at 4% of paediatric male IBD patients, although this is likely to be a significant overestimate due to selection bias-overrepresentation of young, severe patients in studies sequencing exomes/genomes[160]. Here we report 1 patient with a causative *XIAP* mutation (1.2% of frequency in 85 EO males and 7.6% frequency in 13 VEO males), our data are subject to the same selection bias as previous studies.

The other five patients, all harbouring potentially causative mutations presented with a range of symptoms. Patient 5 presented with pemphigus, a skin condition which causes blistering and and lining of the mouth, nose, throat and genitals, and symptoms of IBD. He was identified as being heterozygote for a novel nonsynonymous variant within the same codon as that of a known causative mutation for Kindler's syndrome[336]. Therefore, the mutation we observe in patient 5 affects the same amino acid change as the previously established pathogenic variant which could result in a milder phenotype and less severe symptoms. Kindler's syndrome is a rare autosomal recessive skin disorder characterised by blistering which improves with age,

photosensitivity, skin and mucosa fragility. Whilst this patient's phenotype was not as severe as typical Kindler's syndrome there is reasonable suspicion that the heterozygous mutation will have contributed to the phenotype seen, perhaps in conjunction with the additional common variants observed in the *FERMT1* gene of this patient.

Patient 2, who presented with colitis at the age of 5 was found to have a novel mutation in *SKIV2L*. A very similar homozygote mutation is known to cause trichohepatoenteric syndrome, presenting at 1-12 weeks of age[337]. Whilst this child does not present with classical symptoms of this condition, his disease course has been extremely severe with frequent relapses. Interestingly there have been previous reports of milder phenotype with colitis presenting at the age of 4.5 years[337,338]. Also of note *SKIV2L* is within 1 megabase of the HLA complex genes although we do not detect any variants in this region.

If all identified 'likely pathogenic' and 'pathogenic' variants have contributed to the development of IBD in these patients, we can estimate the prevalence of paediatric IBD contributed to by a 'monogenic' variant at 4%. However, in our study, VEO patients and patients with severe phenotype were preferentially selected for exome sequencing, possibly leading to an overrepresentation of potential monogenic causes in this cohort. IBD has is typically considered a polygenic disorder, however the 201 IBD-associated loci identified by GWAS only account for a small part of the heritability seen in IBD [339]. Rare mutations in genes associated with monogenic IBD would not be detected by GWAS and therefore may account for some of this missing heritability.

Excluding the *XIAP* variant (disease causing in hemizygous state), all of the variants identified are previously reported to cause disease in a homozygous state, however mutations within our remaining five patients are all in heterozygous state. By relaxing our filtering criteria, we identified a second, more common variant in 4 of the 5 genes (see table 5.4). There is precedent in genetic causes of nystagmus and albinism for common variants contributing to disease when they coexist with a highly deleterious heterozygote variant[339].

We postulate that heterozygote 'pathogenic' and 'likely pathogenic' variants seen in genes associated with monogenic IBD may account for an attenuated phenotype (of

the full condition) with variable penetrance of the mutated allele. Our hypothesis that heterozygote variants in genes known to cause monogenic disease in a recessive mode of inheritance, may still effect clinical manifestation to some lesser degree has been observed in other conditions such as adenomatous polyposis coli[340], Parkinsonism and disorders of eye development[341]. Across the 22 patients with VEO-IBD, only two patients harboured a mutation falling within the 'pathogenic' category, suggesting that mutations within functional regions of the genome or variants in genes not associated with IBD yet might induce disease risk in these children. Exome sequencing by definition overlooks non-coding but potentially functional regions of genes; mutations in these regions as well as their intergenic regulatory sequences may be contributing to disease susceptibility, severity and co-morbidities. In addition, it is likely that additional mutations in modifier genes are also determining aspects of clinical presentation. Extensive functional studies are required for definitive interpretation of mutations and for common diseases, it may not be informative to assess single variants in isolation. The bottleneck to clinical translation of personalised genomics imposed by a relative deficiency in functional assessment of variants of unknown significance, is already being observed for rare diseases and the scope of this problem increases in complexity for common disease.

Recent reviews by Uhlig and collaborators have highlighted the importance of considering monogenic causes of IBD, especially in very early onset IBD. It has long been accepted that intractable diarrhoea of infancy (early onset IBD) was likely to have a specific genetic basis in many patients with a significant proportion having underlying IBD[342]. Previous work has highlighted the genetic heterogeneity of CD and hypothesised that some of the missing heritability is associated with monogenic disease[343]. Our study highlights the utility of whole-exome sequencing in identification of novel and known variants in this panel of genes associated with monogenic causes of IBD.

## 5.5 Conclusions

Although all patients presented with IBD it is important to recognise that many of the monogenic conditions associated with IBD have broader phenotypes that may lead to

subsequent development of other problems, most often associated with immune dysfunction or deficiency. We hypothesize that some cases of IBD may be contributed to by heterozygous mutations in genes previously associated with severe monogenic IBD. Further work is needed to functionally examine potentially pathogenic variants in genes associated with monogenic forms of IBD in other large cohorts. Early identification of these conditions, potentially via routine exome sequencing, may be of huge benefit to individual patients, preventing mismanagement and enabling potentially curative treatments.

# Chapter 6 **Thesis summary and future perspective**

## 6.1 Thesis summary

This thesis describes five distinctive projects that share the common theme of unravelling the genetic background of paediatric IBD. The analysis of genetic variation is a crucial step for understanding disease susceptibility and progression. In Chapter 1, I present the IBD phenotype and a brief summary on the techniques used to understand the genetics of complex traits, from linkage studies to the applications of the whole exome and whole genome sequencing technology. In chapters 2 and 3 I describe the application of a gene-based test on known associated genes and on an established IBD pathway using whole exome data from our local cohort of IBD children. The application of gene-based statistical tests using targeted sequencing has allowed us to confirm already known IBD genes detected by GWAS, to show that known genetic loci may be enriched for rare variants with high effect size and to identify *de novo* disease-associated genes. Once a set of significantly associated genes is found, replication analysis, extensive literature review and functional assays are needed in order to fully understand the implication of these genes on the phenotype. As an example of this requirement for functional studies, in Chapter 4 I present our collaboration between the Genomic Informatics group at Southampton University and the University of Stanford (USA) in which functional analysis further proved the significant statistical association for a previously unknown IBD gene suggesting that this gene can render individuals susceptible to disease.

In Chapter 5 I describe the application of the next generation sequencing technology for identifying known and likely pathogenic mutations within known genes associated with monogenic IBD. A summary of all genes nominally significant ($P < 0.05$) resulting from the gene based association analyses are shown in Table 6.1. The p value for the 11 nominally significant genes discussed within this thesis do not withstand the GWAS significant threshold ($1x10^{-8}$) or the exome-wide significance ($1x10^{-6}$). These findings describe the need of a larger sample size (which is often inaccessibly) in order to maximize the statistical power of the test. For this reason, we targeted our analysis to a restricted number of genes and we replicated most significant genes in an independent replication cohort. For this reason, the use of the exome-wide

significance threshold would not had been appropriate in this scenario. Importantly, three (*BIRC2*, *SUGT1* and *HSPA1L*) of the 11 described most significant genes are previously unreported IBD genes. By looking at the frequency of the variations within *BIRC2*, *SUGT1* and *HSPA1L* across cases and controls it appears that mutations within *SUGT1* might have a protective effect, whereas mutations within *BIRC2* and *HSPA1L* might increase disease risk. However, functional studies and association studies in a bigger cohort are needed to fully understand the function of these genes. As next generation sequencing is increasingly applied to Mendelian disorders, our analyses and others demonstrate the potential for genetic stratification of patients with complex disease using this technology and personalized diagnosis, management and prognosis will increase. The relative power of this methodology is likely to be most easily realized in a paediatric population where the genetic contribution to disease is highest.

**Table 6.1 Nominally significant genes (p value < 0.05) from SKAT-O association analyses.**

| | Gene | Chr | Bp position (hg19) | Total number of samples | Fraction of individuals who carry rare variants under the MAF thresholds (MAF < 0.05)* | Number of all variants defined in the group file | Number of variant defined as rare | SKAT-O P value unadjusted | SKAT-O Weighted (W) or Unweighted (UW) p value | Putative direction | P value from IBD GWAS/meta-analysis (reference) | Pathway/function |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **SAMPLE SIZE OF 18 CASES AND 56 CONTROLS** | | | | | | | | | | | | |
| Chapter 2 | *ZPBP2* | 17 | 38024626-38032996 | 74 | 0.027 | 4 | 1 | 0.009 | W | Protective | 6 x10-23 ([344]) | Zona pellucida binding protein 2 |
| | *PYHIN1* | 1 | 158906777-158943483 | 74 | 0.108 | 5 | 5 | 0.025 | UW | Causal | 4 x10-9([251]) | Involved in the transcriptional regulation of genes important for cell cycle control, differentiation, and apoptosis |
| | *IL1R1* | 2 | 102781629-158943483 | 74 | 0.014 | 5 | 2 | 0.037 | W | Causal | 2 x10-16([345]) | Cytokine Cytokine receptor interaction |
| | *IL2RB* | 22 | 37524329-37539651 | 74 | 0.014 | 4 | 1 | 0.037 | UW | Protective | 3 x10-9([346]) | Cytokine Cytokine receptor interaction |
| | *IL18R1* | 2 | 102984279-103001402 | 74 | 0.014 | 3 | 1 | 0.039 | UW | Protective | 2 x10-15([347]) | Cytokine Cytokine receptor interaction |
| | *GSTP1* | 11 | 67352183-67353970 | 74 | 0.014 | 4 | 1 | 0.041 | W | Protective | 1 x 10-8([346]) | Involved in oxygen species detoxification |
| **SAMPLE SIZE OF 136 CASES; 106 CONTROLS** | | | | | | | | | | | | |
| Chapter 3 | *BIRC2* | 11 | 102220918-102248410 | 242 | 0.07851 | 6 | 6 | 0.004 | W | Protective | NA | NOD signalling pathway |
| | *NFKB1* | 4 | 103488139-103537672 | 242 | 0.11983 | 10 | 9 | 0.005 | UW | Causal | 4 x10-12([59]) | NOD signalling pathway |
| | *NOD2* | 16 | 50733392-50763778 | 242 | 0.21488 | 31 | 25 | 0.029 | W | Causal | 6 x10-209 ([59]) | NOD signalling pathway |
| | *SUGT1* | 13 | 53231709-53261936 | 242 | 0.33058 | 6 | 5 | 0.047 | UW | Protective | NA | NOD signalling pathway |
| **SAMPLE SIZE OF 136 CASES; 106 CONTROLS** | | | | | | | | | | | | |
| Chapter 4 | Using non-synonymous and non-frameshift variants, excluding low frequency (0.01 < MAF < 0.05) and common (MAF > 0.05) variants. | | | | | | | | | | | |
| | *HSPA1L* | 6 | 31778076-31779521 | 242 | 0.024793 | 5 | 5 | 0.024 | W | Causal | NA | HSP70 family |
| | Using non-synonymous and non-frameshift variants, excluding common variants (MAF > 0.05). | | | | | | | | | | | |
| | *HSPA1L* | 6 | 31778076-31779728 | 242 | 0.033058 | 7 | 6 | 0.034 | W | Causal | NA | HSP70 family |

Mat analysis GWAS p value taken from the GWAS catalog https://www.ebi.ac.uk/gwas/

## 6.2 Study limitations

Although we successfully applied exome sequencing in our cohort of pIBD patients and identified biologically significant genes, this study was limited by various factors:

- The use of data generated from whole exome sequencing. Multiple studies have shown that noncoding regions of the genome might pay an important function in in complex diseases. However, despite this limitation, whole genome sequencing for a large cohort is still prohibitively expensive and whole exome sequencing is a feasible method for the study of rare variants.
- The use of a modest sample size which might have impacted the power of the statistical association test. Although our data have derived from whole exome sequencing, we did not conduct the gene based association analysis on all genes that underwent whole exome sequencing due to our modest sample size. Although no power test was executed, we did not have enough power to detect an association which satisfy the exome-wide significant threshold ($1x10^{-6}$) and we would have not expected to observe variants with high penetrance within our cohort. For these reasons we decided to adopt an independent replication cohort to validate significant genes.
- The use of a replication cohort gleaned from different sources. Although an established method to take into account such differences (e.g. depth of coverage and capture kit used) is not yet available,[348,349] we minimized bias by analysing only variants that occurred in the regions common to all capture kits and by conducting a joint recalling of the data. Future statistical tests that account for such differences in larger sample sizes are needed to fully assess the contribution of rare variants to disease phenotype.
- The use of an uneven number of cases and controls might have induced bias towards the discovery of putative protective or disease risk associated variants/genes. However, we removed population bias by selecting an ethnically homogeneous cohort and by selecting non-IBD controls without a primary immumodeficiency diagnose. By selecting controls without an autoimmune condition, we removed bias of including diseases with large shared genetic component with IBD.

- The use of Bonferroni correction for multiple testing correction. The Bonferroni method consist in reducing the significance level according to the number of independent tests conducted. However, there might be an overcorrection when the tests are correlated. As we conducted the tests on genes related to the same pathway we could have executed the test on a pathway based instead on a gene by gene method reducing the number of multiple testing correction needed. Alternative methods to the conserved Bonferroni approach is the use of permutation based correction methods. The permutation methods randomly shuffle the data but these methods are computationally intensive as a large number of permutations are needed to accurately estimate p value. However, by limiting our analysis of sequencing data to key pathways implicated by GWAS, we massively reduce the genomic search space for causal variation to concentrate on regions with high prior probability of containing pathogenic mutation while maximising sensitivity to rare and private mutations.
- Although only one functional study is presented within the research projects, by conducting the study in a local cohort with ethical approval for additional biological sample collection, we have biological samples, facilities and staff to confirm immunological functional relevance of precise gene variants by selecting the exact patients in whom these variants occur.
- The use of covariates in the statistical association studies. Within our statistical association analyses we did not apply covariates as a parameter for the association test. We could have used as covariates the different capture kit used within the cohort, age of onset or mean depth of coverage of each individual in order to try to reduce batch effects and make the data more uniform. However, due to our limited sample size little benefit would have been brought to the power of the test.

## 6.3 Future prospects for inflammatory bowel disease genetics

The rise of next generation sequencing technologies has improved our understanding of the genetic pathology of diseases[253]. In the past decade, linkage studies and then, more recently, genome-wide association studies have been the leading approaches for searching for causal genes in IBD. GWAS were often referred to as 'hypothesis-free' because the entire genome is scanned for genetic variants without *a priori* knowledge of a candidate gene[235]. GWAS in IBD have been hugely successful and today there are more than 200 known IBD loci. However, due to GWAS limitations, these loci explain less than 20% of IBD genetic variability[136]. Currently next generation sequencing is used as a diagnostic tool in medical genetics for Mendelian disorders. There is potential to apply this new technology to complex traits such as IBD, diabetes and heart disease in order to identify the importance of low frequency, rare, and private variants. However, to date, complex disease association studies have largely been limited to analysis of common variation and the study of rare variants is practically and computationally challenging. Many obstacles are still present in order to fully incorporate genomic findings into medical practise. As an example, the lack of a single repository database for genomic and clinical data, the absence of electronic health questionnaire, the report of secondary findings and the interpretation of variants of unknown significance are only a few examples of the issues that require further effort to maximise potential. To try to overcome some of these issues, Genomics England, which manages the 100, 000 Genome Project, is currently setting up a secure web-based information system for a comprehensive collection of phenotypes and clinical data which will be easily accessed by researchers and NHS healthcare teams. As next generation sequencing will become routinely applied in a diagnostic settings, methodological and ethical issues need to be addressed to facilitate the application of genomic technologies into clinical practise.

*-To extend the sample size with more collaborations*

Over the past decade genome-wide association studies have dominated the genetic field by successfully identifying a large number of common variants (MAF > 5% within a population) that impact complex disease. However the joint effect of these variations explains only a small proportion of disease heritability leading to speculation that rare

genetic variation might account for part of the "missing heritability"[139]. Today, whole exome and whole genome sequencing have become the most cost-effective methods for identifying common and rare variations. Within Chapter 1, I have described the reasons why rare variations are likely to play an important role in complex diseases and over this thesis I have confirmed the importance of rare variations in pIBD patients. However, recent sequencing studies within known risk loci have identified very few novel independent rare variants highlighting the need of larger cohorts or to select individuals with extreme phenotypes in order to study such rare mutations[348]. Therefore national and international collaborations are necessary to increase sample size and have well-powered studies[348].

-- *To conduct WGS for the study of non-coding variations*

Moreover, given the importance of non-coding variation in complex and Mendelian disease[349] risk there is an increase need for whole genome sequencing approaches. However, high coverage whole genome sequencing is currently still prohibitive therefore low coverage whole genome sequencing might represent a possible approach. In low coverage whole genome sequencing imputation methods are required in order to optimise genotypes calls[350].

-- *To develop and apply new statistical methods for the study of rare variants and to extend this analysis to the other IBD pathways*

The development of new statistical methodology for assessing rare variations will help to identify new genes enriched for rare risk variants. The use of statistical tests that evaluate association of multiple variants in a genomic region help to address the issue of large sample sizes and of the need of rare variants of large effect size. Such tests present challenges that include the choice of the test, defining the appropriate genomic region to analyse and the type of variation to test. As rare variants are less constrained to natural selection, these variations are more subject to population stratification, so that methods such as PCA or MDS are needed to account for this selection bias[351]. Another point to take into consideration is that different sequencing studies might use different sequencing technologies and methods which influence coverage, read lengths and variant calling and might impact variant characteristics.

There is the need of methods that take into account these differences when cases and controls are sequenced separately. Gene-based tests can be applied on known disease pathways (as shown in Chapter 3), for this reason extending such analysis to the other known IBD pathways (e.g. IL10, autophagy, epithelial cells, and tight junction pathway) may help to identify novel disease associated genes.

-- *To conduct more functional and animal studies*

Assessment of the functional significance of mutations will require substantial improvements to *in silico* annotation and the execution of rigorous and extensive functional validation assays. Depending on the nature of the putative causal variant various experimental validation techniques, such as the emerging method CRISPR/Cas for DNA editing, can be applied. The insertion of a genetic construct into the cells of a patient by using either viral or non-viral systems is defined as gene therapy. The term gene therapy describes any procedure intended to treat or alleviate disease by genetically modifying the cells of a patient[352]. Cells are removed from the patients and genetically modified using small DNA, RNA or oligonucleotides[352]. The selected cells are then cultured and returned to the patient. The idea underlying this approach is to deliver genes to appropriate target cells with the aim of obtaining optimal expression of the introduced genes. There are two approaches for transferring the genetic material: *in vivo*; where the genetic material is transferred directly into cells of the patient and *ex vivo,* in which the cells are removed from the patient and the genetic material is inserted into them in vitro. The modified cells are then transplanted back into the patient[353]. Recent advances in stem cells technologies have made it possible to apply gene therapy to human pluripotent stem cells taken from blood or skin of the affected patient[354]. Animal models have been widely used over the years to test novel therapeutic approaches and understand the nature of diseases. Disease models are usually created by inserting a transgene, an external DNA, into the zygote of the animal. This approach is widely used on a range of different animals[355]. Knocking down the relevant genes identified in model organisms will further enable understanding of how these genes affect the organism as a whole. Cell based assays can also be used to measure the parameters of the immune cascade. In our research study cell assays are conducted by the local immunology team extracting peripheral blood mononuclear

cells (PBMCs) from blood of patients of interest. PBMC comprises of monocytes, B and T lymphocytes and natural killer cells[356]. The complex is controlled mainly through the release of a wide variety of cytokines, such as TNF alpha, chemokines and growth factors. The assays are designed to measure the level and activity of every cytokines and cytokines released from the cells of the targeted patient. This analysis will aid to uncover potential defects in anti-inflammatory control mechanisms[86].

-- *To use the local cohort of adult IBD as validation cohort*

Since 2015 adult IBD patients have been recruited to the Southampton research study with an increasing cohort currently consisting of 356 individuals. The results of the statistical tests conducted on the paediatric cohort could be replicated in the adult cohort to better understand the strength of the association and to elucidate if the association is early-onset IBD specific.

-- *To routinely screen monogenic IBD patients*

Although most IBD cases are likely to have a polygenic basis for their condition, there is a group of rare genetic disorders that can contribute to very early-onset IBD (before 6 years). Genetic variants that cause these disorders have a wide effect on gene function. Thanks to the application of NGS technologies, 51 genes have been identified and associated with IBD-like immunopathology[160]. Monogenic defects have been found to alter intestinal immune homeostasis through many mechanisms. Candidate gene screening for the monogenic IBD genes (as shown in Chapter 5) should be routinely carried out in order to quickly identify patients harbouring causative mutations.

-- *To extend pharmacogenomics analysis*

Whole genome and whole exome sequencing permit the study of response to specific individual drugs. To date pharmacogenomics has been widely applied to the study of *TPMT* gene and thiopurine therapy, therefore a wider application of this approach to other widely used IBD drugs such as infliximab the *MDR1* gene, which have been shown being associated with failed medical therapy[357], will help to determine patient drug toxicity and efficacy. However, due to the complex nature of drugs affects and of

IBD pathogenesis, replication studies and extensive clinical trials are needed before being able to translate the genetics findings into effective therapeutics.

-- *To study microRNAs*

The study of microRNA (miRNA) is a research field on expanding. MiRNA are implicated in the pathogenesis of many common diseases due to their regulatory role such as differentiation and apoptosis. MicroRNAs are a class of endogenous small non-coding single-stranded RNA molecules, ~18–24 nucleotides long, which act as post-transcriptional regulators of gene expression. It is estimated that miRNAs regulate more than 60% of protein-coding messenger RNAs and that more than one-third of human genes are targets for miRNA regulation[358]. However, little is known about miRNA specific function. There is evidence supporting their involvement in cancer and autoimmune diseases. In the intestinal tracts miRNA were found to be involved in tissue homeostasis and intestinal cell differentiation[358]. Moreover different miRNA expression has been found between IBD patients and healthy controls[358]. As gene encoding miRNA are captured by the kit used for whole exome sequencing, a deeper study of the variation within miRNA molecules in pIBD might offer the possibility to use such small molecules as biomarkers and therapeutic target in IBD.

-- *To study the epigenome of pIBD*

Although genetics play an important role in IBD, epigenetic mechanism like DNA methylation, histone modification and altered expression of miRNAs could explain the connection between genes and environmental factors in triggering the development of IBD. It is known that variation in DNA methylation is a cause of human disease and is likely to play an important role in the cause of complex disorders. Several well-known disorders of imprinting are known including Temple syndrome, Silver–Russell syndrome and Prader–Willi syndrome. Imprinted genes are thought to play an important role in foetal growth and their carefully regulated expression is important for normal cellular metabolism and human behaviour. A differential methylation status between normal and inflamed CD and UC tissues has been shown[359,360]. Therefore the identification of methylated genes and pattern within pIBD patients might be a valid

approach to identify specific allele expressed in the gut, confirm diagnoses, stratify disease course and lead to new therapeutic strategies.

-- *To study the microbiome*

As described in Chapter 1, the intestinal microbe species plays an important role in the development of disease. Patients with IBD present a different composition in their intestine with an altered balance of microbiota in the gut which might contribute to IBD development[361]. Therefore the study of microbiota changes between IBD and healthy controls and between affected and non-affected individuals of the same family will aid to identify and better characterise the patient specific IBD microbe.

-- *To study the influence of environmental factors such as Vitamin D*

In the literature may studies support the importance of vitamin D in the IBD pathogenesis indicating a high rate of vitamin D deficiency in IBD patients[362,363]. Analysis focused on the association of the month of birth and the incidence of IBD might provide insights into IBD subtypes and month of birth with our local cohort.

-- *To use machine learning for IBD risk prediction*

The combination of information gleaned from genetics, epigenetics, transcriptomics, medical notes and quality of life questionnaire could help to develop methods for an accurate IBD risk prediction making targeted intervention realistic. Prediction methods for complex disease usually deals with the assignments of a score to an individual based on characteristics such as genotypes, type of variation, disease severity and family history. The accuracy of the model can be evaluated with the use of the receiver operating characteristic curve (ROC), which assesses the true and false positive ratio of the model at various cut-off scores[364], and the area under the ROC curve (AUC) which represents the probability of a randomly selected pair of healthy and disease individuals in which the disease person will have a higher AUC value. An AUC value of 1 means that the model is able to perfectly discriminate between diseased and healthy individuals.

Prediction models can be developed using approaches derived from the machine learning field. Machine learning defines a series of approaches for inference and

prediction using mathematical tools for pattern identification within big data[365]. This technique has been applied to a broad range of fields within genetics and genomics. An example is the use of machine learning to annotate a wide variety of genomic sequence elements such as transcription start sites and to understand gene expression[365]. To date there are very few published studies applying machine learning to predict individual risk of complex disease. The study conducted by Wei *et al* used machine learning on the International IBD Genetics Consortium's Immunochip project data to perform risk assessment for CD and UC[366]. The analysis produced a predictive model with an AUC of 0.86 and 0.83 for CD and UC, respectively taking into account only genotype information. The Critical Assessment of Genome Interpretation (CAGI, https://genomeinterpretation.org/) is the community of blinded predicted experiments aiming to predict phenotypes from human variation. The community aims to develop and test new methodologies for disease classification using machine learning approaches. The fourth and latest edition of CAGI has seen the participation of over 100 submissions for the 11 proposed challenges. The challenges of the 2015/2016 CAGI edition entailed in: distinguishing cases from patients diagnosed with bipolar disease and cases from patients diagnosed with Crohn's disease using whole exome data, identifying Walfarine response using whole exome data of a cohort of African-Americans, extrapolating the eQTL causal SNPs, assessing the impact of missense mutations on the NAGLU, NPM-ALK, Pyruvate kinase and SUMO ligase enzymes and predicting individual phenotypes using the Hopkins clinical panel and whole genome data. Up to date, CAGI methodologies have been successful when tailored to the specific challenge and the performance of a model changes between challenges. Although some algorithms showed significant results, these have very little clinical implications due to their low predictive value on such complicated and heterogeneous data.

These data suggest the need of more accurate models based on larger testing cohorts, or on cohort of individuals with extreme phenotypes, and on the integration of clinical and other omics information to improve model's prediction. Therefore, data sharing is becoming fundamental to have powered studies.

--*Conclusion*

With the reducing cost of whole genome and whole exome sequencing, it is almost certain that these techniques will be routinely applied in diagnostic settings and large cohorts of whole genome sequenced cases and controls will be available. This is further supported by the 100,000 whole genome project which the UK Government announced in 2012 to sequence all 100,000 patients with rare disorders and various cancers by the end of 2017. The project is still currently in its beginning phase and aims to bring immediate benefits to the patients and their families. Although challenges will remain in order to translate the genetic discoveries into effective therapeutic approaches, such studies will enable the way to personalized treatments based on an individual's genetic background.

# Appendix I

## Quality control data for the 18 patients affected by IBD and asthma. All exomes were captured with SureSelect Human All Exon51Mb v4 kit (Agilent).

| Sample | Total no. read seqs | Total no. aligned reads | Total no. unique align. | Mapped to target reads +/- 150bp (%) | Mapped to target reads (%) | Target bases with coverage >1 (%) | Target bases with coverage >5 (%) | Target bases with coverage >10 (%) | Target bases with coverage >20 (%) | Mean coverage | Pipeline Gender | X Heterozygosity | % Autosome heterozygosity |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **PR0032** | 61891260 | 61170010 | 60697939 | 88.57 | 79.88 | 99.86 | 99.17 | 97.45 | 91.23 | 78.93 | M | 18 | 59.64 |
| **PR0007** | 61168596 | 60439477 | 59932617 | 86.48 | 77.55 | 99.86 | 99.14 | 97.30 | 90.67 | 76.23 | M | 20 | 59.85 |
| **PR0031** | 54550262 | 53889492 | 53477697 | 87.87 | 78.47 | 99.84 | 99.03 | 96.88 | 89.02 | 68.46 | M | 18 | 59.89 |
| **PR0085** | 58398084 | 57615280 | 57206330 | 91.32 | 81.39 | 99.84 | 99.06 | 97.10 | 89.94 | 71.48 | M | 19 | 60.02 |
| **PR0011** | 69670912 | 69051849 | 68552813 | 91.42 | 82.32 | 99.86 | 99.29 | 97.92 | 92.90 | 88.14 | M | 28 | 60.99 |
| **PR0110** | 65421500 | 64811390 | 64250237 | 91.30 | 81.49 | 99.77 | 99.10 | 97.49 | 91.44 | 78.40 | F | 229 | 61.17 |
| **PR0158** | 51702888 | 51360929 | 50940429 | 89.10 | 80.82 | 99.70 | 98.56 | 95.65 | 86.00 | 67.09 | F | 220 | 60.58 |
| **PR0160** | 54556114 | 54187984 | 53702988 | 88.11 | 80.32 | 99.81 | 98.69 | 95.79 | 86.39 | 70.34 | M | 29 | 61.26 |
| **PR0167** | 55264388 | 54888215 | 54452943 | 90.61 | 82.14 | 99.83 | 98.76 | 95.97 | 86.55 | 68.89 | M | 29 | 60.67 |
| **PR0188** | 62180148 | 61748635 | 61191406 | 87.84 | 79.98 | 99.84 | 99.01 | 96.90 | 89.63 | 80.64 | M | 29 | 60.91 |
| **PR0036** | 43600438 | 43332569 | 42984315 | 89.33 | 81.41 | 99.67 | 98.16 | 94.32 | 82.26 | 57.72 | F | 245 | 60.86 |
| **PR0039** | 50980348 | 50646264 | 50195689 | 86.63 | 79.07 | 99.80 | 98.58 | 95.43 | 85.34 | 65.96 | M | 27 | 60.22 |
| **PR0068** | 39114706 | 38536574 | 38207151 | 87.40 | 79.45 | 99.78 | 99.23 | 97.79 | 92.56 | 98.08 | F | 252 | 61.73 |
| **PR0083** | 39114706 | 42945008 | 42450716 | 85.03 | 77.58 | 99.80 | 99.33 | 98.16 | 93.82 | 106.53 | F | 267 | 61.95 |
| **PR0107** | 28699795 | 28550198 | 28262880 | 84.36 | 76.89 | 99.82 | 98.76 | 95.95 | 86.82 | 72.49 | M | 23 | 60.50 |
| **PR0146** | 37280008 | 37021161 | 36686018 | 86.57 | 78.98 | 99.86 | 99.22 | 97.61 | 91.96 | 94.78 | M | 23 | 59.99 |
| **PR0148** | 42605042 | 40826661 | 40411155 | 84.21 | 77.20 | 99.88 | 99.31 | 97.87 | 92.77 | 102.51 | M | 31 | 60.77 |
| **PR0151** | 36355666 | 36122486 | 35800555 | 88.72 | 83.31 | 99.85 | 99.16 | 97.39 | 91.51 | 97.21 | M | 31 | 61.02 |

Total no. read seqs- total number of reads sequenced; Total no. aligned reads - the total number of reads aligned to the reference sequence; Total no. unique align- the number of reads that uniquely mapped to the reference sequence; Mapped to target reads +/-150bp (%)- the percentage of reads mapped ±150 base pair to the target; Mapped to target reads (%)- the percentage of reads mapped to the target sequence; Target bases with coverage >1.5.10.20- the percentage of targets with 1. 5. 10 and 20 read depth; Mean coverage - the mean of the depth coverage; X heterozygosity - calls mapped to X chromosome heterozygous; Pipeline. Gender – apparent gender based upon X heterozygosity; % Autosome heterozygosity – genome wide % of calls heterozygous.

## Appendix II

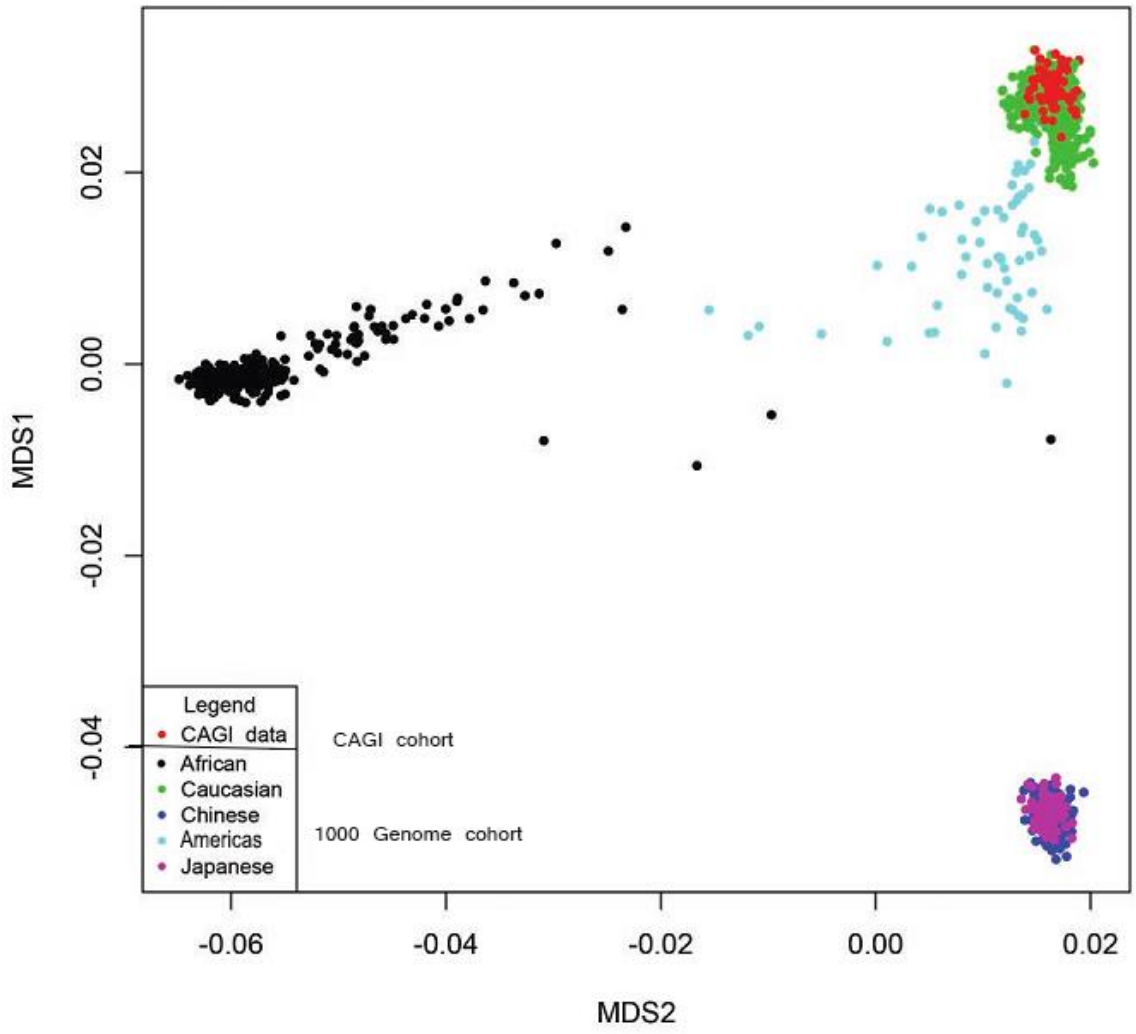
**Similarity matrix: percentage of shared variant across samples sequenced on the same plate.**

| | PR0158 | PR0159 | PR0160 | PR0161 | PR0165 | PR0167 | PR0186 | PR0188 | PR0036 | PR0039 | PR0049 | RL0013 | RL0015 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PR0158 | 100 | 44.27 | 44.02 | 44.11 | 43.26 | 43.26 | 43.26 | 44.18 | 43.92 | 43.91 | 42.88 | 42.32 | 43.39 |
| PR0159 | 44.03 | 100 | 43.96 | 44.01 | 43.3 | 43.08 | 43.38 | 44.39 | 44.3 | 43.82 | 42.84 | 42.7 | 43.06 |
| PR0160 | 43.92 | 44.1 | 100 | 43.34 | 42.83 | 42.79 | 43.24 | 43.55 | 43.17 | 43.89 | 43.07 | 42.86 | 42.65 |
| PR0161 | 44.05 | 44.2 | 43.38 | 100 | 43.04 | 42.22 | 42.91 | 43.46 | 43.48 | 43.97 | 43.22 | 42.18 | 43.19 |
| PR0165 | 43.71 | 43.98 | 43.38 | 43.54 | 100 | 43.74 | 42.8 | 43.52 | 43.21 | 42.52 | 42.98 | 42.4 | 44.21 |
| PR0167 | 43.42 | 43.47 | 43.04 | 42.42 | 43.44 | 100 | 43.65 | 43.58 | 43.25 | 43.24 | 42.56 | 42.19 | 42.47 |
| PR0186 | 43.59 | 43.94 | 43.66 | 43.29 | 42.68 | 43.82 | 100 | 43.66 | 42.9 | 43.18 | 42.9 | 42.05 | 42.88 |
| PR0188 | 44.11 | 44.56 | 43.58 | 43.45 | 43.01 | 43.36 | 43.27 | 100 | 43.91 | 43.65 | 41.37 | 42.73 | 42.76 |
| PR0036 | 43.91 | 44.53 | 43.26 | 43.52 | 42.75 | 43.09 | 42.57 | 43.97 | 100 | 43.59 | 42.31 | 42.49 | 42.77 |
| PR0039 | 44.05 | 44.19 | 44.13 | 44.16 | 42.21 | 43.22 | 43 | 43.86 | 43.74 | 100 | 43.96 | 42.51 | 42.91 |
| PR0049 | 43.65 | 43.84 | 43.94 | 44.05 | 43.3 | 43.17 | 43.34 | 42.18 | 43.08 | 44.6 | 100 | 42.98 | 43.18 |
| RL0013 | 43.9 | 44.53 | 44.55 | 43.8 | 43.52 | 43.6 | 43.29 | 44.38 | 44.08 | 43.95 | 43.8 | 100 | 42.49 |
| RL0015 | 44.01 | 43.91 | 43.36 | 43.87 | 44.38 | 42.93 | 43.17 | 43.44 | 43.39 | 43.39 | 43.03 | 41.55 | 100 |

In each square is the percentage of share variants between samples sequenced on the same plate. Samples are coloured based on the percentage. Usually two unrelated individuals shared approximately 43% of genetic component. Example of similarity matrix for sample IDs PR0158, PR0159, PR0160, PR0161, PR0165, PR0167, PR0186, PR0188, PR0036, PR0039, PR0049, RL0013 and RL0015. The samples were run with other samples not relative to the IBD study.

## Appendix III

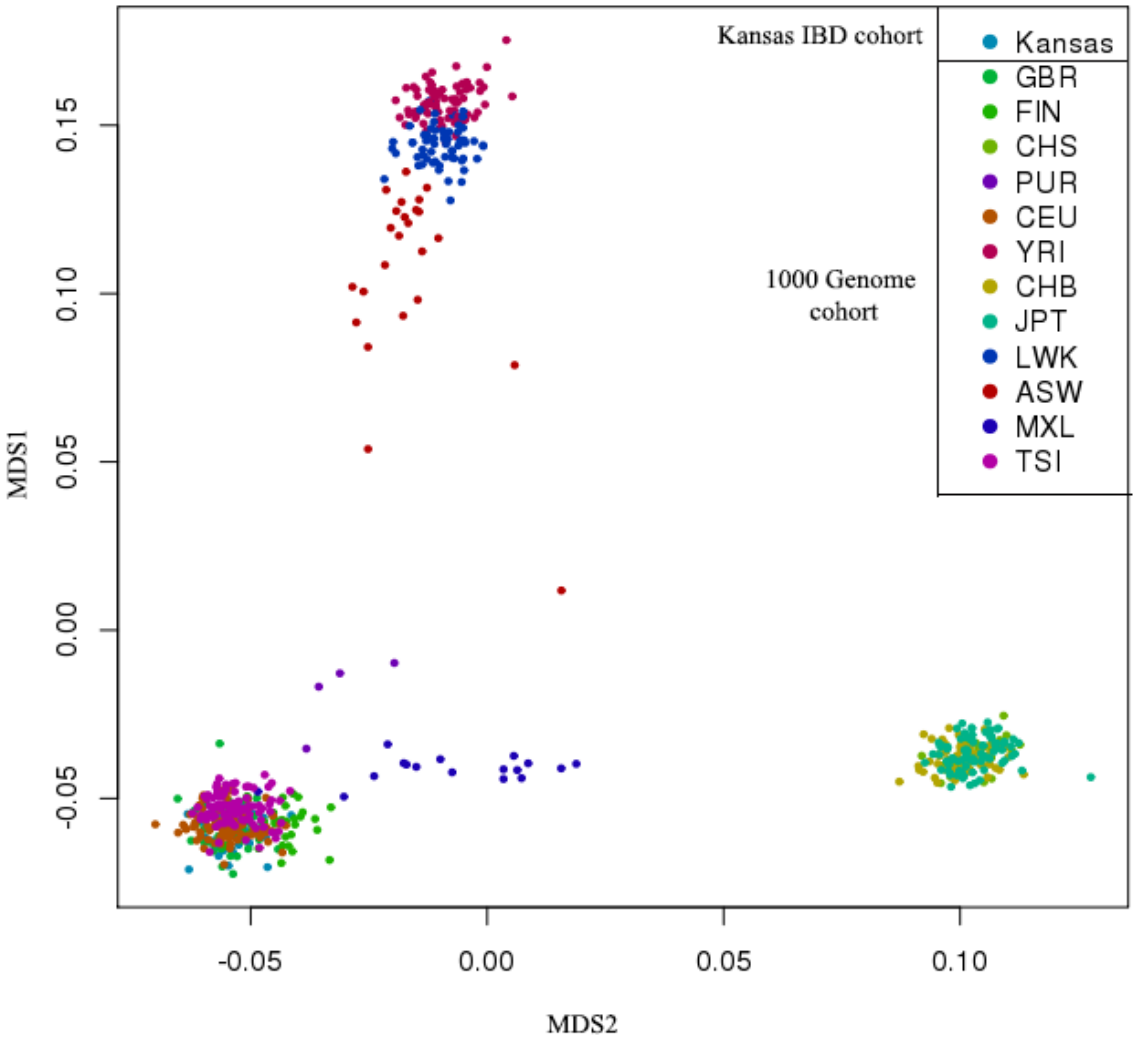
**Principle component analysis (PCA) across five ethnic groups from 1000 genome project and the validation cohort (66 whole-exome data of the CAGI dataset and 89 whole-genome data of British ethnicity of the 1000 genome dataset).**




The five ethnic groups from 1000 Genome are colored as indicated. The CAGI cohort is in red.

## Appendix IV

**Principle component analysis (PCA) across ethnic groups from 1000 genome project and the Kansas subgroup of the validation cohort (43 whole-exome data).**




The ethnic groups from 1000 Genome are coloured as indicated. The CAGI cohort is in blue.

# Appendix V

## List of 250 variants found within the 40 genes of the NOD2 pathway in which variations was found across the entire cohort of 136 cases and 106 controls.

| Gene | Chr | Bp position (hg19) | Variant. | Coding change | Protein change | SIFT | Gerp | MaxEnt score | CADD | dbSNP | Frequency in 1KG | Frequency in NHLBI ESP | HGMD | Frequency in cases (n=136) | | | Frequency in controls (n=106) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | | | | Homozygous reference | Heterozygous | Homozygous alternative | Homozygous reference | Heterozygous | Homozygous alternative |
| **BIRC2** | 11 | 102220918 | fr | c.335_339del | p.112_113del | . | . | . | . | . | . | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **BIRC2** | 11 | 102221045 | ns | c.460T>G | p.S154A | T | 2.14 | . | 0.553713 | . | . | . | not listed | 100 | 0 | 0 | 98.11 | 1.89 | 0 |
| **BIRC2** | 11 | 102221633 | sn | c.954T>C | p.S318S | . | . | . | . | rs182906329 | 0.0014 | 0.001163 | not listed | 99.26 | 0.74 | 0 | 98.11 | 1.89 | 0 |
| **BIRC2** | 11 | 102248377 | ns | c.1517C>T | p.A506V | D | 4.65 | . | 3.506238 | rs34510872 | 0.03 | 0.052117 | not listed | 92.65 | 7.35 | 0 | 100 | 0 | 0 |
| **BIRC2** | 11 | 102248406 | ns | c.1546A>G | p.K516E | D | 5.61 | . | 5.497399 | rs61754131 | 0.0005 | 0.001745 | not listed | 98.53 | 1.47 | 0 | 100 | 0 | 0 |
| **BIRC2** | 11 | 102248410 | ns | c.1550G>A | p.G517E | D | 5.61 | . | 4.831211 | . | . | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **BIRC3** | 11 | 102195882 | sn | c.642T>C | p.N214N | . | . | . | . | . | . | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **BIRC3** | 11 | 102195897 | sn | c.657T>C | p.D219D | . | . | . | . | . | . | . | not listed | 99.26 | 0.74 | 0 | 98.11 | 1.89 | 0 |
| **BIRC3** | 11 | 102196019 | ns | c.779A>G | p.K260R | T | -3.66 | . | -0.04809 | rs2276113 | 0.02 | 0.000814 | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **BIRC3** | 11 | 102201804 | ns | c.1156G>A | p.V386M | D | 2.37 | . | 0.619961 | rs12222256 | 0.0023 | . | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **BIRC3** | 11 | 102201848 | sn | c.1200G>A | p.Q400Q | . | . | . | . | rs17878663 | 0.07 | 0.001163 | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **BIRC3** | 11 | 102201850 | ns | c.1202G>A | p.R401K | T | 4.08 | . | 2.353145 | rs17881197 | 0.0041 | 0.008258 | not listed | 98.53 | 1.47 | 0 | 95.28 | 4.72 | 0 |
| **CARD6** | 5 | 40841561 | ns | c.77C>G | p.P26R | D | 3.66 | . | 1.550447 | . | . | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **CARD6** | 5 | 40841571 | fr | c.88_89del | p.30_30del | . | . | . | . | rs141244584 | 0.01 | . | not listed | 95.59 | 4.41 | 0 | 97.17 | 2.83 | 0 |
| **CARD6** | 5 | 40841741 | ns | c.257C>T | p.S86L | D | 3.74 | . | 1.737134 | rs10512747 | 0.06 | 0.120465 | not listed | 76.47 | 21.32 | 2.21 | 77.36 | 19.81 | 2.83 |
| **CARD6** | 5 | 40843493 | ns | c.523A>T | p.T175S | D | -2.97 | . | -1.70298 | rs61748215 | 0.0037 | 0.01 | not listed | 98.53 | 1.47 | 0 | 99.06 | 0.94 | 0 |
| **CARD6** | 5 | 40843550 | ns | c.580A>G | p.I194V | T | 3.51 | . | 1.02021 | rs61757654 | 0.01 | 0.013023 | not listed | 97.79 | 2.21 | 0 | 100 | 0 | 0 |
| **CARD6** | 5 | 40843735 | ns | c.765C>A | p.F255L | D | 1.41 | . | 2.570366 | rs35188876 | 0.01 | 0.018953 | not listed | 97.06 | 2.94 | 0 | 99.06 | 0.94 | 0 |
| **CARD6** | 5 | 40852317 | ns | c.883A>G | p.M295V | D | -0.882 | . | 1.770901 | rs61748217 | 0.0009 | 0.00186 | not listed | 97.79 | 2.21 | 0 | 100 | 0 | 0 |
| **CARD6** | 5 | 40852902 | ns | c.1468G>A | p.D490N | T | 4.85 | . | . | . | . | . | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **CARD6** | 5 | 40853048 | sn | c.1614C>T | p.S538S | . | . | . | . | rs16870407 | 0.15 | 0.092791 | not listed | 83.82 | 15.44 | 0.74 | 80.19 | 18.87 | 0.94 |
| **CARD6** | 5 | 40853404 | ns | c.1970T>C | p.V657A | D | 4.18 | . | 4.484833 | . | . | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **CARD9** | 9 | 139258779 | ns | c.1586A>C | p.D529A | T | 2.88 | . | 2.038243 | . | . | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **CARD9** | 9 | 139258814 | ns | c.1551G>C | p.Q517H | T | 1.57 | . | 0.959136 | . | . | . | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **CARD9** | 9 | 139259644 | sn | c.1383G>A | p.P461P | . | . | . | . | rs138344913 | 0.0018 | 0.003373 | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **CARD9** | 9 | 139262205 | ns | c.1153G>C | p.V385L | T | 1.37 | . | 0.261356 | rs3124993 | 0.01 | 0.021567 | not listed | 94.85 | 5.15 | 0 | 95.28 | 4.72 | 0 |
| **CARD9** | 9 | 139264888 | ns | c.809A>T | p.E270V | T | 2.85 | . | 0.381144 | rs114895119 | 0.0032 | 0.004307 | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **CARD9** | 9 | 139265088 | sn | c.693G>A | p.T231T | . | . | . | . | rs59902911 | 0.06 | 0.031177 | not listed | 94.12 | 5.88 | 0 | 89.62 | 10.38 | 0 |
| **CARD9** | 9 | 139265801 | sn | c.297G>A | p.P99P | . | . | . | . | rs115131813 | 0.03 | 0.017113 | not listed | 95.59 | 4.41 | 0 | 97.17 | 2.83 | 0 |
| **CARD9** | 9 | 139265810 | sn | c.288C>A | p.G96G | . | . | . | . | rs137986801 | . | 0.002096 | not listed | 99.26 | 0.74 | 0 | 99.06 | 0.94 | 0 |
| **CARD9** | 9 | 139265870 | sn | c.228C>T | p.Y76Y | . | . | . | . | rs11145769 | 0.06 | 0.031585 | not listed | 94.12 | 5.88 | 0 | 91.51 | 8.49 | 0 |
| **CARD9** | 9 | 139266405 | sn | c.126C>T | p.P42P | . | . | . | . | rs10781499 | 0.37 | 0.423372 | not listed | 34.56 | 43.38 | 22.06 | 34.91 | 49.06 | 16.04 |
| **CARD9** | 9 | 139266496 | ns | c.35G>A | p.S12N | T | -0.646 | . | 0.446262 | rs4077515 | 0.37 | 0.421512 | listed | 34.56 | 43.38 | 22.06 | 37.74 | 46.23 | 16.04 |
| **CARD9** | 9 | 139266519 | sn | c.12C>T | p.Y4Y | . | . | . | . | rs35051231 | 0.0005 | 0.001163 | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **CASP8** | 2 | 202122956 | ns | c.2T>C | p.M1T | D | -0.239 | . | 0.151387 | rs3769824 | 0.03 | 0.043705 | not listed | 94.85 | 5.15 | 0 | 88.68 | 11.32 | 0 |
| **CASP8** | 2 | 202122995 | ns | c.41A>G | p.K14R | D | 2.94 | . | 0.164368 | rs3769823 | 0.65 | 0.705882 | not listed | 16.18 | 38.24 | 45.59 | 11.32 | 43.4 | 45.28 |
| **CASP8** | 2 | 202123108 | sp | c.151+3G>A | . | . | . | 0.85 | . | rs202238412 | . | 0.001798 | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **CASP8** | 2 | 202136272 | sn | c.339C>T | p.S113S | . | . | . | . | rs17860422 | 0.0023 | 0.00407 | not listed | 98.53 | 1.47 | 0 | 100 | 0 | 0 |
| **CASP8** | 2 | 202137392 | ns | c.443A>G | p.K148R | T | 5.52 | . | 3.96662 | rs148697064 | . | 0.000698 | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |

| | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **CASP8** | 2 | 202149589 | ns | c.808G>C | p.D270H | T | -4.55 | . | -0.30191 | rs1045485 | 0.07 | 0.129186 | listed | 77.94 | 21.32 | 0.74 | 83.96 | 16.04 | 0 |
| **CASP8** | 2 | 202149696 | sn | c.915G>A | p.K305K | . | . | . | . | rs1045487 | 0.15 | 0.051047 | not listed | 91.91 | 8.09 | 0 | 85.85 | 14.15 | 0 |
| **CASP8** | 2 | 202149737 | ns | c.956A>G | p.Y319C | D | -1.54 | . | 0.964215 | . | . | . | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **CCL2** | 17 | 32583269 | sn | c.105T>C | p.C35C | . | . | . | . | rs4586 | 0.54 | 0.36093 | not listed | 49.26 | 43.38 | 7.35 | 44.34 | 39.62 | 16.04 |
| **CHUK** | 10 | 101964267 | sn | c.1503G>A | p.G501G | . | . | . | . | rs2862988 | 0.0018 | 0.00593 | not listed | 99.26 | 0.74 | 0 | 97.17 | 2.83 | 0 |
| **CHUK** | 10 | 101964312 | sn | c.1458C>T | p.S486S | . | . | . | . | rs17880383 | 0.04 | 0.061395 | not listed | 87.5 | 11.03 | 1.47 | 94.34 | 5.66 | 0 |
| **CHUK** | 10 | 101964847 | sn | c.1341A>G | p.G447G | . | . | . | . | rs34458357 | 0.0014 | 0.005465 | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **CHUK** | 10 | 101964950 | ns | c.1238A>C | p.D413A | T | 5.01 | . | 4.795139 | . | . | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **CHUK** | 10 | 101977883 | ns | c.802G>A | p.V268I | T | 4.69 | . | 2.550085 | rs2230804 | 0.56 | 0.501047 | not listed | 21.32 | 52.21 | 26.47 | 28.3 | 50 | 21.7 |
| **CHUK** | 10 | 101980355 | ns | c.464T>C | p.V155A | T | 4.16 | . | 1.072682 | rs2230803 | 0.02 | 0.001163 | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **CXCL1** | 4 | 74735244 | sn | c.57A>G | p.A19A | . | . | . | . | rs2071425 | 0.34 | 0.186713 | not listed | 54.41 | 40.44 | 5.15 | 72.64 | 26.42 | 0.94 |
| **CXCL1** | 4 | 74736235 | sp | c.309-3C>T | . | . | . | 1.03 | . | rs1814092 | 0.05 | 0.001628 | not listed | 98.53 | 1.47 | 0 | 100 | 0 | 0 |
| **CXCL2** | 4 | 74964625 | ns | c.115A>G | p.T39A | T | -0.319 | . | 0.437763 | rs142264518 | 0.0046 | 0.006977 | not listed | 98.53 | 1.47 | 0 | 99.06 | 0.94 | 0 |
| **CXCL2** | 4 | 74964830 | ns | c.8G>A | p.R3H | T | -5.75 | . | -0.01658 | rs186397980 | 0.02 | 0.015973 | not listed | 97.79 | 1.47 | 0.74 | 97.17 | 2.83 | 0 |
| **ERBB2IP** | 5 | 65307924 | ns | c.355A>G | p.I119V | T | -2.81 | . | 0.260566 | rs61758158 | 0.01 | 0.01143 | not listed | 95.59 | 3.68 | 0.74 | 97.17 | 2.83 | 0 |
| **ERBB2IP** | 5 | 65317181 | sn | c.565C>T | p.L189L | . | . | . | . | rs706679 | 0.67 | 0.856478 | not listed | 11.76 | 12.5 | 75.74 | 10.38 | 31.13 | 58.49 |
| **ERBB2IP** | 5 | 65317206 | ns | c.590C>T | p.T197M | D | 4.58 | . | 5.955216 | rs146136641 | 0.0005 | 0.001512 | listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **ERBB2IP** | 5 | 65321311 | ns | c.821C>T | p.S274L | T | 2.41 | . | 1.987188 | rs3213837 | 0.1 | 0.156315 | not listed | 77.94 | 19.12 | 2.94 | 67.92 | 31.13 | 0.94 |
| **ERBB2IP** | 5 | 65349300 | sn | c.2142A>G | p.E714E | . | . | . | . | . | . | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **ERBB2IP** | 5 | 65349887 | ns | c.2729A>G | p.K910R | D | 4.46 | . | . | rs34521887 | 0.01 | 0.000116 | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **ERBB2IP** | 5 | 65350044 | sn | c.2886A>G | p.Q962Q | . | . | . | . | rs35278406 | 0.02 | 0.039893 | not listed | 90.44 | 9.56 | 0 | 91.51 | 8.49 | 0 |
| **ERBB2IP** | 5 | 65350279 | ns | c.3121C>T | p.H1041Y | D | 4.05 | . | 1.755784 | rs142496054 | 0.01 | 0.008023 | not listed | 99.26 | 0.74 | 0 | 96.23 | 2.83 | 0.94 |
| **ERBB2IP** | 5 | 65350374 | sn | c.3216A>G | p.R1072R | . | . | . | . | rs36303 | 0.23 | 0.138488 | not listed | 80.15 | 17.65 | 2.21 | 65.09 | 31.13 | 3.77 |
| **ERBB2IP** | 5 | 65350481 | ns | c.3323C>T | p.S1108L | D | 5.32 | . | 4.6549 | rs3805466 | 0.1 | 0.047674 | not listed | 91.18 | 6.62 | 2.21 | 76.42 | 22.64 | 0.94 |
| **ERBB2IP** | 5 | 65350527 | sn | c.3369T>G | p.L1123L | . | . | . | 0.161838 | . | . | 0.000581 | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **ERBB2IP** | 5 | 65370927 | ns | c.3697C>G | p.Q1233E | D | 5.34 | . | 2.367375 | rs201285970 | . | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **ERBB2IP** | 5 | 65372200 | sn | c.3885C>T | p.V1295V | . | . | . | . | . | . | . | not listed | 100 | 0 | 0 | 98.11 | 1.89 | 0 |
| **IKBKB** | 8 | 42128942 | sn | c.54C>T | p.F18F | . | . | . | . | rs12545246 | 0.03 | . | not listed | 97.06 | 2.94 | 0 | 99.06 | 0.94 | 0 |
| **IKBKB** | 8 | 42128970 | ns | c.82C>T | p.P28S | . | . | . | 0.730313 | . | . | . | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **IKBKB** | 8 | 42163863 | sn | c.303A>G | p.L101L | . | . | . | . | rs17875704 | 0.0009 | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **IKBKB** | 8 | 42174380 | sn | c.1077G>A | p.L359L | . | . | . | . | rs56230731 | 0.01 | 0.014884 | not listed | 97.06 | 2.94 | 0 | 100 | 0 | 0 |
| **IKBKB** | 8 | 42178343 | ns | c.1663G>A | p.G555R | T | 4.96 | . | 3.028211 | rs149701177 | . | 0.000465 | not listed | 100 | 0 | 0 | 98.11 | 1.89 | 0 |
| **IKBKB** | 8 | 42179427 | sn | c.1696A>C | p.R566R | . | . | . | . | rs151057347 | 0.0023 | 0.003837 | not listed | 99.26 | 0.74 | 0 | 99.06 | 0.94 | 0 |
| **IKBKG** | X | 153780386 | ns | c.169G>A | p.E57K | D | 5.17 | . | 2.125638 | rs148695964 | . | 0.001784 | listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **IL6** | 7 | 22771038 | ns | c.485A>T | p.D162V | T | 0.258 | . | -0.30563 | rs2069860 | 0.0018 | 0.00814 | not listed | 97.79 | 2.21 | 0 | 98.11 | 1.89 | 0 |
| **IL6** | 7 | 22771039 | ns | c.486T>A | p.D162E | T | -2.22 | . | 0.028822 | rs13306435 | 0.03 | 0.008837 | not listed | 99.26 | 0.74 | 0 | 98.11 | 1.89 | 0 |
| **IL6** | 7 | 22771156 | sn | c.603C>T | p.F201F | . | . | . | . | rs2069849 | 0.06 | 0.023605 | not listed | 96.32 | 3.68 | 0 | 92.45 | 6.6 | 0.94 |
| **IL8** | 4 | 74606393 | sn | c.18C>T | p.A6A | . | . | . | . | rs1803205 | . | 0.001977 | not listed | 100 | 0 | 0 | 98.11 | 1.89 | 0 |
| **IL8** | 4 | 74607328 | ns | c.134A>G | p.H45R | D | 3.7 | . | . | rs139503118 | . | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **MAP3K7** | 6 | 91256978 | sn | c.1209A>G | p.T403T | . | . | . | . | . | . | 0.000116 | not listed | 100 | 0 | 0 | 100 | 0 | 0 |
| **MAP3K7** | 6 | 91266350 | sp | c.483-7T>A | . | . | . | 2.01 | . | rs45625637 | 0.06 | . | not listed | 100 | 0 | 0 | 100 | 0 | 0 |
| **MAPK1** | 22 | 22123519 | ns | c.1057A>G | p.R353G | T | 4.04 | . | 2.942178 | . | . | . | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **MAPK1** | 22 | 22142659 | sg | c.743C>A | p.S248X | . | 5.57 | . | 2.569548 | . | . | . | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **MAPK1** | 22 | 22160301 | sn | c.330A>G | p.T110T | . | . | . | . | rs150378600 | 0.0009 | 0.001163 | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **MAPK1** | 22 | 22162126 | sn | c.129T>C | p.Y43Y | . | . | . | . | rs3729910 | 0.04 | 0.060698 | not listed | 88.24 | 11.03 | 0.74 | 85.85 | 14.15 | 0 |

| | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **MAPK1** | 22 | 22221708 | nonfi | c.2_3insGGC | p.M1delinsMA | . | . | . | . | . | . | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **MAPK10** | 4 | 86952589 | sp | c.1111-5GA | . | . | . | 1.08 | . | rs200643314 | . | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **MAPK10** | 4 | 86952590 | sp | c.1111-6CT | . | . | . | 0.61 | . | rs13103861 | 0.14 | 0.179884 | not listed | 64.71 | 33.09 | 2.21 | 66.98 | 30.19 | 2.83 |
| **MAPK11** | 22 | 50703796 | sn | c.969T>C | p.Y323Y | . | . | . | . | rs139548825 | . | 0.000581 | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **MAPK11** | 22 | 50704028 | ns | c.824G>A | p.R275H | T | -3.93 | . | 1.290414 | rs33932986 | 0.02 | 0.019419 | not listed | 97.79 | 1.47 | 0.74 | 98.11 | 1.89 | 0 |
| **MAPK11** | 22 | 50704661 | sn | c.756A>G | p.S252S | . | . | . | . | rs2076139 | 0.7 | 0.762116 | not listed | 5.88 | 37.5 | 56.62 | 18.87 | 27.36 | 53.77 |
| **MAPK11** | 22 | 50705466 | sn | c.507T>C | p.F169F | . | . | . | . | rs760748 | 0.97 | 0.994416 | not listed | 0 | 2.21 | 97.79 | 5.66 | 0.94 | 93.4 |
| **MAPK11** | 22 | 50705821 | sn | c.396C>T | p.Y132Y | . | . | . | . | rs2066762 | 0.03 | 0.002214 | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **MAPK11** | 22 | 50705830 | sn | c.387C>T | p.F129F | . | . | . | . | rs140519122 | . | 0.002098 | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **MAPK11** | 22 | 50706381 | sp | c.117-3C>T | . | . | . | 0.57 | . | rs36083586 | 0.12 | 0.112509 | not listed | 72.06 | 25 | 2.94 | 84.91 | 13.21 | 1.89 |
| **MAPK12** | 22 | 50691870 | ns | c.1064G>A | p.R355Q | T | 0.925 | . | . | rs138582408 | 0.0027 | 0.002749 | not listed | 100 | 0 | 0 | 97.17 | 2.83 | 0 |
| **MAPK12** | 22 | 50691914 | sp | c.1025-5GC | . | . | . | 1.14 | . | . | . | . | not listed | 97.79 | 2.21 | 0 | 97.17 | 2.83 | 0 |
| **MAPK12** | 22 | 50691915 | sp | c.1025-6CT | . | . | . | 0.07 | . | . | . | . | not listed | 100 | 0 | 0 | 98.11 | 1.89 | 0 |
| **MAPK12** | 22 | 50693619 | sp | c.1024+7GT | . | . | . | 0.92 | . | . | . | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **MAPK12** | 22 | 50693705 | sn | c.945C>T | p.H315H | . | . | . | . | rs45606035 | 0.0018 | 0.005349 | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **MAPK12** | 22 | 50693889 | sn | c.843C>T | p.S281S | . | . | . | . | rs2066770 | 0.04 | 0.018721 | not listed | 95.59 | 4.41 | 0 | 97.17 | 2.83 | 0 |
| **MAPK12** | 22 | 50693919 | sn | c.813G>A | p.K271K | . | . | . | . | rs55861809 | 0.01 | . | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **MAPK12** | 22 | 50693934 | sn | c.798C>T | p.P266P | . | . | . | . | rs62239359 | 0.0009 | 0.005116 | not listed | 98.53 | 1.47 | 0 | 100 | 0 | 0 |
| **MAPK12** | 22 | 50694084 | ns | c.731C>T | p.T244M | D | 4.2 | . | . | rs2066776 | . | 0.002209 | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **MAPK12** | 22 | 50694297 | sn | c.633T>C | p.S211S | . | . | . | . | rs1129880 | 0.79 | 0.704281 | not listed | 8.09 | 41.18 | 50.74 | 20.75 | 33.02 | 46.23 |
| **MAPK12** | 22 | 50694542 | sn | c.591C>T | p.I197I | . | . | . | . | . | . | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **MAPK12** | 22 | 50694578 | sn | c.555C>T | p.Y185Y | . | . | . | . | . | . | . | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **MAPK12** | 22 | 50695370 | sn | c.450C>T | p.I150I | . | . | . | . | rs2072876 | 0.04 | 0.023166 | not listed | 94.12 | 5.15 | 0.74 | 95.28 | 4.72 | 0 |
| **MAPK12** | 22 | 50696678 | ns | c.308C>T | p.T103M | T | 2.11 | . | . | rs34422484 | 0.09 | 0.031047 | not listed | 89.71 | 10.29 | 0 | 95.28 | 4.72 | 0 |
| **MAPK12** | 22 | 50699668 | sn | c.183T>C | p.P61P | . | . | . | . | rs2272857 | 0.68 | 0.769446 | not listed | 5.88 | 36.76 | 57.35 | 24.53 | 26.42 | 49.06 |
| **MAPK13** | 6 | 36098410 | sn | c.51A>C | p.T17T | . | . | . | . | rs1059227 | 0.77 | 0.660191 | not listed | 8.09 | 46.32 | 45.59 | 21.7 | 36.79 | 41.51 |
| **MAPK13** | 6 | 36098434 | sg | c.75C>A | p.Y25X | NA | -0.985 | . | 2.114124 | rs151226715 | 0.0014 | 0.000698 | not listed | 99.26 | 0.74 | 0 | 99.06 | 0.94 | 0 |
| **MAPK13** | 6 | 36098481 | sp | c.119+3>GA | . | . | . | 2.78 | . | rs140374075 | 0.01 | 0.001865 | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **MAPK13** | 6 | 36099050 | ns | c.122C>T | p.S41L | D | 2.66 | . | 2.103868 | rs55776345 | 0.01 | 0.012907 | not listed | 99.26 | 0.74 | 0 | 97.17 | 2.83 | 0 |
| **MAPK13** | 6 | 36100425 | ns | c.277C>T | p.P93S | T | 1.67 | . | 2.505966 | rs148256444 | . | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **MAPK13** | 6 | 36104430 | sp | c.496-3>TC | . | . | . | 0.27 | . | rs55732669 | 0.02 | 0.000349 | not listed | 99.26 | 0.74 | 0 | 98.11 | 1.89 | 0 |
| **MAPK13** | 6 | 36104455 | ns | c.518G>A | p.R173Q | D | 4.53 | . | 1.817101 | . | . | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **MAPK13** | 6 | 36104502 | sg | c.565C>T | p.R189X | NA | 4.53 | . | 2.531246 | rs148572287 | . | 0.000116 | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **MAPK13** | 6 | 36107131 | ns | c.1079G>A | p.R360Q | T | 2.53 | . | 2.12748 | rs150915766 | . | 0.001744 | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **MAPK14** | 6 | 36063793 | ns | c.712C>T | p.L238F | D | 4.75 | . | 2.393865 | rs139802452 | . | 0.000465 | not listed | 98.53 | 1.47 | 0 | 100 | 0 | 0 |
| **MAPK14** | 6 | 36068038 | sn | c.756C>T | p.S252S | . | . | . | . | . | . | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **MAPK14** | 6 | 36068041 | sn | c.759T>C | p.H253H | . | . | . | . | rs2815805 | 0.04 | 0.016512 | not listed | 95.59 | 4.41 | 0 | 97.17 | 2.83 | 0 |
| **MAPK14** | 6 | 36075286 | ns | c.896C>T | p.A299V | NA | 5.61 | . | 5.494333 | . | . | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **MAPK3** | 16 | 30128224 | sn | c.1008G>A | p.P336P | . | . | . | . | rs1143695 | 0.0032 | 0.003023 | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **MAPK3** | 16 | 30128580 | ns | c.802G>A | p.D268N | D | 5.35 | . | 2.862978 | . | . | . | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **MAPK3** | 16 | 30129377 | sn | c.651G>T | p.L217L | . | . | . | 2.257118 | rs139957276 | 0.0005 | 0.000465 | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **MAPK3** | 16 | 30134507 | sn | c.24G>A | p.G8G | . | . | . | . | . | . | . | not listed | 100 | 0 | 0 | 98.11 | 1.89 | 0 |
| **MAPK8** | 10 | 49609720 | ns | c.17G>A | p.R6H | T | 4.25 | . | 3.151268 | . | . | . | not listed | 100 | 0 | 0 | 98.11 | 1.89 | 0 |
| **MAPK8** | 10 | 49632183 | sn | c.669C>T | p.I223I | . | . | . | . | . | . | . | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **MAPK8** | 10 | 49642974 | ns | c.1186G>A | p.V396I | T | 5.46 | . | 1.599793 | . | . | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **MAPK9** | 5 | 179665354 | sn | c.1110T>C | p.G370G | . | . | . | . | rs138473736 | . | 0.000465 | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **MAPK9** | 5 | 179676062 | ns | c.527C>T | p.A176V | T | 4.11 | . | 2.462092 | . | . | . | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **NFKB1** | 4 | 103488139 | sp | c.256-5>TC | . | . | . | 0.25 | . | . | . | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **NFKB1** | 4 | 103505961 | sn | c.1047C>T | p.Y349Y | . | . | . | . | rs4648039 | 0.01 | 0.024535 | not listed | 88.97 | 11.03 | 0 | 97.17 | 2.83 | 0 |
| **NFKB1** | 4 | 103514658 | sn | c.1140T>C | p.A380A | . | . | . | . | rs1609993 | 0.96 | 0.919186 | not listed | 0.74 | 16.18 | 83.09 | 2.83 | 16.98 | 80.19 |
| **NFKB1** | 4 | 103516146 | sp | c.1297+8>AG | . | . | . | 0.27 | . | . | . | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **NFKB1** | 4 | 103517301 | ns | c.1304T>C | p.M435T | T | 4.6 | . | 1.092039 | . | . | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **NFKB1** | 4 | 103518700 | ns | c.1516A>G | p.M506V | T | -4.82 | . | -0.21319 | rs4648072 | 0.01 | 0.008023 | not listed | 97.79 | 2.21 | 0 | 99.06 | 0.94 | 0 |
| **NFKB1** | 4 | 103527654 | ns | c.1751C>T | p.T584M | D | 4.59 | . | 5.152581 | . | . | . | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **NFKB1** | 4 | 103527745 | ns | c.1842G>T | p.L614F | T | 1.05 | . | 1.308065 | rs149211506 | 0.0027 | 0.003023 | not listed | 98.53 | 1.47 | 0 | 100 | 0 | 0 |
| **NFKB1** | 4 | 103534701 | sn | c.2709G>A | p.S903S | . | . | . | . | rs4648119 | . | 0.001047 | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **NFKB1** | 4 | 103537672 | ns | c.2828C>A | p.T943N | T | 2 | . | 1.290152 | rs143882681 | . | 0.000581 | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **NFKBIA** | 14 | 35872068 | sp | c.548-3>CT | . | . | . | 0.18 | . | rs2233418 | 0.0018 | 0.008488 | not listed | 98.53 | 1.47 | 0 | 99.06 | 0.94 | 0 |
| **NFKBIA** | 14 | 35872414 | sn | c.489G>A | p.L163L | . | . | . | . | . | . | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **NFKBIA** | 14 | 35872926 | sn | c.306C>T | p.A102A | . | . | . | . | rs1050851 | 0.12 | 0.227791 | not listed | 52.94 | 41.18 | 5.88 | 60.38 | 34.91 | 4.72 |
| **NFKBIA** | 14 | 35873770 | sn | c.81C>T | p.D27D | . | . | . | . | rs1957106 | 0.24 | 0.275911 | not listed | 58.09 | 33.09 | 8.82 | 57.55 | 33.96 | 8.49 |
| **NFKBIB** | 19 | 39395836 | sp | c.28-6>CT | . | . | . | 0.67 | . | rs200550654 | . | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **NFKBIB** | 19 | 39396013 | ns | c.199C>T | p.R67C | T | -7.16 | . | 0.67979 | . | 0.0046 | 0.007286 | not listed | 96.32 | 2.94 | 0.74 | 97.17 | 2.83 | 0 |
| **NFKBIB** | 19 | 39398188 | sn | c.600C>T | p.N200N | . | . | . | . | . | . | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **NFKBIB** | 19 | 39398201 | ns | c.613C>T | p.R205C | D | -0.188 | . | 2.135488 | rs187346322 | 0.0023 | 0.000818 | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **NOD1** | 7 | 30487954 | ns | c.2245A>G | p.S749G | D | -2.15 | . | . | . | . | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **NOD1** | 7 | 30490919 | ns | c.2114G>A | p.R705Q | T | -3.79 | . | 0.13785 | rs144684378 | . | 0.002093 | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **NOD1** | 7 | 30491123 | ns | c.1910G>A | p.R637H | T | 5.26 | . | 3.830579 | rs5743347 | 0.0018 | 0.002791 | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **NOD1** | 7 | 30491143 | sn | c.1890C>T | p.G630G | . | . | . | . | . | . | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **NOD1** | 7 | 30491311 | sn | c.1722G>A | p.A574A | . | . | . | . | rs2075821 | 0.29 | 0.258953 | not listed | 46.32 | 45.59 | 8.09 | 57.55 | 33.96 | 8.49 |
| **NOD1** | 7 | 30491693 | ns | c.1340G>A | p.R447H | T | 3.22 | . | 3.31794 | rs2975634 | 0.02 | 0.000349 | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **NOD1** | 7 | 30491837 | ns | c.1196G>A | p.R399Q | T | 1.4 | . | 3.429291 | rs141422065 | 0.0009 | 0.000465 | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **NOD1** | 7 | 30492086 | ns | c.947A>G | p.N316S | D | -3.6 | . | -1.23484 | . | . | . | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **NOD1** | 7 | 30492142 | sn | c.891C>T | p.R297R | . | . | . | . | rs3020208 | 0.02 | 0.000349 | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **NOD1** | 7 | 30492237 | ns | c.796G>A | p.E266K | D | 5.19 | . | 4.169 | rs2075820 | 0.3 | 0.245349 | listed | 50 | 44.12 | 5.88 | 61.32 | 31.13 | 7.55 |
| **NOD1** | 7 | 30492246 | ns | c.787C>A | p.R263R | . | . | . | 0.862199 | . | . | 0.000116 | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **NOD1** | 7 | 30492550 | sn | c.483C>T | p.D161D | . | . | . | . | rs2235099 | 0.3 | 0.246395 | not listed | 50 | 44.12 | 5.88 | 60.38 | 32.08 | 7.55 |
| **NOD1** | 7 | 30492598 | sn | c.435G>T | p.L145L | . | . | . | . | rs5743340 | 0.0046 | 0.018256 | not listed | 97.06 | 2.94 | 0 | 95.28 | 4.72 | 0 |
| **NOD1** | 7 | 30494802 | sn | c.327C>T | p.F109F | . | . | . | . | . | . | . | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **NOD1** | 7 | 30494866 | ns | c.263A>G | p.Y88C | T | 3.72 | . | 2.589497 | . | . | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **NOD1** | 7 | 30496382 | sn | c.156C>G | p.A52A | . | . | . | . | rs2075818 | 0.32 | 0.248488 | not listed | 48.53 | 45.59 | 5.88 | 62.26 | 31.13 | 6.6 |
| **NOD1** | 7 | 30496518 | ns | c.20G>A | p.S7N | T | 1.17 | . | 0.755481 | rs61757653 | . | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **NOD2** | 16 | 50733392 | sp | c.74-7>TA | . | . | . | 1.83 | . | rs104895421 | 0.0014 | 0.001861 | listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **NOD2** | 16 | 50733423 | ns | c.98C>A | p.A33D | T | 1.13 | . | 1.402571 | . | 0.000008 | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **NOD2** | 16 | 50733661 | sn | c.336C>T | p.A112A | . | . | . | . | . | 0.00002 | . | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **NOD2** | 16 | 50733785 | ns | c.460G>A | p.D154N | T | 0.958 | . | 0.410941 | rs146054564 | . | 0.002093 | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **NOD2** | 16 | 50733859 | sn | c.534C>G | p.S178S | . | . | . | . | rs2067085 | 0.26 | 0.409302 | not listed | 38.97 | 54.41 | 6.62 | 47.17 | 37.74 | 15.09 |
| **NOD2** | 16 | 50741791 | ns | c.566C>T | p.T189M | T | 3.48 | . | 2.334887 | rs61755182 | 0.0014 | 0.004419 | listed | 98.53 | 1.47 | 0 | 100 | 0 | 0 |
| **NOD2** | 16 | 50741800 | ns | c.575C>T | p.A192V | D | 0.916 | . | 0.885309 | rs149071116 | 0.00004 | . | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **NOD2** | 16 | 50741858 | sn | c.633C>T | p.A211A | . | . | . | . | rs5743269 | 0.0009 | 0.001744 | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |

| | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **NOD2** | 16 | 50744624 | ns | c.802C>T | p.P268S | T | -9.98 | . | -0.27189 | rs2066842 | 0.12 | 0.26907 | listed | 42.65 | 47.06 | 10.29 | 55.66 | 34.91 | 9.43 |
| **NOD2** | 16 | 50744688 | ns | c.866A>G | p.N289S | D | 4.56 | . | 0.444188 | rs5743271 | 0.01 | 0.006279 | listed | 98.53 | 1.47 | 0 | 98.11 | 1.89 | 0 |
| **NOD2** | 16 | 50744850 | ns | c.1028T>C | p.L343P | D | 5.4 | . | 0.517926 | . | . | 0.000116 | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **NOD2** | 16 | 50745114 | ns | c.1292C>T | p.S431L | D | 3.64 | . | 0.851472 | rs104895431 | 0.0005 | 0.001395 | listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **NOD2** | 16 | 50745199 | sn | c.1377C>T | p.R459R | . | . | . | . | rs2066843 | 0.13 | 0.270993 | not listed | 41.91 | 47.79 | 10.29 | 50 | 39.62 | 10.38 |
| **NOD2** | 16 | 50745316 | sn | c.1494A>G | p.E498E | . | . | . | . | . | . | . | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **NOD2** | 16 | 50745511 | sn | c.1689C>T | p.Y563Y | . | . | . | . | rs111608429 | 0.0005 | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **NOD2** | 16 | 50745583 | sn | c.1761T>G | p.R587R | . | . | . | . | rs1861759 | 0.25 | 0.402558 | not listed | 39.71 | 53.68 | 6.62 | 47.17 | 38.68 | 14.15 |
| **NOD2** | 16 | 50745655 | sn | c.1833C>T | p.A611A | . | . | . | . | rs61736932 | 0.0046 | 0.010698 | not listed | 98.53 | 1.47 | 0 | 100 | 0 | 0 |
| **NOD2** | 16 | 50745751 | sn | c.1929C>T | p.L643L | . | . | . | . | . | 0.000008 | . | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **NOD2** | 16 | 50745926 | ns | c.2104C>T | p.R702W | D | 2.42 | . | 1.736582 | rs2066844 | 0.02 | 0.043488 | listed | 88.24 | 10.29 | 1.47 | 86.79 | 13.21 | 0 |
| **NOD2** | 16 | 50745929 | ns | c.2107C>T | p.R703C | D | 2.89 | . | 1.788325 | rs5743277 | 0.0023 | 0.006977 | listed | 97.79 | 2.21 | 0 | 100 | 0 | 0 |
| **NOD2** | 16 | 50745960 | ns | c.2138G>A | p.R713H | T | 4.13 | 0.75 | 2.225724 | rs104895483 | . | 0.000233 | listed | 98.53 | 1.47 | 0 | 100 | 0 | 0 |
| **NOD2** | 16 | 50746086 | ns | c.2264C>T | p.A755V | D | 5.12 | . | 1.225314 | rs61747625 | 0.0005 | 0.004651 | listed | 98.53 | 1.47 | 0 | 100 | 0 | 0 |
| **NOD2** | 16 | 50746199 | ns | c.2377G>A | p.V793M | D | 3.51 | . | 1.544959 | rs104895444 | 0.0005 | 0.001628 | listed | 99.26 | 0.74 | 0 | 98.11 | 1.89 | 0 |
| **NOD2** | 16 | 50746228 | sn | c.2406G>T | p.V802V | . | . | . | 1.92838 | rs104895495 | . | 0.00186 | not listed | 98.53 | 1.47 | 0 | 99.06 | 0.94 | 0 |
| **NOD2** | 16 | 50746291 | sp | c.2462+7>GT | . | . | . | 0.83 | . | rs202111813 | 0.0005 | 0.000581 | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **NOD2** | 16 | 50750842 | ns | c.2587A>G | p.M863V | T | -9.48 | . | 0.558526 | rs104895447 | . | 0.00186 | listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **NOD2** | 16 | 50756540 | ns | c.2722G>C | p.G908R | D | 5.56 | . | 5.54325 | rs2066845 | 0.01 | 0.014535 | listed | 96.32 | 3.68 | 0 | 96.23 | 3.77 | 0 |
| **NOD2** | 16 | 50756571 | ns | c.2753C>A | p.A918D | D | 5.56 | . | 5.735298 | rs104895452 | 0.0009 | 0.000814 | listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **NOD2** | 16 | 50757276 | ns | c.2863G>A | p.V955I | T | -9.14 | . | -0.87026 | rs5743291 | 0.05 | 0.096047 | listed | 83.09 | 16.91 | 0 | 81.13 | 18.87 | 0 |
| **NOD2** | 16 | 50759405 | ns | c.288A8>G | p.E963G | T | 5.29 | . | 4.950708 | . | . | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **NOD2** | 16 | 50763778 | fr | c.3019dupC | p.L1007fs | . | . | . | . | rs2066847 | 0.006 | . | listed | 89.71 | 8.82 | 1.47 | 99.06 | 0.94 | 0 |
| **RELA** | 11 | 65422007 | ns | c.1498A>G | p.I500V | D | 3.39 | 1.83 | 2.237479 | . | . | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **RELA** | 11 | 65425764 | ns | c.871G>A | p.D291N | T | 4.76 | . | 3.374902 | rs61759893 | 0.0046 | 0.004655 | not listed | 98.53 | 1.47 | 0 | 99.06 | 0.94 | 0 |
| **RELA** | 11 | 65425804 | sn | c.831C>T | p.D277D | . | . | . | . | rs147357241 | . | 0.000233 | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **RELA** | 11 | 65427183 | ns | c.513G>T | p.R171S | T | 4.36 | . | 3.419675 | . | . | . | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **RIPK2** | 8 | 90770315 | sn | c.27C>T | p.A9A | . | . | . | . | rs2293809 | 0.09 | 0.031789 | not listed | 92.65 | 6.62 | 0.74 | 94.34 | 5.66 | 0 |
| **RIPK2** | 8 | 90784979 | ns | c.776T>C | p.I259T | T | 4.45 | . | 3.648531 | rs2230801 | 0.08 | 0.07907 | not listed | 84.56 | 15.44 | 0 | 78.3 | 21.7 | 0 |
| **RIPK2** | 8 | 90801670 | sn | c.1245T>C | p.S415S | . | . | . | . | rs56109184 | 0.0005 | 0.000698 | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **RIPK2** | 8 | 90802491 | sn | c.1470A>G | p.L490L | . | . | . | . | rs16900617 | 0.08 | 0.002558 | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **RIPK2** | 8 | 90802611 | sn | c.1590A>G | p.P530P | . | . | . | . | rs186397742 | 0.0009 | . | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **SUGT1** | 13 | 53231709 | ns | c.139T>A | p.Y47N | D | 4.12 | . | . | . | . | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **SUGT1** | 13 | 53239767 | sp | c.519-5>AT | . | . | . | 1.07 | . | . | . | . | not listed | 100 | 0 | 0 | 98.11 | 1.89 | 0 |
| **SUGT1** | 13 | 53240958 | ns | c.627G>T | p.L209F | D | 2.93 | . | 2.604733 | rs61756205 | 0.0018 | 0.005119 | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **SUGT1** | 13 | 53254116 | ns | c.822G>T | p.K274N | T | 3.92 | . | 2.817749 | rs202155148 | . | 0.000116 | not listed | 99.26 | 0.74 | 0 | 99.06 | 0.94 | 0 |
| **SUGT1** | 13 | 53254296 | sp | c.996+6>AG | . | . | . | 0.75 | . | rs7986540 | 0.97 | 0.946395 | not listed | 1.47 | 7.35 | 91.18 | 4.72 | 11.32 | 83.96 |
| **SUGT1** | 13 | 53261936 | nonfd | c.1069_1071del | p.357_357del | . | . | . | . | . | . | . | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **TAB1** | 22 | 39795831 | sn | c.24G>A | p.L8L | . | . | . | . | . | . | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **TAB1** | 22 | 39814746 | ns | c.560G>A | p.R187H | D | 4.45 | . | 6.166536 | rs140879164 | 0.0009 | 0.000233 | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **TAB1** | 22 | 39814802 | ns | c.616G>A | p.D206N | T | 4.45 | . | 4.232386 | rs148869940 | . | 0.000698 | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **TAB1** | 22 | 39826049 | ns | c.1337C>T | p.T446I | T | 4.67 | . | 5.706612 | rs118074217 | 0.0005 | 0.001163 | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **TAB1** | 22 | 39826137 | sn | c.1425C>T | p.D475D | . | . | . | . | rs147601362 | 0.0009 | 0.001395 | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **TAB1** | 22 | 39832516 | sn | c.1329C>T | p.S443S | . | . | . | . | . | . | 0.000116 | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **TAB2** | 6 | 149699333 | sn | c.282A>G | p.G94G | . | . | . | . | rs13215304 | 0.02 | 0.031163 | not listed | 100 | 0 | 0 | 100 | 0 | 0 |

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **TAB2** | 6 | 149699483 | sn | c.432T>C | p.S144S | . | . | . | . | . | . | . | not listed | 100 | 0 | 0 | 100 | 0 | 0 |
| **TAB2** | 6 | 149699483 | sn | c.432T>C | p.S144S | . | . | . | . | . | . | . | not listed | 100 | 0 | 0 | 100 | 0 | 0 |
| **TAB2** | 6 | 149700128 | sn | c.1077C>T | p.T359T | . | . | . | . | rs138731123 | . | . | not listed | 100 | 0 | 0 | 100 | 0 | 0 |
| **TAB2** | 6 | 149700491 | sn | c.1440G>A | p.V480V | . | . | . | . | rs3734296 | 0.21 | 0.105698 | not listed | 100 | 0 | 0 | 100 | 0 | 0 |
| **TAB2** | 6 | 149730846 | sn | c.2073G>A | p.R691R | . | . | . | . | rs652921 | 0.21 | 0.105698 | not listed | 100 | 0 | 0 | 100 | 0 | 0 |
| **TAB3** | X | 30849697 | sp | c.1991-5>CT | . | . | . | 0.12 | . | rs202074143 | . | 0.000595 | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **TAB3** | X | 30870971 | ns | c.1634C>G | p.S545C | T | 4.31 | . | 2.179871 | . | . | . | not listed | 100 | 0 | 0 | 99.06 | 0 | 0.94 |
| **TAB3** | X | 30873039 | ns | c.743C>T | p.T248M | D | 2.53 | . | 0.808718 | . | . | . | not listed | 99.26 | 0 | 0.74 | 100 | 0 | 0 |
| **TAB3** | X | 30873245 | sn | c.537G>A | p.P179P | . | . | . | . | rs146319957 | 0.0018 | 0.00431 | not listed | 98.53 | 1.47 | 0 | 100 | 0 | 0 |
| **TNF** | 6 | 31544562 | ns | c.251C>T | p.P84L | T | 0.92 | . | 1.458155 | rs4645843 | . | 0.002953 | not listed | 98.53 | 1.47 | 0 | 100 | 0 | 0 |
| **TNFAIP3** | 6 | 138196066 | ns | c.380T>G | p.F127C | T | 0.836 | . | 1.667118 | rs2230926 | 0.12 | 0.032093 | listed | 100 | 0 | 0 | 100 | 0 | 0 |
| **TNFAIP3** | 6 | 138196817 | sp | c.487-8>CG | . | . | . | 0.45 | . | rs5029947 | 0.02 | 0.00093 | not listed | 100 | 0 | 0 | 100 | 0 | 0 |
| **TNFAIP3** | 6 | 138199644 | sn | c.1062G>A | p.K354K | . | . | . | . | . | . | 0.000233 | not listed | 100 | 0 | 0 | 100 | 0 | 0 |
| **TNFAIP3** | 6 | 138199950 | sn | c.1368G>C | p.G456G | . | . | . | . | rs201600532 | 0.0005 | . | not listed | 100 | 0 | 0 | 100 | 0 | 0 |
| **TNFAIP3** | 6 | 138201240 | ns | c.1939T>C | p.S647P | T | 0.362 | . | 1.04551 | rs142253225 | 0.0009 | 0.002791 | not listed | 100 | 0 | 0 | 100 | 0 | 0 |
| **TNFAIP3** | 6 | 138202258 | sn | c.2175G>A | p.L725L | . | . | . | . | rs140354477 | 0.0018 | 0.002442 | not listed | 100 | 0 | 0 | 100 | 0 | 0 |
| **TNFAIP3** | 6 | 138202378 | sn | c.2295C>T | p.P765P | . | . | . | . | rs5029956 | 0.02 | 0.000465 | not listed | 100 | 0 | 0 | 100 | 0 | 0 |
| **TRAF6** | 11 | 36514122 | ns | c.735T>A | p.S245R | T | 3.02 | . | 2.144458 | . | . | 0.000116 | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **TRIP6** | 7 | 100465128 | sn | c.9G>A | p.G3G | . | . | . | . | . | . | . | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **TRIP6** | 7 | 100465747 | ns | c.255G>T | p.R85S | T | -5.88 | . | 0.195055 | rs139351872 | 0.0014 | 0.000698 | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **TRIP6** | 7 | 100465807 | sn | c.315C>T | p.A105A | . | . | . | . | rs144580285 | 0.0014 | 0.003605 | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **TRIP6** | 7 | 100465824 | ns | c.332G>A | p.R111Q | T | 3.13 | . | 1.472456 | rs2437100 | 0.01 | 0.020465 | not listed | 96.32 | 3.68 | 0 | 95.28 | 4.72 | 0 |
| **TRIP6** | 7 | 100466176 | sn | c.423C>T | p.A141A | . | . | 1.07 | . | . | . | . | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **TRIP6** | 7 | 100466441 | ns | c.688G>A | p.V230I | T | 4.66 | . | 1.751354 | rs2075756 | 0.28 | 0.269014 | not listed | 55.15 | 38.97 | 5.88 | 66.98 | 27.36 | 5.66 |
| **TRIP6** | 7 | 100466457 | ns | c.704G>C | p.G235A | T | 2.36 | . | 1.995768 | . | . | . | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **TRIP6** | 7 | 100468284 | sn | c.918a>G | p.V306V | . | . | . | 2.457742 | rs1054391 | 0.57 | 0.523372 | not listed | 25.74 | 49.26 | 25 | 27.36 | 47.17 | 25.47 |
| **TRIP6** | 7 | 100468345 | ns | c.979T>C | p.Y327H | D | 5.47 | . | 2.389748 | . | . | . | not listed | 98.53 | 1.47 | 0 | 100 | 0 | 0 |
| **TRIP6** | 7 | 100469219 | ns | c.1054C>T | p.R352W | D | 3.04 | . | 2.48622 | . | . | . | not listed | 100 | 0 | 0 | 99.06 | 0.94 | 0 |
| **TRIP6** | 7 | 100469223 | ns | c.1058C>T | p.A353V | D | 4.21 | . | 2.699637 | rs147492293 | . | 0.000349 | not listed | 99.26 | 0.74 | 0 | 100 | 0 | 0 |
| **XIAP** | X | 123034511 | ns | c.1268A>C | p.Q423P | T | 4.26 | . | 0.733377 | rs5956583 | 0.261 | 0.33 | listed | 52.21 | 24.26 | 23.53 | 60.38 | 16.04 | 23.58 |
| **XIAP** | X | 123040945 | ns | c.1408A>T | p.T470S | T | 4.11 | . | 1.624093 | rs143165174 | . | 0.000595 | listed | 99.26 | 0 | 0.74 | 100 | 0 | 0 |

ns, non-synonymous; sn, synonymous; fi, frameshift insertion, fd, frameshift deletion; sp, splicing; nfi, non-frameshift insertion; nfd, non-frameshift deletion,sp, splicing

B, benign; C, Conservative; D, deleterious; MC, moderately Conservative; MR, moderately Radical; NR, not reported; P, possibly damaging; R

# Appendix VI

**Expression of the HSPA1L protein**

The HSPA1L gene consists of a single exon. The coding region was amplified using genomic DNA from the affected patient 12s of Family A by PCR, and the PCR products were cloned into a pCR-Blunt II-TOPO vector (Invitrogen). After cloning, a common single nucleotide variant rs2227956 was reverted to its reference sequence (WT) by using a QuikChange II Site-Directed Mutagenesis Kit (Agilent Technologies, La Jolla, CA), and p.Lys73Ser (c.218A>G, c.219A>C), p.Gly77Ser (c.229G>A), p.Leu172del (c.515-517del), p.Thr267Ile (c.800C>T), p.Ala268Thr (c.802G>A), p.Ser277Lys (c.830C>T), and p.Glu558Asp (c.1674A>T) mutants were generated and subsequently cloned into pGEX-6P-1 vector (GE Healthcare, Waukesha, WI) at the BamHI-NotI restriction site. All sequences were confirmed by Sanger sequencing analysis at the Protein and Nucleic Acid Facility (Stanford University).

The resulting vector was transformed into Escherichia coli strain BL21 (New England Biolabs., Ipswich, MA) and recombinant fusion protein with a glutathione S-transferase (GST) tag was expressed by induction with 0.1 mM of isopropyl-β-thiogalacto-pyranoside (Sigma) for 5-6 hours at 28°C. Cells were pelleted and resuspended in lysis buffer (50 mM pH7.5 Tris-HCl, 150 mM NaCl, 0.05% NP-40) and lysed with 0.25 mg/mL lysozyme (EMD Millipore, Billerica, MA) on ice for 30 minutes. The samples were then sonicated and centrifuged at 20,000 × g for 20 minutes. The resulting supernatants were incubated with Glutathione Sepharose 4B beads (GE Healthcare) for 3 hours at 4°C. Recombinant protein-bound beads were subsequently washed with lysis buffer, and incubated with PreScission Protease (GE Healthcare) overnight at 4°C. Protein concentration was measured by Bradford assay. The eluted protein was concentrated as necessary by using Amicon Ultracel-3K columns (Millipore, MA, USA). The purified protein samples were aliquoted and stored at -80°C.

**In vitro chaperone assay**

In vitro chaperone activity was measured with the HSP70/HSP40 Glow-Fold Protein Refolding Kits (K-290, Boston Biochem, Inc., Cambridge, MA) according to the manufacturer's protocol with modifications. In brief, recombinant HSPA1L protein (4 μM), a 1:1 mixture of recombinant HSPA1L WT protein (2 μM) and HSPA1L mutant protein, or a 1:1 mixture of HSPA1A protein (2 μM, Boston Biochem, Inc.) and recombinant HSPA1L protein (2 μM) was used to test for refolding efficiency of heat-denatured Glow-Fold Substrate protein. Luminescence measurements were taken using a TECAN infinite 200 microplate reader (TECAN Austria GmbH, Salzburg, Austria) at indicated time points within 1 minute of mixing with luciferin reagent. Refolding activity was calculated by subtracting the luminescence at time 0 (before refolding reaction) from that at 120 minutes (after refolding reaction). Refolding activity of each control at 120 minutes was set as 100%. Data were compared between the control and test samples using Dunnett's multiple comparison test.

# Appendix VII

## Summary statistics for exome sequencing – mapping and coverage

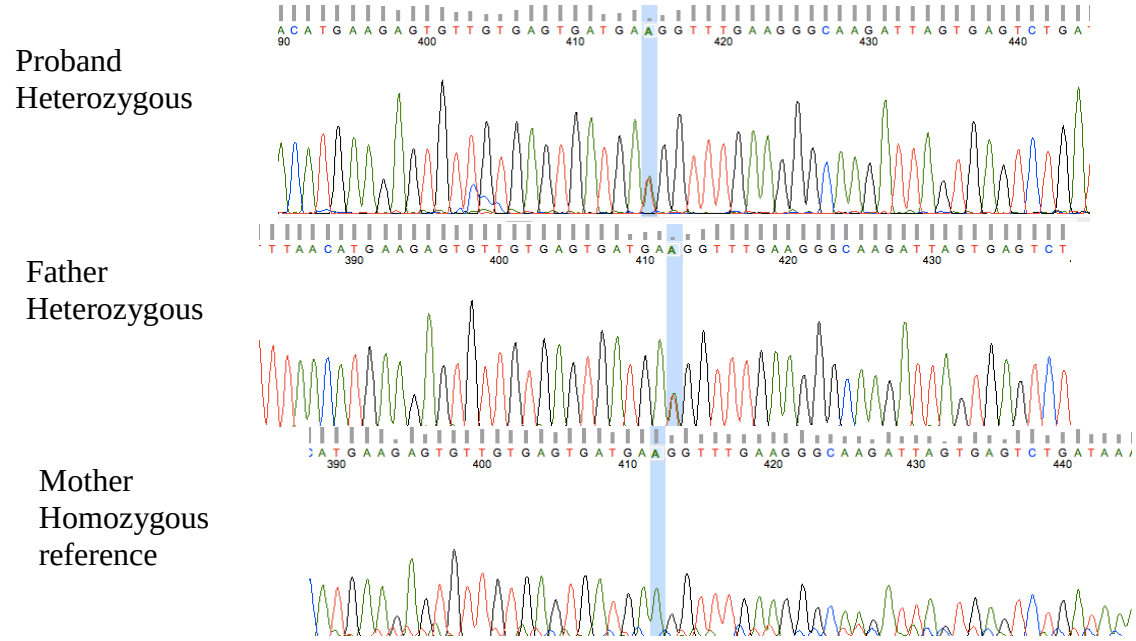
| Sample ID | Agilent Exome capture | Number of sequenced reads | Total no. aligned reads | Total no. unique align. | Mapped to target reads +/-150bp (%) | Mapped to target reads (%) | Target bases with coverage >1 (%) | Target bases with coverage >5 (%) | Target bases with coverage >10 (%) | Target bases with coverage >20 (%) | Mean coverage |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 12s[#] | V5+ UTRs | 87387846 | 86684829 | 66127387 | 75.93 | 70.16 | 99.8 | 99.2 | 97.7 | 92.1 | 65.14 |
| PR0034[*] | V5 | 44626172 | 44309916 | 43565707 | 88.16 | 75.89 | 99.28 | 98.35 | 95.87 | 86.56 | 55.43 |
| PR0151[*] | V4 | 36355666 | 36122486 | 35800555 | 88.72 | 83.31 | 99.85 | 99.16 | 97.39 | 91.51 | 97.21 |
| PR0161[*] | V4 | 50590970 | 49467147 | 49007862 | 84.3 | 76.74 | 99.68 | 98.28 | 94.77 | 83.67 | 62.04 |
| PR0142[*] | V4 | 67676356 | 66652672 | 66223398 | 94.55 | 85.01 | 99.83 | 99.14 | 97.57 | 92.08 | 86.13 |
| PR0156[*] | V5 | 46590950 | 46257701 | 45479121 | 87.42 | 75.31 | 99.26 | 98.37 | 96.08 | 87.26 | 57.04 |
| PR0244[*] | V5 | 48919246 | 48579280 | 47801772 | 82.28 | 99.31 | 99.27 | 98.46 | 97.21 | 91.31 | 58.94 |

#, from Family A; *, from IBD cohort. Number of sequenced reads - total number of reads sequenced; Total no. aligned reads - the total number of reads aligned to the reference sequence; Total no. unique align- the number of reads that uniquely mapped to the reference sequence; Mapped to target reads +/-150bp (%)- the percentage of reads mapped ±150 base pair to the target; Mapped to target reads (%)- the percentage of reads mapped to the target sequence; Target bases with coverage >1,5,10,20- the percentage of targets with 1, 5, 10 and 20 read depth; Mean coverage - the mean of the depth coverage.

## Appendix VIII

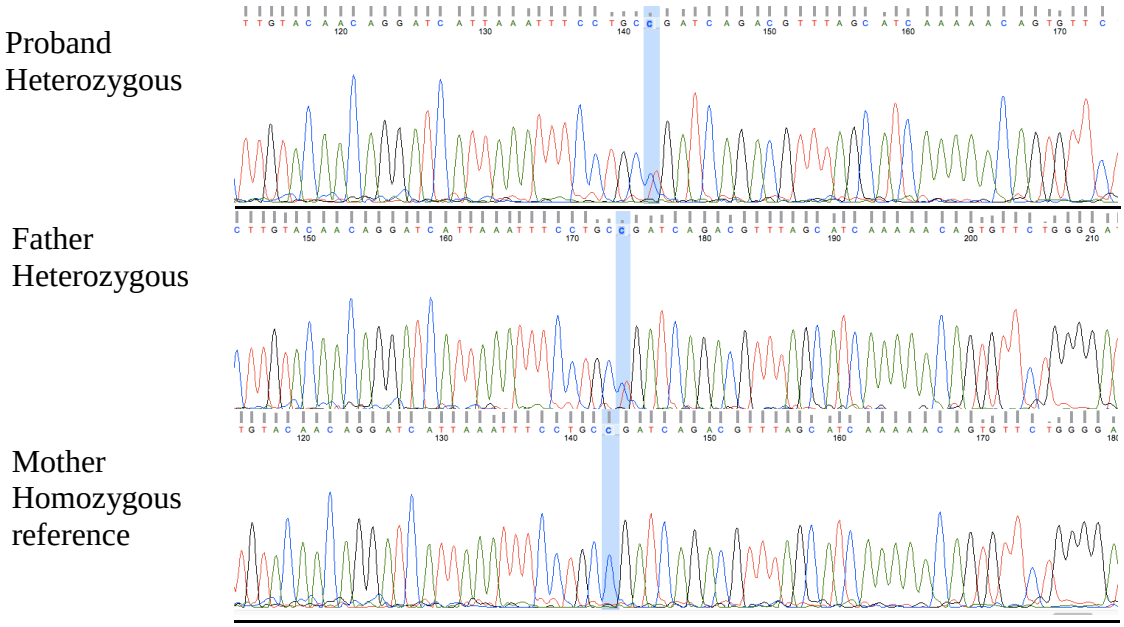
**Figure 1a-f.** Sanger traces of each of the four variants of interest found across six pedigrees (Table 5.4). Genotypic state was confirmed in all available family members.

**Figure 1a. Sanger trace for Pedigree 34**




Variant: HSPA1L:NM_005527:exon2:c.A1674T:p.E558D
Primers used: Forward ACTGCCCTGATAAAGCGCAA; Reverse GGGGCCTAGTTTTCCTGAGTC
Heterozygous status in proband and father (unaffected)

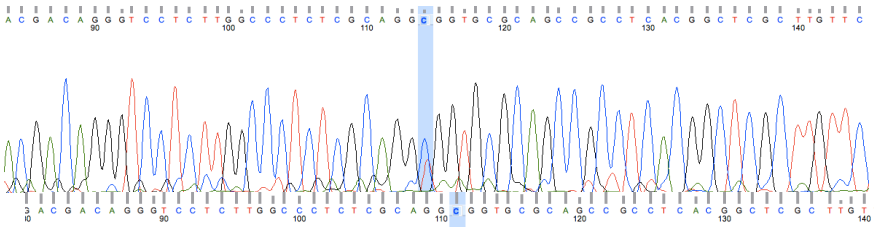
**Figure 1b. Sanger trace for Pedigree 142**



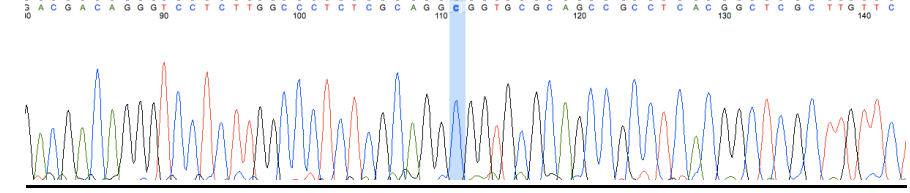

Variant: HSPA1L:NM_005527:exon2:c.G229A:p.G77S
Primers used: Forward TTGACAACAGGCTTGTGAGC;
Reverse AAATCGAGCTCTGGTGATGG
Heterozygous status in proband and father (unaffected)

**Figure 1c. Sanger trace for Pedigree 151**

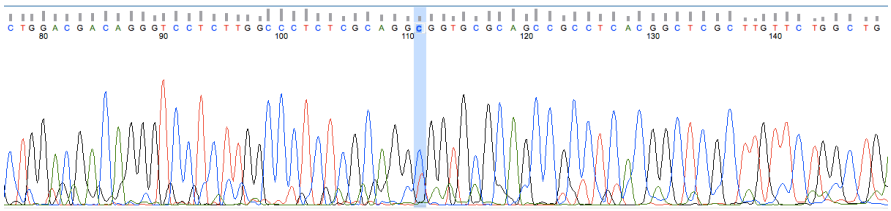
Proband
Heterozygous



Father
Homozygous
reference



Mother
Heterozygous



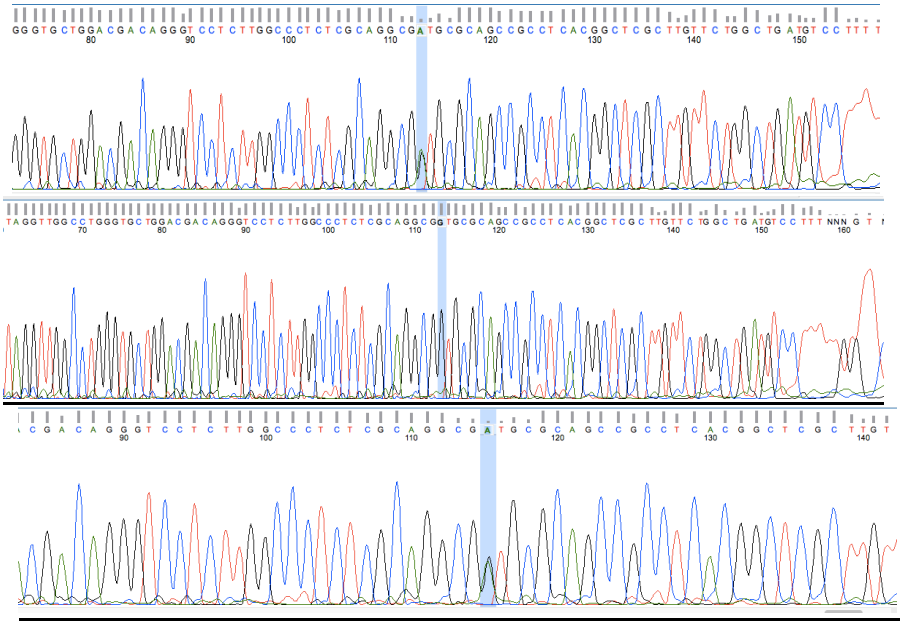
Variant :HSPA1L:NM_005527:exon2:c.G802A:p.A268T
Primers used: Forward TTGACAACAGGCTTGTGAGC;
Reverse AAATCGAGCTCTGGTGATGG
Heterozygous status in proband and mother (unaffected)

**Figure 1d. Sanger trace for Pedigree 156**

Proband
Heterozygous

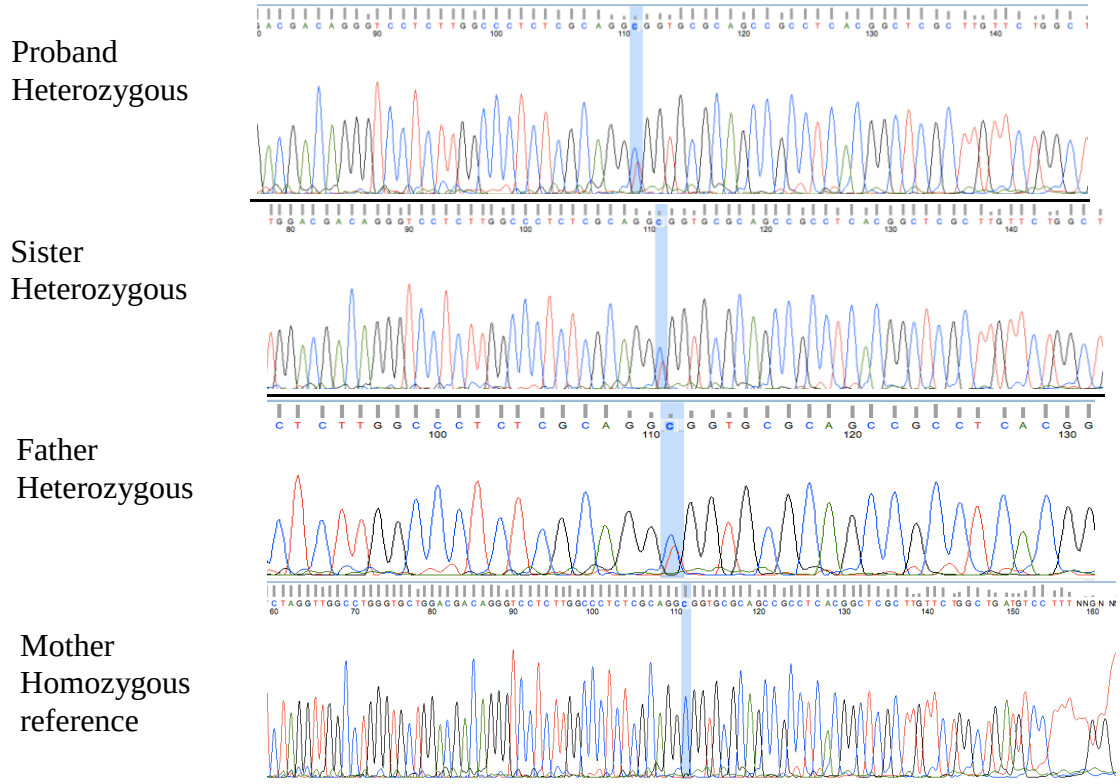
Father
Homozygous
reference

Mother
Heterozygous



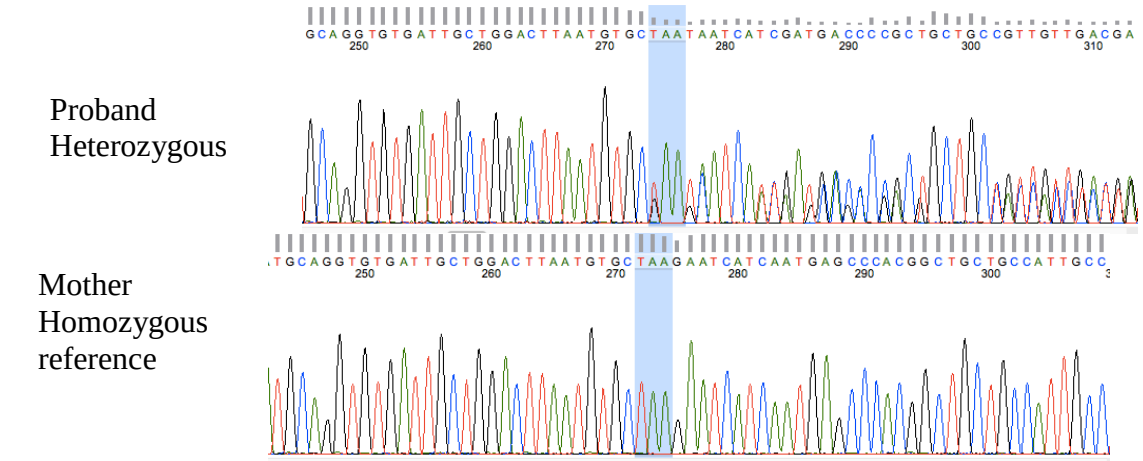
Variant: HSPA1L:NM_005527:exon2:c.C800T:p.T267I
Primer: Forward TTGACAACAGGCTTGTGAGC;
Reverse AAATCGAGCTCTGGTGATGG
Heterozygous status in proband and mother (unaffected)

**Figure 1e. Sanger trace for Pedigree 161**



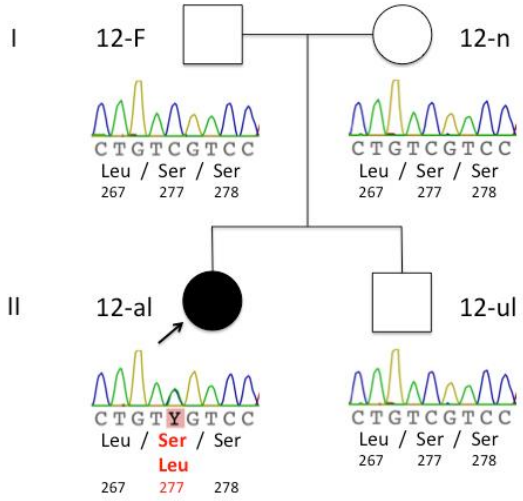

Variant : HSPA1L:NM_005527:exon2:c.G802A:p.A268T
Primers used: Forward TTGACAACAGGCTTGTGAGC;
Reverse AAATCGAGCTCTGGTGATGG
Heterozygous status in proband, unaffected father (unaffected) and sister (ulcerative colitis)

**Figure 1f Sanger trace for Pedigree 244**




Variant: HSPA1L:NM_005527:exon2:c.515_517del:p.172_173del
Primers used: Forward ACTGCCCTGATAAAGCGCAA;
Reverse GGGGCCTAGTTTTCCTGAGTC
Heterozygous status in proband

**Figure 1g Pedigree Family A**




Primers used: Forward TAGATGATGGGATTTTTGAGGTA; Reverse

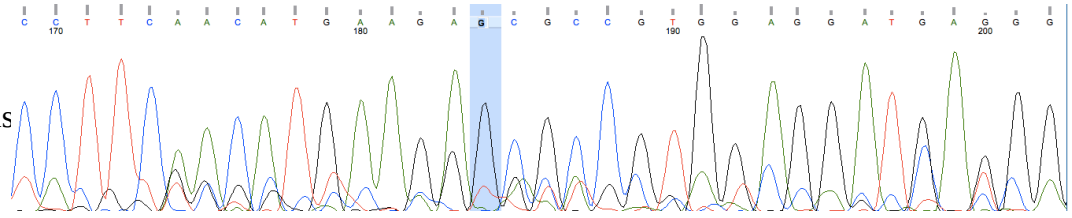
CTACTAAAACAATGTCATGGATTTT

**Figure 1f Sanger trace for Pedigree 111**

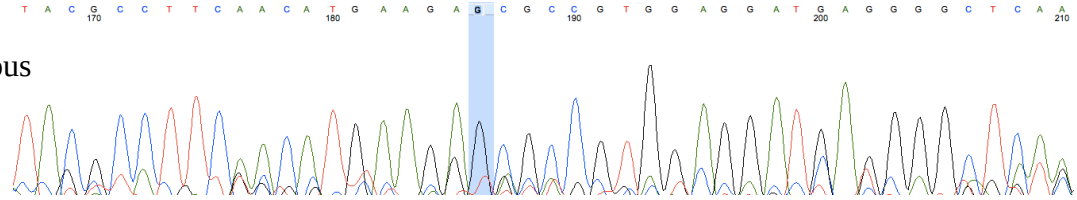
Variant: HSPA1A:NM_005345:exon1:c.G1652T:p.S551I

***HSPA1A***

Proband
Homozygous
reference



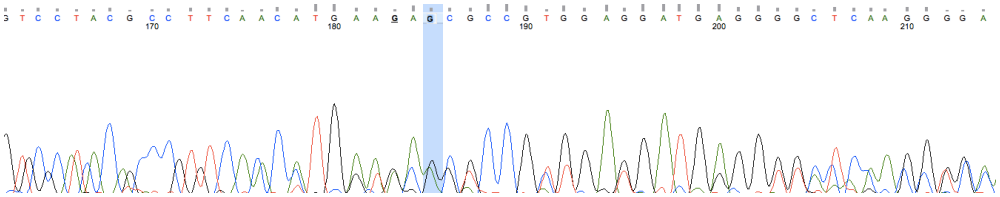
Mother
Homozygous
reference



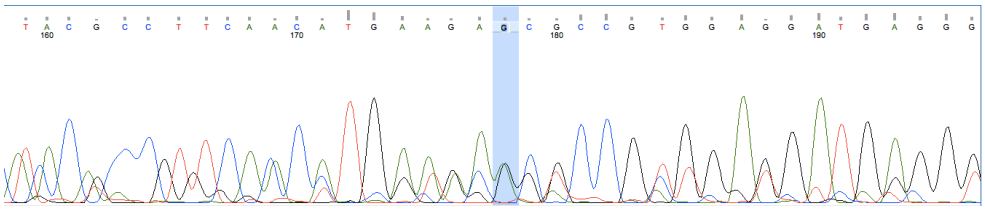
Primers used: forward: TCTCGCGGATCCAGTGTTC, reverse: TCCAAAACAAAAACAGCAATCTTGG

***HSPA1B***

Proband
Homozygous
reference



Mother
Homozygous
reference

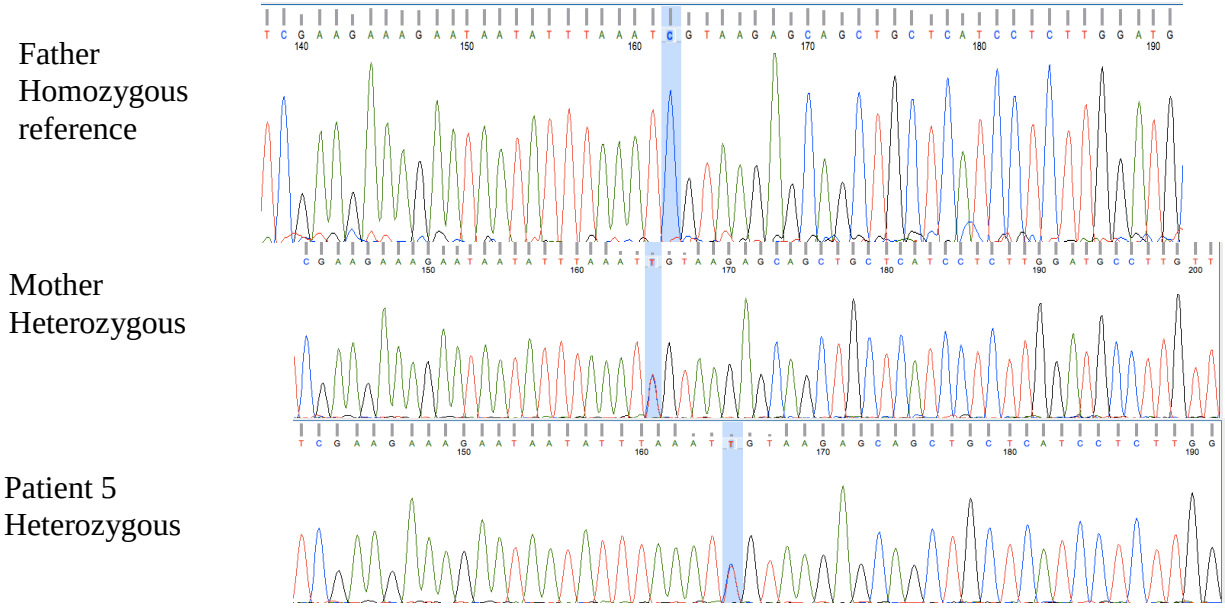


Primers used: forward: TTGTCGCGGATCCCGTCCG,
Reverse: GAAGTGAAGCAGCAAAGAGCTGAAGC

## Appendix IX

**Figure 1a-c.** Sanger traces of each of the three variants of interest found across two pedigrees. Genotypic state was confirmed in all available family members.

**Figure 1a. Sanger trace for Pedigree of patient 5**




Variant: FERMT1:NM_017671:exon6:c.G812A:p.R271Q,

Primers used: Forward TCAGAGACCAGGGTCCATGTAT; Reverse GGCTAGACTCCTCACGCTCC

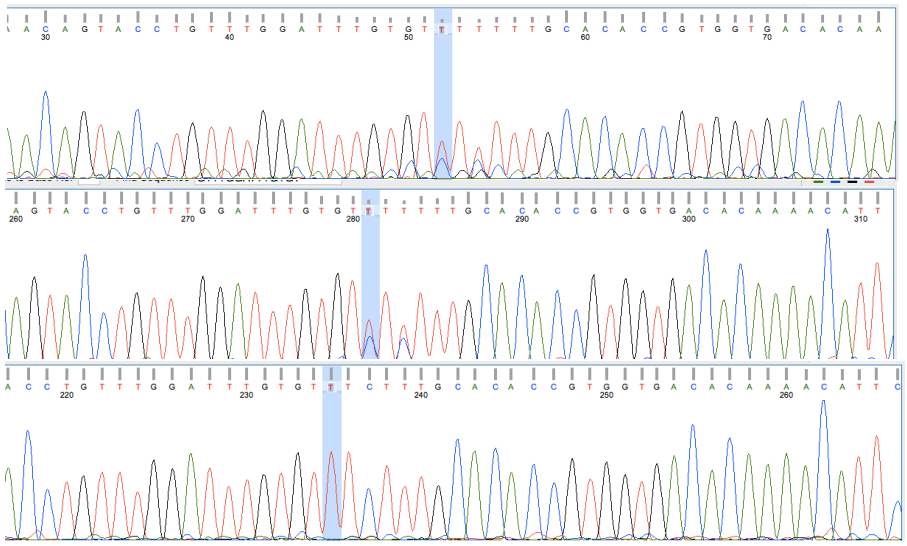
Heterozygous status in proband and mother (unaffected)

**Figure 1b. Sanger trace for Pedigree of patient 5**

Father heterozygous

Mother heterozygous

Patient 5 Homozygous alternative

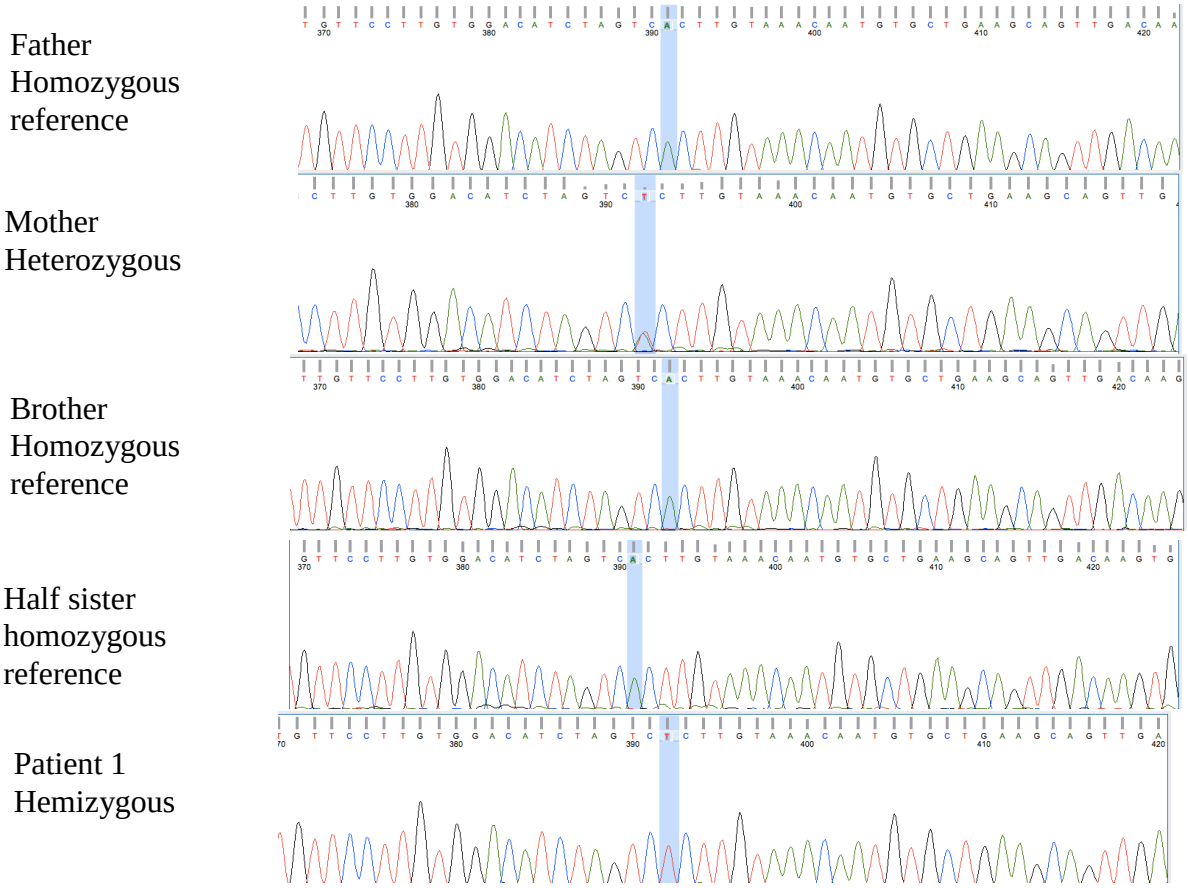


Variant: FERMT1:NM_017671:exon12:c.G1577A:p.R526K,

Primers used: Forward TGCAAACAATTGCCCTAACAAGATT; Reverse ATACGCCCAATGGATGGCTG

Homozygous status in proband and heterozygous status in mother and father (unaffected)

**Figure 1c. Sanger trace for Pedigree of patient 6**




Variant: XIAP:NM_001167:exon7:c.A1408T:p.T470S

Primers used: Forward AACCTGTGAAGCCTTCTCCAACT; Reverse

CTGTGTAGCACATGGGACACTTG

Heterozygous status in proband and mother (unaffected)

## Glossary

Alelle

An allele is one of several alternative forms of a gene or DNA sequence at a specific chromosomal location. At each autosomal location an individual possesses two alleles, one inherited from the father and one from the mother[93–95].

Allele frequency

Allele frequency is the proportion of a particular allele (variant of a gene) among all allele copies being considered in a population[93–95].

Biotin-streptavidin system

Biotin-streptavidin system is a tool for isolating molecules. The streptavidin protein binds biotin with high affinity therefore biotinylated molecules can be isolated using streptavidin coated magnetic beads[93–95].

Case-control study

A case control study is a type of study which compares a group of affected patients, cases and healthy individuals, controls[93–95].

Chromatid

A chromatid is one copy of a duplicated chromosome, which is generally joined to the other copy by a single centromere[93–95].

Coding DNA

The coding region of a gene is the portion of a gene's DNA coding for protein[93–95].

Codon

A codon is a sequence of three nucleotides which together form a unit of genetic code in a DNA or RNA molecule[93–95].

Common mutations

Common mutations are genetic changes occurring in more than 1% of the population[93–95].

Compound heterozygous

Compound heterozygosity is the genotype state of having two heterozygous alleles at a particular locus[93–95].

Conserved sequenced

Conserved sequences are similar or identical sequences that occur within nucleic acid sequences, protein sequences, protein structures or polymeric carbohydrates across species or within different molecules produced by the same organism[93–95].

Copy Number Variant (CNV)

Copy number variations (CNVs) are as important a component of genomic diversity as single nucleotide polymorphisms (SNPs). Redon*et al.* (2006) defined a CNV as a DNA segment of one kilobase (kb) or larger that is present at a variable copy number in comparison with a reference genome[93–95].

Exome

The exome is the part of the genome formed by exons, the sequences which when transcribed remain within the mature RNA after introns are removed by RNA splicing[93–95].

Frameshift

A frameshift mutation (also called a framing error or a reading frame shift) is a genetic mutation caused by indels (insertions or deletions) of a number of nucleotides in a DNA sequence that is not divisible by three[93–95].

Gain/loss of function

Gain/loss of function mutations are a type of mutations in which the altered gene product possesses a new/loses molecular function or a new pattern of gene expression[93–95].

Gene therapy

The therapeutic application of either blocking the unwanted result of a defective gene or repairing a defective gene by inserting a new gene into the chromosome of the cell[93–95].

Genome

Genome is the complete set of genes or genetic material present in a cell or organism.The human genome is made up of $3x10^9$ base pairs of DNA which encode for approximately 20 000-25 000 genes[367].

Genome bowser

A genome browser is a graphical interface for display of information from a biological database for genomic data[93–95].

Genotype

The genetic constitution of an individual at a specific locus[93–95].

Haplotype

Haplotype is the genetic constitution of an individual with respect to one member of a pair of allelic genes. A haplotype can refer to only one locus or to an entire genome[93–95].

Hardy-Weinberg equilibrium

The Hardy–Weinberg principle states that allele and genotype frequencies within a population will remain constant from generation to generation in the absence of other evolutionary influences[93–95].

Hemizygous

Hemizygous refers to a diploid cell with only one copy of a gene instead of the usual two copies (e.g. XY in Homo sapiens males) [93–95].

Homozygous/Heterozygous

Homozygous refers to two identical alleles at a particular chromosomal locus/ in an individual. Heterozygous refers to having two different allele at a particular locus/ in an individual [93–95].

Indel

Indel refers the insertion or the deletion of bases in the DNA of an organism[93–95].

Inflammatory bowel disease (IBD)

Inflammatory bowel disease (IBD) is a group of inflammatory conditions of the colon and small intestine. Crohn's disease and ulcerative colitis are the principal types of inflammatory bowel disease[93–95].

Intron

An intron is a segment of a DNA or RNA molecule which does not code for proteins and interrupts the sequence of genes[93–95].

Induced pluripotent stem cells

Induced pluripotent stem cells (also known as iPS cells or iPSCs) are a type of pluripotent stem cell that can be generated directly from adult cells[93–95].

Linkage disequilibrium (LD)

LD is a statistical association test estimating the non-random associations between alleles at two separate loci[93–95].

Locus

Locus is the position of a chromosome defining a gene or DNA sequence[93–95].

Lod score (Z)

Lod score is a statistical estimate of whether two genes, or a gene and a disease gene, are likely to be located near each other on a chromosome and are therefore likely to be inherited. It is a measure of the likelihood two loci are in genetic linkage[93–95].

MircoRNAs

Small RNA molecules (21-22 nucleotides) that regulate gene expression[93–95].

Meiosis/Mitosis

Mitosis and meiosis are two ways that cells reproduce. By mitosis a cell splits to create two identical copies of the original cell. In meiosis cells split to form new cells with half the usual number of chromosomes, to produce gametes for sexual reproduction[93–95].

Polymorphism

Natural variations in a gene, DNA sequence, or chromosome that have no adverse effects on the individual and occur with fairly high frequency (more than 1%) in the general population[93–95].

Private mutations

A distinct gene alteration observed in a single family[93–95].

Rare mutations

Variations of a gene occurring in less than 1% within a population[93–95].

Variant of unknown significance (VUS)

A variation in a genetic sequence whose association with disease risk is unknown[93–95].

## Bibliography


1. Khor, B., Gardet, A. & Xavier, R. J. Genetics and pathogenesis of inflammatory bowel disease. *Nature* **474,** 307–317 (2011).
2. Baumgart, D. C. & Sandborn, W. J. Crohn's disease. *Lancet* **380,** 1590–605 (2012).
3. Aloi, M. *et al.* Phenotype and Disease Course of Early-onset Pediatric Inflammatory Bowel Disease. *Inflamm. Bowel Dis.* (2014).
4. Baumgart, D. C. & Sandborn, W. J. Crohn's disease. *Lancet* **380,** 1590–1605 (2012).
5. Ordás, I., Eckmann, L., Talamini, M., Baumgart, D. C. & Sandborn, W. J. Ulcerative colitis. *Lancet* **380,** 1606–1619 (2012).
6. Baumgart, D. C. & Sandborn, W. J. Inflammatory bowel disease: clinical aspects and established and evolving therapies. *Lancet* **369,** 1641–57 (2007).
7. Trier, J. S. Studies on small intestinal crypt epithelium. i. the fine structure of the crypt epithelium of the proximal small intestine of fasting humans. *J. Cell Biol.* **18,** 599–620 (1963).
8. Diefenbach, K.-A. & Breuer, C.-K. Pediatric inflammatory bowel disease. *World J. Gastroenterol.* **12,** 3204–12 (2006).
9. Satsangi, J., Silverberg, M. S., Vermeire, S. & Colombel, J.-F. F. The Montreal classification of inflammatory bowel disease: controversies, consensus, and implications. *Gut* **55,** 749–53 (2006).
10. Müller, K. E. *et al.* Incidence, Paris classification, and follow-up in a nationwide incident cohort of pediatric patients with inflammatory bowel disease. *J. Pediatr. Gastroenterol. Nutr.* **57,** 576–82 (2013).
11. Fell, J. M. E. Update of the management of inflammatory bowel disease. *Arch. Dis. Child.* **97,** 78–83 (2012).
12. Okou, D. T. & Kugathasan, S. Role of Genetics in Pediatric Inflammatory Bowel Disease. *Inflamm. Bowel Dis.* **20,** 1878–1884 (2014).
13. Levine, A. *et al.* Pediatric modification of the Montreal classification for inflammatory bowel disease: The Paris classification. *Inflamm. Bowel Dis.* **17,** 1314–1321 (2011).
14. Kundhal, P. S. *et al.* Pediatric Crohn Disease Activity Index: responsive to short-term change. *J. Pediatr. Gastroenterol. Nutr.* **36,** 83–9 (2003).
15. Turner, D. *et al.* Development, validation, and evaluation of a pediatric ulcerative colitis activity index: a prospective multicenter study. *Gastroenterology* **133,** 423–32 (2007).
16. Naspghan Guidelines. at <http://www.naspghan.org/>
17. Loftus, E. V. Clinical epidemiology of inflammatory bowel disease: Incidence, prevalence, and environmental influences. *Gastroenterology* **126,** 1504–1517

(2004).

18. Farrokhyar, F., Swarbrick, E. T. & Irvine, E. J. A critical review of epidemiological studies in inflammatory bowel disease. *Scand J Gastroenterol* **36,** 2–15 (2001).

19. Ashton, J. J. *et al.* Rising incidence of paediatric inflammatory bowel disease (PIBD) in Wessex, Southern England. *Arch. Dis. Child.* (2014).

20. Benchimol, E. I. *et al.* Epidemiology of pediatric inflammatory bowel disease: a systematic review of international trends. *Inflamm Bowel Dis* **17,** 423–439 (2011).

21. Benchimol, E. I. *et al.* Increasing incidence of paediatric inflammatory bowel disease in Ontario, Canada: evidence from health administrative data. *Gut* **58,** 1490–1497 (2009).

22. M'Koma, A. E. Inflammatory bowel disease: an expanding global health problem. *Clin. Med. Insights. Gastroenterol.* **6,** 33–47 (2013).

23. Shivananda, S. *et al.* Incidence of inflammatory bowel disease across Europe: is there a difference between north and south? Results of the European Collaborative Study on Inflammatory Bowel Disease (EC-IBD). *Gut* **39,** 690–7 (1996).

24. Henderson, P. *et al.* Rising incidence of pediatric inflammatory bowel disease in Scotland. *Inflamm. Bowel Dis.* **18,** 999–1005 (2012).

25. Jussila, A. *et al.* High and increasing prevalence of inflammatory bowel disease in Finland with a clear North–South difference. *J. Crohn's Colitis* **7,** e256–e262 (2013).

26. Kaplan, G. G. The global burden of IBD: from 2015 to 2025. *Nat. Rev. Gastroenterol. Hepatol.* **12,** 720–727 (2015).

27. Mahid, S. S., Mulhall, A. M., Gholson, R. D., Eichenberger, M. R. & Galandiuk, S. Inflammatory bowel disease and African Americans: a systematic review. *Inflamm. Bowel Dis.* **14,** 960–7 (2008).

28. Van Limbergen, J. *et al.* Definition of phenotypic characteristics of childhood-onset inflammatory bowel disease. *Gastroenterology* **135,** 1114–22 (2008).

29. Ruel, J., Ruane, D., Mehandru, S., Gower-Rousseau, C. & Colombel, J.-F. IBD across the age spectrum-is it the same disease? *Nat. Rev. Gastroenterol. Hepatol.* **11,** 88–98 (2014).

30. Jedel, S., Hood, M. M. & Keshavarzian, A. Getting personal: a review of sexual functioning, body image, and their impact on quality of life in patients with inflammatory bowel disease. *Inflamm. Bowel Dis.* **21,** 923–38 (2015).

31. Lönnfors, S. *et al.* IBD and health-related quality of life - Discovering the true impact. *J. Crohns. Colitis* (2014).

32. Perrin, J. M. *et al.* Measuring quality of life in pediatric patients with inflammatory bowel disease: psychometric and clinical characteristics. *J. Pediatr. Gastroenterol. Nutr.* **46,** 164–71 (2008).

33. Cosnes, J. Smoking, physical activity, nutrition and lifestyle: environmental

factors and their impact on IBD. *Dig Dis* **28,** 411–417 (2010).

34. Andreoletti, G. *et al.* Exome analysis of patients with concurrent pediatric inflammatory bowel disease (PIBD) and autoimmune disease. *Inflamm. Bowel Dis.* (2015).

35. Sawczenko, A. & Sandhu, B. K. Presenting features of inflammatory bowel disease in Great Britain and Ireland. *Arch Dis Child* **88,** 995–1000 (2003).

36. Pallis, A. G., Vlachonikolis, I. G. & Mouzas, I. A. Assessing health-related quality of life in patients with inflammatory bowel disease, in Crete, Greece. *BMC Gastroenterol.* **2,** 1 (2002).

37. Kalafateli, M. *et al.* Health-related quality of life in patients with inflammatory bowel disease: a single-center experience. *Annals of Gastroenterology* **26,** 243 (2013).

38. Griffiths, A. M. *et al.* Development of a quality-of-life index for pediatric inflammatory bowel disease: dealing with differences related to age and IBD type. *J. Pediatr. Gastroenterol. Nutr.* **28,** S46-52 (1999).

39. Carter, M. J., Lobo, A. J. & Travis, S. P. L. Guidelines for the management of inflammatory bowel disease in adults. *Gut* **53 Suppl 5,** V1-16 (2004).

40. Danese, S. New therapies for inflammatory bowel disease: from the bench to the bedside. *Gut* **61,** 918–932 (2012).

41. Kansal, S., Wagner, J., Kirkwood, C. D. & Catto-Smith, A. G. Enteral Nutrition in Crohn's Disease: An Underused Therapy. *Gastroenterol. Res. Pract.* **2013,** 1–11 (2013).

42. Madsen, K. L. *et al.* Interleukin-10 gene-deficient mice develop a primary intestinal permeability defect in response to enteric microflora. *Inflamm. Bowel Dis.* **5,** 262–70 (1999).

43. Escher, J. C., Taminiau, J. A. J. M., Nieuwenhuis, E. E. S., Büller, H. A. & Grand, R. J. Treatment of inflammatory bowel disease in childhood: best available evidence. *Inflamm. Bowel Dis.* **9,** 34–58 (2003).

44. Uchiyama, K. *et al.* New genetic biomarkers predicting azathioprine blood concentrations in combination therapy with 5-aminosalicylic acid. *PLoS One* **9,** e95080 (2014).

45. Su, C. G., Judge, T. A. & Lichtenstein, G. R. The role of biological therapy in inflammatory bowel disease. *Drugs Today (Barc).* **37,** 121–133 (2001).

46. Denmark, V. K. & Mayer, L. Current status of monoclonal antibody therapy for the treatment of inflammatory bowel disease: an update. *Expert Rev. Clin. Immunol.* **9,** 77–92 (2013).

47. Travassos, W. J. & Cheifetz, A. S. Infliximab: Use in Inflammatory Bowel Disease. *Curr. Treat. Options Gastroenterol.* **8,** 187–196 (2005).

48. Pech, T. *et al.* Combination therapy of tacrolimus and infliximab reduces inflammatory response and dysmotility in experimental small bowel transplantation in rats. *Transplantation* **93,** 249–56 (2012).

49. Scaldaferri, F. *et al.* Gut microbial flora, prebiotics, and probiotics in IBD: their current usage and utility. *Biomed Res. Int.* **2013,** 435268 (2013).

50. Jonkers, D. & Stockbrügger, R. Probiotics and inflammatory bowel disease. *J. R. Soc. Med.* **96,** 167–71 (2003).

51. Virta, L. J. & Kolho, K.-L. Antidepressant use among paediatric patients with recent-onset inflammatory bowel disease: A nationwide case control study in Finland. *J. Paediatr. Child Health* **50,** 562–565 (2014).

52. NICE: National Institute for Health and Care Excellence. at <http://www.nice.org.uk/>

53. Mowat, C. *et al.* Guidelines for the management of inflammatory bowel disease in adults. *Gut* **60,** 571–607 (2011).

54. Jenkins, H. R. Inflammatory bowel disease. *Arch Dis Child* **85,** 435–437 (2001).

55. Halme, L. *et al.* Family and twin studies in inflammatory bowel disease. *World J. Gastroenterol.* **12,** 3668–72 (2006).

56. Spehlmann, M. E. *et al.* Epidemiology of inflammatory bowel disease in a German twin cohort: results of a nationwide study. *Inflamm Bowel Dis* **14,** 968–976 (2008).

57. Aujnarain, A., Mack, D. R. & Benchimol, E. I. The role of the environment in the development of pediatric inflammatory bowel disease. *Curr. Gastroenterol. Rep.* **15,** 326 (2013).

58. Morgan, X. C. *et al.* Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol.* **13,** R79 (2012).

59. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491,** 119–124 (2012).

60. Gearry, R. B. & Dodgshun, A. J. The 'hygiene hypothesis' in IBD. *J. Crohns. Colitis* **6,** 869; author reply 870 (2012).

61. Koloski, N.-A., Bret, L. & Radford-Smith, G. Hygiene hypothesis in inflammatory bowel disease: a critical review of the literature. *World J. Gastroenterol.* **14,** 165–73 (2008).

62. Makola, D., Peura, D. A. & Crowe, S. E. Helicobacter pylori infection and related gastrointestinal diseases. *J. Clin. Gastroenterol.* **41,** 548–58 (2007).

63. Chapman-Kiddell, C. A., Davies, P. S. W., Gillen, L. & Radford-Smith, G. L. Role of diet in the development of inflammatory bowel disease. *Inflamm. Bowel Dis.* **16,** 137–51 (2010).

64. Hou, J. K., Abraham, B. & El-Serag, H. Dietary intake and risk of developing inflammatory bowel disease: a systematic review of the literature. *Am. J. Gastroenterol.* **106,** 563–73 (2011).

65. Amre, D. K. *et al.* Imbalances in dietary consumption of fatty acids, vegetables, and fruits are associated with risk for Crohn's disease in children. *Am. J. Gastroenterol.* **102,** 2016–25 (2007).

66. Ananthakrishnan, A. N. *et al.* Long-term intake of dietary fat and risk of

ulcerative colitis and Crohn's disease. *Gut* **63,** 776–84 (2014).

67. Tedelind, S., Westberg, F., Kjerrulf, M. & Vidal, A. Anti-inflammatory properties of the short-chain fatty acids acetate and propionate: a study with relevance to inflammatory bowel disease. *World J. Gastroenterol.* **13,** 2826–32 (2007).

68. Russel, M. G. *et al.* Appendectomy and the risk of developing ulcerative colitis or Crohn's disease: results of a large case-control study. South Limburg Inflammatory Bowel Disease Study Group. *Gastroenterology* **113,** 377–82 (1997).

69. Feeney, M. A. *et al.* A case-control study of childhood environmental risk factors for the development of inflammatory bowel disease. *Eur J Gastroenterol Hepatol* **14,** 529–534 (2002).

70. Molodecky, N. A. & Kaplan, G. G. Environmental risk factors for inflammatory bowel disease. *Gastroenterol. Hepatol. (N. Y).* **6,** 339–46 (2010).

71. Eysenck, H. J. Meta-analysis and its problems. *BMJ* **309,** 789–92 (1994).

72. Bonen, D. K. & Cho, J. H. The genetics of inflammatory bowel disease. *Gastroenterology* **124,** 521–536 (2003).

73. Parkes, M., Cortes, A., van Heel, D. A. & Brown, M. A. Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nat. Rev. Genet.* **14,** 661–73 (2013).

74. Pontieri, G. *Patologia generale & fisiopatologia generale*. (Piccin-Nuova Libraria, 2007).

75. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am J Hum Genet* **90,** 7–24 (2012).

76. Hampe, J., Wienker, T., Nürnberg, P. & Schreiber, S. Mapping genes for polygenic disorders: considerations for study design in the complex trait of inflammatory bowel disease. *Hum. Hered.* **50,** 91–101

77. Worthey, E. A. *et al.* Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet. Med.* **13,** 255–262 (2011).

78. Kotlarz, D. *et al.* Loss of interleukin-10 signaling and infantile inflammatory bowel disease: implications for diagnosis and therapy. *Gastroenterology* **143,** 347–55 (2012).

79. Dinwiddie, D. L. *et al.* Molecular diagnosis of infantile onset inflammatory bowel disease by exome sequencing. *Genomics* **102,** 442–7 (2013).

80. Kappelman, M. D., Galanko, J. A., Porter, C. Q. & Sandler, R. S. Association of paediatric inflammatory bowel disease with other immune-mediated diseases. *Arch. Dis. Child.* **96,** 1042–6 (2011).

81. Román, A. L. S. & Muñoz, F. Comorbidity in inflammatory bowel disease. *World J. Gastroenterol.* **17,** 2723–33 (2011).

82. de Souza, H. S. P. & Fiocchi, C. Immunopathogenesis of IBD: current state of the art. *Nat. Rev. Gastroenterol. Hepatol.* **13,** 13–27 (2015).

83. Shih, D. Q. & Targan, S. R. Immunopathogenesis of inflammatory bowel disease. *World J. Gastroenterol.* **14,** 390–400 (2008).

84. Baumgart, D. C. & Carding, S. R. Inflammatory bowel disease: cause and immunobiology. *Lancet* **369,** 1627–1640 (2007).

85. Xavier, R. J. & Podolsky, D. K. Unravelling the pathogenesis of inflammatory bowel disease. *Nature* **448,** 427–34 (2007).

86. Begue, B. *et al.* Defective IL10 signaling defining a subgroup of patients with inflammatory bowel disease. *Am. J. Gastroenterol.* **106,** 1544–55 (2011).

87. Matricon, J., Barnich, N. & Ardid, D. Immunopathogenesis of inflammatory bowel disease. *Self. Nonself.* **1,** 299–309 (2010).

88. Murch, S. H., Lamkin, V. A., Savage, M. O., Walker-Smith, J. A. & MacDonald, T. T. Serum concentrations of tumour necrosis factor alpha in childhood chronic inflammatory bowel disease. *Gut* **32,** 913–7 (1991).

89. van Heel, D. A. *et al.* Inflammatory bowel disease is associated with a TNF polymorphism that affects an interaction between the OCT1 and NF(-kappa)B transcription factors. *Hum. Mol. Genet.* **11,** 1281–9 (2002).

90. Breese, E. J. *et al.* Tumor necrosis factor alpha-producing cells in the intestinal mucosa of children with inflammatory bowel disease. *Gastroenterology* **106,** 1455–66 (1994).

91. Bull, L. Genetics, Mutations, and Polymorphisms. (2000). at <http://www.ncbi.nlm.nih.gov/books/NBK6475/>

92. Karki, R., Pandya, D., Elston, R. C. & Ferlini, C. Defining 'mutation' and 'polymorphism' in the era of personal genomics. *BMC Med. Genomics* **8,** 37 (2015).

93. Strachan, T., Goodship, J. & Chinnery, P. *Genetics and Genomics in Medicine*. (Taylor & Francis, 2014). at <http://books.google.com/books?id=oRfhAwAAQBAJ&pgis=1>

94. Korf, B. R. & Irons, M. B. *Human Genetics and Genomics*. **19,** (John Wiley & Sons, 2012).

95. *Genomic and Personalized Medicine, Volumes 1-2*. (Academic Press, 2012). at <http://books.google.com/books?id=brNYrwtJGtYC&pgis=1>

96. Gordon, H., Trier Moller, F., Andersen, V. & Harbord, M. Heritability in inflammatory bowel disease: from the first twin study to genome-wide association studies. *Inflamm. Bowel Dis.* **21,** 1428–34 (2015).

97. Brant, S. R. Update on the heritability of inflammatory bowel disease: The importance of twin studies. *Inflamm. Bowel Dis.* **17,** 1–5 (2011).

98. Risch, N. J. Searching for genetic determinants in the new millennium. *Nature* **405,** 847–856 (2000).

99. Griffiths, A. J., Miller, J. H., Suzuki, D. T., Lewontin, R. C. & Gelbart, W. M. Linkage maps. (2000). at <http://www.ncbi.nlm.nih.gov/books/NBK21827/>

100. MORTON, N. E. Sequential tests for the detection of linkage. *Am. J. Hum. Genet.*

**7,** 277–318 (1955).

101. Maniatis, N. *et al.* The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 2228–33 (2002).

102. Lander, E. & Kruglyak, L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat. Genet.* **11,** 241–7 (1995).

103. Mathew, C. G. Genetics of inflammatory bowel disease: progress and prospects. *Hum. Mol. Genet.* **13,** 161R–168 (2004).

104. Ahmad, T., Tamboli, C. P., Jewell, D. & Colombel, J.-F. Clinical relevance of advances in genetics and pharmacogenetics of IBD. *Gastroenterology* **126,** 1533–49 (2004).

105. Van Limbergen, J., Russell, R. K., Nimmo, E. R. & Satsangi, J. The Genetics of Inflammatory Bowel Disease. *Am. J. Gastroenterol.* **102,** 2820–2831 (2007).

106. Ohmen, J. D. *et al.* Susceptibility locus for inflammatory bowel disease on chromosome 16 has a role in Crohn's disease, but not in ulcerative colitis. *Hum Mol Genet* **5,** 1679–1683 (1996).

107. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet* **42,** 1118–1125 (2010).

108. Hugot, J. P. *et al.* Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411,** 599–603 (2001).

109. Ogura, Y. *et al.* A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* **411,** 603–606 (2001).

110. Inohara, N., Ogura, Y., Chen, F. F., Muto, A. & Nuñez, G. Human Nod1 confers responsiveness to bacterial lipopolysaccharides. *J Biol Chem* **276,** 2551–2554 (2001).

111. Rioux, J. D. *et al.* Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nat. Genet.* **29,** 223–8 (2001).

112. Brescianini, S. *et al.* IBD5 is associated with an extensive complicated Crohn's disease feature: implications from genotype-phenotype analysis. *Gut* **56,** 149–50 (2007).

113. de Ridder, L. *et al.* Genetic susceptibility has a more important role in pediatric-onset Crohn's disease than in adult-onset Crohn's disease. *Inflamm. Bowel Dis.* **13,** 1083–92 (2007).

114. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273,** 1516–7 (1996).

115. Manolio, T. A., Brooks, L. D. & Collins, F. S. A HapMap harvest of insights into the genetics of common disease. *J. Clin. Invest.* **118,** 1590–605 (2008).

116. Cordell, H. J. & Clayton, D. G. Genetic association studies. *Lancet* **366,** 1121–1131 (2005).

117. DeLisi, C. Meetings that changed the world: Santa Fe 1986: Human genome baby-steps. *Nature* **455,** 876–7 (2008).

118. Morton, N. E. Parameters of the human genome. *Proc. Natl. Acad. Sci. U. S. A.* **88,** 7474–6 (1991).

119. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).

120. Becker, K. G. The common variants/multiple disease hypothesis of common complex genetic disorders. *Med. Hypotheses* **62,** 309–17 (2004).

121. Schork, N. J., Murray, S. S., Frazer, K. A. & Topol, E. J. Common vs. rare allele hypotheses for complex diseases. *Curr. Opin. Genet. Dev.* **19,** 212–9 (2009).

122. Morton, N. E. Into the post-HapMap era. *Adv. Genet.* **60,** 727–42 (2008).

123. Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–8 (2010).

124. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–320 (2005).

125. Xavier, R. J. & Rioux, J. D. Genome-wide association studies: a new window into immune-mediated diseases. *Nat. Rev. Immunol.* **8,** 631–43 (2008).

126. Barrett, J. C. & Cardon, L. R. Evaluating coverage of genome-wide association studies. *Nat. Genet.* **38,** 659–62 (2006).

127. Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc Natl Acad Sci U S A* **109,** 1193–1198 (2012).

128. Eichler, E. E. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* **11,** 446–450 (2010).

129. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009).

130. Bonen, D. K. *et al.* Crohn's disease-associated NOD2 variants share a signaling defect in response to lipopolysaccharide and peptidoglycan. *Gastroenterology* **124,** 140–146 (2003).

131. Barrett, J. C. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet* **40,** 955–962 (2008).

132. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447,** 661–78 (2007).

133. Imielinski, M. *et al.* Common variants at five new loci associated with early-onset inflammatory bowel disease. *Nat. Genet.* **41,** 1335–40 (2009).

134. Kugathasan, S. *et al.* Loci on 20q13 and 21q22 are associated with pediatric-onset inflammatory bowel disease. *Nat. Genet.* **40,** 1211–5 (2008).

135. Devkota, S. *et al.* Dietary-fat-induced taurocholic acid promotes pathobiont expansion and colitis in Il10-/- mice. *Nature* **487,** 104–108 (2012).

136. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47,** 979–986 (2015).

137. Teo, Y.-Y., Small, K. S. & Kwiatkowski, D. P. Methodological challenges of genome-wide association analysis in Africa. *Nat. Rev. Genet.* **11,** 149–60 (2010).

138. Van Limbergen, J., Radford-Smith, G. & Satsangi, J. Advances in IBD genetics. *Nat. Rev. Gastroenterol. Hepatol.* **11,** 372–385 (2014).

139. Vineis, P. & Pearce, N. Missing heritability in genome-wide association study research. *Nat Rev Genet* **11,** 589 (2010).

140. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491,** 119–124 (2012).

141. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).

142. Blair, D. R. *et al.* A nondegenerate code of deleterious variants in Mendelian loci contributes to complex disease risk. *Cell* **155,** 70–80 (2013).

143. Casals, F., Idaghdour, Y., Hussin, J. & Awadalla, P. Next-generation sequencing approaches for genetic mapping of complex diseases. *J Neuroimmunol* **248,** 10–22 (2012).

144. Rabbani, B., Tekin, M. & Mahdieh, N. The promise of whole-exome sequencing in medical genetics. *J. Hum. Genet.* **59,** 5–15 (2014).

145. Chong, J. X. *et al.* The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am. J. Hum. Genet.* **97,** 199–215 (2015).

146. Figure 3.24, DNA sequencing by the Sanger procedure - The Cell - NCBI Bookshelf. (2000). at <http://www.ncbi.nlm.nih.gov/books/NBK9950/figure/A462/>

147. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 5463–7 (1977).

148. Metzker, M. L. Emerging technologies in DNA sequencing. *Genome Res* **15,** 1767–1776 (2005).

149. Fullwood, M. J., Wei, C.-L., Liu, E. T. & Ruan, Y. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res.* **19,** 521–32 (2009).

150. Ng, P. C. & Kirkness, E. F. Whole genome sequencing. *Methods Mol. Biol.* **628,** 215–26 (2010).

151. Forrest, A. R. R. *et al.* A promoter-level mammalian expression atlas. *Nature* **507,** 462–70 (2014).

152. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22,** 1760–74 (2012).

153. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).

154. Biesecker, L. G. Exome sequencing makes medical genomics a reality. *Nat. Genet.* **42,** 13–4 (2010).

155. Strachan, T. & Read, A. *Human Molecular Genetics*. (Garland Science, 2010). at <https://books.google.com/books?id=dSwWBAAAQBAJ&pgis=1>

156. Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.* **42,** 30–5 (2010).

157. Johansen, C. T. *et al.* Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nat. Genet.* **42,** 684–7 (2010).

158. Yang, Y. *et al.* Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders — NEJM. *The New England journal of medicine* **369,** 1502–11 (2013).

159. Li, Q. *et al.* Variants in TRIM22 that Affect NOD2 Signaling Are Associated With Very Early Onset Inflammatory Bowel Disease. *Gastroenterology* (2016).

160. Uhlig, H. H. *et al.* The Diagnostic Approach to Monogenic Very Early Onset Inflammatory Bowel Disease. *Gastroenterology* (2014).

161. Mao, H. *et al.* Exome sequencing identifies novel compound heterozygous mutations of IL-10 receptor 1 in neonatal-onset Crohn's disease. *Genes Immun.* **13,** 437–42 (2012).

162. Cardinale, C. J., Kelsen, J. R., Baldassano, R. N. & Hakonarson, H. Impact of exome sequencing in inflammatory bowel disease. *World J. Gastroenterol.* **19,** 6721–9 (2013).

163. Okou, D. T. *et al.* Exome sequencing identifies a novel FOXP3 mutation in a 2-generation family with inflammatory bowel disease. *J. Pediatr. Gastroenterol. Nutr.* **58,** 561–8 (2014).

164. Blaydon, D. C. *et al.* Inflammatory skin and bowel disease linked to ADAM17 deletion. *N. Engl. J. Med.* **365,** 1502–8 (2011).

165. BBC News - 10,000 NHS patients 'to have genes mapped'. at <http://www.bbc.co.uk/news/10367883>

166. Morel, C. F. & Clarke, J. T. R. The use of agalsidase alfa enzyme replacement therapy in the treatment of Fabry disease. *Expert Opin. Biol. Ther.* **9,** 631–9 (2009).

167. Mamanova, L. *et al.* Target-enrichment strategies for next-generation sequencing. *Nat. Methods* **7,** 111–8 (2010).

168. Pabinger, S. *et al.* A survey of tools for variant analysis of next-generation genome sequencing data. *Br. Bioinform* (2013).

169. Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461,** 272–276 (2009).

170. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8,** 186–194 (1998).

171. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-Calling of Automated Sequencer Traces UsingPhred. I. Accuracy Assessment. *Genome Res* **8,** 175–185 (1998).

172. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res* **12,** 996–1006 (2002).

173. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–60 (2009).

174. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10,** R25 (2009).

175. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18,** 1851–1858 (2008).

176. Novoalign. at <http://novocraft.com>

177. Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25,** 1966–1967 (2009).

178. Trapnell, C. & Salzberg, S. L. How to map billions of short reads onto genomes. *Nat. Biotechnol.* **27,** 455–7 (2009).

179. Sims, D., Sudbery, I., Ilott, N. E., Heger, A. & Ponting, C. P. Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.* **15,** 121–32 (2014).

180. Quinlan, A. R. *et al.* Genome sequencing of mouse induced pluripotent stem cells reveals retroelement stability and infrequent DNA rearrangement during reprogramming. *Cell Stem Cell* **9,** 366–73 (2011).

181. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

182. Futema, M. *et al.* Use of targeted exome sequencing as a diagnostic tool for Familial Hypercholesterolaemia. *J Med Genet* **49,** 644–649 (2012).

183. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43,** 491–498 (2011).

184. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27,** 2156–8 (2011).

185. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46,** 310–315 (2014).

186. Goldgar, D. E. *et al.* Integrated evaluation of DNA sequence variants of unknown clinical significance: application to BRCA1 and BRCA2. *Am. J. Hum. Genet.* **75,** 535–44 (2004).

187. McCarthy, D. *et al.* Choice of transcripts and software has a large effect on variant annotation. *Genome Med.* **6,** 26 (2014).

188. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31,** 3812–3814 (2003).

189. Li, M.-X. *et al.* Predicting mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. *PLoS Genet.* **9,** e1003143 (2013).

190. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15,** 901–913 (2005).

191. Grantham, R. Amino Acid Difference Formula to Help Explain Protein Evolution. *Science (80-. ).* **185,** 862–864 (1974).

192. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of

nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20,** 110–121 (2010).

193. Li, M.-X., Gui, H.-S., Kwan, J. S. H., Bao, S.-Y. & Sham, P. C. A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. *Nucleic Acids Res.* **40,** e53 (2012).

194. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38,** e164 (2010).

195. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Br. Bioinform* (2012).

196. Bahcall, O. G. Genetic testing. ACMG guides on the interpretation of sequence variants. *Nat. Rev. Genet.* **16,** 256–7 (2015).

197. Krawitz, P. M. *et al.* Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. *Nat. Genet.* **42,** 827–9 (2010).

198. Ng, S. B. *et al.* Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat. Genet.* **42,** 790–3 (2010).

199. Vissers, L. E. L. M. *et al.* A de novo paradigm for mental retardation. *Nat. Genet.* **42,** 1109–12 (2010).

200. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 961–8 (2010).

201. Roach, J. C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328,** 636–9 (2010).

202. Wright, C. F. *et al.* Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet* **385,** 1305–1314 (2014).

203. Parla, J. S. *et al.* A comparative analysis of exome capture. *Genome Biol.* **12,** R97 (2011).

204. Goh, G. & Choi, M. Application of whole exome sequencing to identify disease-causing variants in inherited human diseases. *Genomics Inf.* **10,** 214–219 (2012).

205. Wang, Y. *et al.* Mutational Analysis of the TYR and OCA2 Genes in Four Chinese Families with Oculocutaneous Albinism. *PLoS One* **10,** e0125651 (2015).

206. Shikhare, G. & Kugathasan, S. Inflammatory bowel disease in children: current trends. *J. Gastroenterol.* **45,** 673–82 (2010).

207. Cosnes, J., Gower–Rousseau, C., Seksik, P., Cortot, A. & Gower-Rousseau, C. Epidemiology and Natural History of Inflammatory Bowel Diseases. *Gastroenterology* **140,** 1785–1794 (2011).

208. Valderas, J. M., Starfield, B., Sibbald, B., Salisbury, C. & Roland, M. Defining comorbidity: implications for understanding health and health services. *Ann. Fam. Med.* **7,** 357–63

209. Lees, C. W., Barrett, J. C., Parkes, M. & Satsangi, J. New IBD genetics: common pathways with other diseases. *Gut* **60,** 1739–1753 (2011).

210. Hemminki, K., Li, X., Sundquist, K. & Sundquist, J. Familial association of inflammatory bowel diseases with other autoimmune and related diseases. *Am. J. Gastroenterol.* **105,** 139–47 (2010).

211. Zhernakova, A., van Diemen, C. C. & Wijmenga, C. Detecting shared pathogenesis from the shared genetics of immune-related diseases. *Nat. Rev. Genet.* **10,** 43–55 (2009).

212. Moffatt, M. F. *et al.* Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* **448,** 470–3 (2007).

213. Hisamatsu, T. *et al.* Immune aspects of the pathogenesis of inflammatory bowel disease. *Pharmacol. Ther.* **137,** 283–97 (2013).

214. Saich, R. & Chapman, R. Primary sclerosing cholangitis, autoimmune hepatitis and overlap syndromes in inflammatory bowel disease. *World J. Gastroenterol.* **14,** 331–7 (2008).

215. Saarinen, S., Olerup, O. & Broomé, U. Increased frequency of autoimmune diseases in patients with primary sclerosing cholangitis. *Am. J. Gastroenterol.* **95,** 3195–9 (2000).

216. Gerich, M. E. & McGovern, D. P. B. Towards personalized care in IBD. *Nat. Rev. Gastroenterol. Hepatol.* **advance on,** (2013).

217. IBD Working Group of the European Society for Paediatric Gastroenterology-Hepatology and Nutrition. Inflammatory bowel disease in children and adolescents: recommendations for diagnosis--the Porto criteria. *J Pediatr Gastroenterol Nutr* **41,** 1–7 (2005).

218. Miller, S. A., Dykes, D. D. & Polesky, H. F. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res.* **16,** 1215 (1988).

219. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–2 (2010).

220. Fuentes Fajardo, K. V *et al.* Detecting false-positive signals in exome sequencing. *Hum. Mutat.* **33,** 609–13 (2012).

221. Pengelly, R. J. *et al.* A SNP profiling panel for sample tracking in whole-exome sequencing studies. *Genome Med.* **5,** 89 (2013).

222. Lockett, G. A. & Holloway, J. W. Genome-wide association studies in asthma; perhaps, the end of the beginning. *Curr. Opin. Allergy Clin. Immunol.* **13,** 463–9 (2013).

223. Van Limbergen, J. *et al.* Filaggrin loss-of-function variants are associated with atopic comorbidity in pediatric inflammatory bowel disease. *Inflamm. Bowel Dis.* **15,** 1492–8 (2009).

224. Andersen, A. B. T., Ehrenstein, V., Erichsen, R., Frøslev, T. & Sørensen, H. T. Parental inflammatory bowel disease and risk of asthma in offspring: a nationwide cohort study in denmark. *Clin. Transl. Gastroenterol.* **4,** e41 (2013).

225. H. M. Kang, X. Zhan, X. Sim, C. Ma Biostatistics Dept, Univ Michigan, Ann Arbor, Ann Arbor, M. EPACTS (Efficient and Parallelizable Association Container Toolbox). at <http://genome.sph.umich.edu/wiki/EPACTS>

226. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science (80-. ).* **327,** 78–81 (2010).

227. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29,** 308–311 (2001).

228. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4,** 44–57 (2009).

229. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28,** 27–30 (2000).

230. Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D. D. & Lin, X. Sequence kernel association tests for the combined effect of rare and common variants. *Am. J. Hum. Genet.* **92,** 841–53 (2013).

231. Anderson, H. R., Gupta, R., Strachan, D. P. & Limb, E. S. 50 years of asthma: UK trends from 1955 to 2004. *Thorax* **62,** 85–90 (2007).

232. Bernstein, C. N., Wajda, A. & Blanchard, J. F. The clustering of other chronic inflammatory diseases in inflammatory bowel disease: a population-based study. *Gastroenterology* **129,** 827–36 (2005).

233. Emerson, R. M., Williams, H. C. & Allen, B. R. Severity distribution of atopic dermatitis in the community and its relationship to secondary referral. *Br. J. Dermatol.* **139,** 73–6 (1998).

234. van Heel, D. A. & West, J. Recent advances in coeliac disease. *Gut* **55,** 1037–46 (2006).

235. Lindkvist, B., Benito de Valle, M., Gullberg, B. & Björnsson, E. Incidence and prevalence of primary sclerosing cholangitis in a defined adult population in Sweden. *Hepatology* **52,** 571–7 (2010).

236. Alkhateeb, A., Fain, P. R., Thody, A., Bennett, D. C. & Spritz, R. A. Epidemiology of vitiligo and associated autoimmune diseases in Caucasian probands and their families. *Pigment Cell Res.* **16,** 208–14 (2003).

237. Anderson, C. A. *et al.* Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat Genet* **43,** 246–252 (2011).

238. Lu, L. *et al.* Role of SMAD and non-SMAD signals in the development of Th17 and regulatory T cells. *J Immunol* **184,** 4295–4306 (2010).

239. Sleiman, P. M. A. *et al.* Variants of DENND1B Associated with Asthma in Children. *N. Engl. J. Med.* **362,** 36–44 (2010).

240. Lee, G. R., Fields, P. E., Griffin, T. J. & Flavell, R. A. Regulation of the Th2 cytokine locus by a locus control region. *Immunity* **19,** 145–53 (2003).

241. Kim, S. H. *et al.* Alpha-T-catenin (CTNNA3) gene was identified as a risk variant

for toluene diisocyanate-induced asthma by genome-wide association analysis. *Clin Exp Allergy* **39,** 203–212 (2009).

242. The Asthma Epidemic — NEJM. at <http://www.nejm.org/doi/full/10.1056/NEJMra054308>

243. Asher, M. I. *et al.* Worldwide time trends in the prevalence of symptoms of asthma, allergic rhinoconjunctivitis, and eczema in childhood: ISAAC Phases One and Three repeat multicountry cross-sectional surveys. *Lancet* **368,** 733–43 (2006).

244. Verlaan, D. J. *et al.* Allele-specific chromatin remodeling in the ZPBP2/GSDMB/ORMDL3 locus associated with the risk of asthma and autoimmune disease. *Am. J. Hum. Genet.* **85,** 377–93 (2009).

245. Wan, Y. I. *et al.* Genome-wide association study to identify genetic determinants of severe asthma. *Thorax* **67,** 762–8 (2012).

246. Wu, H. *et al.* Evaluation of candidate genes in a genome-wide association study of childhood asthma in Mexicans. *J. Allergy Clin. Immunol.* **125,** 321–327.e13 (2010).

247. Zhu, G. *et al.* Interleukin 18 receptor 1 gene polymorphisms are associated with asthma. *Eur. J. Hum. Genet.* **16,** 1083–90 (2008).

248. Michel, S. *et al.* Unifying candidate gene and GWAS Approaches in Asthma. *PLoS One* **5,** e13894 (2010).

249. Zhang, Y., Moffatt, M. F. & Cookson, W. O. C. Genetic and genomic approaches to asthma: new insights for the origins. *Curr. Opin. Pulm. Med.* **18,** 6–13 (2012).

250. PYHIN1 pyrin and HIN domain family, member 1 [ Homo sapiens (human)]. at <http://www.ncbi.nlm.nih.gov/gene/149628>

251. Torgerson, D. G. *et al.* Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations. *Nat. Genet.* **43,** 887–92 (2011).

252. Chisholm, J., Caulfield, M., Parker, M., Davies, J. & Palin, M. Briefing- Genomics England and the 100K Genome Project. *Genomics England* (2013). at <http://www.genomicsengland.co.uk/briefing/>

253. Witte, J. S. Prostate cancer genomics: towards a new understanding. *Nat. Rev. Genet.* **10,** 77–82 (2009).

254. Khor, B., Gardet, A. & Xavier, R. J. Genetics and pathogenesis of inflammatory bowel disease. *Nature* **474,** 307–317 (2011).

255. Majewski, J., Schwartzentruber, J., Lalonde, E., Montpetit, A. & Jabado, N. What can exome sequencing do for you? *J. Med. Genet.* **48,** 580–9 (2011).

256. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease.

257. Ma, S., Kosorok, M. R. & Fine, J. P. Additive risk models for survival data with high-dimensional covariates. *Biometrics* **62,** 202–210 (2006).

258. Philpott, D. J., Sorbara, M. T., Robertson, S. J., Croitoru, K. & Girardin, S. E. NOD

proteins: regulators of inflammation in health and disease. *Nat. Rev. Immunol.* **14,** 9–23 (2013).

259. Strober, W., Murray, P. J., Kitani, A. & Watanabe, T. Signalling pathways and molecular interactions of NOD1 and NOD2. *Nat. Rev. Immunol.* **6,** 9–20 (2006).

260. Werts, C., Girardin, S. E. & Philpott, D. J. TIR, CARD and PYRIN: three domains for an antimicrobial triad. *Cell Death Differ.* **13,** 798–815 (2006).

261. Murray, P. J. Beyond peptidoglycan for Nod2. *Nat. Immunol.* **10,** 1053–4 (2009).

262. Schreiber, S., Rosenstiel, P., Albrecht, M., Hampe, J. & Krawczak, M. Genetics of Crohn disease, an archetypal inflammatory barrier disease. *Nat. Rev. Genet.* **6,** 376–88 (2005).

263. Man, S. M., Kaakoush, N. O. & Mitchell, H. M. The role of bacteria and pattern-recognition receptors in Crohn's disease. *Nat. Rev. Gastroenterol. Hepatol.* **8,** 152–68 (2011).

264. Vallabhapurapu, S. & Karin, M. Regulation and function of NF-kappaB transcription factors in the immune system. *Annu. Rev. Immunol.* **27,** 693–733 (2009).

265. Kobayashi, K. S. *et al.* Nod2-dependent regulation of innate and adaptive immunity in the intestinal tract. *Science* **307,** 731–4 (2005).

266. Economou, M., Trikalinos, T. A., Loizou, K. T., Tsianos, E. V & Ioannidis, J. P. Differential effects of NOD2 variants on Crohn's disease risk and phenotype in diverse populations: a metaanalysis. *Am J Gastroenterol* **99,** 2393–2404 (2004).

267. Török, H.-P., Glas, J., Lohse, P. & Folwaczny, C. Alterations of the CARD15/NOD2 gene and the impact on management and treatment of Crohn's disease patients. *Dig. Dis.* **21,** 339–45 (2003).

268. Rodriguez-Bores, L., Fonseca, G.-C., Villeda, M.-A. & Yamamoto-Furusho, J.-K. Novel genetic markers in inflammatory bowel disease. *World J. Gastroenterol.* **13,** 5560–70 (2007).

269. Rose, C. D., Martin, T. M. & Wouters, C. H. Blau syndrome revisited. *Curr. Opin. Rheumatol.* **23,** 411–8 (2011).

270. Miceli-Richard, C. *et al.* CARD15 mutations in Blau syndrome. *Nat. Genet.* **29,** 19–20 (2001).

271. Kurzawski, G. *et al.* The NOD2 3020insC mutation and the risk of colorectal cancer. *Cancer Res.* **64,** 1604–6 (2004).

272. Huzarski, T. *et al.* The 3020insC allele of NOD2 predisposes to early-onset breast cancer. *Breast Cancer Res. Treat.* **89,** 91–3 (2005).

273. Mayor, N. *et al.* Single nucleotide Polymorphisms in the NOD2/CARD15 gene are associated with an increased risk of relapse and death for patients with acute leukemia after hematopoietic stem-cell transplantation with unrelated donors. *JOURNAL OF CLINICAL ONCOLOGY* (2007). at <http://eprints.ucl.ac.uk/6972/>

274. Lee, S. *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am.*

*J. Hum. Genet.* **91,** 224–37 (2012).

275. CAGI. at <http://genomeinterpretation.org>

276. Christodoulou, K. *et al.* Next generation exome sequencing of paediatric inflammatory bowel disease patients identifies rare and novel variants in candidate genes - Cerca con Google. *Gut* **62,** 977–84 (2012).

277. Zheng, X. *et al.* A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28,** 3326–8 (2012).

278. Anderson, C. A. *et al.* Data quality control in genetic case-control association studies. *Nat. Protoc.* **5,** 1564–73 (2010).

279. Stenson, P. D. *et al.* Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.* **21,** 577–81 (2003).

280. Baldwin, A. S. Series introduction: the transcription factor NF-kappaB and human disease. *J. Clin. Invest.* **107,** 3–6 (2001).

281. Achkar, J.-P. *et al.* Phenotype-stratified genetic linkage study demonstrates that IBD2 is an extensive ulcerative colitis locus. *Am. J. Gastroenterol.* **101,** 572–80 (2006).

282. Anderson, C. A. *et al.* Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat Genet* **43,** 246–252 (2011).

283. Lesage, S. *et al.* CARD15/NOD2 mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. *Am J Hum Genet* **70,** 845–857 (2002).

284. Mayor, A., Martinon, F., De Smedt, T., Pétrilli, V. & Tschopp, J. A crucial function of SGT1 and HSP90 in inflammasome activity links mammalian and plant innate immune responses. *Nat. Immunol.* **8,** 497–503 (2007).

285. Damgaard, R. B., Gyrd-Hansen, M. & B Damgaard, R. Inhibitor of apoptosis (IAP) proteins in regulation of inflammation and innate immunity. *Discov. Med.* **11,** 221–31 (2011).

286. Pedersen, J., LaCasse, E. C., Seidelin, J. B., Coskun, M. & Nielsen, O. H. Inhibitors of apoptosis (IAPs) regulate intestinal immunity and inflammatory bowel disease (IBD) inflammation. *Trends Mol. Med.* **20,** 652–65 (2014).

287. McComb, S. *et al.* cIAP1 and cIAP2 limit macrophage necroptosis by inhibiting Rip1 and Rip3 activation. *Cell Death Differ.* **19,** 1791–801 (2012).

288. Estornes, Y. & Bertrand, M. J. M. IAPs, regulators of innate immunity and inflammation. *Semin. Cell Dev. Biol.* (2014).

289. Seidelin, J. B., Vainer, B., Andresen, L. & Nielsen, O. H. Upregulation of cIAP2 in regenerating colonocytes in ulcerative colitis. *Virchows Arch.* **451,** 1031–8 (2007).

290. Naugler, K. M., Baer, K. A. & Ropeleski, M. J. Interleukin-11 antagonizes Fas ligand-mediated apoptosis in IEC-18 intestinal epithelial crypt cells: role of MEK and Akt-dependent signaling. *Am. J. Physiol. Gastrointest. Liver Physiol.* **294,**

G728-37 (2008).

291. Zeissig, Y. *et al.* XIAP variants in male Crohn's disease. *Gut* **64,** 66–76 (2014).

292. McLean, L. P., Shea-Donohue, T. & Cross, R. K. Vedolizumab for the treatment of ulcerative colitis and Crohn's disease. *Immunotherapy* **4,** 883–98 (2012).

293. Asano, K. *et al.* A genome-wide association study identifies three new susceptibility loci for ulcerative colitis in the Japanese population. *Nat. Genet.* **41,** 1325–9 (2009).

294. Juyal, G. *et al.* Genome-wide association scan in north Indians reveals three novel HLA-independent risk loci for ulcerative colitis. *Gut* 1–9 (2014).

295. Shiina, T., Inoko, H. & Kulski, J. K. An update of the HLA genomic region, locus information and disease associations: 2004. *Tissue Antigens* **64,** 631–49 (2004).

296. Van Molle, W. *et al.* HSP70 protects against TNF-induced lethal inflammatory shock. *Immunity* **16,** 685–695 (2002).

297. Tanaka, K.-I. *et al.* Genetic evidence for a protective role of heat shock factor 1 against irritant-induced gastric lesions. *Mol. Pharmacol.* **71,** 985–93 (2007).

298. Mosser, D. D., Caron, A. W., Bourget, L., Denis-Larose, C. & Massie, B. Role of the human heat shock protein hsp70 in protection against stress-induced apoptosis. *Mol. Cell. Biol.* **17,** 5317–5327 (1997).

299. Barbatis, C. & Tsopanomichalou, M. Heat shock proteins in inflammatory bowel disease. *Ann. Gastroenterol.* **22,** 244–247 (2009).

300. Samborski, P. & Grzymisławski, M. The Role of HSP70 Heat Shock Proteins in the Pathogenesis and Treatment of Inflammatory Bowel Diseases. *Adv. Clin. Exp. Med.* **24,** 525–30

301. Schell, M. T., Spitzer, A. L., Johnson, J. A., Lee, D. & Harris, H. W. Heat shock inhibits NF-kB activation in a dose- and time-dependent manner. *J. Surg. Res.* **129,** 90–3 (2005).

302. Adachi, T. *et al.* Involvement of heat shock protein a4/apg-2 in refractory inflammatory bowel disease. *Inflamm. Bowel Dis.* **21,** 31–9 (2015).

303. Chen, R. *et al.* Whole-exome sequencing identifies tetratricopeptide repeat domain 7A ( TTC7A ) mutations for combined immunodeficiency with intestinal atresias. *J. Allergy Clin. Immunol.* **132,** 656–664.e17 (2013).

304. McKenna, A. *et al.* The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).

305. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4,** 1073–1081 (2009).

306. Adzhubei, I. a *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7,** 248–249 (2010).

307. Christodoulou, K. *et al.* Next generation exome sequencing of paediatric inflammatory bowel disease patients identifies rare and novel variants in

candidate genes. 977–984 (2012).

308. Dubinsky, M. C. *et al.* Genome wide association (GWA) predictors of anti-TNFalpha therapeutic responsiveness in pediatric inflammatory bowel disease. *Inflamm. Bowel Dis.* **16,** 1357–66 (2010).

309. Uhlig, H. H. Monogenic diseases associated with intestinal inflammation: implications for the understanding of inflammatory bowel disease. *Gut* **62,** 1795–805 (2013).

310. Auer, P. L. & Lettre, G. Rare variant association studies: considerations, challenges and opportunities. *Genome Med.* **7,** 16 (2015).

311. Majithia, A. R. *et al.* Rare variants in PPARG with decreased activity in adipocyte differentiation are associated with increased risk of type 2 diabetes. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 13127–32 (2014).

312. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *Am. J. Hum. Genet.* **95,** 5–23 (2014).

313. Wisniewska, M. *et al.* Crystal structures of the ATPase domains of four human Hsp70 isoforms: HSPA1L/Hsp70-hom, HSPA2/Hsp70-2, HSPA6/Hsp70B', and HSPA5/BiP/GRP78. *PLoS One* **5,** e8625 (2010).

314. Kelley, L. A. & Sternberg, M. J. E. Protein structure prediction on the Web: a case study using the Phyre server. *Nat. Protoc.* **4,** 363–371 (2009).

315. Scieglinska, D., Piglowski, W., Chekan, M., Mazurek, A. & Krawczyk, Z. Differential expression of HSPA1 and HSPA2 proteins in human tissues; tissue microarray-based immunohistochemical study. *Histochem. Cell Biol.* **135,** 337–50 (2011).

316. Alangari, A. *et al.* LPS-responsive beige-like anchor (LRBA) gene mutation in a family with inflammatory bowel disease and combined immunodeficiency. *J. Allergy Clin. Immunol.* **130,** 481–8.e2 (2012).

317. Fontalba, A., Gutierrez, O. & Fernandez-Luna, J. L. NLRP2, an Inhibitor of the NF- B Pathway, Is Transcriptionally Activated by NF- B and Exhibits a Nonfunctional Allelic Variant. *J. Immunol.* **179,** 8519–8524 (2007).

318. Lesage, S. *et al.* CARD15/NOD2 mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. *Am J Hum Genet* **70,** 845–857 (2002).

319. van Driel, B. *et al.* Signaling lymphocyte activation molecule regulates development of colitis in mice. *Gastroenterology* **143,** 1544–1554.e7 (2012).

320. Rivas, M. A. *et al.* Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat Genet* **43,** 1066–1073 (2011).

321. NHLBI GO Exome Sequencing Project (ESP). at <http://evs.gs.washington.edu/EVS/>

322. Tang, R., Yang, G., Zhang, S., Wu, C. & Chen, M. Opposite Effects of Interferon Regulatory Factor 1 and Osteopontin on the Apoptosis of Epithelial Cells Induced by TNF-α in Inflammatory Bowel Disease. *Inflamm. Bowel Dis.* **20,**

1950–1961 (2014).

323. Brown, E. M., Sadarangani, M. & Finlay, B. B. The role of the immune system in governing host-microbe interactions in the intestine. *Nat. Immunol.* **14,** 660–667 (2013).

324. Fisher, S. A. *et al.* Sex stratification of an inflammatory bowel disease genome search shows male-specific linkage to the HLA region of chromosome 6. *Eur. J. Hum. Genet.* **10,** 259–265 (2002).

325. Hageman, J. & Kampinga, H. H. Computational analysis of the human HSPH/HSPA/DNAJ family and cloning of a human HSPH/HSPA/DNAJ expression library. *Cell Stress and Chaperones* **14,** 1–21 (2009).

326. Fourie, A. M., Peterson, P. a. & Yang, Y. Characterization and regulation of the major histocompatibility complex–encoded proteins Hsp70-Hom and Hsp70-1/2. *Cell Stress Chaperones* **6,** 282 (2001).

327. Hasson, S. a *et al.* High-content genome-wide RNAi screens identify regulators of parkin upstream of mitophagy. *Nature* **504,** 291–5 (2013).

328. Klucken, J., Shin, Y., Hyman, B. T. & McLean, P. J. A single amino acid substitution differentiates Hsp70-dependent effects on alpha-synuclein degradation and toxicity. *Biochem. Biophys. Res. Commun.* **325,** 367–373 (2004).

329. Bouwmeester, T. *et al.* A physical and functional map of the human TNF-alpha/NF-kappa B signal transduction pathway. *Nat. Cell Biol.* **6,** 97–105 (2004).

330. Zanzoni, A. *et al.* MINT: a Molecular INTeraction database. *FEBS Lett.* **513,** 135–140 (2002).

331. Warrick, J. M. *et al.* Suppression of polyglutamine-mediated neurodegeneration in Drosophila by the molecular chaperone HSP70. *Nat. Genet.* **23,** 425–428 (1999).

332. Auluck, P. K., Chan, H. Y. E., Trojanowski, J. Q., Lee, V. M. Y. & Bonini, N. M. Chaperone suppression of alpha-synuclein toxicity in a Drosophila model for Parkinson's disease. *Science* **295,** 865–868 (2002).

333. Kelsen, J. R. *et al.* Exome Sequencing Analysis Reveals Variants in Primary Immunodeficiency Genes in Patients With Very Early Onset Inflammatory Bowel Disease. *Gastroenterology* (2015).

334. Kelsen, J. R. *et al.* A de novo whole gene deletion of XIAP detected by exome sequencing analysis in very early onset inflammatory bowel disease: a case report. *BMC Gastroenterol.* **15,** 160 (2015).

335. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17,** 405–24 (2015).

336. Techanukul, T. *et al.* Novel and recurrent FERMT1 gene mutations in Kindler syndrome. *Acta Derm. Venereol.* **91,** 267–70 (2011).

337. Fabre, A. *et al.* SKIV2L mutations cause syndromic diarrhea, or trichohepatoenteric syndrome. *Am. J. Hum. Genet.* **90,** 689–92 (2012).

338. Egritas, O., Dalgic, B. & Onder, M. Tricho-hepato-enteric syndrome presenting with mild colitis. *Eur. J. Pediatr.* **168,** 933–5 (2009).

339. Bianco, A. M., Girardelli, M. & Tommasini, A. Genetics of inflammatory bowel disease from multifactorial to monogenic forms. *World J. Gastroenterol.* **21,** 12296–310 (2015).

340. Wang, L. *et al.* MYH mutations in patients with attenuated and classic polyposis and with young-onset colorectal cancer without polyps. *Gastroenterology* **127,** 9–16 (2004).

341. Zaki, P. A. *et al.* Penetrance of eye defects in mice heterozygous for mutation of Gli3 is enhanced by heterozygous mutation of Pax6. *BMC Dev. Biol.* **6,** 46 (2006).

342. Hizarcioglu-Gulsen, H. *et al.* Intractable diarrhea of infancy: 10 years of experience. *J. Pediatr. Gastroenterol. Nutr.* **59,** 571–6 (2014).

343. Bianco, A. M. *et al.* A common genetic background could explain early-onset Crohn's disease. *Med. Hypotheses* **78,** 520–2 (2012).

344. Bønnelykke, K. *et al.* A genome-wide association study identifies CDHR3 as a susceptibility locus for early childhood asthma with severe exacerbations. *Nat. Genet.* **46,** 51–55 (2013).

345. Hinds, D. A. *et al.* A genome-wide association meta-analysis of self-reported allergy identifies shared and allergy-specific susceptibility loci. *Nat. Genet.* **45,** 907–911 (2013).

346. Moffatt, M. F. *et al.* A large-scale, consortium-based genomewide association study of asthma. *N. Engl. J. Med.* **363,** 1211–21 (2010).

347. Ferreira, M. A. R. *et al.* Genome-wide association analysis identifies 11 risk variants associated with the asthma with hay fever phenotype. *J. Allergy Clin. Immunol.* **133,** 1564–71 (2014).

348. Hunt, K. A. *et al.* Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. *Nature* **498,** 232–5 (2013).

349. King, K., Flinter, F. A., Nihalani, V. & Green, P. M. Unusual deep intronic mutations in the COL4A5 gene cause X linked Alport syndrome. *Hum. Genet.* **111,** 548–54 (2002).

350. Huang, J. *et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.* **6,** 8111 (2015).

351. Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.* **44,** 243–6 (2012).

352. Strachan, T. & Read, A. P. Gene therapy and other molecular genetic-based therapeutic approaches. (1999). at <http://www.ncbi.nlm.nih.gov/books/NBK7569/>

353. Box 22.1, General gene therapy strategies (see also Figure 22.1) - Human Molecular Genetics - NCBI Bookshelf. (1999). at <http://www.ncbi.nlm.nih.gov/books/NBK7569/box/A2870/?report=objectonly>

354. Simara, P., Motl, J. A. & Kaufman, D. S. Pluripotent stem cells and gene therapy. *Transl. Res.* **161,** 284–92 (2013).

355. Strachan, T. & Read, A. P. Genetic manipulation of animals. (1999). at <http://www.ncbi.nlm.nih.gov/books/NBK7563/>

356. Giacomelli, R. *et al.* Circulating soluble factor-inhibiting natural killer (NK) activity of fresh peripheral blood mononuclear cells (PBMC) from inflammatory bowel disease (IBD) patients. *Clin. Exp. Immunol.* **115,** 72–7 (1999).

357. Potocnik, U., Ferkolj, I., Glavac, D. & Dean, M. Polymorphisms in multidrug resistance 1 (MDR1) gene are associated with refractory Crohn disease and ulcerative colitis. *Genes Immun.* **5,** 530–9 (2004).

358. Kalla, R. *et al.* MicroRNAs: new players in IBD. *Gut* **64,** 504–17 (2015).

359. Loddo, I. & Romano, C. Inflammatory Bowel Disease: Genetics, Epigenetics, and Pathogenesis. *Front. Immunol.* **6,** (2015).

360. Yi, J. M. & Kim, T. O. Epigenetic alterations in inflammatory bowel disease and cancer. *Intest. Res.* **13,** 112–21 (2015).

361. Matsuoka, K. & Kanai, T. The gut microbiota and inflammatory bowel disease. *Semin. Immunopathol.* **37,** 47–55 (2015).

362. Levin, A. D. *et al.* Vitamin D deficiency in children with inflammatory bowel disease. *Dig. Dis. Sci.* **56,** 830–6 (2011).

363. Ulitsky, A. *et al.* Vitamin D deficiency in patients with inflammatory bowel disease: association with disease activity and quality of life. *JPEN. J. Parenter. Enteral Nutr.* **35,** 308–16 (2011).

364. Lasko, T. A., Bhagwat, J. G., Zou, K. H. & Ohno-Machado, L. The use of receiver operating characteristic curves in biomedical informatics. *J. Biomed. Inform.* **38,** 404–15 (2005).

365. Libbrecht, M. W. & Noble, W. S. Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* **16,** 321–32 (2015).

366. Wei, Z. *et al.* Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. *Am. J. Hum. Genet.* **92,** 1008–12 (2013).

367. Human Genome Sequencing ConsortiumInternational. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–45 (2004).

ORIGINAL ARTICLE

# Exome Analysis of Patients with Concurrent Pediatric Inflammatory Bowel Disease and Autoimmune Disease

*Gaia Andreoletti, MSc,* James J. Ashton, BMBS, BMedSci,*† Tracy Coelho, MRCPCH,*† Claire Willis,‡ Rachel Haggarty,‡ Jane Gibson, PhD,§ John Holloway, PhD,* Akshay Batra, MRCPCH,† Nadeem A. Afzal, MD, MRCPCH,† Robert Mark Beattie,† and Sarah Ennis, PhD**

**Background:** Pediatric Inflammatory Bowel Disease (PIBD) is a chronic condition seen in genetically predisposed individuals. Genome-wide association studies have implicated >160 genomic loci in IBD with many genes coding for proteins in key immune pathways. This study looks at autoimmune disease burden in patients diagnosed with PIBD and interrogates exome data of a subset of patients.

**Methods:** Patients were recruited from the Southampton Genetics of PIBD cohort. Clinical diagnosis of autoimmune disease in these individuals was ascertained from medical records. For a subset of patients with PIBD and concurrent asthma, exome data was interrogated to ascertain the burden of pathogenic variants within genes implicated in asthma. Association testing was conducted between cases and population controls using the SKAT-O test.

**Results:** Forty-nine (28.3%) PIBD children (18.49% CD, 8.6% UC, and 21.15% IBDU patients) had a concurrent clinical diagnosis of at least one other autoimmune disorder; asthma was the most prevalent, affecting 16.2% of the PIBD cohort. Rare and common variant association testing revealed 6 significant genes ($P < 0.05$) before Bonferroni adjustment. Three of these genes were previously implicated in both asthma and IBD (*ZPBP2 IL1R1*, and *IL18R1*) and 3 in asthma only (*PYHIN1, IL2RB*, and *GSTP1*).

**Conclusions:** One-third of our cohort had a concurrent autoimmune condition. We observed higher incidence of asthma compared with the overall pediatric prevalence. Despite a small sample size, SKAT-O evaluated a significant burden of rare and common mutations in 6 genes. Variant burden suggests that a systemic immune dysregulation rather than organ-specific could underpin immune dysfunction for a subset of patients.

*(Inflamm Bowel Dis 2015;21:1229–1236)*

**Key Words:** pediatric inflammatory bowel disease, comorbidity, exome sequencing, autoimmune disorders, asthma, genetics

Pediatric inflammatory bowel disease (PIBD) encompasses Crohn's disease (CD), ulcerative colitis (UC), and inflammatory bowel disease unclassified, a group of complex and multifactorial illnesses. The etiology is complex and likely to comprise 4 key concepts: immune dysregulation, barrier dysfunction, microbial flora, and a genetic predisposition[1,2]; how specific interaction between these factors leads to development of disease is poorly understood.

The incidence of PIBD is increasing in Europe and North America. Recent studies from England,[3] Scotland,[4] and Scandanvia[5] have shown increasing incidence over the last 20 years with incidence as high as 12.8/100,000 person-years in Sweden.[6] This increase may be driven by lifestyle changes. The hygiene hypothesis relating to autoimmune conditions is well established in the literature, and it is conceivable that altered microbial exposure in these children may have a role in development of disease.[7,8]

Over the past decade, genome-wide association studies (GWAS) have substantially advanced the understanding of many complex diseases.[9] Since the discovery in 2001 of *NOD2*, the first genetic susceptibility gene for IBD[10,11] more than 160 distinct loci have been shown to have a robust association with IBD.[12–14] However, common variation in these genes account for only approximately one-fourth of the disease heritability.[15] It is now assumed that rare variants in pathways implicated across various autoimmune conditions may account for some of the missing heritability in IBD.[16] Various studies in adult populations have looked at the incidence of other autoimmune-mediated disorders in patients suffering from IBD.[17–19] Previous studies into the coexistence of PIBD

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.ibdjournal.org).

Received for publication January 14, 2015; Accepted February 3, 2015.

From the *Human Genetics and Genomic Medicine, University of Southampton, Southampton General Hospital, Southampton, United Kingdom; †Department of Paediatric Gastroenterology, University Hospital Southampton NHS Foundation Trust, Southampton General Hospital, Southampton, United Kingdom; ‡NIHR Nutrition Biomedical Research Centre, Southampton Centre for Biomedical Research, University Hospital Southampton NHS Foundation Trust, Southampton General Hospital, Southampton, United Kingdom; and §Centre for Biological Sciences, Faculty of Natural and Environmental Studies, University of Southampton, Southampton, United Kingdom.

Supported by the Crohn's in Childhood Research Association (CIRCA) and the Gerald Kerkut Charitable Trust. J. J. Ashton is supported by a University of Southampton National Institute of Health Research Academic Clinical Fellowship.

The authors have no conflicts of interest to disclose.

G. Andreoletti and J. J. Ashton have contributed equally to this study.

Reprints: Sarah Ennis, PhD, Human Genetics and Genomic Medicine, University of Southampton, Duthie Building (Mailpoint 808), Southampton General Hospital, Southampton SO16 6YD, United Kingdom (e-mail: s.ennis@southampton.ac.uk).



DOI 10.1097/MIB.0000000000000381

Published online 17 April 2015.

and other autoimmune disease have reported strong association between PIBD (both CD and UC) and rheumatoid arthritis, systemic lupus erythematosus, and hypothyroidism with a trend towards increased prevalence of other autoimmune conditions, including asthma and eczema.[20]

Predisposition to IBD and other autoimmune disease has a strong genetic component, and analyses of exomes of these patients may yield variations associated with both groups of disease. This study examines the autoimmune disease burden in patients diagnosed with PIBD and interrogates exome data of a subset of patients.

## MATERIALS AND METHODS

### Recruitment

Recruitment of children diagnosed with PIBD was through services at University Hospital Southampton. All children younger than 18 years at the point of diagnosis were eligible. Diagnosis was established according to the Porto criteria.[21] Clinical data were recorded for each patient including family history of IBD and any history of autoimmune disease.[22]

### Patient Data Extraction

Data from 173 patients prospectively collected at recruitment were interrogated to identify patients with PIBD with: (1) comorbidity of other autoimmune diseases (clinician diagnosed) and (2) positive family history of autoimmune diseases other than IBD. Medical notes were consulted for any ambiguous diagnosis and review allowed for exclusion of any unconfirmed diagnoses.

### DNA Extraction

Genomic DNA was extracted from peripheral venous blood samples collected in EDTA, using the salting out method.[23] DNA concentration was estimated using the Qubit 2.0 Fluorometer (Life Technologies Ltd). and 260:280 ratio calculated using a nanodrop spectrophotometer. The average DNA yield obtained was 150 μg/mL and approximately 20 μg of DNA was used for next generation sequencing for each patient.

### Patient Selection

Of the 28 patients with IBD who had a concurrent diagnosis of asthma, we selected the 18 youngest of these for exome analysis. All 18 were of white British ancestry.

### Exome Data Generation and Processing

Whole exome sequencing was performed using the Agilent SureSelect Human all Exon 51 Mb version 4 capture kit. The fastq raw data generated from Illumina paired-end sequencing were aligned against the human reference genome (hg19) using Novoalign (novoalign/2.08.02). Sequence coverage for each sample was calculated using the BedTools package (v2.13.2) (see Table, Supplemental Digital Content 1, http://links.lww.com/IBD/A809).

SAMtools[24] MPileup tool (SAMtools/0.1.18) was used to detect variation from the mapping information to call SNPs and short INDELs from the alignment file. Variations with read depth <4 were excluded. The Phred software[25] reads DNA sequencing trace files, calls bases, and assigns a quality value to each called base and is powered to discriminate between correct and incorrect base calls. Only good quality bases with a Phred score >20 were retained for analysis (99% base call accuracy). ANNOVAR (annovar/February 21, 2013)[26] was applied for variant annotation against a database of RefSeq transcripts. Resultant variants files for each subject were subjected to further in-house quality control tests to detect DNA sample contamination and ensure sex concordance by assessing autosomal and X chromosome heterozygosity. Variant sharing between all pairs of individuals was assessed to confirm sample relationships. Sample provenance was confirmed by independent genotyping of a validated SNP panel, developed specifically for exome data.[27]

### Gene Selection

The latest genome-wide meta-analysis of IBD reported 193 genes across 163 loci with statistically independent signals of association at genome-wide significance ($P < 5 \times 10^{-8}$).[13] These genes were cross-referenced with 49 genes associated with asthma identified by linkage studies and GWAS.[28] Sixteen of these genes have been associated with both asthma and IBD (Fig. 1). Gene names were cross-referenced with the HUGO web server to confirm the approved gene symbol.

#### *Variant Association Testing*

Our findings and those of others[29,30] indicate asthma is the most common concurrent autoimmune disease in patients with IBD. For this reason, we wanted to further investigate if a subset of patients with a concurrent diagnosis of both IBD and asthma present with a significant burden of mutation within known genes associated with asthma. Our modest sample size was underpowered to extend this analysis to all 193 IBD genes.

To detect association between the genetic component and disease status, first, a single variant test and then a gene-based test (SKAT-O) were performed. To run these tests, genotype information (homozygous alternative, homozygous reference, or heterozygous status) were retrieved using customized scripts applying samtools,[24] vcftools,[31] and bedtools[32] packages. All variant sites across 49 genes (comprising 33 genes specific to asthma and 16 genes common to both diseases) were used to generate the variant call file for each of the 18 exome analyzed patients and 56 unrelated germline controls. Our genomics bioinformatics group has a rolling database of non-IBD clinical exomes. Controls without any clinical diagnosis of autoimmune disease were selected from this in-house database.

Variations were further excluded based on the Hardy–Weinberg equilibrium status ($P < 0.001$) in the control group, by using vcftools.[31] Variant call file files containing genotype information for all cases and controls were merged together and annotated.[33] Both single and joint analysis were carried out using the EPACTS software.[34]

### Single Variant Association Testing

The single variant logistic score test[34] was performed to detect differences in variant frequency between cases and control

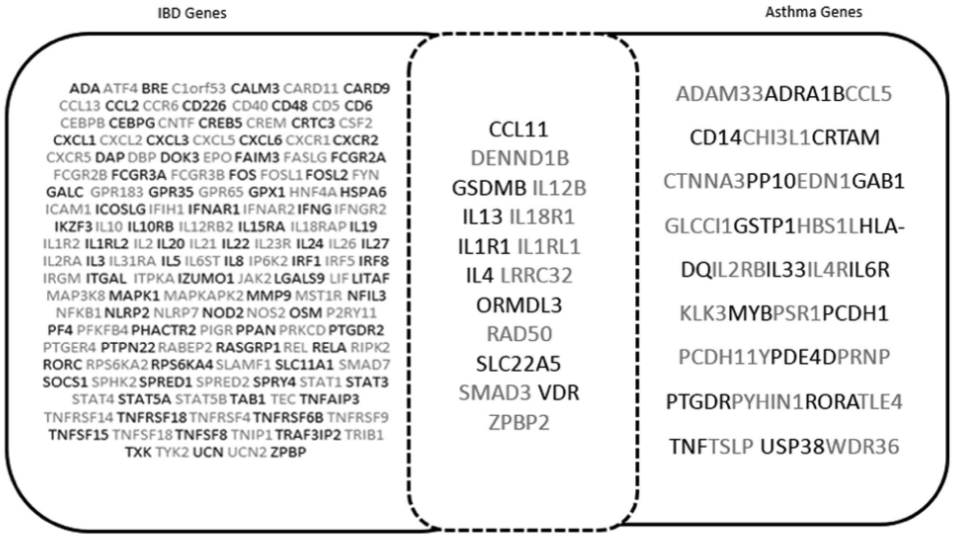



FIGURE 1. Overlap of GWAS significant gene loci in IBD (left) and asthma (right).

group. The test was not performed on mutations occurring in one individual in either case or control group.

## Rare Variant Profile Filtering

The burden of rare and novel damaging variation was described for each of the 18 patients across 49 asthma genes. Synonymous variations were excluded from the analysis on the assumption of their low impact on protein function. All novel to individual, novel to Southampton PIBD cohort, and clinical variants as well as frameshift insertion, frameshift deletion, stop gain, and stop loss mutations were retained for further analysis. Novel to individual denotes variants not previously reported in dbSNP137 database, 1000 Genomes Project, Exome Variants Server (EVS) of European Americans in the NHLI-ESP project with 6500 exomes (http://evs.gs.washington.edu/EVS/) in 46 unrelated human subjects sequenced by Complete Genomics,[35] in other individuals of the Soton IBD cohort, or in the Southampton reference exome database. Novel to Soton cohort denotes variants not previously reported in dbSNP137 database,[36] 1000 Genomes Project,[37] Exome Variants Server of European Americans in the NHLI-ESP project with 6500 exomes (http://evs.gs.washington.edu/EVS/) in 46 unrelated human subjects sequenced by Complete Genomics[35] but has been seen in other individuals of the Soton IBD cohort.

To refine this list to variations most likely to have a biological impact, common variants occurring in ≥5% of individuals from 1000 genomes project[37] were excluded and variants less likely to impact on protein function as expressed by the logit categorical score[38] were excluded (logit = N). Pathways were determined using DAVID (Database for Annotation, Visualization and Integrated Discovery)[39] and KEGG pathway.[40]

## Joint Variant Association Testing

The sequence kernel association testing optimal unified test (SKAT-O)[41] is a gene-based test for assessing the contribution of rare and common variations within a genomic loci with trait.[41] Specifically, SKAT-O encompasses both a burden test and a SKAT[41] test to offer a powerful way of conducting association analysis on combined rare and common variation as single variant tests are often underpowered because of the large sample size needed to detect a significant association.

SKAT-O was executed with the small sample adjustment and by applying an MAF threshold of 0.05 to define rare variations and using default weights. To conduct the test, a group file with mutations of interests (missense, nonsense, splice-site variants, and coding indels) was created for each of the 49 genes.

## Ethical Considerations

The study has ethics approval from Southampton and South West Hampshire Research Ethics Committee (09/H0504/125).

# RESULTS

## Southampton PIBD Cohort

At the time of analysis, the Southampton PIBD study cohort comprised 173 children (98 CD, 55 UC, and 20

**TABLE 1.** PIBD Cohort Demographics

| | CD | UC | IBDU | Total IBD |
|---|---|---|---|---|
| No. patients | 98 | 55 | 20 | 173 |
| Female, % | 44 | 44 | 60 | 46 |
| Median age of onset (25th/75th percentile) | 12.28 (9.16/14.27) | 11.5 (9.61/13.48) | 15.33 (8.72/14.10) | 12.28 |
| Mean age of onset (SD) | 11.58 (3.13) | 10.87 (3.76) | 11.50 (3.31) | 11.35 (3.35) |

inflammatory bowel disease unclassified); demographic data are shown in Table 1.

## Prevalence of Comorbidity

Analysis of the cohort revealed concurrent diagnosis of PIBD with 12 distinct autoimmune-mediated conditions (Table 2).

Asthma (n = 28) and atopic dermatitis (n = 24) represented the conditions with the highest frequency. Additional cases of sclerosing cholangitis (n = 4), coeliac disease (n = 2), and vitiligo (n = 2) were present.

Forty-nine children (28.3%) presented with a second autoimmune condition. Across the cohort, there was a family history of asthma, atopic dermatitis, coeliac disease, sclerosing cholangitis, and vitiligo although no probands had diagnosis of these conditions at the time of analysis.

### *Single Variant Association Test for Variants in Asthma and Dual Susceptibility Genes*

Among the 28 patients affected by asthma, 18 youngest patients were selected for exome sequencing (9 CD and 9 UC). Characteristics for each of the patients that underwent exome sequencing are presented in Table 3. Thirty six of the 49 genes either specific to asthma and common to both asthma and IBD were analyzed, as no coding variants were called in *ADRA1B, CCL5, CD14, HLA-DQ, IL12B, IL13, IL4, ORMDL3, PCDH1, RAD50, TNF, TSLP*, and *SLC22A5* across cases and controls, and these genes were excluded from single variant and joint testing.

A total of 175 different variants were identified across 36 genes in the cases and controls exomes. A total of 73 occurred only in 1 individual (case or control), and these were not analyzed in the single variant test. The single variant test was applied to 102 variants across 33 genes. Three of these variants showed significant association with disease status (association $P < 0.05$; *PYHIN1, ZPBP2*, and *LRRC32*). However, none of these variants would withstand multiple testing corrections (see Table, Supplemental Digital Content 2, http://links.lww.com/IBD/A810). *ZPBP2* and *LRRC32* are known to be involved in both asthma and IBD. Within these genes, *ZPBP2* nonsynonymous variant at position 38027030 bp and *LRRC32* synonymous variant at position 76372052 bp ($P = 0.011$ and $P = 0.043$, respectively) were found with higher frequency in cases compared with controls. *PYHIN1* is known to be involved in asthma pathogenesis. In this gene, the nonsynonymous variant at position 158943483bp ($P = 0.008$) was observed at higher frequency in cases compared with control group. The frequency of these mutations suggests their possible deleterious effect in increasing disease risk in genetically susceptible individuals.

## Individual Profiles of Rare and Deleterious Variants

Individual burden of variation revealed 24 variants (see Table, Supplemental Digital Content 3, http://links.lww.com/IBD/A811). Several dual susceptibility genes (for PIBD and asthma) were identified as harboring one or more variants. Mutations fall within 3 pathways consistently reported in KEGG and DAVID. *ZPBP2* (zona pellucida binding protein 2[12]) and *SMAD3* (involved in the adherens junction pathway[48]) have variants observed across both CD and UC but these variants also occur in 1% and 2% of the 1000 genomes reference population.

**TABLE 2.** Prevalence of Autoimmune Disease in the PIBD Cohort (173 Patients)

| Autoimmune Disease | CD (n = 98) (%) | UC (n = 55) (%) | IBDU (n = 20) (%) | Overall PIBD Cohort Prevalence (n = 173) (%) | Overall Population Pediatric Prevalence, % |
|---|---|---|---|---|---|
| Asthma | 19 (19.40) | 9 (16.40) | 0 | 28 (16.18) | 15.00[30,42,43] |
| Atopic dermatitis | 18 (18.40) | 6 (8.11) | 0 | 24 (13.87) | 16.50[44] |
| Coeliac disease | 1 (1.020) | 1 (1.35) | 0 | 2 (1.15) | 0.99[45] |
| Sclerosing cholangitis | 1 (1.02) | 2 (3.64) | 1 (50) | 4 (2.31) | 0.01[a,46] |
| Vitiligo | 1 (1.02) | 0 | 1 (50) | 2 (1.15) | 1.00[47] |

[a]Data for general population prevalence only and not specific to general IBD.

**TABLE 3.** Clinical Profile of the 18 Patients with Concurrent IBD and Asthma Selected for Exome Analysis

| Study ID | Sex | Age at Diagnosis of IBD | Diagnosis | Paris Classification of IBD | Other Autoimmune Disease Status Other than Asthma |
|---|---|---|---|---|---|
| PR0007 | M | 11.41 | CD | A1L34B1 | |
| PR0011 | M | 15.51 | CD | A1L1B2 | |
| PR0031 | M | 11.43 | CD | A1L3B1p | |
| PR0032 | M | 7.29 | CD | A1L24B1p | Atopic dermatitis |
| PR0036 | F | 9.67 | CD | A1L3B1 | |
| PR0039 | M | 10.30 | UC | E3- | |
| PR0068 | F | 11.23 | UC | E2- | |
| PR0083 | F | 9.68 | UC | E3S3 | |
| PR0085 | M | 13.12 | UC | E3S3 | |
| PR0107 | M | 9.22 | CD | A1L1- | |
| PR0110 | F | 2.98 | UC | E3- | |
| PR0146 | M | 14.52 | CD | A1L3- | |
| PR0148 | M | 9.13 | CD | A1L34B3p | Atopic dermatitis |
| PR0151 | F | 13.30 | CD | A1L24B1 | |
| PR0158 | F | 15.25 | UC | E3S3 | Atopic dermatitis |
| PR0160 | M | 12.55 | UC | E3S1 | Atopic dermatitis |
| PR0167 | M | 13.30 | UC | E2S2 | Atopic dermatitis |
| PR0188 | M | 11.03 | CD | A1L1- | |

- denotes missing classification data.

Patient PR0085 not only carries the nonsynonymous *ZPBP2* mutation but also carries a novel frameshift deletion in *DENND1B* gene expressed by natural killer cells and dendritic cells.[14,49] The same patient harbors a nonsynonymous mutation at position 67353579 bp within *GSTP1*, which is reported to be involved in asthma pathogenesis only. Also of interest among the genes previously implicated in both diseases are 2 distinct and very rare mutations in the *RAD50* gene located within the IBD5 cytokine cluster on chromosome 5q31.[50] This gene contains the locus control region required for the Th2 cytokine gene expression.[51] In asthma specific genes, variations were found in 8 genes. A more common variant (rs3918396) is seen to recur within the *ADAM33* gene, a second variant in the same gene has been identified in PR0158 and other patients within the Southampton PIBD cohort.

PR0110 is a patient diagnosed aged 2 years with severe UC. She is seen to harbor a mutation that could impact splicing at position 8009439 bp in *GLCCI1* and a novel frameshift insertion at position 69407255 within *CTNNA3*. This gene encodes the α–T-catenin protein; a key component of the adherens junctional complex in epithelial cells necessary for cellular adherence.[52]

### Joint Rare Variant Association Test for Variants in IBD and Dual Susceptibility Genes

The joint test for assessing the contribution of private, rare, and common mutation between disease status and genes highlighted 6 genes with a $P < 0.05$ before Bonferroni correction (Table 4).

Of these 6 genes, 3 are known susceptibility genes for both IBD and asthma (*ZPBP2*, $P = 0.009$; *IL1R1*, $P = 0.036$; and *IL18R1*, $P = 0.038$); the remaining genes were asthma specific (*PYHIN*, $P = 0.025$; *IL2RB*, $P = 0.036$; *GSTP1*, $P = 0.040$). These genes are all key determinants of the immune response and have variants observed across both CD and UC.

## DISCUSSION

Our cohort of 173 children with PIBD revealed forty-nine children (28.3%) with a concurrent diagnosis of an autoimmune disease. Asthma and atopic dermatitis occurred with the highest frequency; the prevalence of clinically diagnosed asthma was 19.4% in children with CD and 16.4% in patients with UC, exceeding UK disease estimates (15.3%[53,54]).

Although this study is not powered to demonstrate a statistically significant increase in autoimmune disease burden in children with PIBD, our observations indicate prevalence estimates approaching the upper limit recorded in the literature.[20] Our findings are consistent with literature indicating that children with PIBD are more likely to have other autoimmune conditions, and that a common genetic components etiology may predispose individuals to multiple autoimmune manifestations.[20,29]

Even in a very modest cohort, SKAT-O association analysis revealed 6 genes with significant burden of mutation. Although significant levels would not withstand a Bonferroni correction for 36 genes tested, the strong prior hypothesis to the

**TABLE 4.** Joint Variant Test (SKAT-O) Result for the 36 Known Asthma Genes in Which Variations was Found Across the Entire Cohort

| Gene set | Gene | Chr | Bp Position (hg19) | Total No. Samples, (18 Cases; 56 controls) | Fraction of Individuals Who Carry Rare Variants Under the MAF Thresholds (MAF <0.05)[a] | No. All Variants Defined in the Group File | No. Variant Defined as Rare (MAF <0.05)[a] | Unadjusted *P* |
|---|---|---|---|---|---|---|---|---|
| Asthma/IBD | ZPBP2 | 17 | 38024626-38032996 | 74 | 0.027 | 4 | 1 | 0.009 |
| Asthma | PYHIN1 | 1 | 158906777-158943483 | 74 | 0.108 | 5 | 5 | 0.025 |
| Asthma/IBD | IL1R1 | 2 | 102781629-158943483 | 74 | 0.014 | 5 | 2 | 0.037 |
| Asthma | IL2RB | 22 | 37524329-37539651 | 74 | 0.014 | 4 | 1 | 0.037 |
| Asthma/IBD | IL18R1 | 2 | 102984279-103001402 | 74 | 0.014 | 3 | 1 | 0.039 |
| Asthma | GSTP1 | 11 | 67352183-67353970 | 74 | 0.014 | 4 | 1 | 0.041 |
| Asthma | TLE4 | 9 | 82187750-82336794 | 74 | 0.108 | 4 | 4 | 0.056 |
| Asthma | NPSR1 | 7 | 34698177-34917768 | 74 | 0.135 | 12 | 4 | 0.066 |
| Asthma | CTNNA3 | 10 | 67680203-69407255 | 74 | 0.162 | 6 | 3 | 0.081 |

Only genes with a $P < 0.1$ are shown.
No. variants were found in ADRA1B, CCL5, CD14, HLA-DQ, IL12B, IL13, IL4, ORMDL3, PCDH1, RAD50, TNF, TSLP, and SLC22A5 across cases and controls.
[a]These variants received different weights in the SKAT-O joint test.

analysis of these genes might suggest that such a multiple testing correction would be inappropriate.

*ZPBP2* is located on the chr17q12-q21 region, which has been associated with early-onset asthma, and variants in the same linkage disequilibrium block have been associated with Crohn's disease, type 1 diabetes, and primary biliary cirrhosis.[55]

*IL1R1* encodes for a cytokine receptor that belongs to the interleukin-1 receptor family. The gene was found associated with asthma in a GWAS on 933 European ancestry individuals with severe asthma based on Global Initiative for Asthma criteria.[56] At the same genomic region of *IL1R1*, *IL18R1* was also identified as associated with asthma. Specifically, the gene was evaluated in a GWAS conducted on Mexican pediatric patients.[57] The association was further replicated on a family-based study on Denmark, United Kingdom, and Norway families.[58]

*GSTP1* is involved in the detoxification of a wide variety of exogenous and endogenous compounds, including reactive oxygen species. This gene was discovered by a GWAS conducted on early-onset asthma.[59] *IL2RB* is involved in lymphoid cell differentiation, and it was first discovered by GWAS conducted by the GABRIEL consortium in 2012.[60] *PYHIN1* (Pyrin And HIN Domain Family, Member 1) encodes a protein that belongs to the HIN-200 family of interferon-inducible proteins, important in controlling cell cycle, differentiation, and apoptosis.[61] It has been noted to be an asthma susceptibility locus, specifically in those of African descent.[62] *PYHIN1* was identified as associated with asthma in 2011, through a meta-analysis conducted on 5416 European American, African American, or African Caribbean, and Latino ancestry individuals with asthma. The *PYHIN1* association was specific to the African descent groups.[62]

*PYHIN1* and *ZPBP2* were significantly associated with asthma in both single variant testing and after SKAT-O testing. Variants within these genes were found with higher frequency in cases compared with controls suggesting a deleterious role of the mutations in the pathogenesis of disease. Susceptibility genes for both IBD and asthma are most commonly involved with immune regulation raising the possibility of an overall immune dysregulation underlying both diseases. These genes may be implicated in the same pathways as found in other probands but may not yet have been associated with IBD/asthma or did not hold enough significance power to be included in the GWAS meta-analyses.

This study demonstrates robust data collection, all PIBD diagnoses are made using strict criteria,[21] and autoimmune comorbidity was validated through integration of the medical notes (paper and electronic). This study looked only at genes identified through GWAS (of asthma and IBD); this increased the probability of finding causal, rare, and private mutation within known implicated genes. By design, GWAS are powered only to implicate genes in which common variant alleles are overrepresented in the disease population. It is highly likely that pathogenic coding changes that are either very rare or even private to individuals in other genes have gone undetected by these methods.

Exome sequencing allows capture of extremely large and useful amounts of data. Limitations of this sequencing technique still exist and can have an impact on research data; inefficiencies in the exon targeting process can lead to uneven capture and result in exons with low sequence coverage and off-target hybridizations. Alongside, this unknown or yet-to-be-annotated exons, evolutionary conserved noncoding regions and regulatory sequences (such as enhancers or promoters) involved in IBD and

asthma will not be captured. Exome sequencing is not designed to capture information regarding the methylation state of DNA, and therefore, epigenetic factors in disease are not investigated. Necessary filtering of vast data sets intrinsic to next generation sequencing may lead to missed calling of valid variants.

In this study, we identified the prevalence of concurrent autoimmune diagnoses in a cohort of children with childhood onset IBD. We observe a frequency of asthma and atopic dermatitis at the highest end of the normal range. In children with asthma, we demonstrate that patient-specific mutations in known disease-related genes are extensive and varied, even when restricted to mutations predicted to be pathogenic. Next generation sequencing may be set to become a key routine diagnostic tool of the future, and it is important that we begin to elucidate the role of key genes and pathways already known to us. Improved assessment of true functional significance of mutations will require substantial improvements to in silico annotation informed by rigorous and extensive functional validation of rare variants. However, perfect annotation of single variants in isolation cannot predict outcome in patients who harbor a profile of variants and across genes and pathways. This bottleneck to the interpretation of genomic data may be aided by the assessment of highly selected patient groups.[63] Our study uncovers the patient-specific burden of pathogenic mutations in known disease genes. We find evidence to support causality of key genes such as *ZPBP2* and *PYHIN1* and further postulate that for a subset of patients, the relationship between concurrent PIBD and autoimmune disease lies in systemic immune dysregulation rather than organ-specific immune dysfunction.

## ACKNOWLEDGMENTS

The authors are very grateful to all participants and their families. They thank Matthew Smith for helping with the demographic data and reviewing the clinical notes, Liz Blake for assisting pediatric recruitment, Nikki J Graham for technical assistance in DNA laboratory in Human Genetics & Genomic Medicine, University of Southampton, David Buck & Lorna Gregory from the Wellcome Trust Centre for Human Genetics, the NIHR & the Southampton Centre for Biomedical Research (SCBR).

## REFERENCES

1. Baumgart DC, Carding SR. Inflammatory bowel disease: cause and immunobiology. *Lancet*. 2007;369:1627–1640.
2. Baumgart DC, Sandborn WJ. Crohn's disease. *Lancet*. 2012;380:1590–1605.
3. Ashton JJ, Wiskin AE, Ennis S, et al. Rising incidence of paediatric inflammatory bowel disease (PIBD) in Wessex, Southern England. *Arch Dis Child*. 2014;99:659–664.
4. Henderson P, Hansen R, Cameron FL, et al. Rising incidence of pediatric inflammatory bowel disease in Scotland. *Inflamm Bowel Dis*. 2012;18:999–1005.
5. Hildebrand H, Finkel Y, Grahnquist L, et al. Changing pattern of paediatric inflammatory bowel disease in northern Stockholm 1990-2001. *Gut*. 2003;52:1432–1434.
6. Malmborg P, Grahnquist L, Lindholm J, et al. Increasing incidence of paediatric inflammatory bowel disease in northern Stockholm County, 2002-2007. *J Pediatr Gastroenterol Nutr*. 2013;57:29–34.
7. Gearry RB, Dodgshun AJ. The "hygiene hypothesis" in IBD. *J Crohns Colitis*. 2012;6:869; author reply 870.
8. Aujnarain A, Mack DR, Benchimol EI. The role of the environment in the development of pediatric inflammatory bowel disease. *Curr Gastroenterol Rep*. 2013;15:326.
9. Kilpinen H, Barrett JC. How next-generation sequencing is transforming complex disease genetics. *Trends Genet*. 2013;29:23–30.
10. Lesage S, Zouali H, Cézard JP, et al. CARD15/NOD2 mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. *Am J Hum Genet*. 2002;70:845–857.
11. Hugot JP, Chamaillard M, Zouali H, et al. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature*. 2001;411:599–603.
12. Anderson CA, Boucher G, Lees CW, et al. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat Genet*. 2011;43:246–252.
13. Jostins L, Ripke S, Weersma RK, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*. 2012;491:119–124.
14. Franke A, McGovern DP, Barrett JC, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet*. 2010;42:1118–1125.
15. Rivas MA, Beaudoin M, Gardet A, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat Genet*. 2011;43:1066–1073.
16. Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: genetic interactions create phantom heritability. *Proc Natl Acad Sci U S A*. 2012;109:1193–1198.
17. Gupta G, Gelfand JM, Lewis JD. Increased risk for demyelinating diseases in patients with inflammatory bowel disease. *Gastroenterology*. 2005;129:819–826.
18. Weng X, Liu L, Barcellos LF, et al. Clustering of inflammatory bowel disease with immune mediated diseases among members of a northern California-managed care organization. *Am J Gastroenterol*. 2007;102:1429–1435.
19. Bardella MT, Elli L, De Matteis S, et al. Autoimmune disorders in patients affected by celiac sprue and inflammatory bowel disease. *Ann Med*. 2009;41:139–143.
20. Kappelman MD, Galanko JA, Porter CQ, Sandler RS. Association of paediatric inflammatory bowel disease with other immune-mediated diseases. *Arch Dis Child*. 2011;96:1042–1046.
21. IBD Working Group of the European Society for Paediatric Gastroenterology, Hepatology and Nutrition. Inflammatory bowel disease in children and adolescents: recommendations for diagnosis—the Porto criteria. *J Pediatr Gastroenterol Nutr*. 2005;41:1–7.
22. Levine A, Griffiths A, Markowitz J, et al. Pediatric modification of the Montreal classification for inflammatory bowel disease: the Paris classification. *Inflamm Bowel Dis*. 2011;17:1314–1321.
23. Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res*. 1988;16:1215.
24. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25:2078–2079.
25. Illumina. *Quality Scores for Next-generation Sequencing*. Available at: http://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf. Accessed April 22, 2013.
26. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38:e164.
27. Pengelly RJ, Gibson J, Andreoletti G, et al. A SNP profiling panel for sample tracking in whole-exome sequencing studies. *Genome Med*. 2013;5:89.
28. Lockett GA, Holloway JW. Genome-wide association studies in asthma; perhaps, the end of the beginning. *Curr Opin Allergy Clin Immunol*. 2013;13:463–469.
29. Van Limbergen J, Russell RK, Nimmo ER, et al. Filaggrin loss-of-function variants are associated with atopic comorbidity in pediatric inflammatory bowel disease. *Inflamm Bowel Dis*. 2009;15:1492–1498.
30. Andersen ABT, Ehrenstein V, Erichsen R, et al. Parental inflammatory bowel disease and risk of asthma in offspring: a nationwide cohort study in Denmark. *Clin Transl Gastroenterol*. 2013;4:e41.

31. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. *Bioinformatics*. 2011;27:2156–2158.
32. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26:841–842.
33. Zhan X, Liu DJ. TaSer (TabAnno and SeqMiner): a toolset for annotating and querying next-generation sequence data.2013. Available at: http://arxiv.org/abs/1306.5715v1. Accessed January 22, 2014.
34. Kang HM, Zhan X, Sim X, et al. Biostatistics Dept, Univ Michigan, Ann Arbor, Ann Arbor, M. EPACTS (Efficient and Parallelizable Association Container Toolbox). Available at: http://genome.sph.umich.edu/wiki/EPACTS. Accessed January 22, 2014.
35. Drmanac R, Sparks AB, Callow MJ, et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*. 2010;327:78–81.
36. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 2001;29:308–311.
37. 1000 Genomes Project Consortium, Abecasis GR, Auton A, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491:56–65.
38. Li MX, Kwan JS, Bao SY, et al. Predicting mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. *PLoS Genet.* 2013;9:e1003143.
39. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc.* 2009;4:44–57.
40. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 2000;28:27–30.
41. Ionita-Laza I, Lee S, Makarov V, et al. Sequence kernel association tests for the combined effect of rare and common variants. *Am J Hum Genet.* 2013;92:841–853.
42. Anderson HR, Gupta R, Strachan DP, Limb ES. 50 years of asthma: UK trends from 1955 to 2004. *Thorax*. 2007;62:85–90.
43. Bernstein CN, Wajda A, Blanchard JF. The clustering of other chronic inflammatory diseases in inflammatory bowel disease: a population-based study. *Gastroenterology*. 2005;129:827–836.
44. Emerson RM, Williams HC, Allen BR. Severity distribution of atopic dermatitis in the community and its relationship to secondary referral. *Br J Dermatol.* 1998;139:73–76.
45. Van Heel DA, West J. Recent advances in coeliac disease. *Gut*. 2006;55:1037–1046.
46. Lindkvist B, Benito de Valle M, Gullberg B, Björnsson E. Incidence and prevalence of primary sclerosing cholangitis in a defined adult population in Sweden. *Hepatology*. 2010;52:571–577.
47. Alkhateeb A, Fain PR, Thody A, et al. Epidemiology of vitiligo and associated autoimmune diseases in Caucasian probands and their families. *Pigment Cell Res.* 2003;16:208–214.
48. Lu L, Wang J, Zhang F, et al. Role of SMAD and non-SMAD signals in the development of Th17 and regulatory T cells. *J Immunol.* 2010;184:4295–4306.
49. Sleiman PMA, Flory J, Imielinski M, et al. Variants of DENND1B associated with asthma in children. *N Engl J Med.* 2010;362:36–44.
50. Rioux JD, Daly MJ, Silverberg MS, et al. Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nat Genet.* 2001;29:223–228.
51. Lee GR, Fields PE, Griffin TJ, Flavell RA. Regulation of the Th2 cytokine locus by a locus control region. *Immunity*. 2003;19:145–153.
52. Kim SH, Cho BY, Park CS, et al. Alpha-T-catenin (CTNNA3) gene was identified as a risk variant for toluene diisocyanate-induced asthma by genome-wide association analysis. *Clin Exp Allergy* 2009;39:203–212.
53. The asthma epidemic. *N Engl J Med.* Available at: http://www.nejm.org/doi/full/10.1056/NEJMra054308. Accessed December 2, 2014.
54. Asher MI, Montefort S, Björkstén B, et al. Worldwide time trends in the prevalence of symptoms of asthma, allergic rhinoconjunctivitis, and eczema in childhood: ISAAC Phases One and Three repeat multicountry cross-sectional surveys. *Lancet*. 2006;368:733–743.
55. Verlaan DJ, Berlivet S, Hunninghake GM, et al. Allele-specific chromatin remodeling in the ZPBP2/GSDMB/ORMDL3 locus associated with the risk of asthma and autoimmune disease. *Am J Hum Genet.* 2009;85:377–393.
56. Wan YI, Shrine NR, Soler Artigas M, et al. Genome-wide association study to identify genetic determinants of severe asthma. *Thorax*. 2012;67:762–768.
57. Wu H, Romieu I, Shi M, et al. Evaluation of candidate genes in a genome-wide association study of childhood asthma in Mexicans. *J Allergy Clin Immunol.* 2010;125:321–327.e13.
58. Zhu G, Whyte MK, Vestbo J, et al. Interleukin 18 receptor 1 gene polymorphisms are associated with asthma. *Eur J Hum Genet.* 2008;16:1083–1090.
59. Michel S, Liang L, Depner M, et al. Unifying candidate gene and GWAS Approaches in Asthma. *PLoS One.* 2010;5:e13894.
60. Zhang Y, Moffatt MF, Cookson WOC. Genetic and genomic approaches to asthma: new insights for the origins. *Curr Opin Pulm Med.* 2012;18:6–13.
61. PYHIN1 pyrin and HIN domain family, member 1 [ Homo sapiens (human)]. Available at: http://www.ncbi.nlm.nih.gov/gene/149628. Accessed January 5, 2015.
62. Torgerson DG, Ampleford EJ, Chiu GY, et al. Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations. *Nat Genet.* 2011;43:887–892.
63. Chisholm J, Caulfield M, Parker M, et al. Briefing genomics England and the 100K genome project. *Genomics Engl.* 2013. Available at: http://www.genomicsengland.co.uk/briefing/. Accessed January 5, 2015.

# Immuno-Genomic Profiling of Patients with Inflammatory Bowel Disease: A Systematic Review of Genetic and Functional In Vivo Studies of Implicated Genes

*Tracy Coelho, MRCPCH,*[*,†] Gaia Andreoletti, MSc,*[*] James J. Ashton, BM, BS,*[†] Reuben J. Pengelly, MBiol,*[*] Yifang Gao, PhD,*[‡] Ananth RamaKrishnan, MD,*[‡] Akshay Batra, MD,*[†] Robert M. Beattie, MRCP, FRCPCH,*[†] Anthony P. Williams, PhD,*[‡] and Sarah Ennis, PhD**

**Background:** Over the last 2 decades, there has been an ever-expanding catalog of genetic variants implicated in inflammatory bowel disease (IBD) through genome-wide association studies and next generation sequencing. In this article, we highlight the remarkable developments in understanding the genetic and immunological basis of IBD. The main objective of the study was to perform a systematic review of published literature detailing functional/immunological studies in patients known to harbor genetic variations in the implicated genes.

**Methods:** A panel of 71 candidate genes implicated in IBD was prioritized using 5 network connectivity in silico methods. An electronic search using MEDLINE and EMBASE from 1996 to February 2014 for each of the selected genes was conducted. Only studies describing genotyped IBD cohorts with concurrent in vivo functional studies were included.

**Results:** Between the reviewers, a total of 35,142 potentially eligible publications were identified. Only 8 genes had publications meeting the inclusion criteria. A total of 67 studies were identified across the selected genes. The *NOD2* gene had the most number with 41 studies followed by *IL-10* with 11 eligible studies. A meta-analysis was not practical given the heterogeneity of the study design and the number of implicated genes with diverse immunological and physiological functions.

**Conclusions:** There is a clear lack of functional studies in humans to assess the in vivo impact of the various genetic variants implicated. A collaborative approach merging genomics and functional studies will help to unravel the obscure mechanisms involved in IBD.

*(Inflamm Bowel Dis 2014;20:1813–1819)*

**Key Words:** inflammatory bowel disease, Crohn's disease, ulcerative colitis, functional studies, genetic studies

Inflammatory bowel disease (IBD), like most other common diseases, has a complex pathobiology involving multiple genetic, immunological, and environmental factors. Crohn's disease (CD) and ulcerative colitis (UC) are the 2 main phenotypes of IBD, which can present with a diverse but quite often overlapping symptomatology. Genetics plays a major role in IBD. Over the last decade or so, with the advent of genome-wide association studies (GWAS) and next generation sequencing, a more defined albeit complex interplay between genes, host immunity, and the resident microbiota has emerged.[1,2] This has prompted a proliferation of research studies refining genetic loci to identify causal variants, functional studies to assess the complexities of the immune networks, and more recently a hugely topical subject of the role of microbiome in the pathogenesis of IBD.[3,4] The current hypothesis is that in health, there is a well-balanced homeostasis between the gut immune system and the resident microbiota of the gut. Any break down of this homeostasis which can occur in genetically susceptible individuals due to inherent "weaknesses" in their immune check points can lead to inflammatory changes in the gut wall as seen in IBD.[1] In this article, we present an overview of the genetic milestones that underpin the substantial contribution of the immune sensing and response in the pathogenesis of IBD. We then provide a systematic review of available literature highlighting studies with a combined approach of in vivo assessment of "aberrant" pathways suspected on the basis of genetic variations in human cohorts with IBD.

## A HISTORICAL JOURNEY THROUGH GENETICS OF IBD

Since the original description of CD by Crohn in 1932,[5] there have been several lines of epidemiological pointers

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.ibdjournal.org).

Received for publication June 10, 2014; Accepted July 3, 2014.

From the *Human Genetics and Genomic Medicine, Faculty of Medicine, University of Southampton, Southampton, United Kingdom; †Pediatric Gastrointestinal Unit, University Hospital Southampton, Southampton, United Kingdom; and ‡Cancer Sciences Division, Faculty of Medicine, University of Southampton, Southampton, United Kingdom.

T. Coelho is funded by CICRA (Crohn's in Childhood Research Association, United Kingdom). The remaining authors have no conflicts of interest to disclose.

Reprints: Tracy Coelho, MRCPCH, Human Genetics and Genomic Medicine, Duthie Building (Mp 808), University Hospital Southampton Foundation Trust, Southampton SO16 6YD, United Kingdom (e-mail: t.f.coelho@soton.ac.uk).



DOI 10.1097/MIB.0000000000000174

Published online 28 August 2014.

implicating genetic predisposition in the pathogenesis of IBD. In the 1980s, several studies confirmed the early findings of familial clustering of CD and suggested a positive family history ranging between 8% and 25%.[6] Large studies were carried out in the late 1980s in the United Kingdom and Scandinavia, which showed an increased concordance in twin studies.[7,8] The search for a long-known or suspected heritability in IBD, prompted by these early studies was then driven toward family-based linkage analysis in the 1990s to identify specific susceptibility genes.[9] At least 6 chromosomal regions were identified through linkage studies as linked to IBD and were named from IBD locus 1 to 6 according to their date of reporting and independent replication. The discovery of IBD-1 locus on chromosome 16 was a major breakthrough and till date is regarded as a single largest genetic risk factor for CD.[9,10]

Over the last 2 decades, gene discovery in complex diseases such as IBD has advanced rapidly through genome-wide scans. These studies have made a substantial impact in providing insights into the understanding of the disease and its complex biology. Until now, there have been at least 15 GWAS of IBD and 3 meta-analyses, which have successfully identified a total of 163 loci for IBD.[11] However, a major disadvantage of GWAS has been their intrinsic limitation to detection of common disease variation and so at best, this approach can only account for a portion of the heritability of the disease. Rare variants missed by GWAS may contribute significantly toward the missing heritability of IBD.[1,12] Technological advancements in recent years such as next generation sequencing now offer a feasible modality for studying rare and novel variation in disease causality.[13]

## GENES AND GASTROINTESTINAL HOMEOSTASIS

Candidate genes implicated in IBD highlight the interplay between several cellular mechanisms and immune pathways that are crucial for maintaining gastrointestinal homeostasis. These mechanisms broadly include the barrier function of the epithelium, innate immune regulation of microbial invasion, and the various effectors and regulators of adaptive immune response. Mutations in the key regulatory genes result in perturbations in the carefully balanced homeostasis that exists between the gastrointestinal immune system and the complex microbial milieu of the gut. The adverse outcome of this imbalance is inflammation of the gut resulting in IBD. We briefly present an overview of the complex levels of defense constantly on the role in the gut and their role in the pathogenesis of IBD.

## EPITHELIAL BARRIER AND OTHER LUMINAL ELEMENTS OF DEFENSE

The goblet cells of the intestinal epithelium secrete glycosylated mucins that form a mucus matrix over the epithelium, forming the first level of defense against any microbial invasion. The colon has a dual mucus layer; the inner layer has properties that restrict bacterial motility and adhesion to the epithelium.[14,15] The inner mucus layer is formed from sheets of *Muc2* mucin, which physically separates the epithelium from the bacteria. *Muc2*-null (*Muc2*$^{-/-}$) mice do not have any protective mucus and develop spontaneous colitis. Interestingly, it has also been observed that patients with active UC can have defective and penetrable mucus overlying the epithelium.[16] Epithelial cells associate with each other through a series of intercellular junctions, the most important being apical junction complex, which consists of tight junctions and adherens junctions. Epithelial barrier integrity is crucial for intestinal homeostasis in the context of IBD, and several genes associated with epithelial integrity are now implicated in IBD. Some of these genes include *CDH1*, *GNA12*, *HNF4A*, *ERRRFI1*, *MUC19*, *ITLN1*, and *PTPN2*.[1]

## NOD2 AND AUTOPHAGY

NOD2, also called CARD15 is an intracellular pathogen recognition receptor that recognizes N-acetyl muramyl dipeptide (MDP), derived from bacterial cell wall degradation.[17] One of the most remarkable achievements in the genetics of IBD came in 2001 when fine mapping of the IBD-1 locus on chromosome 16 by a French group identified the leucine-rich repeat variants of the *NOD2* gene (Nucleotide-binding oligomerization domain-containing-2) as conferring susceptibility to CD. The variants commonly found in the *NOD2* gene as described in the original study include single-nucleotide polymorphism (SNP) 8 (R702W), SNP 12 (G908R), and SNP 13 (1007 fs).[10] A number of rare genetic variants have been described subsequently, all of which almost exclusively localize to the leucine-rich repeat region. NOD2 plays a critical role in the induction of autophagy, which is a process whereby the cell tends to autodigest damaged intracellular organelles or intracellular bacteria by formation of an "isolation membrane" which is sequestered and marked for degradation. NOD2 recruits the autophagy protein ATG16L1 to the plasma membrane at the bacterial entry site. Cells with mutant NOD2 are incapable of this directive ATG16L1 recruitment and consequently fail to entrap pathogens through autophagy.[18]

## TH17 CELL SIGNALING

Until recently, it was believed that intestinal inflammation in CD is mediated by a Th1 response (tumor necrosis factor-α [TNF-α], interleukin [IL]-12, interferon-γ) and in UC by Th2 cytokine pathways (e.g., IL-4, IL-5, IL-13). Evidently, with the emerging role of other T-cell lineages such as regulatory T cells (Tregs) and Th17 cells, the concept of Th1 response and Th2 response in CD and UC, respectively, as the primary pathways of inflammation has been largely superseded. In health, a finely tuned homeostasis exists between the Tregs and the proinflammatory T-helper cells. An overzealous Th17 response or an inadequate Treg response can tip the balance toward an undesired inflammatory response. Th17 cells, which produce highly potent proinflammatory cytokines such as IL-17, are abundantly found in the mucosa of patients with active IBD.[19,20] Several genes in the Th17 pathway have been linked with IBD susceptibility, including *IL-23R*, *TNFSF15*, *STAT3*, *IL-12B*, *CCR6*, and *JAK2*.[21]

## IL-10 SIGNALING PATHWAY IN IBD

IL-10 is secreted by a wide variety of cells and over the last many years, it has been identified as crucial anti-inflammatory cytokine essential for maintaining gut homeostasis. IL-10 restricts the secretion of proinflammatory cytokines such as TNF-α and IL-12.[22] IL-10 signaling is required for limiting the expansion of Th17 cells, which are proinflammatory mercenaries.[23] The IL-10 receptor (IL-10R) has 2 alpha subunits and 2 beta subunits, which are encoded by *IL-10RA* and *IL-10RB*, respectively. Homozygous or compound heterozygous mutations in IL-10 or its receptor subunits have been well described in literature. Patients usually present with a very severe form of IBD at a very young age (<1 yr), with poor response to conventional therapy. These studies have highlighted the successful role of hematopoietic stem cell transplant in bringing about a sustained remission in these patients.[22,24,25]

## JAK-STAT PORTAL IN IBD

The Janus kinase/signal transducer and activator of transcription (JAK/STAT) pathway constitute a major portal for vital cellular processes including cell growth, differentiation, proliferation, and several immune mechanisms. Various cytokines and effectors communicate through this pathway to orchestrate an appropriate cellular response through target gene expression.[26] GWAS have implicated several genes in JAK-STAT pathway as candidate genes for IBD. Some of these include *JAK2*, *TYK2*, *STAT3*, and *STAT4*, with genetic variants associated with an increased risk of developing IBD.[11,21]

## SYSTEMATIC REVIEW OF IMMUNE FUNCTION STUDIES IN GENOTYPED INDIVIDUALS

Clear interpretation of the functional relevance of the many newly discovered genomic loci in IBD and associated variants is limited by the palpable lack of functional studies to characterize the molecular aberrations caused by these variants. Although many candidate genes implicated by genetic studies can be mapped to known pathways, a substantial fraction of these genes (>40%) are poorly understood at the functional level.[27] Establishing the causal effect of a genetic variant requires a mechanistic insight by the way of functional studies. With this in mind, we performed a systematic search of a selected panel of genes implicated in IBD through GWAS and other genetic studies. We specifically searched for studies in patients with IBD, where functional/immunological assessment of genotyped individuals was carried out to assess the in vivo functional impact of the genetic variants. This is a unique systematic review of literature aimed at exploring the strengths of integrating genetics/genomics with mechanistic studies in establishing causality of the multitude of variants implicated in IBD.

## METHODS

### Selection of Genes

Based on the largest and the most recent meta-analysis of IBD genome-wide association scans, the overall number of IBD loci is estimated as 163.[11] In this study, causal genes within the IBD loci were prioritized using 5 network connectivity tools such as Gene Relationships Across Implicated Loci, Disease Association Protein–Protein Link Evaluator, 3 different sources of expression quantitative trait locus, coding SNPs, and Coexpression network analysis (Fig. 1). Given the individual limitations of each of the tools used for functional clustering of genes, we adopted a consensus approach, requiring genes to be simultaneously implicated by at least 2 of the in silico methods. This approach focused our analysis on 71 candidate genes (see Table, Supplemental Digital Content 1, http://links.lww.com/IBD/A555).

### Literature Search

We conducted an electronic search through OvidSP using MEDLINE and EMBASE from 1996 to February 2014 for each of the 71 selected genes, looking specifically for genetic studies in IBD. A uniform search strategy was developed using a structured approach for every selected gene combining key words, gene symbol, and name as approved on "HUGO Gene Nomenclature Committee," using Boolean terms enabling multiple combinations of terms to be searched at once. We used the approved gene symbol, approved name, previous names, and synonyms as on HGNC (HUGO Gene Nomenclature Committee) by standardizing syntax as and where necessary, during the conduct of the search. An example of the search on *ATG16L1* gene is shown in Table 1.

The initial search intentionally focused on the genetic aspects and not the functional component, so as to be able to retrieve all the relevant articles. A study was eligible if it reported in vivo immunological/functional assessment of patients with genotyped IBD, relevant to the genetic variants detected. No language restrictions were applied. The titles and abstracts of articles obtained from electronic searches were screened and assessed for their relevance. Selected articles identified after the initial screen were retrieved in full text, and the reference lists were scanned to look for literature that had not been obtained by searches. At each stage of the study selection, eligibility of publications was assessed independently by 2 reviewers, and discrepancies if any, were resolved through discussion.

### Inclusion Criteria

Only studies describing genotyped cohorts with concurrent in vivo functional/immunological studies to assess the functional impact of genetic variants in individuals with IBD were included.

### Exclusion Criteria

1. Murine models and other animal studies
2. Functional studies carried out on cell lines
3. Review articles with no original data
4. Functional studies in healthy volunteers (and not IBD patients), bearing genetic variants implicated in IBD
5. Studies on same/overlapping sets of patients published by same groups in different journals
6. Conference abstracts.

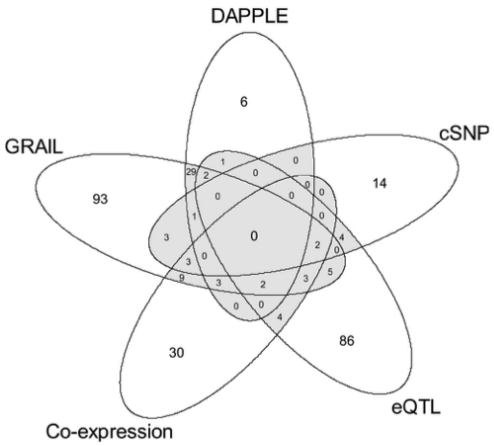



FIGURE 1. Shows the method used for prioritizing the genes implicated in IBD. The 5 network connectivity tools used in the third meta-analysis[11] to identify candidate genomic loci shown in the Venn diagram, namely Gene Relationships Across Implicated Loci (GRAIL), Disease Association Protein-Protein Link Evaluator, cSNP, Coexpression, and expression quantitative trait locus (eQTL). GRAIL uses text searching of abstracts in the scientific literature to identify linked genes; Disease Association Protein-Protein Link Evaluator (DAPPLE) interrogates known protein–protein interactions to identify proteins which are likely to physically interact; cSNP indicates where associated SNPs directly alter the protein, or are in strong linkage disequilibrium with known SNPs which alter the protein; eQTL denotes genes where associated SNPs are correlated with an alteration in protein expression levels and coexpression refers to genes for which expression patterns are linked to known inflammatory processes. For further details, see Jostins et al.[11] Genes identified by at least 2 connectivity tools were selected for the systematic search (71 genes out of 300; shaded region).

## RESULTS

A structured search across the 71 genes yielded a large number of studies involving genetic and functional studies on individuals with IBD. Between the reviewers, we identified 35,142 potentially eligible publications, which were assessed through titles, abstracts, and full text where appropriate. Of the 71 genes interrogated, only 8 genes had publications meeting our inclusion criteria. A total of 67 studies were identified across the selected genes, which met the criteria for inclusion. The *NOD2* gene had the most number of immuno-genomic studies, 41 studies meeting the eligibility criteria, followed by *IL-10* with 11 eligible studies. We also included *IL-10RA* and *IL-10RB* (genes encoding the receptor subunits), as most genetic studies on *IL-10* invariably included the receptor genes as well. Other genes investigated using functional studies on genotyped cohorts of IBD included *ATG16L1* (7 studies), *SLC22A4* (3 studies), *IL-23R* (2 studies), and *SLC11A1*, *CCL2*, and *STAT3* had one each (see Table, Supplemental Digital Content 2, http://links.lww.com/IBD/A556).

**TABLE 1.** Example of a Systematic Search on *ATG16L1*

| | |
|---|---|
| 1 | IBD*.tw. |
| 2 | IBD*.tw. |
| 3 | Crohn*.tw. |
| 4 | CD.tw. |
| 5 | UC.tw. |
| 6 | UC.tw. |
| 7 | 1 or 2 or 3 or 4 or 5 or 6 |
| 8 | ATG16L1.tw. |
| 9 | Autophagy related 16-like 1.tw. |
| 10 | "APG16 autophagy 16-like (*Saccharomyces cerevisiae*)".tw. |
| 11 | APG16L.tw. |
| 12 | "ATG16 autophagy related 16-like (*S. cerevisiae*)".tw. |
| 13 | "ATG16 autophagy related 16-like 1 (*S. cerevisiae*)".tw. |
| 14 | ATG16L.tw. |
| 15 | ATG16A.tw. |
| 16 | FLJ10035.tw. |
| 17 | WDR30.tw. |
| 18 | 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17 |
| 19 | 7 and 18 |

Numbers 1 to 7 to include all publications with IBD, number 8 includes approved gene symbol, number 9 approved gene name, 10 to 14 previous names of the gene, and 15 to 17 include synonyms for the gene.

For brevity, here, we present an overview of the studies retrieved through our search strategy on *NOD2* and *IL-10* genes (including *IL-10RA* and *IL-10RB*).

### Review of Selected Studies on the *NOD2* Gene

A structured search on the *NOD2* gene retrieved a total of 4994 publications between the 2 reviewers. Details of genotyping and functional methods on the 41 selected articles are given in Table, Supplemental Digital Content 2, http://links.lww.com/IBD/A556.

It is understood that NOD2 after the recognition of MDP triggers a defense response through NF-κB pathway. Hence, most of the functional studies focus on the assessment of the read-outs of the NF-κB pathway such as IL-8, TNF-α, IL-1β, and other cytokines for mechanistic assessment of the NOD2 protein.[28–34] Some studies have shown impaired NF-κB activation in response to MDP by detecting reduced phosphorylation levels in nuclear extracts of cells stimulated by MDP.[35,36] Rahman et al performed studies on Treg cells isolated from patients with CD with *NOD2* mutations, patients with wild-type allele, and healthy volunteers. They demonstrated that MDP-stimulated Tregs were normally protected from apoptosis; however, this protection was not evident in patients with *NOD2* polymorphisms.[37]

Hedl et al in 2007 found that pretreatment with MDP, significantly decreased production of the proinflammatory cytokines TNF-α, IL-8, and IL-1β on NOD2, TLR4, and TLR2 restimulation in primary human monocyte-derived macrophages. MDP-stimulated macrophages from CD-relevant Leu1007insC

*NOD2* homozygote individuals were deficient in their ability to cross-tolerize to subsequent treatment with TLR2 and TLR4 ligands.[33] On the same lines, NOD2/TLR cross talk has also been implicated by work from other groups.[29,38–44] Some studies proposed a cross talk and synergy between NOD1 AND NOD2 pathways, suggesting that *NOD2* mutations could lead to down regulation of NOD1 signaling.[45]

More recently, studies have also concentrated on the role of NOD2 in autophagy induction and the impaired bacterial handling consequent to *NOD2* mutations. Cooney et al showed that in cells homozygous for the CD-associated *NOD2* frame-shift mutation, mutant NOD2 fails to recruit ATG16L1 to the plasma membrane, resulting in an impaired engulfment of invading bacteria by autophagosomes. Their findings revealed that dendritic cells from patients with CD bearing *NOD2* mutations were defective in autophagy induction, bacterial trafficking, and antigen presentation.[46]

In terms of the methods used for genotyping of individuals, 38 of the selected studies performed genotyping for *NOD2* variants, 3 studies performed candidate gene sequencing, and none employed exome sequencing. Blood specimens were analyzed for functional work in 34 studies, ileal/colonic tissue in 5 studies, and a combination of peripheral blood cells and gut tissue in 2 studies. The synopsis of excluded studies on *NOD2* gene is given in Table 2.

## Review of Studies on the *IL-10*, *IL-10RA*, and *IL-10RB* Genes

Our search retrieved a total of 7300 publications, 11 of which met the selection criteria (see Table, Supplemental Digital Content 2, http://links.lww.com/IBD/A556). Immunological studies have focused on the functional capacity of IL-10 to limit the secretion of TNF-α and other proinflammatory cytokines. The IL-10 receptor consists of 2 alpha molecules (IL-10R1 encoded by *IL-10RA*) and 2 beta molecules (IL-10R2 encoded by *IL-10RB*). Mutations abrogate IL-10–induced signaling, causing disruption of the IL-10 dependent "negative feedback" regulation. Koss et al studied the influence of biallelic polymorphisms in TNF-α, lymphotoxin-α, and IL-10 genes on TNF-α and IL-10 production. Three haplotypes of IL-10 were identified. The functional effect of IL-10 haplotypes on IL-10 protein production in whole blood samples stimulated with lipopolysaccharide (LPS) was analyzed.[47] Van der Linde et al identified a point mutation (Gly15Arg) in the leader sequence of IL-10 following genotyping of IL-10 alleles in 17 sibling pairs with CD and 75 healthy controls. The functional consequences of this genetic variation were tested by stimulating peripheral blood mononuclear cells of patients bearing this mutation with LPS or phorbol ester, and then assessing the cellular supernatants for IL-10 production by enzyme-linked immunosorbent assay or Western blotting. The activity of recombinant immature wild-type or mutated IL-10 was also tested in vitro in a proliferation assay with LPS-stimulated human monocytic cell line (HL60 cells).[48] In keeping with previous reports, Glocker et al, in 2009, proposed that the pathophysiology of a deficiency in the IL-10 receptor involves undue and prolonged activation of mononuclear cells on exposure to bacterial particles, resulting in an exaggerated efflux of inflammatory cytokines such as TNF-α causing mucosal damage. To test this hypothesis, TNF-α secretion of monocytes and monocyte-derived macrophages was analyzed on exposure to LPS or LPS plus IL-10 in patients with the receptor *IL-10RA* mutations and in healthy controls. Patients in this group had a very early onset IBD (in the first year of life) with a very severe disease refractory to conventional treatment. Mutations were identified through genetic-linkage analysis and candidate gene sequencing on samples from 2 unrelated consanguineous families with children who were affected by very early onset IBD and from 6 additional patients with very early onset IBD. IL-10 substantially reduced the release of TNF-α in cells from control subjects; however, this inhibitory effect was absent in cells from patients with the receptor mutations. The impairment in the capacities of mononuclear cells with IL-10 or IL-10 receptor deficiencies to inhibit TNF-α production have been reproduced through studies by other groups subsequently.[49–52]

Of the 11 selected studies, whole exome sequencing was used in 1 study,[51] candidate gene sequencing in 6 studies[22,48–50,52,53] and SNP genotyping in 4 studies.[47,54–56] Specimens analyzed for functional studies were blood samples in 8 studies, and in 3 studies, ileal/colonic tissue was assessed in addition to blood samples. The synopsis of excluded studies on *IL-10* is given in Table 2.

**TABLE 2.** A Synopsis of All Excluded Studies on *NOD2* Gene and *IL-10/RA/RB* Gene Retrieved Through Initial Search (Relevant Studies but not Meeting Eligibility Criteria)

| Studies | *NOD2* Gene | *IL-10/IL-10RA/ IL-10RB* Gene |
|---|---|---|
| Total no. publications screened | 1670 | 2354 |
| Functional studies in humans (no genotyping), n (%) | 62 (3.7) | 289 (12.3) |
| Human genetic studies including GWAS and meta-analysis (no functional studies), n (%) | 315 (18.8) | 55 (2.3) |
| Studies on animal models (genetic or functional), n (%) | 98 (5.8) | 845 (35.9) |
| Studies on cell lines (genetic or functional), n (%) | 22 (1.3) | 22 (1) |
| Reviews (no original data), n (%) | 81 (4.8) | 59 (2.5) |
| Selected studies, n (%)[a] | 41 (2.5) | 11 (0.46) |
| Publications not pertinent to our review and duplicate studies, n (%) | 1051 (63) | 1073 (45.5) |

Assessed by a Single Reviewer TC on EMBASE.
[a]Details of selected articles included in Supplemental Digital Content.

## DISCUSSION

During our structured search for immuno-genomic studies in IBD, we found a large number of functional studies conducted

in humans, focusing on the implicated immunological pathways. A meta-analysis was not practical given the heterogeneity of the study design and the number of implicated genes with diverse immunological and physiological functions. For the *NOD2* gene, 3.7% of the studies (66 studies out of 1670 publications retrieved) were functional studies carried out in humans, and for the *IL-10* gene (including *IL-10/IL-10RA/IL-10RB*), human functional studies were 12.3% (289 studies out of 2354 publications). However, these studies were conducted on cohorts without a genotyped profile. Conversely, we found an ever-expanding number of studies with genotyped cohorts, but without the functional element (19% for *NOD2* and 2.5% for *IL-10*). There could be several reasons for this discrepant observation. One of the key factors may be due to the fact that IBD in a given individual may be due to multiple hits at the implicated loci. Several inflammatory mechanisms and pathways may be involved with varying degrees of contributory impact, thereby making it extremely difficult to design functional studies bespoke to the genetic variants identified in a given individual with IBD. Yet, another reason for the paucity of functional studies to backup implicated genetic variants is the relative lack of enthusiasm to invest in conducting functional work in a GWAS variant where there is no proven causality.

An extensive number of murine models have contributed significantly toward understanding the mechanistic basis with both induced and spontaneous mutations in a diversity of genes. This was clearly obvious during our search, which generated a vast number of experimental studies in knockout models. We identified 98 (6%) of 1670 publications, conducted as preclinical experiments in animal models for the *NOD2* gene and 845 (36%) of 2354 publications for the *IL-10/IL-10RA/IL-10RB* gene. As is obvious, the search retrieved a significantly larger number of animal model studies for *IL-10* as compared with *NOD2* because of the extensive use of *IL-10* knockout models for experimental colitis. These experimental models have enhanced our understanding of the functional impact of specific genes in a defined biological process; however, it does not effectively map out how genetic variants in human population will impact immune function. Animal models, although crucial for identifying pathways of susceptibility, are not an ideal platform to establish and define the pathogenetic associations in humans. Similarly, transformed cell lines are good experimental models for transfecting in a gene when the behavior of the cell itself is not of interest. However, this can be a major pitfall whilst trying to establish the actual functional impact caused by genetic variants in human disease. Also, transfection of cell lines can lead to overexpression of genes resulting in an unpredictable outcome. Therefore, in this context, to determine and decipher the role of candidate genes in the causation of disease, in vivo studies in human cohorts may be more meaningful.

Our study was limited to only 71 genes based on our inclusion strategy. It is possible that the immuno-genomic landscape is different for the other genes identified in IBD pathogenesis. Some of the commonly implicated genes such as *IRGM*, *ERAP2*, *MUC19*, *CDH1*, and others were not identified for our review based on the nature of our methodology. Given the heterogeneity of the functional methods used, selection of publications was subject to reviewer bias. This was however kept to the minimum through a standardized assessment, discussion, and consensus at all stages of the review. A substantial number of studies evaluated a panel of genes for genetic profiling, rather than a single gene resulting in an unavoidable overlap between studies across the genes under assessment. For example, some of the selected studies were included for both *ATG16L1* and *NOD2* genes, as the studies evaluated both the genes from an immunogenomic angle given that the 2 genes work closely together.[46,57]

Human genomics has provided key insights into the complexities of the biology of IBD. To consolidate and take it to the next level of understanding, there is now a clear pressing need for more collaborative approach between human genomics and immunology. Immuno-genomic profiling can possibly inform a risk-prediction model for complicated IBD progression. This will enhance the prospects of a more refined tailor-made diagnostic and therapeutic approach in IBD for the foreseeable future.

## REFERENCES


1. Khor B, Gardet A, Xavier RJ. Genetics and pathogenesis of inflammatory bowel disease. *Nature*. 2011;474:307–317.
2. Lees CW, Barrett JC, Parkes M, et al. New IBD genetics: common pathways with other diseases. *Gut*. 2011;60:1739–1753.
3. Hansen R, Berry SH, Mukhopadhya I, et al. The microaerophilic microbiota of de-novo paediatric inflammatory bowel disease: the BISCUIT study. *PLoS One*. 2013;8:e58825.
4. Qin J, Li R, Raes J, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*. 2010;464:59–65.
5. Crohn BB, Ginzburg L, Oppenheimer GD. Landmark article Oct 15, 1932. Regional ileitis. A pathological and clinical entity. By Burril B. Crohn, Leon Ginzburg, and Gordon D. Oppenheimer. *JAMA*. 1984;251:73–79.
6. Peeters M, et al. Familial aggregation in Crohn's disease: increased age-adjusted risk and concordance in clinical characteristics. *Gastroenterology*. 1996;111:597–603.
7. Henderson P, Satsangi J. Genes in inflammatory bowel disease: lessons from complex diseases. *Clin Med*. 2011;11:8–10.
8. Tysk C, Lindberg E, Järnerot G, et al. Ulcerative colitis and Crohn's disease in an unselected population of monozygotic and dizygotic twins. A study of heritability and the influence of smoking. *Gut*. 1988;29:990–996.
9. Cardinale CJ, Kelsen JR, Baldassano RN, et al. Impact of exome sequencing in inflammatory bowel disease. *World J Gastroenterol*. 2013;19:6721–6729.
10. Hugot JP, Chamaillard M, Zouali H, et al. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature*. 2001;411:599–603.
11. Jostins L, Ripke S, Weersma RK, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*. 2012;491:119–124.
12. Van Limbergen J, Radford-Smith G, Satsangi J. Advances in IBD genetics. *Nat Rev Gastroenterol Hepatol*. 2014;11:372–385.
13. Rivas MA, Beaudoin M, Gardet A, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat Genet*. 2011;43:1066–1073.
14. Cader MZ, Kaser A. Recent advances in inflammatory bowel disease: mucosal immune cells in intestinal inflammation. *Gut*. 2013;62:1653–1664.
15. Laukoetter MG, Nava P, Nusrat A. Role of the intestinal barrier in inflammatory bowel disease. *World J Gastroenterol*. 2008;14:401–407.
16. Johansson ME, Gustafsson JK, Holmén-Larsson J, et al. Bacteria penetrate the normally impenetrable inner colon mucus layer in both murine colitis models and patients with ulcerative colitis. *Gut*. 2014;63:281–291.
17. Fritz T, Niederreiter L, Adolph T, et al. Crohn's disease: NOD2, autophagy and ER stress converge. *Gut*. 2011;60:1580–1588.

18. Muzes G, Tulassay Z, Sipos F. Interplay of autophagy and innate immunity in Crohn's disease: a key immunobiologic feature. *World J Gastroenterol.* 2013;19:4447–4454.
19. Fujino S, Andoh A, Bamba S, et al. Increased expression of interleukin 17 in inflammatory bowel disease. *Gut.* 2003;52:65–70.
20. Rovedatti L, Kudo T, Biancheri P, et al. Differential regulation of interleukin 17 and interferon gamma production in inflammatory bowel disease. *Gut.* 2009;58:1629–1636.
21. Franke A, McGovern DP, Barrett JC, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet.* 2010;42:1118–1125.
22. Glocker EO, Kotlarz D, Boztug K, et al. Inflammatory bowel disease and mutations affecting the interleukin-10 receptor. *N Engl J Med.* 2009;361:2033–2045.
23. Chaudhry A, Samstein RM, Treuting P, et al. Interleukin-10 signaling in regulatory T cells is required for suppression of Th17 cell-mediated inflammation. *Immunity.* 2011;34:566–578.
24. Engelhardt KR, Shah N, Faizura-Yeop I, et al. Clinical outcome in IL-10- and IL-10 receptor-deficient patients with or without hematopoietic stem cell transplantation. *J Allergy Clin Immunol.* 2013;131:825–830.
25. Moran CJ, Walters TD, Guo CH, et al. IL-10R polymorphisms are associated with very-early-onset ulcerative colitis. *Inflamm Bowel Dis.* 2013;19:115–123.
26. Coskun M, Salem M, Pedersen J, et al. Involvement of JAK/STAT signaling in the pathogenesis of inflammatory bowel disease. *Pharmacol Res.* 2013;76:1–8.
27. Graham DB, Xavier RJ. From genetics of inflammatory bowel disease towards mechanistic insights. *Trends Immunol.* 2013;34:371–378.
28. Li J, Moran T, Swanson E, et al. Regulation of IL-8 and IL-1beta expression in Crohn's disease associated NOD2/CARD15 mutations. *Hum Mol Genet.* 2004;13:1715–1725.
29. van Heel DA, Ghosh S, Butler M, et al. Muramyl dipeptide and toll-like receptor sensitivity in NOD2-associated Crohn's disease. *Lancet.* 2005;365:1794–1796.
30. van Heel DA, Hunt KA, Ghosh S, et al. Normal responses to specific NOD1-activating peptidoglycan agonists in the presence of the NOD2 frameshift and other mutations in Crohn's disease. *Eur J Immunol.* 2006;36:1629–1635.
31. van Heel DA, Hunt KA, King K, et al. Detection of muramyl dipeptide-sensing pathway defects in patients with Crohn's disease. *Inflamm Bowel Dis.* 2006;12:598–605.
32. Peeters H, Bogaert S, Laukens D, et al. CARD15 variants determine a disturbed early response of monocytes to adherent-invasive Escherichia coli strain LF82 in Crohn's disease. *Int J Immunogenet.* 2007;34:181–191.
33. Hedl M, Li J, Cho JH, et al. Chronic stimulation of Nod2 mediates tolerance to bacterial products. *Proc Natl Acad Sci U S A.* 2007;104:19440–19445.
34. Lappalainen M, Paavola-Sakki P, Halme L, et al. Novel CARD15/NOD2 mutations in Finnish patients with Crohn's disease and their relation to phenotypic variation in vitro and in vivo. *Inflamm Bowel Dis.* 2008;14:176–185.
35. Seidelin JB, Broom OJ, Olsen J, et al. Evidence for impaired CARD15 signalling in Crohn's disease without disease linked variants. *PLoS One.* 2009;4:e7794.
36. Kuuliala K, Lappalainen M, Turunen U, et al. Detection of muramyl dipeptide-sensing pathway defects in monocytes of patients with Crohn's disease using phospho-specific whole blood flow cytometry. *Scand J Clin Lab Invest.* 2013;73:494–502.
37. Rahman MK, Midtling EH, Svingen PA, et al. The pathogen recognition receptor NOD2 regulates human FOXP3+ T cell survival. *J Immunol.* 2010;184:7247–7256.
38. van Heel DA, Ghosh S, Hunt KA, et al. Synergy between TLR9 and NOD2 innate immune responses is lost in genetic Crohn's disease. *Gut.* 2005;54:1553–1557.
39. Braat H, Stokkers P, Hommes T, et al. Consequence of functional Nod2 and Tlr4 mutations on gene transcription in Crohn's disease patients. *J Mol Med (Berl).* 2005;83:601–609.
40. Canto E, Moga E, Ricart E, et al. MDP-Induced selective tolerance to TLR4 ligands: impairment in NOD2 mutant Crohn's disease patients. *Inflamm Bowel Dis.* 2009;15:1686–1696.
41. Brosbol-Ravnborg A, Hvas CL, Agnholt J, et al. Toll-like receptor-induced granulocyte-macrophage colony-stimulating factor secretion is impaired in Crohn's disease by nucleotide oligomerization domain 2-dependent and -independent pathways. *Clin Exp Immunol.* 2009;155:487–495.
42. Butler M, Chaudhary R, van Heel DA, et al. NOD2 activity modulates the phenotype of LPS-stimulated dendritic cells to promote the development of T-helper type 2-like lymphocytes—possible implications for NOD2-associated Crohn's disease. *J Crohns Colitis.* 2007;1:106–115.
43. Vissers M, Remijn T, Oosting M, et al. Respiratory syncytial virus infection augments NOD2 signaling in an IFN-beta-dependent manner in human primary cells. *Eur J Immunol.* 2012;42:2727–2735.
44. Kullberg BJ, Ferwerda G, de Jong DJ, et al. Crohn's disease patients homozygous for the 3020insC NOD2 mutation have a defective NOD2/TLR4 cross-tolerance to intestinal stimuli. *Immunology.* 2008;123:600–605.
45. Netea MG, Ferwerda G, de Jong DJ, et al. The frameshift mutation in Nod2 results in unresponsiveness not only to Nod2- but also Nod1-activating peptidoglycan agonists. *J Biol Chem.* 2005;280:35859–35867.
46. Cooney R, Baker J, Brain O, et al. NOD2 stimulation induces autophagy in dendritic cells influencing bacterial handling and antigen presentation. *Nat Med.* 2010;16:90–97.
47. Koss K, Satsangi J, Fanning GC, et al. Cytokine (TNF alpha, LT alpha and IL-10) polymorphisms in inflammatory bowel diseases and normal controls: differential effects on production and allele frequencies. *Genes Immun.* 2000;1:185–190.
48. van der Linde K, Boor PP, Sandkuijl LA, et al. A Gly15Arg mutation in the interleukin-10 gene reduces secretion of interleukin-10 in Crohn disease. *Scand J Gastroenterol.* 2003;38:611–617.
49. Begue B, Verdier J, Rieux-Laucat F, et al. Defective IL-10 signaling defining a subgroup of patients with inflammatory bowel disease. *Am J Gastroenterol.* 2011;106:1544–1555.
50. Kotlarz D, Beier R, Murugan D, et al. Loss of interleukin-10 signaling and infantile inflammatory bowel disease: implications for diagnosis and therapy. *Gastroenterology.* 2012;143:347–355.
51. Mao H, Yang W, Lee PP, et al. Exome sequencing identifies novel compound heterozygous mutations of IL-10 receptor 1 in neonatal-onset Crohn's disease. *Genes Immun.* 2012;13:437–442.
52. Marcuzzi A, Girardelli M, Bianco AM, et al. Inflammation profile of four early onset Crohn patients. *Gene.* 2012;493:282–285.
53. Galatola M, Miele E, Strisciuglio C, et al. Synergistic effect of interleukin-10-receptor variants in a case of early-onset ulcerative colitis. *World J Gastroenterol.* 2013;19:8659–8670.
54. Gasche C, Grundtner P, Zwirn P, et al. Novel variants of the IL-10 receptor 1 affect inhibition of monocyte TNF-alpha production. *J Immunol.* 2003;170:5578–5582.
55. Wang AH, Lam WJ, Han DY, et al. The effect of IL-10 genetic variation and interleukin 10 serum levels on Crohn's disease susceptibility in a New Zealand population. *Hum Immunol.* 2011;72:431–435.
56. Wagner J, Skinner NA, Catto-Smith AG, et al. TLR4, IL10RA, and NOD2 mutation in paediatric Crohn's disease patients: an association with Mycobacterium avium subspecies paratuberculosis and TLR4 and IL10-RA expression. *Med Microbiol Immunol.* 2013;202:267–276.
57. Glubb DM, Gearry RB, Barclay ML, et al. NOD2 and ATG16L1 polymorphisms affect monocyte responses in Crohn's disease. *World J Gastroenterol.* 2011;17:2829–2837.

Pengelly et al. Genome Medicine 2013, 5:89
http://genomemedicine.com/content/5/9/89

Genome Medicine

METHOD Open Access

# A SNP profiling panel for sample tracking in whole-exome sequencing studies

Reuben J Pengelly[1], Jane Gibson[1], Gaia Andreoletti[1], Andrew Collins[1], Christopher J Mattocks[2] and Sarah Ennis[1*]

**Abstract**

Whole-exome sequencing provides a cost-effective means to sequence protein coding regions within the genome, which are significantly enriched for etiological variants. We describe a panel of single nucleotide polymorphisms (SNPs) to facilitate the validation of data provenance in whole-exome sequencing studies. This is particularly significant where multiple processing steps necessitate transfer of sample custody between clinical, laboratory and bioinformatics facilities. SNPs captured by all commonly used exome enrichment kits were identified, and filtered for possible confounding properties. The optimised panel provides a simple, yet powerful, method for the assignment of intrinsic, highly discriminatory identifiers to genetic samples.

## Background

Whole-exome sequencing (WES) is presently one of the most efficient means of identifying aetiological genetic mutations [1], minimising some of the challenges associated with whole-genome sequencing, such as high cost and data processing burden, analysis and interpretation. In WES, protein-coding regions of the genome are targeted and enriched via specific hybridisation of genomic fragments with complementary oligonucleotides, or 'baits'. These targeted regions are then sequenced using high throughput next-generation sequencing (NGS) technologies [2].

The high start-up investment required for in-house WES is currently prohibitive to many groups so sample preparation and/or sequencing is commonly outsourced. This transference of sample custody, combined with the complex sample preparation workflow, makes sample mix-ups possible, and difficult to detect. In both clinical and research contexts, ensuring provenance of data is essential to allow the accurate assignment of clinical details to sequence data. It is possible that samples may be misidentified at any stage of the analytical process, both *in vitro* and *in silico*. Therefore, sample tracking must be contiguous throughout both data generation and analysis.

Consequent to sample mix-ups in a research setting, erroneous data and sample matching may result in a loss of power for identification of causal variants [3]. In a clinical setting, this may lead to delayed or inaccurate reporting of results to patients. Whilst good practice in the handling of samples and increased laboratory automation minimises potential for error, additional checkpoints are still required to support quality control [4]. A method for *post hoc* confirmation of sample identity is therefore highly desirable.

Genetic sample identification methods have an advantage over alternative sample management systems in that the genetic 'label' is intrinsic to the biological sample itself, removing the possibility of manual labelling errors. Single nucleotide polymorphisms (SNPs) are increasingly utilised for DNA-based identification of human samples, with several benefits compared to standard forensic methods [5-7]. Existing SNP panels for human forensic identification and commercial SNP panels for sample identification, such as the iPLEX Sample ID Plus panel (Sequenom, San Diego, CA, USA), utilise pan-genome SNPs, the majority of which are non-exonic, and are therefore not useful for WES studies, as the majority of markers will not lie within the enriched regions of the genome. In addition to existing SNP panels, short tandem repeat markers can be used for genetic sample tracking. However, again, markers applied are frequently outside exomic regions and, if captured, will be prone to erroneous NGS genotyping using standard pipelines due to the repetitive nature of the markers [8].

Several methods for genetic tracking of human biological samples have been previously described, some of

* Correspondence: s.ennis@soton.ac.uk
[1]Human Genetics and Genomic Medicine, Faculty of Medicine, University of Southampton, Duthie Building (MP 808), Southampton General Hospital, Tremona Road, Southampton SO16 6YD, UK
Full list of author information is available at the end of the article

which are application specific - for example, for transcriptome microarray studies [3,9,10]. Although software for the validation of NGS (including WES) sample identity, such as verifyBamID, exist [11], for the detection of sample misidentifications external array-based genotypes of the samples are required, without which only contamination of the samples can be assessed.

Here we describe an optimised panel of SNPs for which WES data are typically informative, the genotypic profile of which can be utilised to extract intrinsic identifiers from human genomic DNA. These SNP profiles have high discriminatory power, even in large datasets. The profile derived from this panel can be compared to an independently genotyped profile for the same individual, allowing accurate validation of data and sample pairings, at a modest cost per sample.

## Methods

### Candidate identification and panel selection

Regions of overlap between three current commonly used whole-exome enrichment kits, (namely Agilent SureSelect Human All Exon V4, Illumina TruSeq Exome Enrichment and Nimblegen SeqCap EZ Human Exome Library V3.0 kits), and common SNPs (dbSNP 137, [12]) were established using BEDTools [13]. SNPs were further filtered for inclusion based upon their presence in genes targeted by the Illumina TruSight Exome kit, which targets only genes of clinical interest.

Primary candidate selection criteria required SNPs to: 1) represent bi-allelic substitutions, excluding substitutions of complementary bases, that is, A↔T and G↔C transversions; 2) be technically amenable to both accurate WES and orthogonal genotyping, that is, not present in large-scale genomic repeats [14], or homopolymeric tracts of ≥5 bp, GC content for the flanking 250 bp was restricted to a range of between 40% and 55% and no other variant within 50 bp with an alternative-allele frequency (AF) ≥0.01 was permitted; 3) conform to desirable phase 3 HapMap AFs across several populations, explicitly AFs of between 0.2 and 0.8 in: CEPH (Utah residents with ancestry from northern and western Europe; CEU), Japanese in Tokyo, Japan (JPT), Han Chinese in Beijing, China (CHB) and Yoruba in Ibidan, Nigeria (YRI) [15] and; 4) not alter the primary sequence of the encoded protein or have an associated Online Mendelian Inheritance in Man (OMIM) record [16].

Following primary candidate identification steps, SNPs were further optimised by the following requirements: 1) be located at least 10 bp from exon boundaries; 2) not be situated in regions with a high sequence similarity to non-target regions, that is, no non-target BLAT score >100 [17], as this could result in non-specific genotyping; and 4) be outside of linkage disequilibrium with all other selected SNPs.

Finally, SNPs were prioritised for inclusion in the panel by proximity of the AFs to 0.5, across HapMap populations, in order to maximise discriminatory power.

### SNP coverage in whole-exome sequencing data

A set of 91 in-house exome samples was evaluated for depth of sequence coverage for the candidate SNPs. Exome capture was performed using Agilent SureSelect Human All Exon V3 (n = 22) and V4 (n = 55), Illumina TruSeq Exome Enrichment (n = 9) and Nimblegen SeqCap EZ Human Exome Library V3.0 (n = 5). Exome enrichment, sequencing and *in silico* analysis of samples was performed as previously described [18,19].

### Optimised panel validation

The power of sample resolution for the panel was validated using data from phase 1 of the 1000 Genomes Project (n = 1,092) [20] and the UK10K project (n = 2,688; 2,432 of which are whole-genome data) [21]. Genotypes were extracted from data using custom scripts and Tabix [22]. Quantification of mismatches between samples was performed using MEGA5 [23].

Simulated datasets were generated by taking individual population AFs for each SNP as input and generating random SNP profiles in accordance with Hardy-Weinberg equilibrium based upon this; the randomisation of each SNP was independent of all other SNPs. We then quantified the rate of non-unique profiles per simulated dataset. We performed 20,000 independent replicates of dataset generation in all cases.

### Panel application

We applied the panel to a batch of 48 samples exome sequenced by an external service provider, for which orthogonal genotypes were obtained concurrently through an independent genotyping provider using KASP genotyping (LGC Genomics, Hoddeston, UK). Following plating of DNA samples for dispatch, a replicate plate was made directly from the primary plate, to be dispatched for the orthogonal genotyping. Genotypes derived from exome data and orthogonal genotyping assays were compared using PLINK [24] and custom scripts.

### Ethics

This study was approved by the Southampton and South West Hampshire Research Ethics Committee (09/H0504/125). Informed consent was obtained for all participants.

## Results

In total, 26.2 Mbp of genome sequence was found to overlap all three commonly applied whole-exome capture kits, containing 9,493 common SNPs (Figure 1A,B). Of these, 1,662 SNPs are additionally covered by the Illumina TruSight Exome kit. Within this

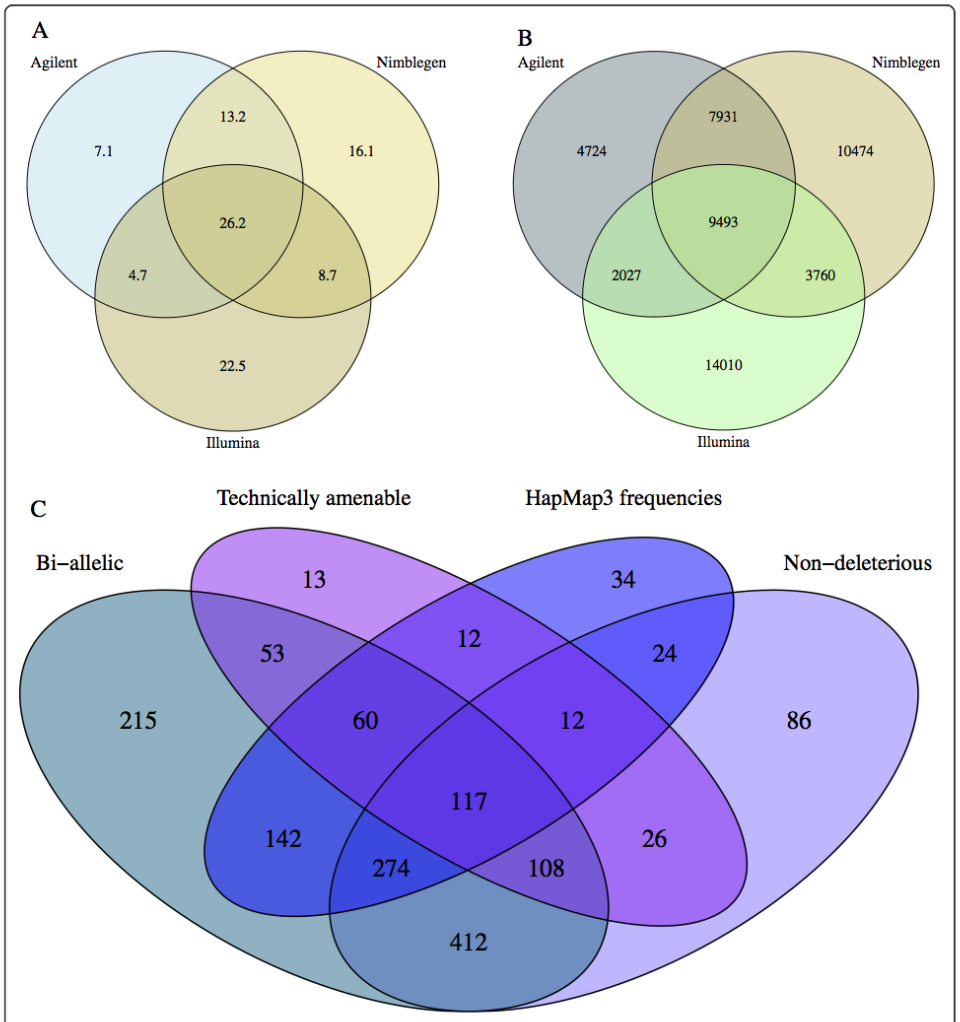



**Figure 1 Venn diagrams showing commonality of targeting between capture kits (A,B) and properties of encompassed SNPs (C).** Overlap between exome capture kits is presented in Mbp **(A)** and number of SNPs with an AF ≥0.3 **(B)**. Agilent - SureSelect Human All Exon V4; Illumina - TruSeq Exome Enrichment; Nimblegen - SeqCap EZ Human Exome Library V3.0. For a subset of SNPs present in both the intersection of the three kits shown, and the Illumina TruSight Exome kit, a breakdown of fulfilment of the four classes of candidate filtering criteria is shown **(C)** (see the main text for details of filtering criteria); 117 SNPs exhibited all desired characteristic; 74 SNPs exhibited none of the desired characteristics.

subset, following the filtering for all primary candidate criteria, 117 candidate SNPs were identified (Figure 1C; Additional file 1), from which the optimised panel of 24 SNPs was selected (Table 1). Within the set of 91 in-house exome samples, all 24 SNPs were sequenced at sufficient read-depth for accurate genotype calling, across all capture kits.

The 24 biallelic SNPs afford 48 points of allelic comparison. Testing the optimised panel in the 1000 Genomes Project data (n = 1,092) [20], an average of

**Table 1 Optimised panel of identifying SNPs**

| Chromosome | Position[a] | dbSNP rsID | Gene | Alleles | HapMap 3 AF | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | CEU | CHB | JPT | YRI |
| 1 | 179520506 | rs1410592 | *NPHS2* | A/C | 0.59 | 0.62 | 0.54 | 0.53 |
| 1 | 67861520 | rs2229546 | *IL12RB2* | A/G | 0.64 | 0.36 | 0.44 | 0.58 |
| 2 | 169789016 | rs497692 | *ABCB11* | A/G[b] | 0.55 | 0.65 | 0.51 | 0.22 |
| 2 | 227896976 | rs10203363 | *COL4A4* | C/T | 0.46 | 0.44 | 0.36 | 0.57 |
| 3 | 4403767 | rs2819561 | *SUMF1* | A/G[b] | 0.56 | 0.73 | 0.73 | 0.72 |
| 4 | 5749904 | rs4688963 | *EVC* | A/G[b] | 0.33 | 0.65 | 0.67 | 0.52 |
| 5 | 82834630 | rs309557 | *VCAN* | A/G[b] | 0.49 | 0.34 | 0.52 | 0.50 |
| 6 | 146755140 | rs2942 | *GRM1* | C/T | 0.54 | 0.49 | 0.55 | 0.47 |
| 7 | 48450157 | rs17548783 | *ABCA13* | C/T | 0.46 | 0.72 | 0.53 | 0.48 |
| 8 | 94935937 | rs4735258 | *PDP1* | C/T | 0.40 | 0.64 | 0.66 | 0.46 |
| 9 | 100190780 | rs1381532 | *TDRD7* | A/G[b] | 0.48 | 0.59 | 0.50 | 0.58 |
| 10 | 100219314 | rs10883099 | *HPSE2* | A/G | 0.52 | 0.52 | 0.53 | 0.62 |
| 11 | 16133413 | rs4617548 | *SOX6* | C/T | 0.52 | 0.65 | 0.61 | 0.51 |
| 12 | 993930 | rs7300444 | *WNK1* | A/G | 0.46 | 0.55 | 0.48 | 0.28 |
| 13 | 39433606 | rs9532292 | *FREM2* | A/G | 0.29 | 0.41 | 0.44 | 0.54 |
| 14 | 50769717 | rs2297995 | *L2HGDH* | A/G | 0.55 | 0.65 | 0.67 | 0.59 |
| 15 | 34528948 | rs4577050 | *SLC12A6* | C/T | 0.68 | 0.75 | 0.63 | 0.32 |
| 16 | 70303580 | rs2070203 | *AARS* | A/G[b] | 0.53 | 0.28 | 0.51 | 0.49 |
| 17 | 71197748 | rs1037256 | *COG1* | C/T | 0.50 | 0.67 | 0.65 | 0.56 |
| 18 | 21413869 | rs9962023 | *LAMA3* | A/G | 0.67 | 0.81[c] | 0.75 | 0.51 |
| 19 | 10267077 | rs2228611 | *DNMT1* | C/T[b] | 0.47 | 0.73 | 0.56 | 0.48 |
| 20 | 6100088 | rs10373 | *FERMT1* | G/T[b] | 0.54 | 0.31 | 0.35 | 0.58 |
| 21 | 44323590 | rs4148973 | *NDUFV3* | C/T | 0.65 | 0.33 | 0.38 | 0.73 |
| 22 | 21141300 | rs4675 | *SERPIND1* | A/C | 0.46 | 0.62 | 0.51 | 0.57 |

[a]Position as defined in genome reference assembly GRCh37 (hg19).
[b]SNP is defined on the negative strand.
[c]AF marginally outside target range for candidate selection. Selected due to paucity of candidates on chromosome 18.

18.0 (standard deviation = 3.3) allelic differences between all pairwise combinations was observed, with a range of 3 to 34. As such, there will be, on average, 18 differential alleles between any two samples, enabling discrimination.

On addition of the UK10K data (n = 2,688) to the 1000 Genomes Project data ($n_{combined}$ = 3,780), there remained an average of 17.8 allele mismatches across the profiles. Eighteen UK10K sample pairs produced duplicate profiles. On investigation of these pairs, they were found to share >98% genotypic concordance across an extended panel of 1,662 SNPs in all cases, compared to an average of 42%, with a range of 27 to 77%, for all sample pairs with unique SNP profiles (Additional file 2). As such, these pairs represent extreme outliers, and are derived from genetically identical biological samples, either from the same individual or monozygotic twins, and were therefore excluded from the mismatch average.

### Simulated data

The discriminatory power of the panel was evaluated by dataset simulation. We simulated datasets of 10,000 individuals, that conformed to AF distributions for investigated HapMap populations (CEU, CHB, JPT and YRI),

**Table 2 Profile collisions per simulated dataset of 10,000 individuals with various population AFs**

| AF source | Average collisions per dataset (±SD) |
|---|---|
| **1000 Genomes average** | 0.0039 (0.062) |
| **HapMap phase 3:** | |
| CEU | 0.0064 (0.079) |
| CHB | 0.0239 (0.154) |
| JPT | 0.0082 (0.090) |
| YRI | 0.0076 (0.086) |
| **Theoretical perfect[a]** | 0.0031 (0.056) |

[a]All 24 SNPs assigned an AF of 0.5, which will give the most even trifurcation per SNP, and thus discriminatory power. SD, standard deviation.

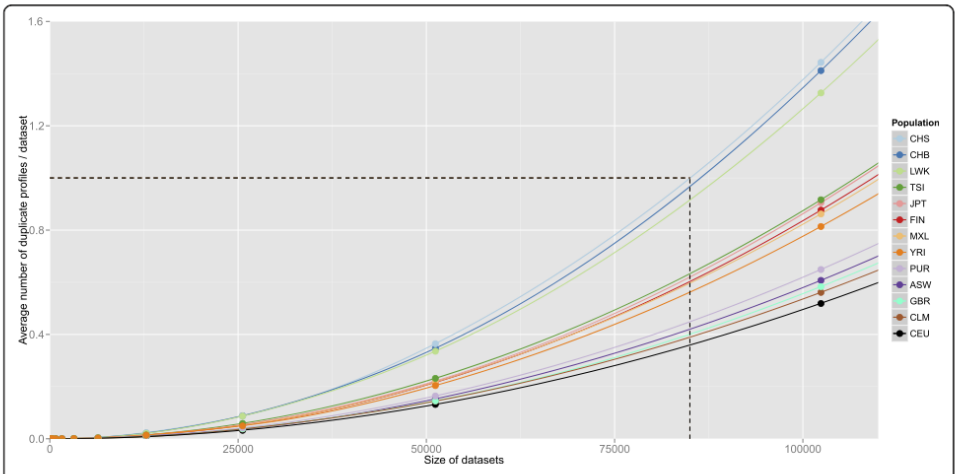



**Figure 2 Relationship between size of simulated datasets and the occurrence of non-unique profiles.** Thirteen 1000 Genomes Project populations were simulated [20]. Datasets were simulated as described in Methods. With increasing dataset size, the probability of repeat profiles increases. Only populations with a sample size of >50 individuals in the dataset were simulated. Additional populations are Americans of African ancestry in Southwest USA (ASW), Columbians from Medellin, Colombia (CLM), Finnish in Finland (FIN), British in England and Scotland (GBR), Luhya in Webuye, Kenya (LWK), Mexican ancestry from Los Angeles, USA (MXL), Puerto Ricans from Puerto Rico (PUR) and Toscany in Italia (TSI).

1000 Genomes Project pilot average [25], as well as for a hypothetical perfect allele distribution (AF = 0.5 for all SNPs) (Table 2). In all simulated populations, <2.5% of simulated datasets of 10,000 contained any repeat SNP profiles (henceforth termed 'collisions'). This translates approximately into less than 1 in every 40 independent datasets of 10,000 individuals containing a single matching pair of profiles.

The effect of dataset size on the frequency of collisions was investigated for populations present in 1000 Genomes Project phase 1 data [20]. An exponential increase in the frequency of collisions was observed with increasing dataset size, though the panel continued to have high power for the discrimination of samples. For instance, were we to have 85,000 Southern Han Chinese (CHS) samples, (the worst performing 1000 Genomes population evaluated, due to the AF distribution for SNPs within this panel), we would expect the dataset to contain, on average, a single duplicate SNP profile (Figure 2). In addition, total SNP absence - for example, through technical failure of orthogonal genotyping - was modelled. For each SNP that entirely failed to provide data, a less than three-fold drop in discriminatory power was observed in all cases (data not shown). This suggests that our approach is robust against technical failure.

Application of the SNP panel to our batch of 48 samples revealed a discrepancy between exome and orthogonal genotypes for two samples dispatched in adjacent wells, suggesting a reciprocal transposition (Figure 3). The occurrence of this error in the exome data was also

| Sample | | rs1410592 | rs2229546 | rs497692 | rs10203363 | rs2819561 | rs4688963 | rs309557 | rs2942 | rs17588783 | rs4705288 | rs1381532 | rs10883099 | rs4617548 | rs7300444 | rs9532292 | rs2297995 | rs4577080 | rs2070203 | rs1187256 | rs9962023 | rs2228611 | rs10373 | rs4148973 | rs4675 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Exome | A | A | G | A | C | C | C | T | G | G | T | T | T | C | G | A | T | C | C | T | A | A | A | A | A | G | C | T | G | A | G | A | A | A | G | A | G | G | T | C | C | C | G | G | T | G | T | T |
| | Geno | N | N | G | A | C | C | C | T | G | G | T | T | T | C | G | A | T | C | C | T | A | A | A | A | A | G | C | T | G | A | G | A | A | A | G | A | G | G | T | C | C | C | G | G | T | G | T | T |
| 2 | Exome | C | A | A | A | C | C | C | T | A | G | T | T | T | C | G | A | T | C | C | C | A | A | G | G | G | G | C | T | A | A | G | G | G | A | G | A | G | G | C | C | C | C | G | G | G | G | T | T |
| | Geno | C | A | G | A | T | C | C | T | G | G | T | T | T | C | A | A | C | C | C | C | G | G | A | G | A | G | C | C | A | A | A | A | G | A | A | G | A | C | C | T | C | A | G | G | G | C | T |
| 3 | Exome | C | A | G | A | T | C | C | T | G | G | T | T | T | C | A | A | C | C | C | C | G | G | A | G | A | G | C | C | A | A | A | A | G | A | A | A | G | A | C | C | T | C | A | G | G | G | C | T |
| | Geno | C | A | A | A | C | C | C | T | A | G | T | T | T | C | G | A | T | C | C | C | A | A | G | G | G | G | C | T | A | A | G | G | G | A | G | A | G | G | C | C | C | C | G | G | G | G | T | T |
| 4 | Exome | C | A | A | A | T | T | C | T | G | G | T | T | T | C | G | A | T | C | C | T | A | A | G | G | A | A | C | T | A | A | A | A | A | A | G | A | G | A | C | C | T | C | A | A | G | G | C | T |
| | Geno | C | A | A | A | T | T | C | T | G | G | T | T | T | C | G | A | T | C | C | T | A | A | G | G | A | A | C | T | A | A | A | A | A | A | G | A | G | A | C | C | T | C | A | A | G | G | C | T |

**Figure 3 Exome derived and orthogonal genotypes (Geno) for four samples, showing a sample-switch between samples 2 and 3.** Informative markers for the resolution of this switch are highlighted in yellow.

supported by interrogation of X-chromosome heterozygosity to confirm sample gender. In addition to the identification of the switch, the panel allowed for expeditious resolution of the error, permitting the continued use of the data in downstream analyses.

## Discussion

Validation of sample identity is essential in order to ensure data integrity and validity of conclusions drawn from data. We have described a powerful tool for the identification and validation of data provenance throughout the workflow of WES data collection and analysis. The power of discrimination, that is, the precision with which samples can be uniquely identifiable, is sufficient and robust for most projects on the current scale of up to 10,000 samples, with inbuilt redundancy of SNPs to protect against technical failures. In WES, the exome enrichment process provides the limiting step for the availability of data on SNPs for use in sample identification. As such, this panel will also be of utility for whole-genome sequencing data, where there is no such limitation on SNP coverage. This will be beneficial where there are mixed datasets of both whole-genome sequence and WES data.

NGS is now developing as the diagnostic methodology of choice across a range of applications, including mutation scanning in targeted gene panels and WES for congenital disorders, as well as high depth analysis for tumour profiling. Whilst the service model for delivery of these tests is not fully resolved at this stage, there will certainly be economic arguments for centralising certain tests. This will have the effect of increasing the throughput requirements as well as physically moving samples between labs. Both of these factors will increase the opportunity for sample misidentification.

Even for testing within a single lab, the use of inherent sample and data identification methods, as described in this study, seems a robust approach to fulfil the regulatory requirement for providing a full audit trail and ensuring data provenance [26,27]. The SNP panel presented here is immediately usable across all commonly used exome capture kits, and would be equally applicable to any gene panel by incorporating, or 'spiking', the SNP regions into the custom capture kit at the design stage. Where it can be shown that there are no expected repeat profiles (that is, no paired samples from the same individual are being analysed), it may even be beneficial from a process perspective to use the SNP profile as the primary method for sample tracking.

The discriminatory power of the panel may be reduced for various reasons, such as geographically localised variation in AFs, and degradation of DNA samples, resulting in incomplete data. We have shown our panel to have a high discriminatory power across a diverse range of populations. Additionally, the discriminatory power will be marginally reduced where many relatives are sequenced. In the case of highly consanguineous families, sample tracking methods such as barcoding will afford optimal certainty in these particular cases. Should concerns over insufficient discriminatory power arise, additional SNPs may be added to the panel from the existing list of candidates (Additional file 1), also allowing the tailoring of an enhanced panel to the population(s) of interest, should this be desired. Nevertheless, we have demonstrated our panel to be sufficiently robust to withstand power reductions without loss of utility for most purposes.

We have also presented a recent case in which use of this panel has allowed us to identify, confirm, and resolve a sample switch, highlighting the importance of using such a tool. Monetary cost will vary with the technology used for orthogonal genotyping and sample throughput. We have intentionally designed the panel to be platform non-specific, allowing for the establishment of in-house assays using preferred genotyping methodology or outsourced where required. Our own chosen methodology costs approximately £5 GBP per sample, representing a small fraction of the cost of exome data generation.

## Conclusions

The size of held NGS datasets continues to increase, with the UK Government recently committing to the sequencing of 100,000 samples as part of healthcare provisions [28]. As such, the demand for the development of effective tools for bioinformatic analysis, data compression, mutation effect prediction and quality control is high. We have described a panel of SNPs for the discrimination of human biological samples on the basis of data intrinsic to WES data derived from samples processed using common capture kits. We recommend the routine use of this panel to maintain data integrity and protect sample provenance.

## Additional files

**Additional file 1: List of all candidate SNPs with evaluated properties.**

**Additional file 2: Distribution of pairwise genotype concordance between samples.** Pairs resulting in duplicate SNP profiles (n = 18) and pairs between samples with unique SNP profiles (n = 7,142,293) within the combined dataset of 3,780 samples are shown. Concordance across the 1,662 SNPs detailed in Figure 1C was evaluated. All pairs resulting in duplicate profiles have >98% concordance, well separated from the distribution of samples with unique profiles. Note the logarithmic scale.

### Abbreviations

AF: alternative-allele frequency; bp: base pair; CEU: CEPH (Utah residents with ancestry from northern and western Europe); CHB: Han Chinese in Beijing, China; CHS: Southern Han Chinese; JPT: Japanese in Tokyo, Japan; Mbp: megabase pair; NGS: next-generation sequencing; SNP: single nucleotide polymorphism; WES: whole-exome sequencing; YRI: Yoruba in Ibidan, Nigeria.

**Competing interests**
The authors declare that they have no competing interests.

**Authors' contributions**
RJP performed analysis and interpretation of data, and drafted the manuscript, JG, GA and AC contributed to analysis, CJM contributed to data interpretation and manuscript preparation and SE conceived and supervised the project, and contributed to manuscript preparation. All authors read and approved the final manuscript.

**Acknowledgements**
The authors thank the Technology Strategy Board and the University of Southampton for funding, and Dr Dietrich Lueerssen for discussion. The authors also acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work. This study makes use of data generated by the UK10K Consortium. A full list of the investigators who contributed to the generation of the data is available from www.UK10K.org. Funding for UK10K was provided by the Wellcome Trust under award WT091310.

**Author details**
[1]Human Genetics and Genomic Medicine, Faculty of Medicine, University of Southampton, Duthie Building (MP 808), Southampton General Hospital, Tremona Road, Southampton SO16 6YD, UK. [2]National Genetics Reference Laboratory (Wessex), Salisbury District Hospital, Salisbury SP2 8BJ, UK.

**Received: 22 July 2013 Accepted: 16 September 2013**
**Published: 27 September 2013**

**References**

1. Need AC, Shashi V, Hitomi Y, Schoch K, Shianna KV, McDonald MT, Meisler MH, Goldstein DB: **Clinical application of exome sequencing in undiagnosed genetic conditions.** *J Med Genet* 2012, **49:**353–361.
2. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J: **Exome sequencing as a tool for Mendelian disease gene discovery.** *Nat Rev Genet* 2011, **12:**745–755.
3. Westra H-J, Jansen RC, Fehrmann RSN, te Meerman GJ, van Heel D, Wijmenga C, Franke L: **MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects.** *Bioinformatics* 2011, **27:**2104–2111.
4. Lam CW, Jacob E: **Implementing a laboratory automation system: experience of a large clinical laboratory.** *J Lab Autom* 2012, **17:**16–23.
5. Pakstis AJ, Speed WC, Fang R, Hyland FC, Furtado MR, Kidd JR, Kidd KK: **SNPs for a universal individual identification panel.** *Hum Genet* 2010, **127:**315–324.
6. Zietkiewicz E, Witt M, Daca P, Zebracka-Gala J, Goniewicz M, Jarzab B, Witt M: **Current genetic methodologies in the identification of disaster victims and in forensic analysis.** *J Appl Genet* 2012, **53:**41–60.
7. Freire-Aradas A, Fondevila M, Kriegel AK, Phillips C, Gill P, Prieto L, Schneider PM, Carracedo A, Lareu MV: **A new SNP assay for identification of highly degraded human DNA.** *Forensic Sci Int Genet* 2012, **6:**341–349.
8. Highnam G, Franck C, Martin A, Stephens C, Puthige A, Mittelman D: **Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles.** *Nucleic Acids Res* 2013, **41:**e32.
9. Castro F, Dirks WG, Fähnrich S, Hotz-Wagenblatt A, Pawlita M, Schmitt M: **High-throughput SNP-based authentication of human cell lines.** *Int J Cancer* 2013, **132:**308–314.
10. Xu W, Gao H, Seok J, Wilhelmy J, Mindrinos MN, Davis RW, Xiao W: **Coding SNPs as intrinsic markers for sample tracking in large-scale transcriptome studies.** *Biotechniques* 2012, **52:**386–388.
11. Jun G, Flickinger M, Hetrick KN, Romm JM, Doheny KF, Abecasis GR, Boehnke M, Kang HM: **Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data.** *Am J Hum Genet* 2012, **91:**839–848.
12. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation.** *Nucleic Acids Res* 2001, **29:**308–311.
13. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics* 2010, **26:**841–842.
14. *Repeat Masker.* http://www.repeatmasker.org/.
15. International HapMap Consortium: **Integrating common and rare genetic variation in diverse human populations.** *Nature* 2010, **467:**52–58.
16. *OMIM - Online Mendelian Inheritance in Man.* http://www.omim.org/.
17. Kent WJ: **BLAT–The BLAST-Like Alignment Tool.** *Genome Res* 2002, **12:**656–664.
18. Christodoulou K, Wiskin AE, Gibson J, Tapper W, Willis C, Afzal NA, Upstill-Goddard R, Holloway JW, Simpson MA, Beattie RM, Collins A, Ennis S: **Next generation exome sequencing of paediatric inflammatory bowel disease patients identifies rare and novel variants in candidate genes.** *Gut* 2013, **62:**977–984.
19. Gibson J, Tapper W, Ennis S, Collins A: **Exome-based linkage disequilibrium maps of individual genes: functional clustering and relationship to disease.** *Hum Genet* 2013, **132:**233–243.
20. 1000 Genomes Project Consortium: **An integrated map of genetic variation from 1,092 human genomes.** *Nature* 2012, **491:**56–65.
21. *UK10K Study Samples.* http://www.uk10k.org/studies/.
22. Li H: **Tabix: fast retrieval of sequence features from generic TAB-delimited files.** *Bioinformatics* 2011, **27:**718–719.
23. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods.** *Mol Biol Evol* 2011, **28:**2731–2739.
24. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet* 2007, **81:**559–575.
25. 1000 Genomes Project Consortium: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467:**1061–1073.
26. International Organization for Standardization: *Medical Laboratories - Requirements for Quality and Competence*; 2012. ISO 15189:2012.
27. Rehm HL, Bale SJ, Bayrak-Toydemir P, Berg JS, Brown KK, Deignan JL, Friez MJ, Funke BH, Hegde MR, Lyon E: **ACMG clinical laboratory standards for next-generation sequencing.** *Genet Med* 2013, **15:**733–747.
28. *DNA tests to revolutionise fight against cancer and help 100,000 NHS patients.* https://www.gov.uk/government/news/dna-tests-to-revolutionise-fight-against-cancer-and-help-100000-nhs-patients.

doi:10.1186/gm492
**Cite this article as:** Pengelly *et al.*: A SNP profiling panel for sample tracking in whole-exome sequencing studies. *Genome Medicine* 2013 **5**:89.