University of Southampton

Faculty of Medicine

Human Development and Health

Integration of genomic variation, ileal transcriptomics and longitudinal clinical data in paediatric inflammatory bowel disease

Volume 1 of 1

by

James J Ashton, BMedSci (1st class), BMBS (Hons), MRCPCH

ORCID 0000-0003-0348-8198

Thesis for the degree of Doctor of Philosophy

Funded by an Action Medical Research, research training fellowship

October 2020

University of Southampton

Abstract

Faculty of Medicine

Human Development and Health

Thesis for the degree of Doctor of Philosophy

Integration of genomic variation, ileal transcriptomics and longitudinal clinical data in paediatric inflammatory bowel disease

by

James J Ashton

Inflammatory bowel disease (IBD) is a chronic relapsing and remitting condition characterised by intestinal inflammation. IBD is considered a complex condition, arising from interaction between host genetic susceptibility, immune dysregulation and environmental stimuli in the form of intestinal bacteria. Paediatric onset disease is heterogenous and patients often follow an unpredictable severe course requiring immunosuppression, biological therapy or surgery. Utilising multi-omic and clinical data to predict disease course and complicated phenotypes, such as stricturing and fistulating disease, has the potential to provide a route to personalised therapy and improved patient outcomes.

This thesis describes the integration of genomic, targeted terminal ileal transcriptomics and longitudinal clinical data in paediatric patients with IBD. These chapters utilise a cohort of 96 individuals with intestinal biopsies including treatment naïve patients, established disease patients and control recruited during this PhD, alongside data derived from 501 patients with whole exome sequencing. This thesis identifies novel groupings of patients determined by blood results at diagnosis. We identify a precise molecular diagnosis in 8% of patients through interrogation of variation in monogenic IBD genes, and directly link monogenic *NOD2*-disease to a stricturing phenotype. Utilising the whole gene deleteriousness score, GenePy, we provide evidence for a digenic risk of development of fistulating disease in small subset of patients with

high burden of variation in *NCF4* and *NOD2,* or in the *NOX4* NAPDH complex. Through targeted autoimmune transcriptomic analysis of terminal ileal biopsies we identify an upregulation of IL17 and NOD-signalling genes in treatment naïve Crohn's disease patients. Through single cell sequencing of two individuals we determine a small population of specialised ileal epithelial cells driving the IL17-signalling signature. Finally, through integration of exome and transcriptomic data we identify variation across the NOD-signalling pathway, including in *NOD2*, *ATG16L1* and the TAK1-TAB complex that directly impacts on ileal transcription, leading to an overall hypoinflammatory response.

It is increasingly clear that IBD pathogenesis is private to an individual, or a cluster of individuals. Whilst activated inflammatory pathways demonstrate commonality, the underlying cause of disease may relate to deleterious variation in many genes across implicated pathways. This genetic variation leads to an inability to clear bacteria, resulting in chronic activated inflammation through alternative immune pathways. Genetic stratification also provides the ability to predict phenotypes. The next step is translating these findings into clinical practice.

Table of Contents

Table of Tables

Table of Figures

Table of Figures

List of Accompanying Materials

All accompanying materials can be found at the following DOI-

https://doi.org/10.5258/SOTON/D1657

Supplementary data

- Microbiome
  - Quality control
  - Taxonomy
  - Phylogenetic trees
- Chapter 3
  - All blood results
- Chapter 4
  - Data available for all patients
  - Segregation analysis
  - Monogenic genes in analysis
  - Functional evidence table
  - Correct zygosity variants
  - NOD2 variants
  - Compound heterozygous variants
  - Ileal disease correction analysis
  - Surgical resection analysis
- Chapter 5
  - Raw data NADPH oxidase complex genes CADD 1.6
  - Regression analysis
- Chapter 6
  - RNA extraction quality control
  - Genes in WGCNA modules
  - Module 3 pathways
  - DEGs controls vs patients
  - DEGs ED patients vs controls
  - WGCNA gene module membership
  - Blue module enrichment analysis
- Chapter 7
  - Autoimmune gene list- LOEUF correction
  - Annotated NOD pathway VCF
  - Genomic and RNA integration raw data
  - Module hub genes variation

Standard operating procedures

Collection of samples-

1. Biopsy samples
2. Blood collection

3. Faecal collection
4. Plasma collection

Storage of samples

1. Biopsy storage
2. Faecal sample storage

Processing-

1. Biopsy processing
    a. Bacterial DNA
    b. Human RNA
2. DNA
    a. From blood
    b. From saliva
3. Plasma
    a. From whole blood
4. Faecal samples
    a. Bacterial DNA
    b. Faecal calprotectin

Genetics of PIBD study paperwork

1. Consent forms
2. Letters for results
3. Patient information sheets
4. Data collection proformas
5. Study protocol

Research Thesis: Declaration of Authorship

Print name: **JAMES J ASHTON**

Title of thesis: **Integration of genomic variation, ileal transcriptomics and longitudinal clinical data in paediatric inflammatory bowel disease**

I 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.

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. Parts of this work have been published as:

   Ashton JJ, Boukas K, Davies JD, Stafford ISS, Vallejo AF, Haggarty R, et al Ileal transcriptomic analysis in paediatric Crohn's disease reveals IL17- and NOD-signalling expression signatures in treatment-naïve patients and identifies epithelial cells driving differentially expressed genes, Journal of Crohn's and Colitis, Nov 2020, ePub ahead of print

   Ashton JJ, Mossotto E, Stafford IS, Haggarty R, Coelho TAF, Batra A, et al. Genetic Sequencing of Pediatric Patients Identifies Mutations in Monogenic Inflammatory Bowel Disease Genes

that Translate to Distinct Clinical Phenotypes. Clin Transl Gastroenterol. 2020;11:e00129. doi:10.14309/ctg.0000000000000129.

Ashton JJ, Mossotto E, Beattie RM, Ennis S. TTC7A Variants Previously Described to Cause Enteropathy Are Observed on a Single Haplotype and Appear Non-pathogenic in Pediatric Inflammatory Bowel Disease Patients. Journal of Clinical Immunology. 2020;40:245–7. doi:10.1007/s10875-019-00726-0.

Ashton JJ, Latham K, Beattie RM, Ennis S. Review article: the genetics of the human leucocyte antigen region in inflammatory bowel disease. Aliment Pharmacol Ther. 2019;50:885–900. doi:10.1111/apt.15485.

Mossotto E, **Ashton** JJ, O'Gorman L, Pengelly RJ, Beattie RM, MacArthur BD, et al. GenePy - a score for estimating gene pathogenicity in individuals using next-generation sequencing data. BMC Bioinformatics. 2019;20:254. doi:10.1186/s12859-019-2877-3.

Ashton JJ, Borca F, Mossotto E, Coelho T, Batra A, Afzal NA, et al. Increased prevalence of anti-TNF therapy in paediatric inflammatory bowel disease is associated with a decline in surgical resections during childhood. Aliment Pharmacol Ther 2019;**49**(4):398–407. doi: 10.1111/apt.15094.

Ashton JJ, Borca F, Mossotto E, Phan HTT, Ennis S, Beattie RM. Analysis and Hierarchical Clustering of Blood Results Before Diagnosis in Pediatric Inflammatory Bowel Disease. Inflamm Bowel Dis. 2020; 26(3):469-475. doi:10.1093/ibd/izy369.

Signature: Date:

Acknowledgements

My supervisory team of Prof. Sarah Ennis, Prof. Mark Beattie, Dr. Marta Polak and Dr. David Cleary.

My colleagues working on the Genetics of PIBD study, Imogen Stafford, James Davies and Dr. Enrico Mossotto.

The research nursing team who have contributed to patient recruitment and data collection; Rachel Haggarty, Rachel Brampton, Genevieve Roberts and Gabby Price.

The laboratory teams who have aided with sample processing and extraction including Nikki Graham and Konstantinos Boukas.

The Southampton BRC data science team of Florina Borca and Hang Phan who have helped with clinical data extraction.

Collaborating researchers including Muise lab (Toronto) and Uhlig lab (Oxford).

The UHS clinicians and specialist nurses who have facilitated patient recruitment and sample acquisition; Dr. Akshay Batra, Dr. Tracy Coelho, Dr. Nadeem Afzal, Mick Cullen, Claire Barnes, Rachel Russell and Jo Himsworth. Specifically, I would like to thank Dr. Bhumita Vadgama for help with histological analysis.

Commonly used abbreviations

5-ASA- 5-aminosalycylic acid

ACMG- American College of Medical Genetics

AI- autoimmune

ALT- alanine transaminase

BSPGHAN- British society of paediatric gastroenterology, hepatology and nutrition

CD- Crohn’s disease

CDED- Crohn's disease exclusion diet

CGD- chronic granulomatous disease

CLR- C-type lectin receptors

CPM- counts per million

CRP- C-reactive protein

CT- computorised tomography

DNA- deoxyribonucleic acid

EBV- epstein barr virus

ECCO- European crohn's + colitis organisation

ED- established disease

EEN- exclusive enteral nutrition

ESPGHAN- European society of paediatric gastroenterology, hepatology and nutrition

ESR- erythrocyte sedimentation rate

EWAS- epigenome-wide association studies

FBC- full blood count

FCp- faecal calprotectin

gVCF- genomic variant call file

GWAS- genome-wide association study

Hb- haemoglobin

HGMD- human genetic mutation database

HLA- human leucocyte antigen

IBD- inflammatory bowel disease

Commonly used abbreviations

IBDU- inflammatory bowel disease unclassified

IL- interleukin

INF- interferon

LRR- leucine-rich repeat receptors

MN- median normalisation

MRI- magnetic resonance imaging

NGS- next generation sequencing

NOD- nucleotide-binding oligomerization domain

OUT- operational taxonomic unit

PAMP- pathogen-associated molecular protein

PCR- polymerase chain reaction

PCV- packed cell volume

PIBD- paediatric inflammatory bowel disease

QN- quantile normalisation

RNA- ribonucleic acid

ROS- reactive oxygen species

SCID- severe combined immunodeficiency

SDS- standard deviation score

SNP- single nucleotide polymorphism

Th- T-helper cell

TLR- toll-like receptors

TN- treatment naïve

TNF- tumour necrosis factor

TPMT- Thiopurine S-methyltransferase

UC- ulcerative colitis

UHS- university hospital Southampton

VCF- variant call file

WES- whole exome sequencing

WGS- whole genome sequencing

Chapter 1 Background and introduction

Chapter summary- *This chapter provides a background on clinical inflammatory bowel disease, alongside the role of genomics, transcriptomics and the microbiome in IBD pathogenesis. The concept of personalised medicine based on a precise diagnosis is introduced as the overall theme of the thesis.*

1.1 Background

Paediatric inflammatory bowel disease (PIBD) is a chronic, relapsing and remitting condition characterised by intestinal inflammation leading to abdominal pain, diarrhoea, bloody stools and a range of other intestinal and extra-intestinal symptoms and complications[1]. PIBD is comprised of Crohn's disease (CD), ulcerative colitis (UC) and inflammatory bowel disease unclassified (IBDU) and is defined as IBD presenting before the 18th birthday.

Around 20-25% of IBD cases will present during childhood, offering a distinct set of issues, management challenges and complications when compared to adult-onset disease[2,3]. Characteristic of paediatric-onset disease is the relatively severe presenting phenotype, additional challenges of growth (including nutrition) and, often, the rapid progression of disease requiring careful immunosuppression to avoid complications and maintain remission[3,4].

Over the last 25 years huge progress has been made in diagnosis and management in PIBD, despite the challenges to patients, families and healthcare professionals, outcomes have improved and there is a strong emphasis on evidence-based medicine with research providing scientific and clinical progress[5,6]. However there is much that remains poorly understood,

including the exact disease pathogenesis, predictors of disease progression/response to medication and how best to manage patients with the current medications/surgeries available to us[7].

1.2 Disease incidence, prevalence, age of onset

The incidence of PIBD is increasing, over the last 20 years multiple studies from around the world have detailed the rise in the numbers of children being diagnosed with IBD[8–12]. The main driver behind this increase is higher numbers of CD cases, although UC has also increased during this period. Since 1999 the incidence of PIBD has increased by approximately 50% within the United Kingdom, with the incidence now standing at over 10/100,000/year locally in Wessex, similar to that seen in Canada 9.68/100,000/year[8,13,14]. When considering the highest worldwide incidence of IBD in children a study from Finland in 2017 has estimated incidence at up to 23/100,000/year[15]. Local data from Wessex can be seen in figure 1.

The median age of disease onset is between 11-14 years, but some children will present much earlier[8,11]. Very-early onset disease (<6 years) is rare and is a trigger for further investigation for a monogenic cause of IBD [16,17]. Early-onset (6-10 years) and paediatric (10-17 years) IBD is less likely to represent single gene defects but present their own distinct challenges in diagnosis and management, particularly related to puberty and growth[6].

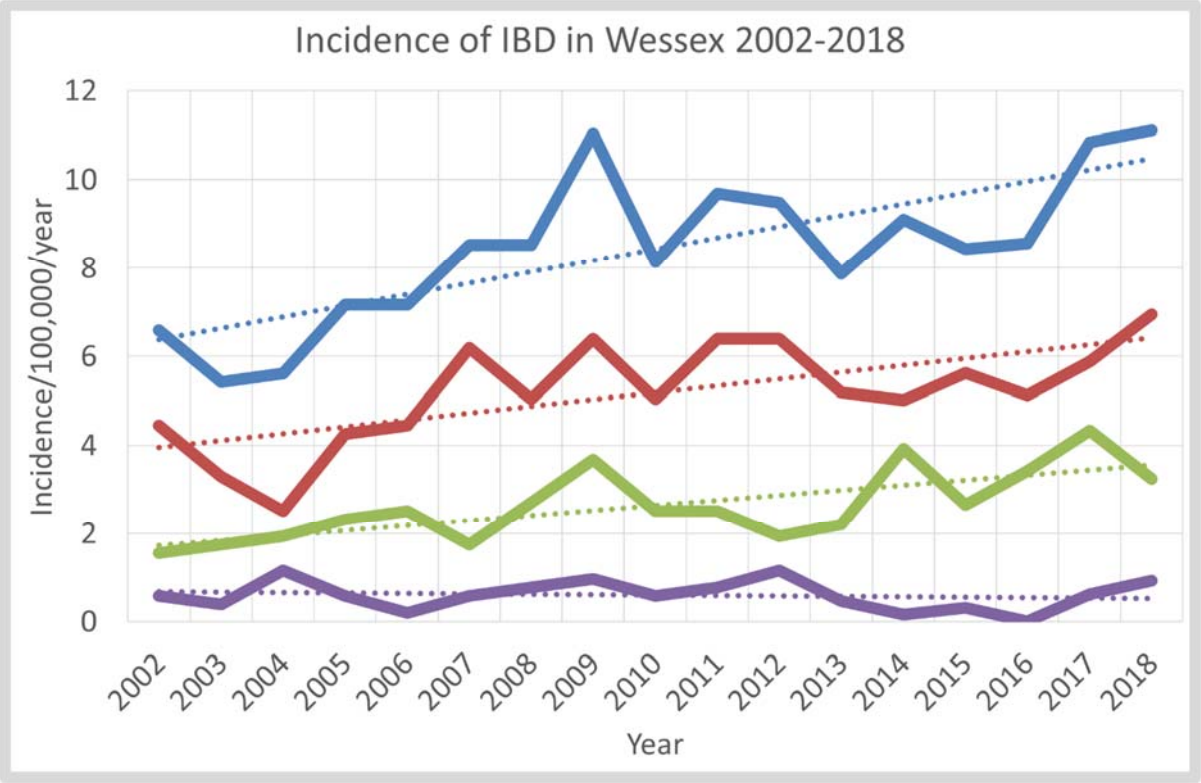

Figure 1- Incidence of paediatric inflammatory bowel disease in Wessex over a 17-year period (2002-2018). Data taken from Ashton JJ et al 2014 8. All PIBD, blue (R2=0.464, p=0.004), CD, red (R2=0.314, p=0.024), UC, green (R2=0.490, p=0.003), IBDU, purple (R2=0.103, p=0.224)

1.3 Clinical features and diagnosis

Data from the last 40 years has consistently reported abdominal pain (>85%), diarrhoea (>75%) and weight loss (>55%) to be the most common presenting features of CD, whilst abdominal pain (>85%), bleeding per rectum (>90%) and diarrhoea (>90%) were the most common in UC[18,19]. The paediatric IBD phenotype is characterised by extensive intestinal (endoscopic and histological) disease alongside rapid disease progression[20–22]. The absence of typical features should not preclude further investigation and patients may present with all, some or none of the common symptoms. Family history of IBD is a common feature (15-25%) and should trigger a lower threshold for referral and investigation[18,23]. Diagnosis of another gastrointestinal disease such as coeliac disease or functional abdominal pain does not preclude a diagnosis of IBD[24].

It is important to consider other presentations of IBD including extraintestinal manifestations (including arthritis/arthropathy, extra-intestinal Crohn's- genital/isolated orofacial, dermatological- erythema nodosum/pyoderma gangrenosum and eye disease- uveitis), growth failure (up to 20% of Crohn's disease cases) and isolated perianal disease in Crohn's[3,5,6,18].

The Porto criteria, and subsequently the modified Porto criteria, have been the diagnostic standard for PIBD since 2005[2,25]. These publications detail the conditions that must be met for diagnosis to be made, including specific histological features.

1.4 Diagnostic testing and classification

PIBD diagnosis requires all patients to undergo upper and lower gastrointestinal endoscopy with histological examination (to include oesophagus, stomach, duodenum, terminal ileum, colonic series and rectum)[2]. This should be under the care and guidance of a specialist paediatric gastroenterologist[26,27]. Small bowel imaging (MRI, CT, contrast study or ultrasound) is recommended in all suspected cases but may be deferred in ulcerative colitis depending on the clinical presentation[2].

The diagnosis of IBD must be confirmed histologically. Features of disease include presence of inflammatory changes in the mucosa (acute or chronic gastritis/duodenitis/oesophagitis, cryptitis, crypt abscesses, and granulomas- in Crohn's disease only), architectural abnormalities (crypt distortion, crypt branching, and crypt atrophy), and epithelial abnormalities such as mucin depletion and metaplasia, alongside surface irregularities (epithelial active/regenerative changes)[2,20]. Differentiation of Crohn's disease and ulcerative colitis can be difficult in some cases and a diagnosis of inflammatory bowel disease unclassified (IBDU) should be made based on published guidance[2,28]. Grouping of disease by site of inflammation is through the Paris classification and is based on endoscopic and radiological disease extent[29]. Due to more extensive histological disease (compared to endoscopic), modification of the classification to incorporate this has been discussed by several groups[20,21,30].

1.5 Management strategies

The treatment of paediatric IBD is discussed in multiple national and international guidelines[26,27,31,32]. It is important to understand treatment options as some patients will fail to respond to some medications and the ability to predict which patients will respond, and who will need treatment escalation forms a key part of personalised therapy.

1.5.1 Induction of remission

Inducing remission requires a treatment choice based on disease severity, location, additional features (perianal disease etc.) and discussion with the child and family.
ECCO/ESPGHAN/BSPGHAN guidelines on treatment of IBD (Crohn's disease and ulcerative colitis) have been published within the last 6 years and provide a framework and guidance for managing disease[31–34].

1.5.1.1 Exclusive enteral nutrition (EEN)

EEN can be divided into polymeric feeds (such as Modulen, Nestle©) or elemental feeds (such as E028, Nutricia©) and is taken in the form of a liquid drink for 6-8 weeks with complete exclusion of all other foods and drinks. EEN is the first line treatment in non-complicated paediatric Crohn's disease, having a response rate of up to 80%, there is no role for EEN in treatment of ulcerative colitis[31,32]. There is currently no role for partial enteral nutrition as an induction agent for Crohn's disease[35,36]. Early and ongoing work on exclusion diets, including the CD-TREAT and CDED diet are potentially encouraging but further work is required to confirm initial findings[37,38].

1.5.1.2 Corticosteroids

In moderate/severe Crohn's disease steroids are commonly used (consider intravenous steroids in severe pan-enteric or severe perianal disease)[32]. Oral steroids are recommended for induction of remission in ulcerative colitis presenting with moderate/severe disease, they have up to a 90%

response rate[31]. Steroids should not be used for maintenance therapy and steroid dependence should trigger escalation of maintenance therapy.

1.5.1.3 Induction in refractory disease

Disease not responding to steroids or EEN can be described as refractory to conventional induction agents. Anti-TNF monoclonal therapy (infliximab, adalimumab) are effective induction agents in Crohn's disease (up to 88% response rate) and in ulcerative colitis (up to 73% response rate) resistant to initial therapy[39,40]. In some centres there is an emerging 'top-down' therapy approach, which began in adult IBD. This is increasingly common in paediatric-onset disease, however the long term impact (good and bad) of the 'top-down' approach are not yet known, currently UK guidance does not recommended this other than in severe disease[26,27,41]. Contemporary guidelines from ECCO and ESPGHAN now recommend up front anti-TNF use in perianal disease, stricturing and penetrating phenotypes, and severe growth retardation[42]. Treating with anti-TNF therapy before complications occur appears to promote improved remission rates but the longer-term remission profile is not yet established.

1.5.2 Maintenance therapy

The aim of maintenance therapy is to prevent relapse of active disease (maintain remission), safely and effectively, with the ultimate goal to achieve mucosal healing (or in Crohn's disease, transmural healing) of the gastrointestinal tract[43,44].

1.5.2.1 5-ASA

Maintenance therapy with 5-ASA is effective in mild/moderate paediatric ulcerative colitis and should be continued long term for the cancer protective effect[31,45]. There is no evidence of efficacy in Crohn's disease[32].

1.5.2.2 Thiopurines

Thiopurines (azathioprine and 6-mercaptopurine) are first or second-line agents for maintaining remission in both Crohn's disease and ulcerative colitis with maintained remission rates of 60-90%[46,47]. Thiopurines can be combined with other agents[31,32]. Prior to starting therapy Thiopurine S-methyltransferase (TPMT) enzymatic activity should be checked, patients with no activity should not be started on thiopurines as they are extremely likely to develop liver or bone marrow toxicity. The typical dose is 2-2.5mg/kg for azathioprine and 1-1.5mg/kg for 6-mercaptopurine, but this can be increased based on drug metabolite blood levels and response. Relatively frequent side effects of thiopurines include liver toxicity, bone marrow suppression and pancreatitis. Regular monitoring of full blood count and liver function is necessary (more frequent during initial use) and it may take 8-14 weeks for response[26].

There is some concern over the safety of thiopurine use, with some studies describing an increased long term malignancy risk (lymphoproliferative disorders, specifically hepatosplenic T-cell lymphoma), and some concerns over EBV infection in naïve patients[48,49]. It is important to remember that the absolute risk of lymphoproliferative disorders remains small whilst the risk of malignancy with uncontrolled inflammation in IBD is increased[50].

1.5.2.3 Anti-TNF (monoclonal antibody) therapy

Use of monoclonal therapy in paediatric IBD, especially in previously refractory disease has revolutionised care, providing steroid-sparing and highly effective therapy leading to prolonged remission, improved growth and mucosal healing [65]. Anti-TNF therapy is recommended in chronic luminal Crohn's disease and fistulating disease[32]. It is also recommended for steroid-dependant ulcerative colitis or disease refractory to immunomodulation[31]. Remission rates are slightly less than in Crohn's disease (38-64%) and long-term use is associated with antibody formation and loss of efficacy[51].

As with thiopurine use there is some concern that long-term anti-TNF therapy increases the risk of malignancy, specifically lymphoproliferative disorders. Whilst some recent data appears to contradict this any absolute risk is very small, although the risks and benefits for use must be considerd[48].

Other treatment options

In refractory disease several other agents such as tacrolimus and thalidomide, and especially methotrexate should be considered[31,32]. There is little safety data in paediatric patients for many additional immunomodulators and risk/benefit must be discussed with the child and family.

1.5.3 Contemporary and future treatments

Use of new therapies in paediatrics lags behind adults, due to safety concerns and drug trial evidence, in addition to complex regulatory requirements. Newer monoclonal therapies such as vedolizumab (an α4β7 integrin antibody, blocking immune cell migration into the intestine) has now shown in excess of 40% efficacy in children for ulcerative colitis, and may also be useful in paediatric onset Crohn's disease[52,53]. Ustekinumab (anti-IL12/IL23) has previously been used in psoriasis but has and is now widely used in adult-onset and paediatric-onset Crohn's disease[54][55]. These medications are now routinely used in children in specific situations, with efficacy and safety data both being very encouraging[55,56].

A newer potential class of medications, 'small-molecule drugs' have been under development and have targets such as JAK (Tofacitinib), S1P (Ozanimod) and anti-inflammatory pathways (Mongersen, Laquinimod), these would potentially provide additional therapeutic options for future treatment of IBD[57]. Tofacitinib is now routinely used in adult practice for ulcerative colitis, but is not yet licenced for use in children[58].

1.6 Disease pathogenesis- genetic and environmental interaction

The exact disease pathogenesis for individuals with PIBD is unclear. In comparison to adult-onset disease there is a larger genetic component and studies have implicated over 230 genes to date, with rare variation in additional genes also playing a role[59,60]. Genes associated with IBD are broadly related to: innate or adaptive immunity (cell recruitment, regulation and immune tolerance, bacterial recognition and response); epithelial barrier function (tight junctions); intracellular downstream signalling (*NFKB, JAK-STAT*); cellular death (apoptotic, autophagy; reactive oxygen species production) and antigen presentation (including dendritic cell activation)[59]. Epigenetic modification of genes has a probable role in disease pathogenesis in some cases, and provide a way for the environment to interact with the genome[61]. Additionally, non-coding regulatory genomic regions are likely to contain areas of interest not yet discovered.

The role of the microbiome is of great interest, with studies reported an altered gut flora (dysbiosis) at diagnosis in IBD, compared to controls[62–64]. It appears that a single bacterial species is not responsible for disease, rather the entire bacterial community, including the functional role of the microbiome, play a more important role in disease development[65]. Further work is required on the interaction of the host (genetic-susceptibility) with environmental factors (such as the intestinal microbiome and nutrition) to provide novel therapeutic strategies and precise molecular diagnosis[66,67].

The role of nutrition, including diet and overall nutritional status (at diagnosis and during recovery), is likely to play a key role in shaping the intestinal microbiome seen in disease. Gene-nutrition interaction has recently been discussed as having a role in cancer pathogenesis, with a potential role in IBD[68,69].

The interaction between genes, the immune system and the microbial environment is important in the development and relapse of inflammation seen in IBD.

1.7 Genomics of IBD

1.7.1 Polygenic IBD

Paediatric-onset IBD appears to have a large genetic component and genome-wide association studies/monogenic IBD studies have implicated over 230 genes to date[60,70,71]. Despite this only around 25% of the heritability is accounted for by these genes and it is presumed rare, private, variation must also play an important role[59]. GWAS studies are limited by the need for variation to be present in many IBD patients, with rare or private variation, unable to be detected by this methodology. Twin studies estimate heritability at 0.75 in CD and 0.67 in UC, compared to 0.37 and 0.27 from GWAS data[72]. There is a 30% concordance in monozygotic for Crohn's disease compared to 13% in ulcerative colitis[72]. Utilising GWAS and single gene studies it we are able to establish that genes associated with IBD are broadly related to a number of pathways:

- Innate or adaptive immunity (cell recruitment, regulation and immune tolerance, bacterial recognition and response)
- Epithelial barrier function (tight junctions)
- Intracellular downstream signalling (*NFKB, JAK-STAT*)
- Cellular death (apoptotic, autophagy; reactive oxygen species production)
- Antigen presentation (including dendritic cell activation)[59]

The most important genes within these pathways can be seen in table 1. The best known IBD risk gene is *NOD2,* conferring increased risk for Crohn's disease only[59,73]. *NOD2* was the first genetic locus identified in IBD and was designated the IBD1 locus through linkage studies in the 1990s[74]. These analyses were focused on pedigrees with early-onset and severe disease and suggested an AR inheritance pattern[75,76]. *NOD2* is required for bacterial recognition and immune stimulation. *NOD2* mutations (including frameshift, non-synonymous and a range of allelic variants, such as Arg702Trp, Gly908Arg, and Leu1007C frame shift insertion) are well documented to be associated with Crohn's disease occurring at any age. Significant progress in

identification of IBD-related genetic variation has been made through next-generation sequencing. Whole exome sequencing, employed within this study, targets all coding parts of the genome, giving sequencing data for exonic areas of genes. Some genes are difficult to accurately sequencing or map to the genome, these reasons include GC rich areas, triplet repeats, presence of pseudogenes and high sequencing homology.

Non-coding DNA and epigenetic modifications (including methylation, histone modification and microRNAs) are other genomic areas associated with inflammatory bowel disease[59,61]. The non-coding regulatory regions of known risk genes are very likely to harbour important genetic variation which have not yet been identified. Epigenetics provides a mechanism whereby the environment can impact on long-term gene expression and cellular function. Initial epigenetic studies in IBD attempted to determine the risk of development ulcerative colitis-associated colorectal carcinoma, with epigenetic changes within the *SLIT2* gene being replicated on a number of occasions[77]. Epigenome-wide association studies (EWAS) have consistently identified differentially methylated DNA between Crohn's disease patients and controls. These sites include inflammatory cytokines such as *CXCL14, CXCL5* and *IFN-Y* but patient numbers are generally small within individual studiesl[77]. Studies have detailed epigenetic changes in both circulating mononuclear cells, alongside intestinal tissue.

Many of the GWAS positive hits for IBD are located in intergenic regions, and may reflect promotor or regulatory areas of the genome[59]. The complex, and multifactorial, nature of IBD makes the interpretation of potentially deleterious variants (including non-synonymous, splicing and stop-gain) in known genes difficult without functional immunological work. Additionally, the role of synonymous variants and less common single-nucleotide polymorphisms (SNPs) is uncertain. Improving the understanding of how subtle genetic changes influence disease is vitally important. An example of where improved understanding of nucleotide changes would be hugely beneficial includes those alterations that do not truncate proteins or impact amino acid

Table 1- Specific IBD risk genes in the context of physiological function sequences

Physiological function	Gene	Gene name	Disease	Normal gene function and associated pathways (underlined)
Innate mucosal defence	***NOD2***	Nucleotide-binding oligomerization domain-containing protein 2	Crohn's disease	Bacterial recognition and response, NFKB activation and autophagy/apoptosis
Immune tolerance	***IL10***	Interleukin 10	Crohn's disease	Anti-inflammatory cytokine, NFKB inhibition, JAK/STAT regulation
	IL10RA	Interleukin 10 receptor A	Crohn's disease	Anti-inflammatory cytokine receptor, NFKB inhibition, JAK/STAT regulation
	Il10RB	Interleukin 10 receptor B	Crohn's disease	Anti-inflammatory cytokine receptor, NFKB inhibition, JAK/STAT regulation
IL-23/T_H17	***IL23R***	Interleukin 23 receptor	Crohn's disease and ulcerative colitis	Immune regulation, pro-inflammatory pathways- JAK/STAT regulation
	TKY2	Tyrosine Kinase 2	Crohn's disease and ulcerative colitis	Inflammatory pathway signalling (IL10, IL6 etc.) through intracellular activity
Autophagy	***IRGM***	Immunity related GTPase M	Crohn's disease	Autophagy/apoptosis in cells infected with bacteria
	ATG16L1	Autophagy related 16 like 1	Crohn's disease	Autophagy/apoptotic pathways
Solute transporters	***SLC22A4***	Solute carrier family 22 member 4	Crohn's disease	Cellular antioxidant transporter
Immune cell recruitment	***CCL2***	C-C motif chemokine ligand 2	Crohn's disease	Cytokine involved in chemotaxis for monocytes
Oxidative stress	***CARD9***	Caspase Recruitment Domain Family Member 9	Crohn's disease and ulcerative colitis	Apoptosis regulation and NFKB pathway activation
T-cell regulation	***IL2***	Interleukin 2	Ulcerative colitis	Cytokine involved in immune cell activation
Epithelial barrier	***MUC19***	Mucin 19	Crohn's disease and ulcerative colitis	Gel-forming mucin protein

(synonymous), improved interpretation of splicing variation and nucleotide changes in intergenic regions that harbour potentially important regulatory switching.

1.7.2 Monogenic causes of disease

Over recent years NGS has identified Mendelian causes of patients presenting with particularly severe IBD-like phenotypes[78,79]. A rare subset of around 100 single-gene conditions may present as IBD, most often in very early childhood[16]. These monogenic conditions should be considered in those presenting with atypical features (such as frequent infections, < age of 6 years, skin manifestations etc.) and patients refractory to conventional treatment (even when older), these patients should be considered for further investigation including genetic testing with next-generation sequencing panels[80]. Identification of these high risk patients is important as many require need specific surveillance (for malignancy, infection etc.) and some may require specific treatments (such as bone marrow transplant)[17]. The clinical features of 73 established causes of monogenic IBD are summarised in table 2.

Table 2- Monogenic inflammatory bowel disease, clinical phenotypes associated with genetic defects for 73 established causes of disease. Adapted from Uhlig et al[16].

Group		Syndrome/disorder	Gene	Inheritance	Intestinal findings								Extraintestinal findings			
					CD-like	Granuloma	UC-like	Epithelial defect (apoptosis)	Disease location (1–5)	Perianal fistula/abscess	Penetrating fistulas	Strictures	Skin lesions	Autoimmunity, inflammation	HLH/MAS	Neoplasia
Epithelial barrier	1	Dystrophic bullosa	*COL7A1*	AR				+	3			+	+			
	2	Kindler syndrome	*FERMT1*	AR			+	+	5			+	+			
	3	X-linked ectodermal immunodeficiency	*IKBKG*	X	+			+	3				+	+		
	4	*TTC7A* deficiency	*TTC7A*	AR				+	3			+				
	5	*ADAM17* deficiency	*ADAM17*	AR			(+)	+	3				+			
	6	Familial diarrhoea	*GUCY2C*	AD	+				3			+				
	7	Congenital secretory sodium diarrhoea	*SLC9A3*	AR	+				3							
	8	Tufting enteropathy	*EPCAM*	AR	+				2							
	9	Enteropathy	*SLCO2A1*	AR	+				4			+				
Phagocyte defects	10	Chronic granulomatous disease (CGD)	*CYBB*	X	+	+			1,3	+			+			
	11	CGD	*CYBA*	AR	+	+			3	+			+			
	12	CGD	*NCF1*	AR	+	+			1,3	+			+			
	13	CGD	*NCF2*	AR	+	+			1,3	+			+			
	14	CGD	*NCF4*	AR	+	+			1,3				+			
	15	Glycogen storage disease type lb	*SLC37A4*	AR	+	+			1,3	+		+	+			

	16	Congenital neutropenia	*G6PC3*	AR	+				1,3	+	?	(+)	+			
	17	Leukocyte adhesion deficiency 1	*ITGB2*	AR	+				1,3	+		+	+			
Hyperinflammatory and autoinflammatory disorders	18	Mevalonate kinase deficiency	*MVK*	AR					3			+	+	+	+	
	19	Phospholipase *C-γ2* defects	*PLCG2*	AD			+		3				+	+		
	20	Familial Mediterranean fever	*MEFV*	AR			+		5				+	+		
	21	Familial hemophagocytic lymphohistiocytosis type 5	*STXBP2*	AR					3							
	22	X-linked lymphoproliferative syndrome 2 (XLP2)	*XIAP*	X	+	+			3	+	+	(+)	+	?	+	
	23	X-linked lymphoproliferative syndrome 1 (XLP1)	*SH2D1A*	X					3						+	+
	24	Hermansky–Pudlak 1	*HPS1*	AR	+	+			3	+		(+)	+			
	25	Hermansky–Pudlak 4	*HPS4*	AR	+	+			3	+		(+)	+			
	26	Hermansky–Pudlak 6	*HPS6*	AR					3				+			
	27	Human ITCH E3 Ubiquitin Ligase Deficiency	*ITCH*	AR					2					+		
	28	X-linked reticulate pigmentary disorder	*POLA1*	X					3					+		
	29	Systemic autoimmunity	*TNFAIP3*	AR					1,3	(+)			+	+		
	30	TRNT1 deficiency	TRNT1	AR			+		5				+			
	31	Auto-inflammatory	NLRC4	AD					3				+	+		

T- and B-cell defects	32	Combined variable immunodeficiency (CVID) 1	*ICOS*	AR					5				+	+		
	33	CVID 8	*LRBA*	AR	+				3				+	+		
	34	IL-21 deficiency (CVID-like)	*IL21*	AR	+	+										
	35	Agammaglobulinemia	*BTK*	X	+				5					+		
	36	Agammaglobulinemia	*PIK3R1*	AR					5				+	+		
	37	Hyper IgM syndrome	*CD40LG*	X					1,5	+				+		
	38	Hyper IgM syndrome	*AICDA*	AR	+				1,3					+		
	39	Hyper IgM syndrome	*PIK3CD*	AR	+									+		+
	40	WAS	*WAS*	X			+		5				+	+		
	41	Wiskott–Aldrich syndrome-like phenotype	*ARPC1B*	AR			+		5				+	+		
	42	Omenn syndrome	*DCLRE1C*	AR	+				1,3							
	43	Severe combined immunodeficiency (SCID)	*ZAP70*	AR			+		5				+			
	44	SCID/hyper IgM syndrome	*RAG2*	AR					5				+	+		
	45	SCID	*IL2RG*	X					3							
	46	SCID	*LIG4*	AR		No further information							+	+		
	47	SCID	*ADA*	AR		No further information							+	+		
	48	SCID	*CD3γ*	AR	+				5	+			+			
	49	SCID	*ZBTB24*	AR	+	+			3,6	+	+					
	50	Hoyeraal–Hreidarsson S.	*DKC1*	X				(+)	1,3			+	+			+
	51	Hoyeraal–Hreidarsson S.	*RTEL1*	AR				+	5			+	+			+
	52	Hyper IgE syndrome	*DOCK8*	AR		+			1,5				+	+		

	53	ANKZF1 deficiency	*ANKZF1*	AR	+			+	1,5,6	+			+			
	54	Hamartoma tumour syndrome	*PTEN*	AD	+				3							+
	55	CARMIL2-deficient	CARMIL2	AR	+		+			(+)			+			
Immunoregulation	56	IPEX	*FOXP3*	X					3				+	+		
	57	IPEX-like	*IL2RA*	AR					2				+	+		
	58	IPEX-like	*STAT1*	AD					2							
	59	IPEX-like	*MALT1*	AR					2					+		
	60	IPEX-like	*CTLA-4*	AR	+				3					+		
	61	IL-10 signalling defects	*IL10RA*	AR	+	(+)			3	+	+		+	+		+
	62	IL-10 signalling defects	*IL10RB*	AR	+	(+)			3	+	+		+	+		+
	63	IL-10 signalling defects	*IL10*	AR	+				3	+	+					
	64	NOD2 deficiency	*NOD2*	AR	+	+			3,6	+						
	65	TRIM22 deficiency	*TRIM22*	AR	+	(+)			3,6	+						
	66	CARD8 deficiency	*CARD8*	AD	+											
	66	CARD9 deficiency	*CARD9*	AR	+											
	67	Loeys-Dietz syndrome	*TGFBR1*	AR			+		5							
	68	Loeys-Dietz syndrome	*TGFBR2*	AR			+		5							
Others	69	*MASP* deficiency	*MASP2*	AR			+						+	+		
	70	Trichohepatoenteric syndrome	*SKIV2L*	AR					3				+			
	71	Trichohepatoenteric syndrome	*TTC37*	AR					3				+			
	72	Heat-shock protein	*HSPA1L*	AD	+		+		3							

	73	Autophagy defect	*NPC1*	AR	+	+			3,6	+	+					

CD- Crohn's disease, UC- ulcerative colitis, disease location- 1= mouth; 2= enteropathy; 3= enterocolitis; 4= isolated ileitis; 5= colitis; 6= perianal disease, HLH, hemophagocytic lymphohistiocytosis; AR, autosomal recessive; eb, epidermolysis bullosa; X, X-linked; A, arthritis; vase, vasculitis; n, nail; h, hair; AD, autosomal dominant; e, eczema; f, folliculitis/pyoderma; SJ, Sjögren syndrome; p, psoriasis; AIHA, autoimmune hemolytic anemia; AN, autoimmune neutropenia; PSC, primary sclerosing cholangitis; HT, Hashimoto thyroiditis; AIH, autoimmune hepatitis; TID, type 1 diabetes mellitus; MAS, macrophage activation syndrome; NSIP, non-specific interstitial pneumonitis; S, serositis

1.7.3 HLA and IBD

The human leukocyte antigen (HLA) complex plays a key role in the disease pathogenesis of inflammatory bowel disease (IBD)[81]. Located on chromosome 6 (6p21.3) the region encodes 'classical' HLA genes (HLA-A, HLA-B, HLA-C, HLA-DR, HLA-DQ and HLA-DP) and over 130 other proteins enriched for roles in, and regulation of, the immune system. Whilst *NOD2* was designated as the 'IBD1' locus in GWAS the only region to reach genome wide significance in a meta-analysis of IBD GWAS studies was 'IBD3', corresponding to the HLA complex on chromosome 6[82]. This association was stronger for Crohn's disease than ulcerative colitis, but still significant for both diseases[82,83]. Whilst GWAS can detect association between a specific single nucleotide polymorphisms (SNPs) lying within the HLA complex, much of the data generated from these studies cannot be used to determine which HLA genes, and certainly not which genotypes, were associated with disease due extensive linkage disequilibrium within this region. Some association with haplotypes typified by specific SNPs has been conducted but much of the earlier data is from simple association studies rather than GWAS. Furthermore, in contrast to the progress in understanding of the non-HLA IBD genes since the advent of NGS technologies, research into the role of the HLA complex in IBD has lagged behind. Most data comes from GWAS although specific HLA genotypes/haplotypes/serotypes have been associated with IBD[84–87]. Despite this, the molecular mechanisms and exact HLA polymorphisms associated with disease, alongside the functional impact of variation has been problematic to elucidate. A specific issue which is hugely prevalent in HLA genomics relates to high levels of linkage disequilibrium between genotypes (genotypes co-existing in combinations in non-random proportions).

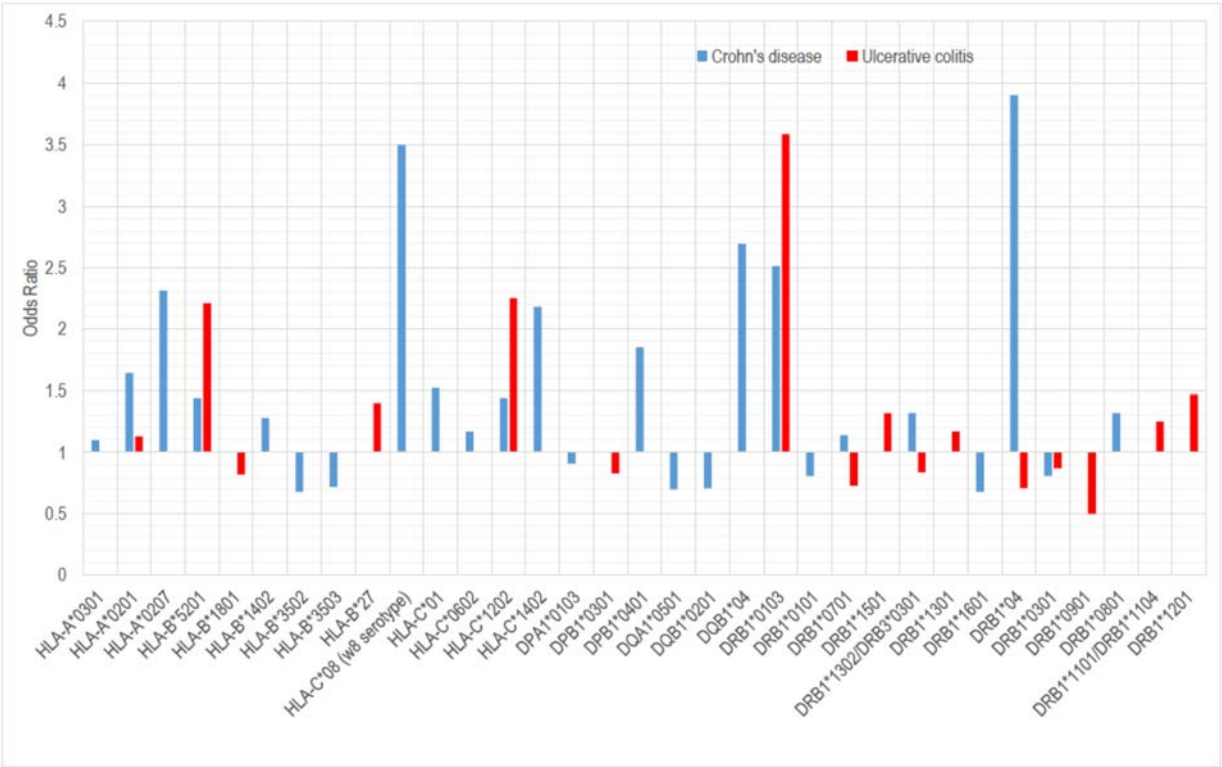

Figure 2- Graphical representation of the odds ratio for each independently associated HLA genotype. Un-replicated studies on <500 patients have been excluded from the graph. Where >1 study has implicated a genotype the odds ratio from the larger study has been used to represent the risk.

The role of HLA in recognition of 'self' and immune tolerance is clearly important in the association with autoimmune conditions[88]. The crucial part that HLA class II heterodimers play in presentation of bacterial antigens and stimulation of downstream immune response, through cellular activation, pro-inflammatory cytokine production and stimulation of T-helper/B-cell proliferation, is important within IBD pathogenesis. These data have fuelled speculation that there is an aberrant response to bacteria (including commensal) mediated by the classical HLA genes[62,89,90]. The strongest association, for both UC and CD, is with the HLA-DRB1*01:03 genotype, the odds ratio for developing IBD with this allele is approximately 2.5[86]. The mechanisms by which different HLA genotypes predispose to development of disease is poorly understood, with changes in the antigen binding cleft, molecular mimicry and altered interaction

for T-cells all muted as possible mechanisms. The risk associated with selected HLA genotypes can be seen in figure 2.

1.8 Key immune pathways in inflammatory bowel disease

There is general consensus that most patients with inflammatory bowel disease develop inflammation due to an acute hypoimmune response, preventing clearance of bacteria and resulting in chronic activation of alternative inflammatory pathways[91]. In a smaller proportion of patients there may be a primary hyperinflammatory response, which can be observed in monogenic conditions such X-linked lymphoproliferative disorders related to variation in *XIAP*[92]. An addition risk factor appears to be poor mucosal barrier function, allowing invasion of bacteria and activation of chronic inflammation[93]. The key immune pathways within IBD are described below, with the host-microbe interaction represented by figure 3.

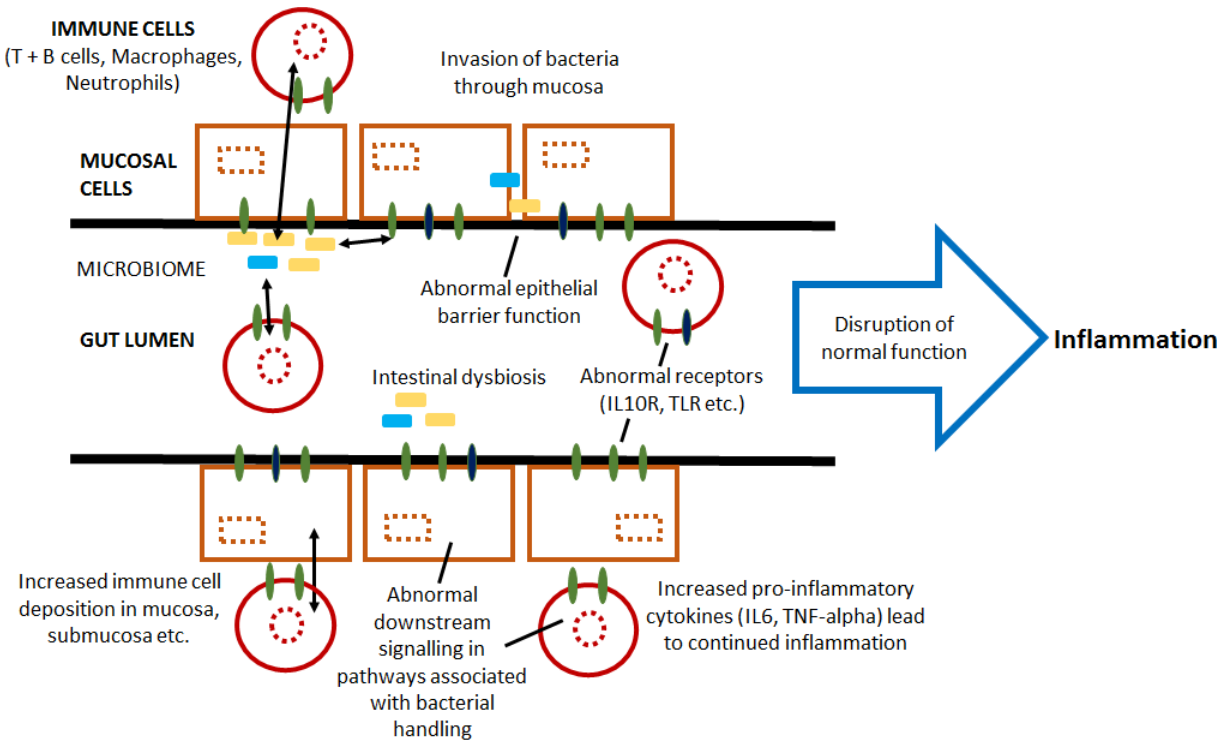

Figure 3- Schematic representation of potential causes of IBD. Host genetic, immune and microbiome interaction in inflammatory bowel disease. Disruption of epithelial barrier

function and invasion of bacteria into the mucosa, abnormal immune receptors (such as IL10R), an increased inflammatory response (mediated through pro-inflammatory cytokines), disrupted downstream immune signalling and abnormal handling of bacteria may all contribute to disease pathogenesis. Environmental factors (such as diet and medication) influence intestinal microbiota composition, in active Crohn's disease there is an abnormal ratio of beneficial and harmful bacterial species (dysbiosis). The complex interaction between these factors underlies disease process; in healthy patients there is a normal synergy between diverse, immune tolerated bacteria and the host immune system.

1.8.1 Recognition of bacterial pathogens

The most implicated pathways in Crohn's disease are related to pattern-recognition receptors (PRRs) which may be intracellular or extracellular. These receptors have well recognised risk variants for development of Crohn's disease and pathways related to these receptors have additional risk genes within them.

1.8.1.1 NOD-signalling

The most recognised risk gene for Crohn's disease is *NOD2. NOD1* and *NOD2* function as intracellular receptors and are present in macrophages, dendritic cells, gut epithelium and intestinal fibroblasts. They recognise gram-negative and gram-positive bacterial products, including muramyl dipeptide (MDP) and diaminopimelic acid (DAP) and trigger downstream inflammatory signalling resulting in inflammasome activation, autophagy and NFKB transcription[94].

1.8.1.2 Toll-like receptor (TLR) signalling

TLRs are membrane bound bacterial receptors recognising pathogen-associated molecular proteins (PAMPs). They recognise a wide range of microbial products and can function as

extracellular and intracellular recognition receptors. TLR2, 3, 4 and 5 are expressed in the intestine. TLR4 is very important for recognition of lipopolysaccharide, a key component of gram-negative bacterial cell membranes. All downstream signalling is pro-inflammatory and results in NFKB activation. There is tight control of TLR activity in the presence of commensal gut bacteria, in IBD there can be chronic activation of TLRs resulting in uncontrolled and inappropriate inflammation[94].

1.8.2 Innate pro-inflammatory pathways

There is activation of pro-inflammatory innate immune pathways in IBD, although the precise pathway activated in specific patients may vary.

1.8.2.1 JAK-STAT pathway

JAK-STAT is a highly conserved signal transduction pathway. There are four JAK and seven STAT genes, important for a range of cytokine signal transduction elements. JAK-STAT functions in conjunction with cytokine receptors (including for IL2, IL6, IL10, IL-23 and interferons), activation of the receptors results in relocation of STAT to the nucleus where it functions as we a transcription factor, promoting inflammatory responses[95].

1.8.2.2 NLRP3 inflammasome

There are a number of NOD-like receptor related inflammasomes, with NLRP3 being the best recognised in IBD. These innate immune pathways recognise and transduce antibacterial signals, resulting in inflammation via activation of IL1B and IL18. They function as a combination of priming and activation stages. Primed NLRP3 is induced by inflammatory stimuli, often through TLR activation and NFKB transcription. Subsequent activation is through recognition of PAMPs, often through presence of reactive oxygen species and NOD-signalling[96].

1.8.2.3 NADPH oxidase activation

Production of reactive oxygen species (ROS) in response to bacterial invasion is largely mediated through NADPH within neutrophils. Activation of TLRs, G-protein coupled receptors and cytokine receptors on neutrophils trigger the release of ROS into phagosomes (containing bacteria), resulting in bacterial killing. Neutrophils will also release ROS at the site of active inflammation, following chemotaxis to the area. This results in direct damage to bacteria but can also damage tissue leading to chronic inflammation[97].

1.8.3 Adaptive immune response

Antigen-presenting cells including dendritic cells, macrophages and B-cells, have pattern-recognition receptors and are vital for sensing bacterial antigens. Subsequent activation of the adaptive immune response through presentation of bacterial antigens to T-cells via MHC class 2 receptors triggers production of proinflammatory cytokines such as IL1, TNF-a and IL6[98].

1.8.3.1 Th17 cells and IL17-signalling

Th17 cells are increased in the inflamed guts of patients with IBD, and are induced by IL6 and TGF-b, and their expansion also appears to be dependent on the presence of specific bacterial communities, including Bacteroides. Th17 cells produce IL17A, IL17F, IL21, and IL22. IL17A is a key cytokine mediating innate and adaptive mucosal immunity, leading to increased inflammatory cell recruitment and epithelial cell production of proinflammatory IL8.

1.8.3.2 T-regulatory, Th1 and Th2 cells

T-regs are key in controlling inflammatory processes, and often secrete anti-inflammatory cytokines such as IL10 and TGF-b. Historically, Th1 cells are thought to be integral to the inflammation seen in Crohn's disease and release IFN-γ, TNF-a and IL2, while conversely, Th2 cells are thought to be involved in ulcerative colitis pathogenesis and release IL4, IL5, and IL13. As additional T-helper subsets have emerged this distinction is largely felt to be outdated. The

regulation of inflammation in response to bacteria in the normal gut is key to maintaining homeostasis. In IBD there is adaptive immune imbalance shifting to a chronic pro-inflammatory state, with a probable increase in the Th1/2 to T-reg ratio[99].

1.9 Transcriptomics of IBD

The interaction of environmental factors (such as the intestinal microbiome and nutrition) and the host, through effects on immune cells and intestinal barrier function, can be studied through gene expression in a target tissue, the transcriptome. Recent, large-scale, studies have detailed transcriptomic profiles in both treatment-naïve paediatric Crohn's disease and ulcerative colitis. One of the most well replicated genes to have differential expression in Crohn's disease is *DUOX2*, encoding for an antimicrobial peptide involved in phagocytosis and the NADPH complex. This upregulated gene characterises an ileal signature in recent studies by Haberman *et al* (2015) and Lloyd-Price *et al* (2019): identified an upregulation of *DUOX2* in the ileum of active Crohn's disease, which was replicated in several cohorts[67,100]. Specifically Lloyd-Price *et al* identified a significant correlation of increased *DUOX2* expression with a reduction of Ruminococcaceae operational taxonomic units (OTUs), implying a direct impact of the host on development of microbial dysbiosis, or visa versa[100]. When subjected to pathway analysis the genes within the '*DUOX2* signature' characterise an activated NFκB innate antimicrobial response pathway. The DUOX2 gene signature included upregulation of Mucin genes.

The signature identified by Haberman *et al* was characterised by differential expression of the *APOA1* gene, encoding for apolipoprotein A-I, part of high-density lipoprotein. This gene was differentially expressed between all Crohn's disease (including colonic and ileal) and controls. The gene signature was classified, through pathway analysis, as leading to increased *STAT1*-dependant signalling. Similarly in paediatric ulcerative colitis the upregulated gene expression signature in rectal tissue was characterised by enrichment for genes in the JAK-STAT pathway, alongside integrin signalling and TNF-α production[101]. This pathway has been identified as a potential

target for medications in IBD, and specifically ulcerative colitis, including the JAK-inhibitors (Tofacitinib)[58,102]. The major additional finding in paediatric ulcerative colitis was of a gene signature typifying decreased mitochondrial activity within the colon of ulcerative colitis patients, but not controls or ileocolonic Crohn's disease. This downregulation of all 13 genes involved in adenosine triphosphate (ATP) production was functionally validated through assessment of electron transport chain in fresh biopsies. The impact of reduced mitochondrial dysfunction appears to exacerbate poor barrier function[101]. It is possible that this mitochondrial dysfunction is a response to an inflammatory state.

These data indicate an upregulation of genes in active Crohn's disease involved with microbial sensing and downstream inflammatory response, largely through NFκB and JAK-STAT signalling. There is increased oxidative stress and activation of innate immune pathways involved in bacterial sensing and clearance. Additionally, Th1 cell polarisation, a process of transformation into a pro-inflammatory, IFN-γ, TNF-α and IL-12 producing immune cell, appears to be increased. Th1 cells are principally involved in the targeted immune response in order to facilitate destruction of cells infected with intracellular bacteria and viruses[103]. Improper, or ineffective, activation of Th1 cells may lead to active and chronic inflammation through over reaction to gut commensals, or through inappropriate clearance of pathogenic bacteria.

Whilst analyses have identified specific gene-expression signatures in patients with PIBD, the interpretation of whole-tissue gene expression should be interpreted with some caution as specific expression in cells of interest (dendritic cells, monocytes, neutrophils) cannot be assessed. Additionally, studies detailing the transcriptome in treatment-naïve Crohn's disease are single time point studies influenced by age-specific changes, do not follow patients through remission and are therefore subject to transient factors and are unable to account for temporal changes.

1.10 Microbiome in IBD

The microbiome refers to the genetic material derived from bacteria (within a specific location). In treatment-naïve patients there is disruption of the normal microbiota leading to an alteration of normal intestinal microflora, termed intestinal dysbiosis. It is not clear whether this is the cause of intestinal inflammation or the effect of acute or chronic inflammation. As discussed above, there is evidence that the host immune response can correlate, and appear to impact, on the intestinal microbiome. In contrast there is emerging evidence that microbial profiles from patients with IBD are sufficient to exacerbate intestinal inflammation in murine models, when compared to microbes taken from healthy individuals [104]. IBD-associated microbiota induced increased pro-inflammatory Th17 and Th2 cells, and resulted in reduced RORγt+ Tregs, regulatory cells implicated in intestinal pathogen response and inflammatory homeostasis[104,105]. There is emerging evidence that transplanted microbes, through faecal microbiota transplant (FMT), can impact on the host metabolism leading to obesity or weight loss[106]. Results of FMT in IBD have not yet proven reproducible or particularly effective, despite significant interest[107].

Prospective characterisation of patients with inflammatory bowel disease, both prior to diagnosis and through treatment consistently report a reduction in alpha and beta diversity in active disease, which return to more normal state, with higher diversity, when patients are in remission[62,108]. The faecal microbiome is a relatively poor reflection of the mucosal associated bacteria, with the latter displaying a significantly more perturbed dysbiosis and being more likely to interact with the host immune system[62]. However, individual variation, rather than disease state, is the main contributor to separation of microbial profiles through taxonomy driven analysis, meaning individual factors, such as diet, antibiotics and climate, are stronger drivers of microbial composition than disease state[100]. Faecal profiles are typically dominated by Firmicutes, Actinobacteria and Bacterioidetes in both disease and health[109]. Despite this, there are specific, replicated, disease associated profiles, typified by increases in genera such as Enterobacteriaceae, Pasteurellacaea and Fusobacteriaceae, with decreased abundance of

Bacteroidales, Erysipelotrichales and Clostridiales[62,110]. Despite these data, studies associating single bacterial genera with IBD have not found replicated, substantiated evidence of causation, other than perhaps for Faecalibacterium prausnitzii, and it is likely that it is the functional potential of whole bacterial communities that would be important in induction of inflammation and subsequently in remission.

Newer *in silico* technologies, such as gene count imputation (PICRUSt, HUMAnN), metagenomics, metatranscriptomics and concurrent metabolomic profiling are now providing insight into the functional potential of microbial communities. The concurrent metagenomic and metatranscriptomic functional profiling of patients with IBD has revealed dominant transcriptional pathways driven by specific bacterial species, such as Faecalibacterium prausnitzii[111]. Similarly, these data allowed identification of bacterial genera that had high abundance but appeared to be functional inactive. Bacteria, such as Alistipes putredinis and Bacteroides vulgatus, were able to be identified at very low relative abundances but appear to transcribe biologically important genes. Additionally identification of active pathways in more common bacteria, such as E. coli results in evidence for production of bacterial proteins, such as phosphate containing antigens, and metabolites, such dimethylallyl pyrophosphate, associated with immune activation and induction of inflammation respectively[111]. Longitudinal profiling has now identified transcriptional activation of specific bacterial groups, such as clostridia species, associated with active disease, that occur alongside alterations in the relative abundance of other species specifically depletion of obligate anaerobes (Faecalibacterium prausnitzii) and the enrichment of facultative anaerobes such as E. coli[100].

Finally, there appears to be a significant impact of the host genome on microbiota. Data on individual risk variants and bacterial signatures (including specific species) is now beginning to emerge. These include *ATG16L1* T300A, which was shown to be associated with increased Bacteriodes ovatus in a murine model and a composite of *NOD2* risk variants was associated with decreased abundance of Roseburia genera and Faecalibacterium prausnitzii in human IBD

patients[112,113]. However more extensive analysis, through selection of 11 IBD-risk SNPs in NOD2, CARD9, ATG16L1, IRGM and FUT2 and the HLA-DRB1*01:03 genotype, did not find any significant correlation with microbial relative abundance [109]. Clearly there is a need to assess the impact of genomic variation on microbial profiles, through WES or WGS, rather than selected genotypes. However, the power to detect significant associations will prove difficult due to the need for multiple testing correction and machine learning approaches will be required.

1.11 Personalised Therapy in Paediatric Inflammatory Bowel Disease

Personalised (or precision) medicine is 'a form of illness management that uses information about a person's genes, proteins, and environment to prevent, diagnose, and treat disease'[114]. The slow rate of uptake of this approach in most areas of medicine is largely related to the complexity of biomarker identification in complex, multi-faceted disease. There are additional challenges in the ability to collect, collate and analyse large data sets involving data with high dimensionality. Personalised medicine has been brought to the fore in the United Kingdom by the 100,000 genomes project and the potential to apply genomic data to cancer and rare disease, giving patients a more personalised diagnosis[115].

In a complex disease, such as IBD, the application of personalised medicine can be considered in a variety of areas[116]:

- Diagnostics- providing rapid and accurate diagnosis of disease, including rare disease subtypes. Application of next-generation sequencing (NGS) and other biological data (such as metabolomics) with specific emphasis on genomics.
- Stratification- use of biomarkers and application of artificial intelligence (including machine learning), leading to the creation of new subgroupings of disease through unsupervised and supervised approaches

- Prognostication- providing accurate information based on data available at diagnosis (clinical and scientific) to provide an outlook for disease severity, complications and co-morbidities
- Medication/treatment response- supervised stratification of patients into groups based on likelihood of response to therapy or likelihood of side effects based on data available at diagnosis and during follow-up. Development of new therapies based on unsupervised grouping of patients and biomarkers identification.

1.11.1 How is personalised medicine developing in inflammatory bowel disease?

Several studies have already applied stratification techniques to patients with IBD based on clinical data (including disease location, hospitalisation records, medication usage), biomarkers (immune and molecular) and next-generation sequencing in an effort to classify patients into multiple groupings based on characteristics beyond CD and UC using machine learning approaches[64,117–119]. A recent and large prospective study from North America applied a multi-omic approach, including clinical data to predict complications in Crohn's disease (stricturing and penetrating disease), with an additional attempt to include treatment outcomes in the model[120]. Whilst the competing-risk model had a specificity of 71% for predicting complications there is room for significant improvement, with an accompanying editorial discussing the limitations but also the optimism associated with the potential of personalised medicine in IBD[121].

Marigorta *et al* (2017) sought to forecast complicated disease based on a transcriptional risk score (a score created by the authors based on gene expression profiles) in the same RISK inception cohort. The authors had a degree of success, allowing researchers to distinguish indolent versus complicating disease (stricturing or penetrating) based on the scores from 29 genes[122]. A more specific approach was adopted by Denson *et al* (2018) where genomic data was integrated for significant mutations in genes associated with neutrophil reactive oxygen species production. Patients with these mutations had significantly increased risk of perianal and stricturing disease,

suggesting more aggressive treatment might be appropriate from diagnosis in this patient group[123]. Lee *et al* (2011) provided the first example of stratifying patients using an immune biomarker (transcriptional characteristics of CD8+ T cells), allowing stratification of adult patients in a high risk and low risk group for relapse[124]. However, these data have not been replicated in children or additional adult cohorts[125].

Predicting response or complications related to medication is also a key area of personalised medicine, enabling reduction in side effects and improved patient outcomes[126]. Two studies from Arijs *et al* (2009 and 2010) analysed gene expression profiles in the mucosa of patients with Crohn's disease (n= 37 patients) and ulcerative colitis (n= 46 patients). The authors found predictive, differentially expressed genes, related to response to infliximab therapy, separating responders versus non-responders with up to 95% accuracy, but these results have not since been replicated[127,128]. Whilst there are fewer data on response to medications in paediatric practice, development of these models is important to predict both response to medication and equally as important the probability of developing side-effects and complications.

There is considerable interest in the predictive role of the microbiome, and it has also been analysed in an effort to assess factors predictive of response. An early example from Kolho *et al* demonstrated microbiota returning to a similar composition to controls in paediatric patients who responded to anti-TNF therapy, but not in those who were non-responders (total number of patients treated= 32). Response to anti-TNF was predicted by 6 bacterial groups[129]. Recently Douglas *et al* used both 16S and metagenomic sequencing to build a predictive model to identify responders to induction therapy in paediatric Crohn's disease, with an accuracy of 94.4%, albeit in a cohort of only 19 paediatric patients[64]. The authors found metagenomic features, in comparison to 16S sequencing, were better at classifying patients into treatment response or non-response. Shaw *et al* (2016) focused on mucosal healing as an outcome for treatment, but failed to formulate a predictive model that could differentiate between 'responders' and 'non-responders' however they did identify significant differences between groups at a bacterial genus

level[110]. Doherty *et al* (2018) observed baseline differences in the faecal microbiome (increased Faecalibacterium and Bacteroides) between Crohn's disease patients responding and not responding to induction with ustekinumab therapy. They were able to predict response to therapy with an accuracy of 84.4% and concluded that microbiota may be a useful biomarker for response[130].

More recently application of clinical data has been used to predict disease outcome, complications (fibrotic stricturing disease, penetrating disease and perianal complication) and early relapse. Ziv-Baran *et al* (2018) reported that the best clinical predictor of complications was early relapse (seen in 29%, compared to 9.7% who did not relapse), whereas subsequent relapse (within 1 year) was associated with disease activity at week 12 (including raised inflammatory markers and raised faecal calprotectin), more so than at diagnosis[131].

1.12 Machine Learning- Supervised and Unsupervised approaches

The analytical routes to clinical application can be broadly divided into two distinct approaches- supervised and unsupervised classification of disease (based on multi-omic or clinical data), figure 4. Stratification of disease subtypes based on response to treatment, severity and outcomes through the underlying genetic (and microbiome/gene expression) variation provides a potential methodology to drive a shift in the way patients are treated. The input to this modelling can be from a variety of numerical data derived from multi-omic analysis, including gene pathogenicity scores, relative bacterial abundances, functional pathway scores and gene expression counts.

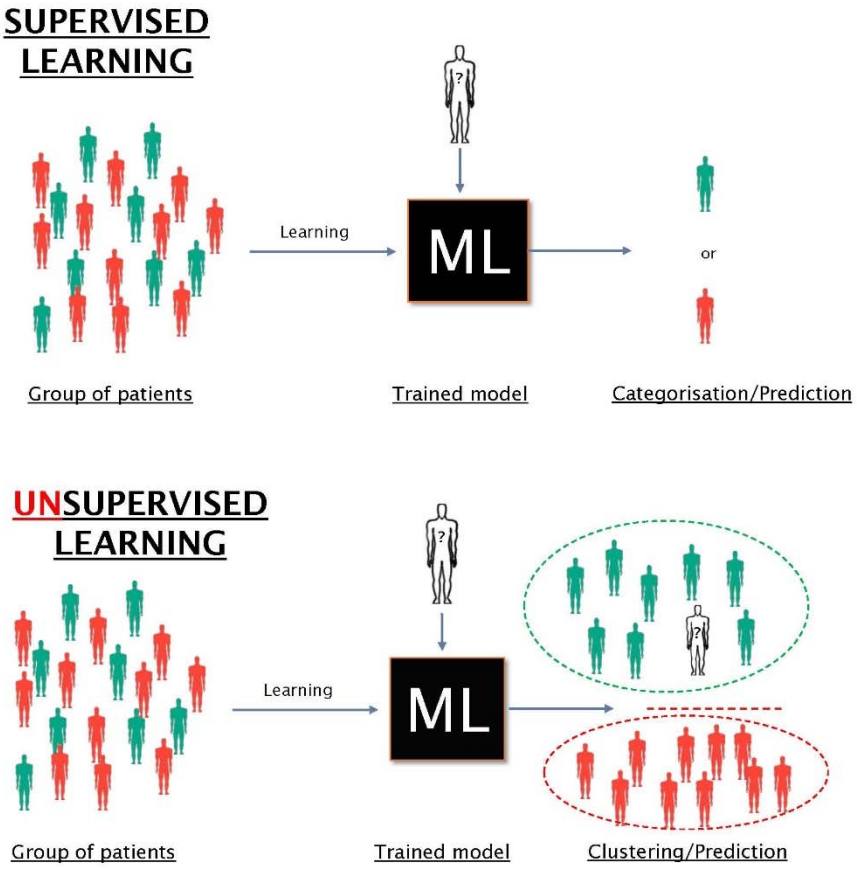

Figure 4- Machine learning schematic drawing, showing supervised and unsupervised approaches. In supervised machine learning the trained model uses the characteristics of the item (patient) to place them in the most appropriate group (diagnosis, outcome etc.). This is used for classification modelling. In unsupervised machine learning the model clusters patients together based on how similar (due to their characteristics) they are, without knowledge of the diagnosis, outcome etc. This is used for novel cluster/group discovery.

Machine learning (ML) algorithms have the ability to providing the means for efficient analysis and interpretation of complex data and is currently applied across different areas of medicine and biology including cancer research, drug discovery, genomics and proteomics[132–134]. Regardless of the specific application, it is possible to identify two very distinct and common tasks for which ML algorithms are employed; classification (supervised) and class discovery (unsupervised).

For classification purposes, supervised machine learning algorithms are trained against labelled (known) data and, as a result of successful learning, identify patterns that match the provided

stratification. This approach has been widely chosen for the classification of patients in known disease subtypes (such as CD vs UC), discrimination of pathogenic and benign variants or in prognosis prediction[118]. Concurrently, several applications of ML in cancer science have successfully classified cancer subtypes depending on histological features, genomic markers or proteomic features[132–134]. In the context of personalised therapy, a supervised approach stratifies patients by outcomes such as response to treatment, development of complications (such as stricturing or penetrating disease), growth outcomes or requirement for escalation to monoclonal therapy/need for surgery. This model could then be used to classify patients into different risk stratifications (outcomes) at diagnosis based on patient characteristics, thus impacting on medication choices, nutritional intervention and management.

Alternatively an unsupervised approach may be taken, allowing a machine learning model to group patients by how similar they are based on underlying features (such as genetic, gene expression or microbiome signatures) and fuelling development of specific treatments for these groups based on these characteristics (including development of new medication targets). These groups may be interrogated post-hoc for enrichment of patient outcomes, such as medication response, adverse events or complications. An example of unsupervised machine learning can be seen in work from Weiser *et al* (2016). They report detailed gene expression profiles of adult and paediatric intestinal tissue, with two distinct clusters of CD emerging for each group. The authors performed post-hoc assessment of clinical outcomes with specific gene expression patterns, finding that colonic-type gene expression profiles were at increased risk of colectomy, whereas ileal-type were at increased risk of biological therapy[117].

There are various mathematical tools available for ML of which support vector machines (SVMs) and random forest classifiers (RFCs) are amongst the most popular models for classification tasks. The fact that these models are relatively easy to interpret and are calculated with comparative computing efficiency has made SVM and RFCs the preferred choice for modelling biological phenomena. When the analytical question is the identification of novel patient strata,

unsupervised machine learning tools, such as principal component analysis (PCA), multidimensional scaling (MDS) and the t-SNE algorithm, are the preferred models currently applied. An additional, and less sophisticated, tool is hierarchical clustering, where samples (such as patients) can be grouped by their similarity, based on constituent features/characteristics (such as genetic or clinical data). This may reveal novel groups enriched for patient outcomes.

Despite the performance achievable by modelling single data types, there is still room for improvement by merging diverse data, achieving superior power in detecting existing and novel disease subtypes. With each data type representing a different characteristic of a single patient/individual, machine learning algorithms can compute simplified representations of this higher complexity. As a direct consequence, machine learning approaches ease the identification of novel strata, which might reflect important clinical outcomes and enable personalisation of therapy based on new groups and distinct multi-omic features.

1.12.1 Pathway analysis

Genes implicated in IBD have been well established by GWAS alongside NGS looking at monogenic IBD and polygenic IBD. Many of these genes are implicated in innate and adaptive immunity, including cell surface receptors, downstream signalling, lymphocyte activation and cytokine response. Genetic variation in pathways associated with bacterial recognition and response (antigen presentation, NOD pathway, TLR pathways) appears to be a key driver for some cases, with *NOD2* mutations being the best-known risk for development of Crohn's disease. Other implicated genes include tight junction regulation, solute carriers, autophagy and cell response to stress[91].

The laying of transcriptomic data on genetic data provides the ability to determine specific impacts of genetic variation on gene expression in the terminal ileum and rectum of patients with IBD. Specific interrogation of genes associated with pathways implicated in IBD may provide insights into differences between controls and IBD patients. The use of the Reactome database

enables mapping of candidate genes (due to variation or differences in expression, up or downregulation) to specific pathways, including disease processed[135].

An additional role in pathway analysis is present for microbiome data. In patients with aberrant bacterial handling/differential gene expression in implicated pathways the impact on the mucosal microbiome can be assessed. Differences or emergence of specific species, alongside functional alterations in the metabolic profile of the community may provide insights into the role the microbiome plays, or the role the host has in shaping the microbiome in disease.

1.13 Thesis outline and work plan

Within this thesis I present work detailing the integration of clinical and multi-omic data in order to provide a personalised approach to diagnosis and management of children with inflammatory bowel disease. We hypothesise that through application and integration of genomic, transcriptomic and clinical data, we will identify novel patient groups related to clinical outcomes and provide the template to personalise diagnosis and prognosis for patients. Utilising monogenic and oligogenic classifications, identification of transcriptomic subgroups and specific impacts of genomic variation on RNA transcription we enable new patterns to emerge. Beyond this, analysis of genomic and transcriptomic data in conjunction with clinical outcomes provides clear clinical phenotypes associated with molecular signatures.

Chapter 2 Methods

Chapter summary- *This chapter describes recruitment, sample storage, sample processing and sequencing methodology. Analytical techniques are described. Clinical data extraction is described.*

Chapter contributions- *Recruitment was performed by James Ashton and Rachel Haggarty. Sampling and recruitment were aided directly by Mark Beattie, Tracy Coelho, Akshay Batra, Nadeem Afzal, Mick Cullen, Rachel Russell and Claire Barnes*

Sample extraction was by Nikki Graham (DNA) and Konstantinos Boukas (RNA). DNA sequencing was outsourced to third party service providers. RNA sequencing was performed by James Ashton and Konstantinos Boukas.

Clinical data extraction was performed by James Ashton, Florina Borca and Hang Phan.

Exome sequencing data were processed by Enrico Mossotto and Imogen Stafford, variant call files and GenePy scores were analysed by James Ashton. RNA sequencing was processed and analysed by James Ashton. Preliminary microbiome data was processed and analysed by James Ashton. Data integration was performed by James Ashton (statistical modelling, hierarchical clustering) and Imogen Stafford (machine learning algorithms).

2.1 Genetics of Paediatric inflammatory bowel disease study and cohort

'Genetics of Paediatric Inflammatory Bowel Disease' is a single centre longitudinal cohort study, with over 500 paediatric IBD patients recruited, all with whole exome sequencing data. Additionally, we have in excess of 1700 relatives. The study is based at Southampton Children's Hospital/University of Southampton. Patients are recruited at Southampton Children's Hospital. All patients are diagnosed with IBD, in line with the Porto criteria, and are aged less than 18 years at diagnosis. Southampton Children's hospital covers a paediatric population of around 650,000

children and is the tertiary referral centre for 12 district general hospitals including the Isle of Wight and the Channel Islands.

2.2 Treatment naïve and established disease patient recruitment + biopsy samples

A cohort of patients was recruited specifically for this PhD. These patients were recruited at the time of endoscopy and consisted of treatment naïve patients and established disease patients. A cohort of controls were also recruited. All of these patients had biopsies collected for analysis. A summary of recruitment can be seen in table 3, alongside the number of biopsies and blood samples collected.

Table 3- Patient recruitment and sample acquisition.

	Treatment Naïve	Established	Controls	Follow-up patients	Total
Patients recruited	47	46	23	10	126
Ileal Biopsies					
RNA	38	34	17	7	96
Microbiome	38	34	17	7	96
Rectal Biopsies					
RNA	45	43	22	10	120
Microbiome	45	43	22	10	120
Blood for genetics	46	45	N/A	N/A	91

2.2.1 Biopsy acquisition and storage

Biopsies were obtained from patients during routine endoscopy. Patients were either established inflammatory bowel disease patients, or were suspected to have inflammatory bowel disease.

Ileal and rectal tissue was extracted using endoscopy biopsy forceps under normal endoscopy conditions.

Biopsies were immediately placed into a cryovial containing 1ml of RNA later. The diameter of each biopsy was estimated to 2.5mm (range 1-4.5mm) with the mean volume estimated to be 27mm^3 (equivalent of 30mg). As per manufacturer instructions each sample should be stored in >10 volume (10µl per 1mg tissue) of RNAlater (Sigma Aldrich) equating to a mean of 300µl per sample, with an upper limit required of 1ml for the largest biopsies. Biopsies were then frozen at -80 within 30 minutes of being placed into RNAlater.

For treatment naïve patients, following histological examination of tissue by a consultant paediatric histopathologist biopsy samples were classified as treatment-naïve inflammatory bowel disease (Crohn's disease, ulcerative colitis or IBDU). If there were no histological features of disease the sample was classified as a control sample.

2.3 Clinical data extraction

2.3.1 Automated data extraction

Where possible clinical data were automatically retrieved from the University Hospital Southampton (UHS) systems by the data science team at UHS/Southampton Biomedical Research Centre. These included laboratory results system, electronic patient record, endoscopy records, histopathology reports and radiology records. Medications (including thiopurine, 5-ASA, infliximab and adalimumab) were sourced from the Wessex paediatric IBD database (2007-2012) or recorded from pharmacy or electronic patient records (2013-2020).

The following data were extracted-

- Demographics- Sex, age at diagnosis, maximal follow-up time

- Treatment and surgery- Any IBD related surgery (intra-abdominal and perianal), abdominal surgery (subtotal colectomy, hemicolectomy), perianal surgery (seton placement, fistula drainage), monoclonal use ever, thiopurine use ever, 5-aminosalycylate use ever,
- Blood results- all automatically extracted
- Endoscopic disease extent
- Histological disease extent
- Complications- fistulating disease ever, stricturing disease ever (histologically proven or narrowing demonstrable on two consecutive MRIs, with prestenotic dilatation).
- Growth- all height SDS measures, all weight SDS measures

All operative procedures (stricturoplasty, small or large bowel resections, primary stoma formation) and all perianal procedures (at any time) conducted from 2007-2020 were recorded either on the Wessex paediatric IBD database (2007-2012) or sourced automatically from the electronic patient record (2013-2020). The date of procedure was also retrieved. Data on surgical procedures occurring between 2002 and 2012 have been previously published for both Crohn's disease and ulcerative colitis[136,137].

2.3.2 Manual data curation

Electronic records were manually reviewed for surgical outcomes and medication use for the whole cohort to ensure completeness. Where results were not available from UHS electronic patient record, the referring hospital records, sent to Southampton at diagnosis, were manually searched. Data on stricturing and penetrating complications, not recorded in a standardised fashion, were extracted by review of electronic records. Overall patient outcomes, including early relapse, were collected by manual searching of the electronic clinical records.

2.4 Ethical considerations and approvals

Patients, and parents/guardians, gave informed consent prior to participation in research. All participants were given adequate time to review age-specific literature on the study and to ask questions.

During the study we undertook a substantial amendment to enable direct feedback of genetic results to patients and clinicians.

2.4.1 Ethical approval

The study had category A ERGO II ethics approval (30630) from the University of Southampton and a REC approval from Southampton and South West Hampshire Research Ethics Committee (09/H0504/125).

2.5 Genomic sequencing and analysis

2.5.1 DNA extraction

Patient DNA was extracted from peripheral venous blood samples collected in EDTA using the salting-out method[138]. DNA concentration was estimated using the Qubit ® 2.0 Fluorometer and 260:280 ratio calculated using a nanodrop spectrophometer.

2.5.2 Whole exome sequencing

2.5.2.1 Library preparation

Approximately 20ug of each patient DNA was extracted and sent for whole exome sequencing (external- Novogene or Macrogen). Sequencing was performed at Novogene or Macrogen sequencing companies. Briefly, a total amount of 1μg genomic DNA per sample was used as input material for the DNA sample preparation. Genomic DNA was enriched with Agilent SureSelect All

Exon capture kit (version 4, 5 or 6). PCR was used to add tags to libraries. Products were purified and quantified using the Agilent Bioanalyzer system.

2.5.2.2 Illumina sequencing

The libraries were loaded onto Illumina platforms after pooling according to concentration and expected data volume.

2.5.2.3 Bioinformatic analysis

Raw fastq sequencing data were processed using the same custom, in house, pipeline. VerifyBamID was utilised to check the presence of DNA contamination across the cohort[139]. Alignment was performed against the human reference genome (GRCh38/hg38 Dec. 2013 assembly) using BWA [140] (version 0.7.12). Aligned BAM files were sorted and duplicate reads were marked using Picard Tools (version 1.97). Following GATK v3.7[141] best practice recommendations[142], base qualities were recalibrated in order to correct for systematic errors produced during sequencing. Finally, variants were called using GATK HaplotypeCaller was applied to produce a gVCF file for each sample. Samples were processed on the University of Southampton IRIDIS cluster requiring an average of 4 hours run time per sample on a 16-processor node.

While the standard VCF format reports only alternative calls, the gVCF format identifies non-variant blocks of sequencing data and returns reference calls for loci therein. This enables affirmative calling of homozygous reference loci when combining call sets from multiple samples, this difference makes calling homozygous reference loci possible when combining multiple call sets, which is vital in downstream analysis, such as GenePy or allele specific expression analysis. Multi-sample variant calling was achieved through calling each individual sample separately and then merging all gVCFs using GATK GenotypeGVCFs. Annotation of this composite file applied Annovar v2016Feb01 using default databases refSeq gene transcripts (refGene), deleteriousness scores databases (dbnsfp33a) and dbSNP147) and the human genetic mutation database (HGMD)

flat file[143]. Variant allele frequencies were sourced through the genome aggregation database (gnomAD)[144]. Following GATK best practice guidelines, HaplotypeCaller default settings were utilised. Only variants with a minimum Phred base quality score of 20 were called.

2.5.3 Application of GenePy score to genomic data

GenePy is a per gene, per patient score for combining the effect of multiple variants into single gene scores for each individual, developed in Southampton. Within this project GenePy is used to integrate genomic data into machine learning analysis and for gene prioritisation. GenePy integrates elements such as variant zygosity, rarity and deleteriousness assessed through a chosen *in-silico* deleteriousness metric. As a result, it is possible to estimate each gene pathogenicity on a per-patient basis where large GenePy scores correspond to a greater mutation burden. The GenePy score S_{gh} for a given gene (*g*) in individual (*h*) is

$$S_{gh} = -\sum_{i=1}^{k} D_i \log_{10}(f_{i1} \cdot f_{i2})$$

Where for each gene (g), for each patient (h) the biallelic mutated locus (i) in a gene is weighted according to its predicted allele deleteriousness (Di), zygosity and allelic frequency (fi). In order to assure compatibility, GenePy scores were generated for regions covered by all WES enrichment chemistries (Agilent SureSelect versions 4, 5 and 6) by intersecting the respective bed files. Moreover, variants with a genotype quality (GQ) lower than 20 and with more than 30% missing genotypes across the cohort were excluded.

GenePy scores were computed by implementing the CADD or DANN deleteriousness metrics (details within each chapter). For specific analyses GenePy scores were further corrected for gene length (covered by enrichment chemistries) and the Gene Damage Index to account for difference in transcript length and gene mutability[145].

2.6 RNA sequencing and transcriptomic analysis

2.6.1 RNA extraction

Frozen biopsies were transferred to the WISH laboratory to be extracted. Biopsies were transported on dry ice and remained frozen in RNAlater. Biopsies were processed in batches of 12 using the Maxwell RSC simply RNAtissue kit.

Maxwell homogenisation solution was prepared using DX reagent to avoid foaming at 0.5%.

- Maxwell Homogenisation solution: homogenisation solution was chilled on ice. Using 2,985µl of chilled Maxwell homogenisation solution, 15 µL of reagent DX was added alongside 60µl of 1-Thioglycerol.

2.6.1.1 TissueLyser

Biopsies stored in RNAlater were thawed to room temperature. A sterile filter tip was used to transfer the biopsies into the pre-cooled tubes containing a TissueLyser bead and incubated at room temperature for 2 min to avoid freezing of lysis buffer. 200µl of the prepared Maxwell homogenisation solution was added in the appropriate bead containing tubes (tube 1-12). The tubes containing the bead, the biopsy and the homogenisation solution were placed into the insert of the TissueLyser LT adapter. Samples underwent lysing for 5 min. Tubes were then removed and centrifuged for 3min at full speed. Supernatant was then transferred into RNAse-free fresh tubes.

2.6.1.2 Maxwell processing

Maxwell processing

Solution preparation:

- 1-Thioglycerol/Homogenization Solution: 20µl of 1-Thioglycerol per millilitre of Homogenization Solution. A volume of 200µl of 1-Thioglycerol/ Homogenization Solution for each sample.
- DNase I Solution: 275µl of Nuclease-Free Water to lyophilized DNase I. 5µl of Blue Dye to the reconstituted DNase I as a visual aid for pipetting. Dispense the DNase I Solution into single-use aliquots in nuclease-free tubes.
- Cartridge Preparation (maximum of 16 samples. 1sample per cartridge). A Plunger was placed in well 8 of each cartridge. 0.5ml Elution Tubes were placed in the Deck Tray. 50µl of Nuclease-Free Water was added to the bottom of each Elution Tube.

200µl of Lysis buffer was added to the 200µl of supernatant from the Tissuelyser and vortexed vigorously, 400µl was then transferred to well 1 of each of the 12 Maxwell RSC cartridges and 5µl of DNase I solution was added to well 4. The Maxwell instrument was run on the simplyRNA Tissue method. The resulting solution was transferred into 2 aliquots of 20µl and stored at -80^{o}C.

2.6.2 RNA quantification and quality check

The Bioanalyzer was used in conjunction with the Agilent RNA 6000 Nanokit. The samples were loaded onto the Gel-Dye mix RNA chip following priming. The chip was vortexed and run on the Agilent Bioanalyzer to give RIN values and RNA concentrations for each extracted sample. Traces and values for each batch of samples can be seen in supplementary data.

2.6.3 Targeted RNA sequencing

The HTG EdgeSeq Autoimmune Panel was chosen for this experiment. This contemporary assay is used to measure mRNA expression levels in 2002 genes associated with autoimmune disease, specifically including inflammatory bowel disease [146]. Previously HTG developed gene panels have been utilised to determine differential gene expression to guide treatment in oncology patients[147]. This technology utilises custom nuclease protection probes to target mRNAs of

interest (2002 autoimmune genes), following amplification, gene expression in samples is determined using next-generation sequencing. The chemistry provides the advantage of specifically targeting genes of interest in inflammatory bowel disease, thus negating highly expressed 'housekeeping' genes and providing increased biological insight within pathways of interest. Lowly expressed genes may not be allocated reads in conventional RNA sequencing, however using a targeted approach expression of these genes, such as *NOD2* can be determined and differences between groups more accurately analysed.

An RNA sample was thawed for each patient, 95 samples passed the quality control for RIN (>9) and RNA concentration (>50ng/µl). The decision was made to run the sample with the lowest concentration in duplicate in order to maximise the chance of useful data. Using the HTG EdgeSeq sample sheet software 96 samples were annotated and assigned to wells on a 96 well plate. RNA was transferred from each sample to the corresponding plate location using a two-person check. Each sample was then diluted to a required concentration determined by the HTG EdgeSeq software and loaded onto the HTG EdgeSeq 96 well plate. The sample plate and the autoimmune panel reagents were loaded onto the HTG EdgeSeq processor, and the 12-hour nuclease protections chemistry run commenced. The sample plate was then frozen at -20^{o}C for 24 hours prior to further processing.

2.6.4 PCR of autoimmune panel product

A PCR reaction using primers designed for the sequencing adaptors and sample barcodes was then commenced. Samples were barcoded using 12 forward (F1-F12) and 8 reverse primers (RA-RH) by well location. The PCR was run for 20 cycles resulting in amplified autoimmune gene targets.

2.6.5 PCR Clean-up

PCR product underwent a clean-up procedure whereby PCR product, magnetic beads and buffer were added to a 96 well plate. Using a magnetic plate stand, PCR product was separated from the solution through binding to magnetic beads. All the solution was then removed, leaving only the beads and attached PCR product. Each sample then underwent two ethanol washes, prior to resuspension of beads and product in Tris buffer. The resuspended PCR product was then separated from the beads through further use of the magnetic stand. Sample was stored at 4^{o}C prior to quantification.

2.6.6 Library quantification

All samples were then quantified using both Qubit fluorometry and qPCR techniques.

2.6.6.1 Qubit

5µL of PCR product was mixed with 195µL of Qubit buffer and reagent. The Qubit instrument was calibrated using 10µL of Qubit concentration standards. Each sample was then analysed on the instrument in duplicate to give two measures of concentration for each sample. A mean of these samples was used to calculate a molarity (in picomoles) for comparison with qPCR quantification. Each measurement can be seen in supplementary data.

2.6.6.2 qPCR

The KAPA SYBR FAST qPCR quantification method and kit was used in line with HTG EdgeSeq recommendations. Briefly, KAPA MasterMix was used in conjunction with forward and reverse primers and ROX high. Each sample was run in triplicate on the sample plate (72 quantifications from 24 samples per plate), with quantification standards and no-template controls run for each qPCR experiment. A total of 4 qPCR runs were performed to quantify samples. Each qPCR plate was run on the Applied Biosystems StepOne qPCR instrument. Data from each run can be seen in supplementary data.

2.6.6.3 Analysis of quantification and standardisation

As per HTG EdgeSeq recommendations quantification of each PCR product sample was finally determined using the qPCR methodology. Output from the Applied Biosystems StepOne qPCR instrument was collated into a single file. Each sample was manually checked for quality assurance 'flags' produced by outlying samples. Were a flag was identified, data for that sample was manually curated to ensure outlying quantification measures did not lead to mis-quantification. A total of 9 adjustments were needed due to outlier flags, where an outlier was found the mean quantification was based on two qPCR runs for that sample, rather than three. Quantification was based on 2 measures for 3 samples on run 1, on 2 measures for 0 samples on run 2, on 2 measures for 2 samples on run 3 and on 2 measures for 4 samples on run 4.

Finally, qPCR quantification data were correlated with Qubit quantification data to ensure there were no outliers requiring further analysis or re-quantification, figure 5.

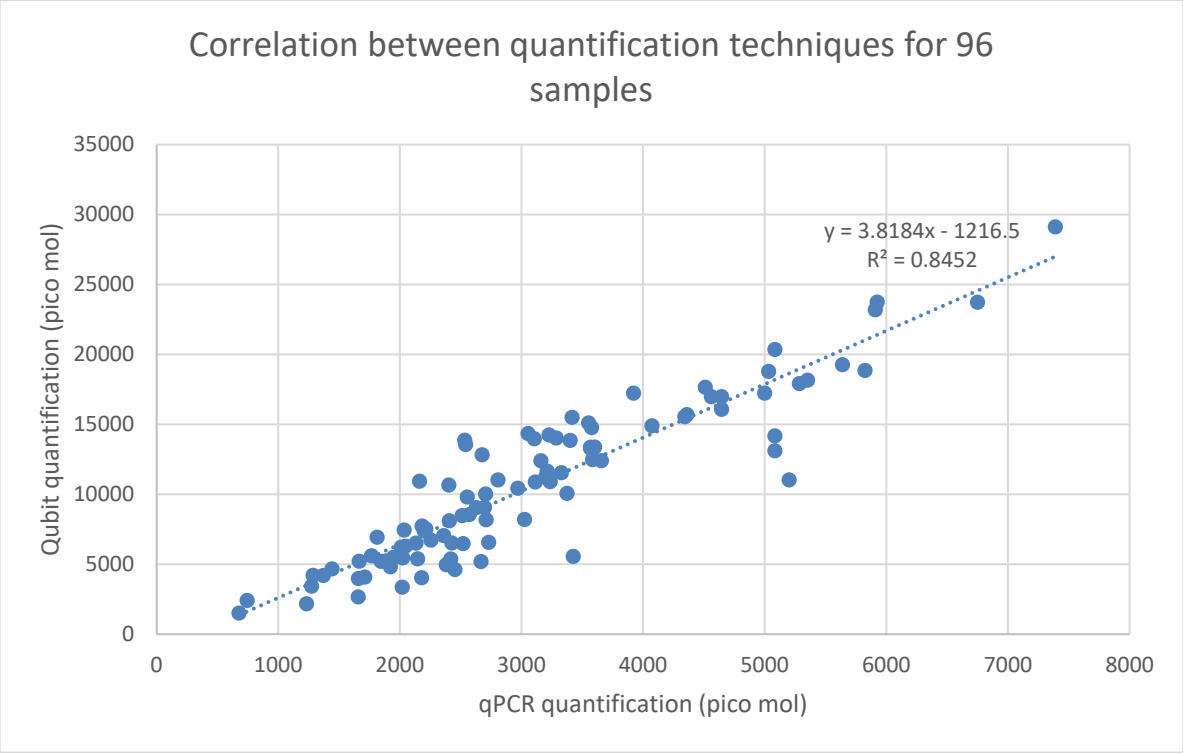

Figure 5- Correlation of Qubit and qPCR RNA quantification techniques demonstrating excellent concordance between samples, $R^2 = 0.08452$.

2.6.7 Sequencing preparation

Based on quantification results each sample underwent a normalisation process to ensure the sample quantity of PCR product is sequenced for each sample. This ensures that each sample has represented by a suitable (>750,000) number of reads. The samples are normalised through a process of dilutions. The quantity of PCR product was very high for our samples, we therefore undertook a 1/10 dilution in Tris for all samples. Following this we used HTG EdgeSeq calculation software to determine a further dilution specific for each sample. Following this further dilution, a specific volume for each samples, as determined by the HTG EdgeSeq calculation software was taken and added to the library for sequencing. These data can be seen in supplementary.

2.6.8 Sequencing

The sample library was prepared for sequencing as directed by HTG EdgeSeq software. In order to denature the library 14.8µl of 2N sodium hydroxide was added to the tube, following by vortexing and spinning down. Following this 1122µl of HT1 buffer was added, alongside 14.8µl of 2N hydrochloric acid. Sample was vortexed. 15.6µl of denatured PhiX was added to the library tube. The library was heat denatured at 98^{o}C for 4 minutes and immediately chilled on ice for 5 minutes. Sequencing was performed on the Illumina NextSeq platform.

2.6.9 Conversion from BCL files to Fastq files

The output files from the local NextSeq run were converted from BCL format to Fastq files using the bcl2fastq software on Iridis 4. Standard parameters were used. Barcode mismatch filter was set to 0 to ensure reads were linked to the correct patient.

Following conversion fastq files were loaded into the HTG EdgeSeq parsing software. Probes sequences were used to identify specific genes. Using barcodes identifying specific patients a gene expression count matrix was constructed for each gene and each patient. These were merged to form a single output file containing all genes and all counts.

2.6.10 Normalisation of data

Three data normalisation methods were employed. Counts per million (CPM), median-ratio normalisation (MN), and quantile normalisation (QN). Briefly CPM employs a simple methodology for standardizing RNA sequencing data through conversion of HTG probe counts into CPM. The CPM method is applied to each individual and does not require inclusion of additional data. MN utilises the median of the ratios of each gene count over the mean across all samples, thus providing a normalisation methodology accounting for variance across multiple samples. QN is the most aggressive normalisation technique employed and removes most technical variation. QN assumes the same distribution of gene expression across samples and therefore if there is expected to be significantly different expression between samples, such as for different tissues, this normalisation is not appropriate. QN uses common all samples distribution determined as the mean of the ordered expression values.

Based on best practice guidelines utilising the HTGEdge sequencing technology, specifically applying the immune-oncology targeted panel we opted for median-ratio normalisation and quantile normalisation [148,149].

2.6.11 Analysis using Reveal software

Downstream analyses of RNA data were performed using HTG Reveal software, detailed description is in chapter 6. Integration of genomic data through regression analyses were performed using SPSS (v25, IBM).

2.7 Microbiome sequencing and analysis

2.7.1 Microbial DNA extraction from intestinal biopsies

Biopsies were collected and stored as previously described. Following appraisal of techniques for extraction from biological samples with low concentration of bacterial DNA we chose the

ultradeep microbiome extraction kit (Molzym). Briefly this technique utilises removal of >95% of host (eukaryote) DNA whilst retaining DNA from bacteria and fungi (prokaryote). This process was assessed on eight rectal biopsies from patients recruited in 2016 using an identical protocol and sample storage method as described above. The SOP for this process can be seen in accompanying materials.

2.7.2 16S sequence analysis

There are two major analysis platforms available for 16S microbiome data. QIIME (now QIIME 2.0) and Mothur. Previous analysis with QIIME 1 was been performed on preliminary data[150], in 2018 the software and analysis pipeline was updated (to QIIME 2.0)-

1. Importing data into an QIIME 'artifact'
2. Upload metadata file (containing patient and sample details)
3. Demultiplex sequences (if required, done by barcoding)
4. Quality control (sequence issues, base calling issues) with DAD2 or Beblur
5. Production of FeatureTable (equivalent of OTU file) and FeatureData (equivalent of RepSeq file)
6. Downstream analysis- phylogenetic trees, diversity analysis, taxonomic classficiation, rarefaction, differential abundance testing

Due to the impact of the COVID-19 pandemic on this project we were unable to sequence the ileal biopsies of recruited patients. We had conducted preliminary analysis on rectal biopsies, confirming the extraction techniques, sequencing methods and data analysis protocols were suitable and yielded valuable results. These initial analyses are available as supplementary data but are not included in this thesis (Supplementary data -> Microbiome, visualisation using https://view.qiime2.org/). We plan to perform the microbiome analyses on the ileal biopsies and integrate with transcriptomic and genomic data, as part of post-doctoral work.

2.8 Data integration- multi-omics

Integration of transcriptomic and genomic data occurred only in Chapter 7 and is detailed here.

2.9 Bioinformatic tools

Analysis has utilised bioinformatic software tools. These are detailed and summarised in table 4.

Table 4- Bioinformatic software tools utilised for analyses of data. Tools are grouped by data analysis type and a brief summary of the function of the software is given.

	Software	**Function**
Genomic	gnomAD v2.1.1[151]	Frequency database used to annotate variants with allele frequency
	CADD v1.5[152]	*In silico* deleterious metric
	GATK v3.7[141]	Analytical software utilised within Southampton WES processing pipeline
	GenePy[153]	*Per gene, per individual*, gene pathogenicity score. Used for integration into downstream analysis
	Integrated genome viewer (IGV) [141]	Tool for visualisation of aligned genomic sequencing data in the form of a BAM file
	ACMG criteria[154]	American College of Medical Genetics guidelines for annotation of variants according to pathogenicity
Transcriptomic	REVEAL[146]	HTG software for analysis of targeted RNA sequencing
	WGCNA (R package)[155]	Weight gene co-expression analytical software for assessment of continuous variables
	CEMItools[156]	Weight gene co-expression analytical software for assessment of categorical variables

	DESeq2[157]	Differential gene expression analysis package
	Kallisto-bustools[158]	Alignment, read filtering, barcode and UMI counting for single cell sequencing
	Scanpy[159]	Single cell sequencing data analyses package
	Scran[160]	Single cell sequencing data normalisation
	BBKNN[161]	Generation of single-cell neighbourhood vidualisation, with data integrated from separate tissue samples
	SingleR[162]	Cell type annotation using gene expression profiles
Other	KEGG[163]	Database of biological systems, including pathways and biological processes
	Reactome[135]	Database providing visualisation, interpretation and analysis of pathways
	HitPredict[164]	Database of protein-protein interactions
	ToppFun[165]	Functional enrichment database integrating transcriptomic, ontological, protein and phenotypic annonation
	EnRichR[166]	Functional enrichment database for analysis of gene sets
	BioPlanet[167]	Software for analysis of pathways utilising gene sets
	Morpheus[168]	Hierarchical clustering software for visualisation of data
Statistical analysis	SPSS	Established statistical software package (IBM)
Microbiome	QIIME 2.0[169]	See microbiome workflow above

	GreenGenes[170]	Bacterial 16S reference database
	iTOL[171]	Phylogenetic tree visualisation software

Chapter 3 Hierarchical Clustering of Clinical Data

Chapter summary- *In this chapter we report and analyse baseline blood results of a cohort of patients with a subsequent diagnosis of PIBD, including the prevalence of normal bloods in CD, UC and IBDU. These data provide an example of the translation of automated and detailed clinical phenotyping to novel findings. The methodology for processing these data has been applied in future chapters to apply clinical data to multi-omic analysis.*

Chapter contributions- *Data were extracted by James Ashton and Florina Borca. All analyses were performed by James Ashton. Hierarchical clustering metrics were calculated by Enrico Mossotto.*

Supplementary data can be found at https://doi.org/10.5258/SOTON/D1657

3.1 Background

Blood tests are a standard, well established and accessible part of the diagnostic work up in children with chronic abdominal symptoms[33,34]. Patients with normal tests are often not referred onto specialist care, and specifically normal inflammatory markers have been considered reassuring[172]. Whilst normal blood tests may be reassuring in children presenting with common, chronic abdominal symptoms, a proportion of patients diagnosed with IBD will present with some or all normal laboratory values[1]. Mack *et al* (2007) described up to 9% of CD patients and 19% of UC patients presenting with normal haemoglobin, erythrocyte sedimentation rate (ESR), platelet count and albumin[173].

Faecal calprotectin (FCp) is a very useful test, although there is not yet a definitive consensus for an abnormal cut-off value[174]. The specificity of FCp ranges from 0.59 at >100µg/g to 0.95 at >800µg/g, with sensitivity at corresponding values displaying an inverse trend, 0.97 at >100µg/g

to 0.73 at >800µg/g[175]. The use of laboratory markers in addition to symptoms for diagnosis of PIBD has been analysed, with FCp proving the most useful and blood tests providing only some additional benefit in a single study[176]. Despite this, blood results remain a vital part of the work-up of children with possible IBD.

We hypothesise that identification of different patient strata based on patterns of blood results alongside observation of enrichment for specific abnormal tests may lead to novel patient subgrouping. We apply hierarchical clustering techniques to patients in order to identify: i) novel subgroupings of children beyond the traditional CD/UC classification; ii) blood test results associated with subsequent diagnosis (CD vs UC vs IBDU).

3.2 Methods

All patients (n=275) diagnosed with PIBD from January 1st 2013 to December 31st 2017 at Southampton Children's hospital were eligible for inclusion in the study. Patients and dates of diagnosis were identified from the prospectively maintained Southampton PIBD database. All patients included were diagnosed in line with the modified Porto criteria.

3.2.1 Clinical data extraction

Results for erythrocyte sedimentation rate (ESR), C-reactive protein (CRP), albumin, haemoglobin, platelets, packed cell volume (PCV), white cell count (WCC) and alanine transferase (ALT) were automatically retrieved from the University Hospital Southampton (UHS) laboratory results system by the data science team at UHS/Southampton Biomedical Research Centre. Where results were not available from UHS the referring hospital records were searched. Where data were missing from both sources patients were included if they had 1 or more blood result. Results were obtained with age and gender specific normal reference ranges as provided by the UHS clinical pathology service or by the referring hospital laboratory, these references ranges are dependent on the assay used within the laboratory and are therefore not applicable to results

beyond the analysing laboratory. There were no differences in the Southampton laboratory gender reference values for the age groups studied (<17 years).

Imaging and faecal calprotectin results were excluded from the analysis as they have not been routinely available to primary or secondary care physicians across the region over the study period.

Blood results were retrieved from 0 to 100 days prior to the day of diagnostic endoscopy (supplementary data). Where multiple blood tests per patient were available the blood results from the first presentation to the paediatric gastroenterology service were used. Where no results were accessible during the allocated time (n=19, 6.9%), patients were excluded from further analysis. All excluded patients were referred directly for endoscopy (from regional clinics run by the Southampton team).

For each blood test result the proportion of patients with an abnormal result was calculated and expressed as a percentage. Statistical analysis was performed using Fisher's exact test, Mann-Whitney U-test and simple linear regression was performed to analyse age at diagnosis (SPSS v24 IBM). Sensitivity for each test to identify patients with IBD was calculated for all blood tests (true positives/(true positives + false negatives).

3.2.2 Statistical analysis and clustering

In order to identify novel groups and enrichment of patients, blood test result data were normalised to a mean value of zero using the standardise function in excel (Microsoft) (based on calculations of standard deviation and mean). These data were then used for production of heatmap and box-whisker plots. The heatmap was grouped using average linkage and Euclidean distance to identify the most similar groups of patients as determined by blood test results using the Morpheus software. Briefly, all normalised data were entered into the Morpheus online application-

1. Similarity (or dissimilarity) between every pair of patients was calculating based on the 'distance' between patients. Here we employed the default Euclidean distance measure. All entered variables were employed to characterise an object prior to clustering.
2. Patients are clustered into a hierarchical tree. Pairs of patients that are close in proximity based on the linkage function, based on distances generated in step one, are placed into binary clusters, these clusters are then grouped into larger clusters until a hierarchical tree is formed. The distance or similarity, between patients is represented by data merges, with higher nodes representing more dissimilar objects.
3. Clusters are then determined by the level at which the hierarchical tree is 'cut' into the groups.

Patients were labelled post-hoc by diagnosis (CD vs UC vs IBDU). Novel groups were interrogated for enrichment of patient diagnosis and underlying drivers of similarity. Statistical analysis for enrichment used Fisher's exact test.

3.3 Results

Two-hundred and fifty-six patients were included in the analysis, 151 with CD, 95 with UC and 10 with IBDU. Median age at diagnosis was 13.48 years, 36.7% (n=94) were female. Not all blood tests results were available for all patients. The mean number of tests per patients was 7.5 (range 2-8). As expected, the CD group was enriched for male patients compared to the UC group (p=0.0092). There was no difference in median age of diagnosis between CD and UC (13.46 years vs 13.61 years). The median time from date of blood result to diagnosis was 8 days (range 0-99 days).

3.3.1 Normal blood tests

When considering all patients with PIBD, 9% presented with all normal blood tests (Table 5), 21.9% of patients presented with normal inflammatory markers (ESR, CRP) and 19.1% of patients presented with a normal full blood count (FBC, consisting of haemoglobin, platelets, WCC, PCV).

	Abnormal CRP	Abnormal ESR	Abnormal WCC	Abnormal Haemoglobin	Abnormal Platelets	Abnormal ALT	Abnormal PCV	Abnormal Alb		Abnormal Inflammatory markers*	Abnormal FBC**	All normal bloods
All PIBD	53.41%	56.44%	22.71%	61.90%	55.56%	7.17%	64.63%	27.98%		78.13%	80.86%	8.98%
Crohn's disease	71.14%	60.33%	24.50%	69.54%	66.89%	2.88%	71.81%	40.69%		84.21%	87.50%	5.26%
Ulcerative Colitis	25.81%	48.05%	16.13%	47.87%	35.11%	14.29%	56.04%	8.70%		65.98%	69.07%	14.43%
P value CD vs UC	**0.00001**	0.11	0.148	**0.0011**	**0.00001**	**0.0017**	**0.017**	**0.00001**		**0.0035**	**0.0005**	**0.02**
Sensitivity for PIBD	*53.41%*	*56.44%*	*22.71%*	*61.90%*	*55.56%*	*7.17%*	*64.63%*	*27.98%*		*78.13%*	*80.86%*	*91.02%*

Table 5- Percentage of patients presenting with abnormal blood tests for all IBD, Crohn's disease and ulcerative colitis. Sensitivity of each blood test for being abnormal in a patient with IBD in this cohort.

3.3.2 Abnormal blood tests

For individual results the most likely tests to be abnormal in PIBD were haemoglobin (61.9%), PCV (64.6%), ESR (56.4%) and platelets (55.6%). CRP was high in 53.4% of patients. Albumin was low in 28% of patients and ALT was high in 7.2% of patients.

The proportion of patients presenting with abnormal blood tests for PIBD, CD and UC can be seen in figure 6.

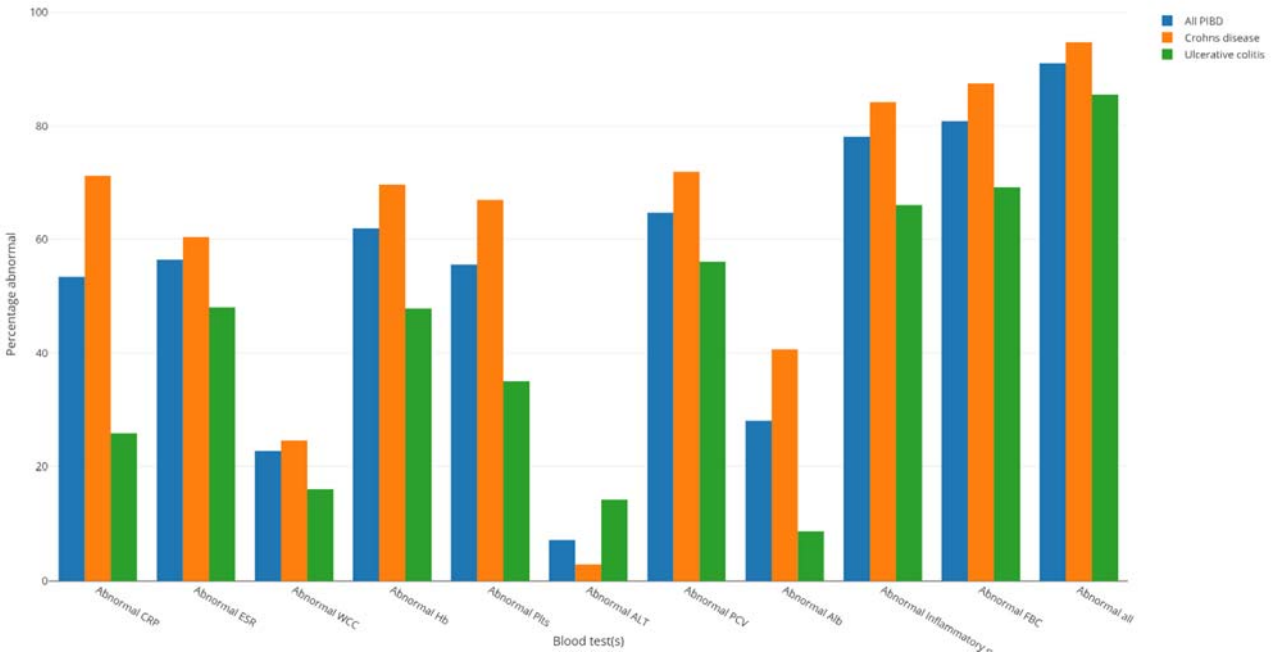

Figure 6- Percentage of patients presenting with abnormal blood tests for all IBD, Crohn's disease and ulcerative colitis. For abnormal inflammatory markers- either CRP or ESR, or both was abnormal. For abnormal FBC- either WCC, Hb, Plts or PCV, or a combination were abnormal. Abnormal all indicates the patient had at least 1 abnormal blood result.

3.3.3 Comparison of Crohn's disease vs Ulcerative colitis

Patients diagnosed with CD were significantly more likely to have abnormal inflammatory markers when compared to UC (UC= 34% normal, CD= 15.8% normal, p= 0.0035). When considering FBC, CD patients were significantly more likely to have abnormal results when compared to UC patients (UC= 30.9% normal, CD= 12.5% normal, p= 0.0005). Considering all blood tests, UC patients were significantly more likely to present with all normal results than CD patients (UC= 14.4% normal, CD= 5.3% normal, p= 0.02). Table 6.

	CRP (mg/L)	ESR (mm/Hr)	WCC (10*9/L)	Haemoglobin (g/dL)	Platelets (10*9/L)	ALT (U/L)	PCV (%)	Albumin (g/dL)
All PIBD including IBDU	13	21	9.2	115	424	13	0.359	35
Crohn's disease	24.5	27	9.2	115	445	12	0.36	32
Ulcerative Colitis	4	12	9.0	117	381.5	16	0.353	38
P value CD vs UC	**0.00001**	**0.0001**	0.168	0.596	**0.0001**	**0.00001**	0.502	**0.00001**

Table 6- Median results for each blood test for all IBD, Crohn's disease and ulcerative colitis.

3.3.4 Blood test median values

Median values were calculated for all tests, and for both CD and UC. Inflammatory markers (CRP and ESR) were significantly higher in CD rather than UC (CRP-median value CD= 24.5mg/L, UC= 4mg/L, p=0.00001, ESR-median value CD= 27mg/L, UC= 12mg/L, p=0.0001).

Platelets were significantly higher in CD compared to UC (median value- CD= 445mg/L, UC= 381.5mg/L, p=0.0001). Albumin was significantly lower in CD (median value- CD= 32mg/L, UC= 38mg/L, p=0.00001. ALT was significantly higher in UC compared to CD (median value- CD= 12U/L, UC= 16U/L, p=0.00001). There was no significant difference for haemoglobin, WCC or PCV. See table 6.

3.3.5 Normalised data analysis and hierarchical clustering

Data for all blood tests was normalised and used to construct a heatmap (figure 7) and box-whisker plot (figure 8). Hierarchical clustering automatically grouped patients by the overall similarity of their blood results, 12 groups were identified, assigned as clusters A-K. Patients were labelled by their diagnosis post-hoc. IBDU occurs throughout the heatmap and does not cluster

together. CD typically clustered together in the presence of low albumin or high inflammatory markers, with UC less likely to form distinctly clusters.

We identified distinct outlying groups, clusters J + K, cluster stability 0.71 and 0.47 respectively, which is highlighted in yellow and characterised by high CRP and low albumin. Cluster B, cluster stability 0.68 of figure 7 and highlighted in green, characterised by normal albumin and low haemoglobin of the heatmap. These groups were enriched for Crohn's disease (85.7%, 12/14 cases, p=0.049) and ulcerative colitis (66.7%, 12/18 cases, p=0.01) respectively.

Interestingly an additional novel group of nine patients (cluster H- cluster stability 0.63), highlighted in pink comprised of 4 Crohn's disease, 3 ulcerative colitis and 2 IBDU patients. This group was characterised by an isolated increase in white blood cell count, representing a novel grouping. See following page.

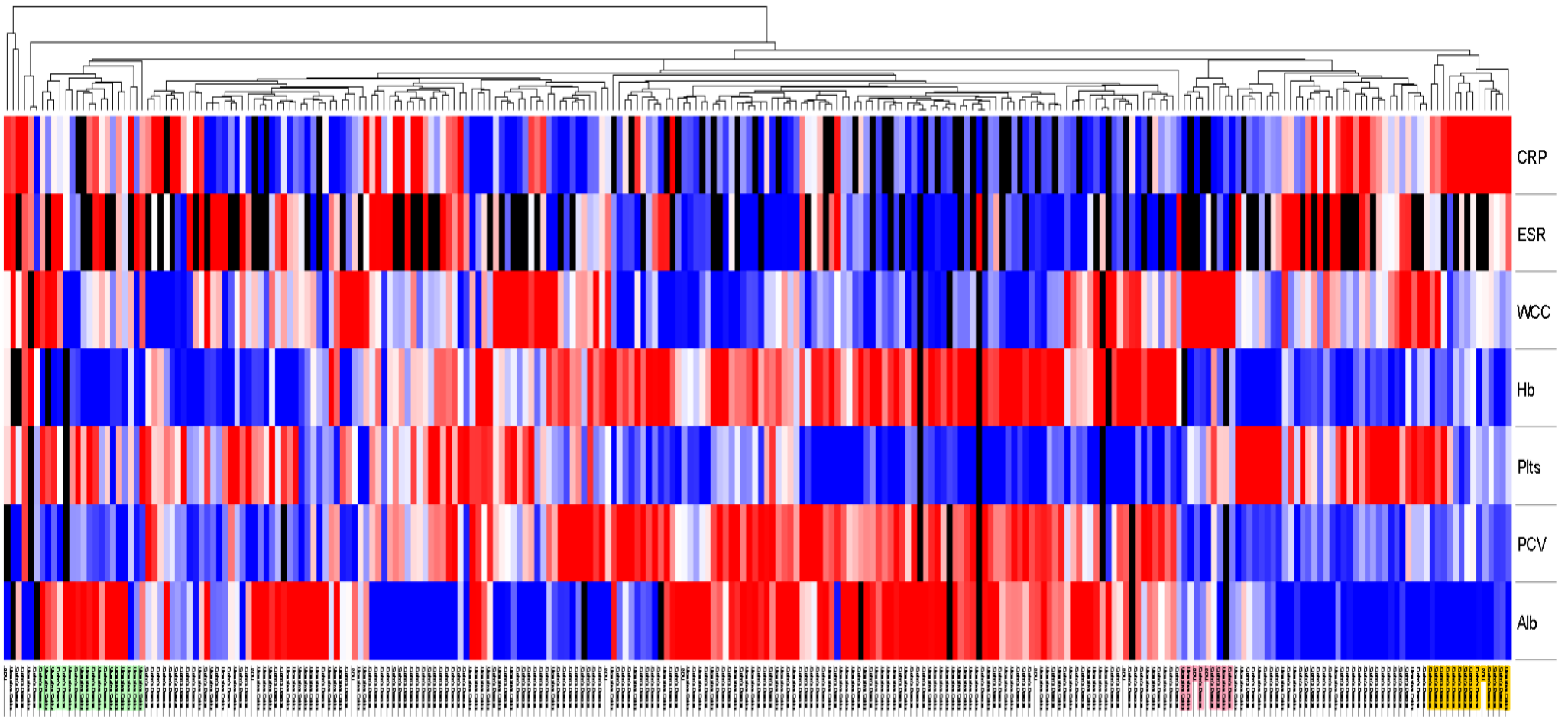

Figure 7- *Normalised blood result data for all 256 patients presenting with IBD. Red indicates a higher value, blue indicates a lower value and white indicates a mean value of 0. Black represents missing data.*

The diagnosis of the patient is annotated on the Y axis. Shorter distances indicate a more similar blood result profile at diagnosis. Groups at the top, clusters J + K, highlighted in yellow, and bottom, cluster B (highlighted in green, characterised by normal albumin and low haemoglobin) of the heatmap. These groups were enriched for Crohn's disease (85.7%, 12/14 cases, p=0.049) and ulcerative colitis (66.7%, 12/18 cases, p=0.01) respectively.

Interestingly an additional novel group of nine patients (cluster H, cluster stability 0.63) was observed (highlighted in pink) comprised of 4 Crohn's disease, 3 ulcerative colitis and 2 IBDU patients. IBDU occurs throughout the heatmap and does not cluster together. Nine patients highlighted in pink cluster due to an isolated increase in white cell count and represent a mix of Crohn's disease, ulcerative colitis and IBDU.

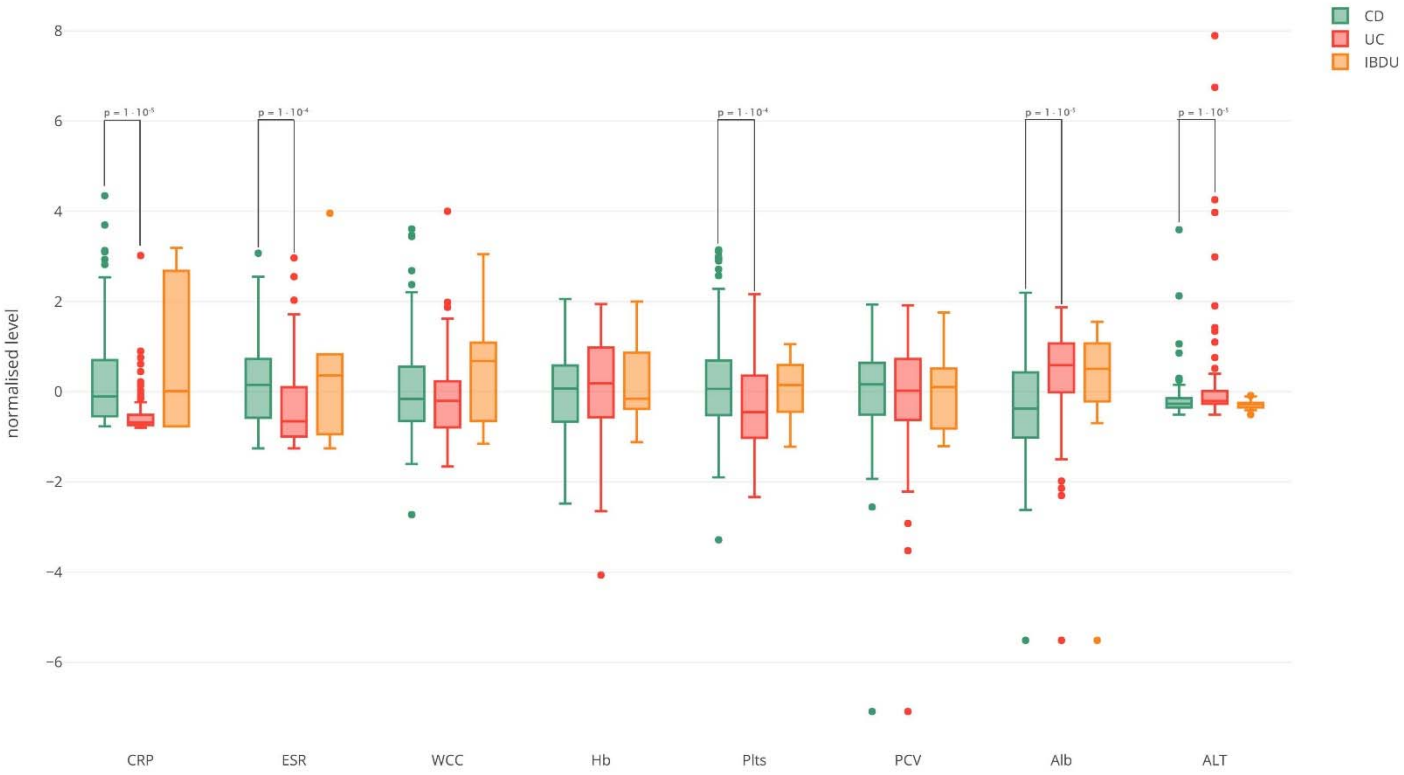

Figure 8- Normalised blood data for 256 patients presenting with IBD. Significant differences between Crohn's disease and ulcerative colitis are indicated on the graph

3.3.6 Sensitivity of blood results

Sensitivity as the proportion of patients presenting with abnormal tests was calculated for all blood tests. Table 5. It can be inferred that in children presenting with chronic abdominal symptoms, individual blood tests have limited utility in the diagnostic work up (sensitivity ranging from 22.71-64.6%). However, the pooled results have a sensitivity of 91%, meaning that 1/11 patients with a subsequent diagnosis of PIBD will present with all normal results, equating to 23 patients in this cohort, over 4 per year.

3.3.7 Age at diagnosis and gender

Blood results with known variation by age or gender (normal values are different for different ages or gender), including haemoglobin, PCV, WCC and albumin were not investigated for association with age at diagnosis or gender.

In all PIBD cases neither CRP nor ESR were significantly correlated with age at diagnosis (CRP- R2=0.011, p=0.070, ESR- R2=0.001, p=0.686). For Crohn's disease age at diagnosis was significantly correlated with CRP value (R2 = 0.031, p = 0.022), older children presented with higher CRP values. This was not seen with ESR. Neither CRP nor ESR significantly correlated with age at diagnosis in UC.

Platelet count was not correlated with age at diagnosis in either PIBD, CD or UC.

Inflammatory markers were compared between males and females, neither CRP nor ESR were significantly different between groups (CRP- male 14mg/L, female = 9mg/L, p=0.055, ESR- male 20.5mm/hr, female 22mm/hr, p=0.72).

3.4 Discussion

These data demonstrate that most patients over a five-year period with a subsequent diagnosis of PIBD will present with at least one abnormal blood test, however one in five patients will present with normal inflammatory markers and 1/11 will present with all normal blood results. Patients with a subsequent diagnosis of CD are significantly more likely to have abnormal results at presentation compared to UC. Normal blood tests, especially normal inflammatory markers, should not preclude from further investigation and referral in children with significant chronic symptoms or a high index of suspicion.

Hierarchical clustering of normalised data revealed novel groups. The main outlying clusters are significantly enriched for either Crohn's disease or ulcerative colitis. An additional cluster of nine patients clustered based on an isolated increase in white cells, this group contained CD, UC and

IBDU and represented a novel cluster, distinct from traditional diagnostic subtypes. Most patients do not cluster into distinctive subgroups and occur throughout the heatmap, with low cluster stability (<0.6). IBDU patients do not cluster together. These data provide an additional framework to help with early differentiation between CD and UC based on both number/pattern of abnormal blood results, the degree of abnormality and enrichment of CD/UC in novel grouping of patients through hierarchical clustering.

This study provides a large cohort of patient blood results at diagnosis, with an extensive number of tests analysed. Other studies have focused on fewer tests, Sabery *et al* (2007) described the proportion of patients with a subsequent diagnosis of PIBD, with abnormal haemoglobin or ESR, as up to 83% of patients[177]. Mack *et al* (2007) detailed blood results at diagnosis of PIBD (ESR, haemoglobin, platelets and albumin) and demonstrated all normal results in 9% of CD and 19% of UC patients, with individual tests, ESR, haemoglobin, platelets and albumin, being normal for all IBD in 18%, 24%, 43% and 50% respectively[173]. Our data show a higher proportion of patients presenting with normal values (seen in 44%, 38%, 54% and 72% of patients) for the respective blood tests. In 1995, Beattie *et al,* in a cohort of 91 children referred with chronic gut symptoms to paediatric gastroenterology, described no CD patients presenting with all normal blood tests and 100% of patients presenting with a raised CRP[172]. In our study, 5% of CD patients had normal results however, 29% had normal CRP and over 1/7 had a normal CRP and ESR. Recent data from Day *et al* detailed ESR, CRP, platelet count and albumin blood results at diagnosis, reporting all normal tests in 13% of patients with CD and 41% of those with UC, highly comparable to our data- 15.8% and 34% respectively[178]. This apparent change in presenting phenotype, with more patients with normal results, may reflect a change in disease type. However it is more likely to reflect the significant increase in incidence and improved identification of disease at an earlier stage[179].

The patient groupings determined by hierarchical clustering in this study present a potential clinical application for these data. Patients with high CRP and low albumin were heavily enriched

for Crohn's disease, whilst those with normal albumin and low haemoglobin were enriched for ulcerative colitis. This allows clinicians to discuss the probable subtype of disease, including treatment implications with families prior to endoscopy, based on blood results alone. However the sensitivity and specificity of these data is not sufficient to use in isolation. It would be interesting to scrutinise patients with these specific patterns of blood markers at diagnosis in a novel independent group of paediatric patients.

Data from 2003 comparing CD with UC showed a similar effect to that seen in our data, ESR and platelets were significantly higher whilst haemoglobin and albumin where significantly lower in CD compared to UC[180]. Our data demonstrated that CRP is significantly higher in CD, but haemoglobin was similar in CD and UC. It is well established that CD is more likely to have a systemic inflammation, which in turn is likely to be associated with more severe blood abnormalities. Younger children presenting with CD were more likely to have a normal/lower CRP value, this may increase the difficulty in making a prompt diagnosis in early-onset PIBD[181].

Significant advances have been made in the diagnosis and treatment of PIBD over the last 20 years. However accessible diagnostic models to classify patient risk have lagged behind. The use of machine learning, multi-omics and artificial intelligence to aid physician diagnosis and stratification of patients is now beginning to occur but only in a research environment[64,118]. Despite this progress the eventual utility to a general paediatrician or primary care physician may be limited, with simple and accessible strategies required to stratify patients for referral. A recent systematic review and meta-analysis detailing the utility of laboratory tests demonstrated that by adding an FCp value to symptoms of PIBD led to a 26% improvement of diagnostic accuracy. In comparison the best blood marker, ESR, was less useful increasing accuracy by only 16%[176]. There is clearly a place for blood results in the diagnostic work up of children with chronic gut symptoms however multiple normal blood tests and a history consistent with IBD should not be ignored. This is especially true in the presence of features such as a family history of IBD, and high

clinical suspicion should trigger the use of additional investigations such as faecal calprotectin or ultrasound.

This study has several limitations; we were unable to include patients who were referred with a possible diagnosis of PIBD whose endoscopy and histology were then normal. The study would benefit from these patients as a control group; however they cannot reliably be identified from the patient record leading to significant concerns over introducing bias. The decision was made to exclude them. We were therefore unable to calculate specificity in addition to sensitivity. In addition, patients without bloods at UHS prior to diagnosis were excluded, reducing the sample size and therefore statistical power. This study benefits from standardised automated blood result data collection and uses age and gender specific normal ranges to interpret the data.

3.4.1 Conclusions

There may be a trend towards patients presenting with more normal blood results, perhaps driven by earlier identification of disease. It is important for general practitioners, general paediatricians and specialist services to keep a diagnosis of IBD in mind even if blood tests, including inflammatory markers, are normal. Use of hierarchical clustering identified groups enriched for Crohn's disease and ulcerative colitis characterised by specific blood results. Clinicians can use this model to help identify the sub-diagnosis (CD vs UC) in patients with PIBD. Future development of a simple risk stratification model, based on symptoms and accessible tests (bloods, FCp etc.) could provide a framework to reduce the diagnostic delay seen in PIBD.

Chapter 4 Monogenic inflammatory bowel disease

Chapter summary- *This chapter summarises the use of whole exome sequencing to diagnose inflammatory bowel disease patients with single-gene disorders. This utilises standardised criteria for classification of variants, Sanger confirmation and segregation analysis. Phenotype-Genotype analysis is conducted utilising GenePy.*

Chapter contributions- *Whole exome sequencing data were processed by Enrico Mossotto. Monogenic IBD genes were identified by James Ashton. Variant call files were analysed by James Ashton alongside application of American college of medical genetic criteria for classification of variants. Genotype-Phenotype analyses were conducted by James Ashton with help from Enrico Mossotto (GenePy score generation).*

Supplementary data can be found at https://doi.org/10.5258/SOTON/D1657

4.1 Background- single gene causes of inflammatory bowel disease

Inflammatory bowel disease (IBD) is a chronic, relapsing and remitting disease characterised by intestinal inflammation. Most patients with IBD harbour an underlying genetic risk impacted upon by environmental factors, including the microbiome[182]. To date, in excess of 230 genes have been associated with IBD, mostly through genome-wide association studies (GWAS)[59,71]. The first locus implicated in the risk of developing disease was on chromosome 16 and was mapped to *NOD2* in the early 2000s[73,74,183]. There is limited data implicating homozygote and compound

heterozygote (the presence of two different alleles of a gene, one on the maternal chromosome and one on the paternal chromosome) *NOD2* variants as disease-causing in an autosomal recessive inheritance pattern[85,184]. The success of prospective projects based on microbiome and RNA sequencing data, such as PROTECT, has brought into focus the need for improving predictive algorithms by also utilising precise genetic diagnoses[120,185].

High throughput next generation sequencing (NGS) technologies are powerful for the detection of genetic conditions. NGS is already being routinely exploited in mainstream diagnostics of rare disease to substantial patient benefit[115]. As yet, NGS technology has seen little clinical implementation in complex diseases such as IBD [181]. However it has aided discovery of IBD risk genes and identified precise causative variants, alongside informing genotype-phenotype correlations[186–188]. Molecular diagnoses using NGS relies on accurate clinical phenotyping and functional assessment of mutations. *De novo* and homozygous recessive inheritance is most easily detected but detection of compound heterozygosity is more difficult[189].

The vanguard of NGS application in IBD is in the identification of a rare subset of conditions that are Mendelian disorders, masquerading as IBD[17,190]. These are a group of diseases (currently underpinned by variation in 68 genes) typically detected in very early onset IBD (VEOIBD) with severe and atypical features[16,17]. Monogenic forms of IBD are often the manifestation of an underlying immune deficiency or epithelial barrier dysfunction, and have specific management considerations[16].

This chapter aims to apply exome sequencing to a cohort of typical paediatric IBD patients, to identify clinically relevant variants within monogenic IBD genes utilising standard guidelines and correlate with patient phenotype. Furthermore, we apply a novel *per gene* deleteriousness score to assess the contribution of monogenic variation to disease phenotype.

4.2 Methods

Patients were recruited from the Wessex regional paediatric IBD service at Southampton Children's Hospital to the genetics of paediatric IBD study (2010 to present). The eligibility criteria for recruitment was a confirmed histological diagnosis of either Crohn's disease (CD), ulcerative colitis (UC) or IBD unclassified (IBDU), in line with the Porto criteria, and age less than 18 years[32].

4.2.1 DNA Extraction

Patient DNA was extracted from peripheral venous blood samples collected in EDTA using the salting-out method, or from saliva, as previously described[138].

4.2.2 Whole Exome Sequencing (WES) Data Processing

Raw fastq sequencing data from patients with paediatric-onset IBD were processed using our in-house pipeline[153]. VerifyBamID was utilised to check the presence of DNA contamination across the cohort[139]. Alignment was performed against the human reference genome (hg19 assembly) using BWA-mem[140] (version 0.7.12). Aligned BAM files were sorted and duplicate reads were marked using Picard Tools (version 1.97). Following GATK v3.7[141] best practice recommendations[142] variants were called using GATK HaplotypeCaller to produce a gVCF file for each sample and later jointly genotyped.

Annotation of this composite file applied Annovar v2016Feb01 using default databases refSeq gene transcripts (refGene), deleteriousness scores databases (dbnsfp33a, CADD 1.3 and DANN), dbSNP147 and the human genetic mutation database (HGMD Pro 2018) flat file[143]. Variant allele frequencies were sourced through the genome aggregation database (gnomAD)[144]. HaplotypeCaller default settings were utilised corresponding to variants with a minimum Phred base quality score of 20 being called.

4.2.3 Monogenic IBD Gene List

A list of 68 genes previously implicated in monogenic IBD was established, table 7. This list combined genes reported by Uhlig et al (2014) (n=50), Uhlig et al (2017) (n=15), Girardelli et al (2018) (n=1) and through direct correspondence with the International Early-Onset Paediatric IBD Cohort Study consortium (NEOPICS) (n=2)[16,190,191]. The reported inheritance pattern for each monogenic disease gene was determined as either autosomal dominant (AD), autosomal recessive (AR) or X-linked (XL). The *NOD2* inheritance pattern was treated as autosomal recessive (AR)[184,191,192].

Table 7- Monogenic IBD genes used in the analysis. GenePy scores were generated for all but one gene, NCF1.

Gene name	GenePy score generated?
ADA	Yes
ADAM17	Yes
AICDA	Yes
ANKZF	Yes
ARPC1B	Yes
BTK	Yes
CARD9	Yes
CD3y	Yes
CD40LG	Yes
COL7A1	Yes
CYBA	Yes
CYBB	Yes
DCLRE1C	Yes
DKC1	Yes
DOCK	Yes
EPCAM	Yes
FERMT1	Yes
FOXP3	Yes
G6PC3	Yes
GUCY2C	Yes
HPS1	Yes
HPS4	Yes
HPS6	Yes
HSPA1L	Yes
ICOS	Yes
IKBKG	Yes

IL10	Yes
IL10RA	Yes
IL10RB	Yes
IL21	Yes
IL2RA	Yes
IL2RG	Yes
ITCH	Yes
ITGB2	Yes
LIG4	Yes
LRBA	Yes
MASP2	Yes
MEFV	Yes
MVK	Yes
NCF1	No*
NCF2	Yes
NCF4	Yes
NOD2	Yes
PIK3CD	Yes
PIK3R1	Yes
PLCG2	Yes
POLA1	Yes
PTEN	Yes
RAG2	Yes
RTEL1	Yes
SH2D1A	Yes
SKIV2L	Yes
SLC37A4	Yes
SLC9A3	Yes
SLCO2A1	Yes
STAT1	Yes
STXBP2	Yes
TGFBR1	Yes
TGFBR2	Yes
TNFAIP3	Yes
TRIM22	Yes
TRNT1	Yes
TTC37	Yes
TTC7A	Yes
WAS	Yes
XIAP	Yes
ZAP70	Yes
ZBTB24	Yes

* Scores were not generated as gene is not represented by one or more exon-enrichment capture kits. GenePy score calculation can be applied only to variants/genes equally covered by all exon-enrichment kit to ensure a fair comparison of individuals.

4.2.4 Variant Filtering

A total of 1405 high quality variants (PHRED >20) were called across the 68 monogenic IBD genes in 401 paediatric IBD patients. A crude preliminary filter was applied in order to exclude variants with no prior evidence for causality in publicly available databases (HGMD Pro 2018 and ClinVar 2018), or those that are common and have *in silico* evidence of being benign[143,193]. Variants with the following annotation in HGMD and/or ClinVar were retained for further investigation:

1. HGMD- Disease-associated polymorphism with supporting functional evidence- *DFP* or disease causing mutation-*DM* or probable/possible pathological mutation- *DM?*
2. ClinVar- *Pathogenic, Likely Pathogenic*

Any variants fulfilling these criteria were scrutinised to confirm their pathogenic status was in the context of IBD, or monogenic disorders with bowel inflammation, while variants achieving pathogenic status due to an unrelated clinical phenotype were excluded.

As HGMD and ClinVar fail to annotate a subset of variants, a second filtering strategy was applied. Variants without any HGMD or ClinVar annotation were retained based on the following criteria: 1) Coding context - (ExonicFunc.knownGene) 'Exonic' or 'Splicing' AND; 2) CADD Phred score > 20 AND; 3) gnomAD 'all genomes' frequency <0.01 or Novel.

Variants withstanding the filtering strategies above were only retained if they were inherited in the correct zygosity to be disease causing. For genes reported as AD, heterozygous variants were retained; for genes reported as AR, homozygous variants were retained and; for genes reported as XL, hemizygous males (one allele on the X chromosome in males) or homozygous females were retained. Patients harbouring two or more different variants within the same gene through were tested for compound heterozygosity using Sanger sequencing of the proband and parental DNA. Confirmed compound heterozygous variants were retained (see supplementary dataset 2).

This substantially reduced the number of patients and variants that warranted close scrutiny for consistency with American College of Medical Genetics (ACMG) guidelines[154].

4.2.5 Literature review for functional validation

An independent literature review was conducted to collate validated functional evidence for all 35 variants (supplementary data). Validated functional evidence was defined as one or more report(s) describing reduced/absent protein function including impact on downstream signalling/protein expression, nonsense mediated decay or deletions. These data were used to annotate each variant according to ACMG criteria for pathogenicity. Each patient underwent final classification to determine if their variant profile fulfilled criteria for 'pathogenic' or 'likely pathogenic' according to ACMG rules for combing criteria to classify sequence variants[154].

4.2.6 Phenotypic Characterisation

In depth, longitudinal, clinical phenotyping was extracted for all patients in the study including diagnostic and follow-up information. Phenotypic characteristics were transformed to binary or continuous data for use in regression analyses. Follow-up duration was calculated for each patient based on date of diagnosis and last recorded clinical contact.

4.2.7 Application of GenePy *in-silico* Score

GenePy was utilised in these analyses[153]. GenePy incorporates biological information on variant deleteriousness, for this study DANN deleteriousness metric was used[145]. All variants meeting minimum genotyping quality (GQ > 20) were retained for GenePy using vcftools. As GenePy scores can be applied in a case-control comparison within ethnic subgroups, *Peddy* software was used to infer relatedness and ethnicity (with a probability >90%) for all IBD patients[194]. A cohort of 173 non-IBD Caucasian individuals (from the EUCLIDS consortium) for whom WES data was available were utilised as controls.

GenePy scores are quantitative values that follow a *Poisson* distribution whereby for any given gene, most patients have a score close to zero and high scores are rare. It is expected that most patients will have scores in the same range as controls with a small subset of patient incurring

high scores. It is possible to assess evidence for gene causality by selecting the most extreme scores in right tail of the GenePy distributions in cases and compare to the same proportion in controls using a one-tailed Mann-Whitney U-test.

Statistical significance was corrected for multiple testing using the false discovery rate (FDR). Enrichment for a diagnosis of either CD or UC was assessed using Fisher's exact test. Forward stepwise linear regression was performed using R (v3.6.0) and SPSS (v24, IBM) software.

4.2.8 Tetratricopeptide Repeat Domain 7A (*TTC7A*) gene

As part of the overall monogenic project we identified several patients harbouring Tetratricopeptide Repeat Domain 7A (*TTC7A*) gene variants on a single haplotype, confirmed by segregation analysis. A recent report identified these exact *TTC7A* variants in compound heterozygosity in a patient presenting with enteropathy and a common variable immunodeficiency (CVID) phenotype[195]. TTC7A deficiency has been described as a cause of monogenic IBD in over 50 patients, in an autosomal recessive inheritance pattern[196]. Multiple variants have been demonstrated as causal, presenting with a varied phenotype involving inflammatory, gastrointestinal and immunological manifestations[196]. Typically, nonsense variants are fatal within a year of birth, whereas missense mutations may lead to milder disease and patients surviving into adulthood[196]. The role of variation within TTC7A in non-monogenic forms of IBD is less clear. TTC7A is functionally key as a scaffolding and chaperone protein in processes crucial for normal intestinal epithelial cell development[196]. It may prove that milder missense mutations are risk variants for polygenic forms of IBD, although this has not been substantiated.

In this part of study, we describe the two variants observed in *trans* by Lawless et al in five of our patients, always observed in *cis*. Additionally, through application of GenePy, we assess the evidence for variation in TTC7A contributing to polygenic IBD.

4.3 Results

Four-hundred and one patients were included in the analysis. Mean age at diagnosis was 11.92 years (range 1.3-17.39), 40.9% were female and 64.8% had a diagnosis of CD. Children diagnosed before the age of 6 years (VEOIBD) accounted for 7.5% of patients (n=30) and a further 17.2% (n = 69) were diagnosed with early-onset IBD (EOIBD), aged 6 or older and less than 10 years. The remaining 75.3% patients (n=302) were diagnosed between the age of 10 and 18 years and designated paediatric onset IBD (POIBD) (Table 8). The median follow-up time for the entire cohort was 4.6 years (range 0.15 - 17.7).

Due to the COVID-19 pandemic we were unable to fully apply monogenic diagnostic criteria to all 501 IBD patients within our cohort. Due to laboratory closures and staff reassignment we were unable to perform Sanger confirmation and segregation analysis on 100 patients. We therefore report the data for 401 patients for which full analysis was undertaken.

Table 8- Demographic characterisation of patient cohort. VEOIBD- very early onset inflammatory bowel disease, <6 years. EOIBD- early onset inflammatory bowel disease, ≥6 <10 years. POIBD- paediatric onset inflammatory bowel disease, ≥10 <18 years

	All patients (%)	VEOIBD (%)	EOIBD (%)	POIBD (%)
Number of Patients	401	30	69	302
Mean Age at Diagnosis (range)	11.92 (1.3-17.39 years)	NA	NA	NA
Number female (%)	164 (40.9%)	12 (40%)	35 (50.7%)	117 (38.7%)
Crohn's disease (%)	259 (64.6%)	13 (43.3%)	48 (69.6%)	198 (65.6%)
Ulcerative colitis (%)	125 (31.2%)	15 (50%)	17 (24.6%)	93 (30.8%)
IBDU (%)	17 (4.2%)	2 (6.7%)	4 (5.8%)	11 (3.6%)

4.3.1 ACMG 'Pathogenic' or 'Likely Pathogenic' Monogenic IBD Gene Variants

Initial filtering excluded 1,345 variants across 312 patients. Subsequent variant confirmation by zygosity and Sanger sequencing excluded 27 patients and application of ACMG guidelines excluded a further 16 patients, figure 9. Twenty-nine variants fulfilled ACMG standards to be classified as 'Pathogenic' or 'Likely Pathogenic' across 46 patients (11.5% of the cohort) and are discussed in detail below (Table 9).

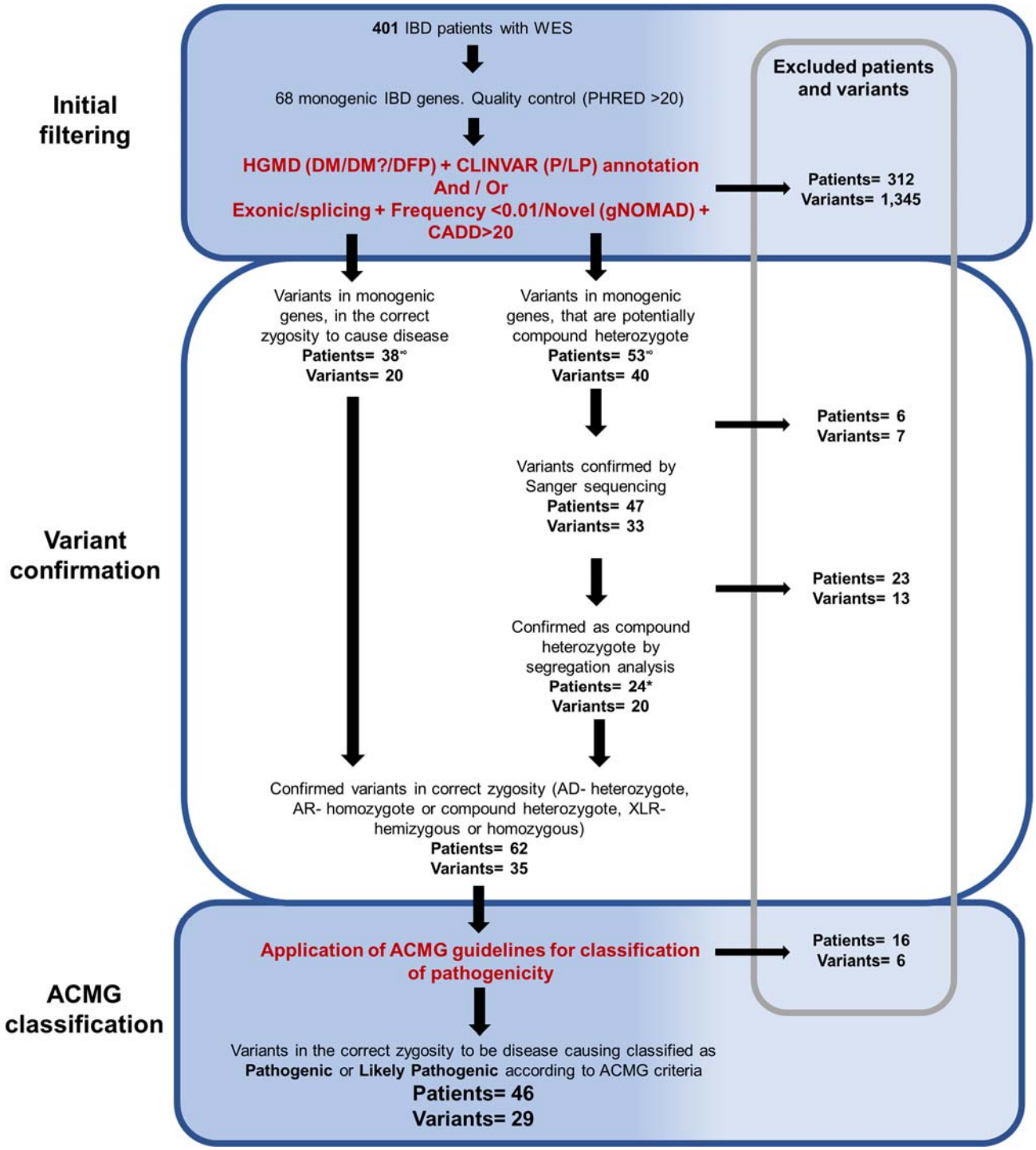

*Figure 9- Flowchart of variant filtering detailing variant/patient exclusions at each filtering stage. Two patients appear in both variant confirmation pathways (correct zygosity and potential compound heterozygote) of the flowchart. *Includes One patient (harbouring TRIM22 R317K/R442K) who was assumed to be compound heterozygote but segregation analysis was not possible due to lack of parental DNA*

Blank page

Table 9- Genetic and phenotypic characterisation of 29 variants across 46 patients with 'Pathogenic' or 'Likely Pathogenic' monogenic IBD gene variants.

Patient	Type of defect	Gene	Chromosome	Variant(s)	ACMG individual variant classification	ACMG pathogenicity classification	Sex	Age at diagnosis	Follow-up time (years)	Intestinal findings					Treatments required			Extraintestinal/Additional findings		
										CD	UC	Disease location at diagnosis (Paris Classification)	Stricturing disease	Perianal fistula/abscess	Thiopurine use	Monoclonal therapy	Surgery	Skin lesions	Liver disease	Other
1	T + B cell defects	*CD40LG*	X	G219R	S	LPa	M	15.3	1.25	X		L2 + L4a	-	-	X	-	-	-	-	-
2		*CD40LG*	X	G219R	S	LPa	M	8.0	6.5	X		L2	-	-		-	-	-	-	Mild disease- no medications
3		*CD40LG*	X	G219R	S	LPa	M	12.2	6.15	X		L3	-	-	X	I, A	-	-	-	-
4		*CD40LG*	X	G219R	S	LPa	M	14.0	2.05		X	E2	-	-		-	-	Eczema	-	-
5		*CD40LG*	X	G219R	S	LPa	M	13.3	5.3		X	E4	-	-	X	I	-	Eczema	-	Asthma
6		*WAS*	X	P460S	S	LPa	M	11.5	7.95		X	E4	-	-	X	I, A	SC	-	PSC	Recurrent infections, intermittent thrombocytopenia
7		*WAS*	X	P460S	S	LPa	M	14.9	5.6		X	E4	-	-	X	-	SC	-	PSC	-
8[Ŧ]		*WAS*	X	E131K	S	LPa	M	14.3	10.4	X		L2 + L4a	-	X	X	-	P	-	-	Also has STAT1 variant
9		*WAS*	X	E131K	S	LPa	M	13.5	5.4	X		L3	X	-	X	I	-	-	-	-
10		*DKC1*	X	UTR5 142C>G	S	LPa	M	13.6	10.1		X	E4	-	-	X	-	-	-	-	-
11		*DKC1*	X	UTR5 142C>G	S	LPa	M	9.2	6.4	X		L1	-	-	X	-	-	-	-	Asthma, oral disease
12		*DCLRE1C*	10	G153R	M	LPa	M	10.7	5.95	X		L3	-	-	X	I	-	-	-	-
				P171R	S															
13		*DCLRE1C*	10	G153R	M	LPa	F	13.2	1.9		X	E4	-	-	X	-	-	-	-	-
				P171R	S															
14	Auto-inflammatory	*XIAP*	X	T470S	S	LPa	M	4.9	4.35	X		L1 + L4a	-	X	X	-	P	-	-	-
15	Phagocytic	*NCF1*	7	R90H	S	Pa	M	2.9	0.9		X	E4	-	-	X	-	-	-	-	-

				R90H	S														
16		*NCF2*	1	H389Q	S	Pa	M	11.0	4.15	X		L3 + L4a	-	-	X	I	-	-	-	Aplastic anaemia, initially abnormal liver function- now resolved
				H389Q	S															
17		*NCF2*	1	H389Q	S	LPa	F	13.2	9.9		X	E2	-	-	X	-	-	-	-	Coeliac disease
				R395Q	M															
18		*NCF2*	1	H389Q	S	Pa	M	14.8	2.0	X		L3 + L4a	-	-	X	I	-	-	-	-
				N419I	S															
19	Immunoregulation	*TRIM22*	11	R321K	S	Pa	M	15.3	7.45	X		L3 + L4a	-	-	X	A	-	-	-	-
				R321K	S															
20		*TRIM22*	11	R321K	S	Pa	M	12.2	6.15	X		L3 + L4a	X	-	X	I	-	-	-	-
				R442C	S															
21*		*TRIM22*	11	R321K	S	Pa	F	14.1	1.75		X	E4	-	-	X	-	-	-	-	-
				R442C	S															
22		*TRIM22*	11	R321K	S	Pa	M	2.4	5.15	X		L2	-	X	X	I	P	-	-	-
				S244L	S															
23		*TRIM22*	11	R321K	S	Pa	F	9.6	9.3	X		L2	-	-	X	-	SC	-	-	Epilepsy
				P484S	M															
24		*STAT1*	2	V266I	S	LPa	M	5.5	10.1		X	E2	-	-	X	-	-	-	-	-
8[F]		*STAT1*	2	V266I	S	LPa	M	14.3	10.4	X		L2 + L4a	-	X	X	-	P	-	-	Also has WAS variant
25		*STAT1*	2	V266I	S	LPa	F	12.6	4.8	X		L3	-	-	-	-	-	-	-	Hypothyroidism, Mild disease- no medication
26		*MASP2*	1	D120G	S	Pa	F	10.7	1.7		X	E1	-	-	X	I	-	-	-	Mild left sided disease consistent with MASP2 deficiency
				D120G	S															
27		*NOD2*	16	R702W	S	Pa	F	16.0	3.42	X		L1	X	-	X	-	RH	-	-	-
				R702W	S															
28		*NOD2*	16	R702W	S	Pa	F	14.5	6.35	X		L3	X	-	X	-	RH	-	-	-
				R702W	S															

29		*NOD2*	16	R702W	**S**	**Pa**	**F**	7.4	4.3	**X**		**L1**	**X**	-	**X**	**I**	**RH**	-	-	-
				R702W	**S**															
30		*NOD2*	16	R702W	**S**	**Pa**	**F**	9.8	1.6	**X**		**L3 + L4a**	**X**	-	**X**	**I**	**RH**	-	-	-
				R702W	**S**															
31		*NOD2*	16	R702W	**S**	**Pa**	**M**	5.8	6.1	**X**		**L3 + L4a**	-	-	**X**	**I**	-	-	-	-
				A755V	**S**															
32		*NOD2*	16	R702W	**S**	**Pa**	**F**	15.5	0.3	**X**		**L1**	**X**	-	**X**	**I**	-	**Eczema**	-	-
				A755V	**S**															
33		*NOD2*	16	R702W	**S**	**Pa**	**F**	9.7	10.7	**X**		**L3**	**X**	-	**X**	-	**RH**	-		**Asthma**
				V955I	**S**															
34		*NOD2*	16	R702W	**S**	**Pa**	**M**	5.9	5.8	**X**		**L2**	-	-	-	-	-	-	-	**Mild disease**
				V955I	**S**															
35		*NOD2*	16	R702W	**S**	**Pa**	**F**	9.4	6.05	**X**		**L3**	-	-	**X**	**I**	-	**Psoriasis**	-	-
				V955I	**S**															
36		*NOD2*	16	R702W	**S**	**LPa**	**M**	15.7	1.4	**X**		**L3 + L4a**	**X**	-	**X**	**I**	**RH**	-	-	**Asthma**
				N852S	**Su**															
37		*NOD2*	16	R702W	**S**	**Pa**	**F**	14.5	2.7	**X**		**L3 + L4a**	**X**	-	**X**	**I**	**RH**	-	-	
				A1007fs	**VS**															
38		*NOD2*	16	R702W	**S**	**Pa**	**M**	11.5	2.55	**X**		**L1**	**X**	-	**X**	**I**	**RH**	**Eczema**	-	
				A1007fs	**VS**															
39		*NOD2*	16	V955I	**S**	**Pa**	**F**	15.1	8.0	**X**		**L3 + L4a**	**X**	-	**X**	-	**RH**	-	-	
				A1007fs	**VS**															
40		*NOD2*	16	R702W	**S**	**LPa**	**M**	13.3	1.85	**X**		**L3**	**X**	-	**X**	**I**	**RH**	**Eczema**	-	**Asthma**
				H352R	**Su**															
41		*NOD2*	16	A755V	**S**	**Pa**	**M**	15.0	5.9	**X**		**L3**	**X**	-	**X**	-	**RH**	-	-	
				G908R	**S**															

42		*NOD2*	16	V955I	**S**	**Pa**	**M**	13.1	2.6		**X**	**E3**	-	-	**X**	**I**	-	-	-	-
				G908R	**S**															
43		*NOD2*	16	V955I	**S**	**LPa**	**F**	11.7	3.45	**X**		**L1**	-	-	**X**	-	-	**Eczema**	-	**Asthma**
				R744W	**Su**															
44		*NOD2*	16	V955I	**S**	**LPa**	**M**	13.6	5.7	**X**		**L3**	**X**	**X**	-	-	**RH, P**	-	-	-
				E963G	**M**															
45		*NOD2*	16	D824N	**Su**	**LPa**	**M**	9.6	7.0	**X**		**L1**	-	-	**X**	-	-	-	-	-
				G908R	**S**															
46		*NOD2*	16	G908R	**S**	**LPa**	**M**	15.0	1.7	**X**		**L2 + L4a**	-	-	-	-	-	-	-	-
				R708H	**Su**															

*Segregation not performed due to lack of parental DNA; ₸ Patient #8 is hemizygous for both a *WAS* and *STAT1* variant

I- infliximab, A- adalimumab, SC- subtotal colectomy, P- perianal procedure (drainage or seton), RH- right hemicolectomy, dash (-) signifies absence of that feature, data were available for all patients.

ACMG individual variant classification- evidence of pathogenicity- VS- very strong, S- strong, M- moderate, Su- supporting

ACMG classification of pathogenicity- Pa- pathogenic, LPa- likely pathogenic

Pathogenic' or 'Likely Pathogenic' variants were observed in 16.7% of patients with VEOIBD, 11.6% of EOIBD patients and in 10.9% of those with POIBD (Supplementary data provides precise variant annotation). Recurrent variants were observed in *NOD2* (20 patients), *TRIM22* (5 patients), *CD40LG* (5 patients), *WAS* (4 patients), *NCF2* (3 patients), *STAT1* (3 patients), *DKC1* (2 patients) and *DCLRE1C* (2 patients). One patient was identified with variant(s) in each of *XIAP*, *NCF1* and *MASP2.* A single patient harboured a hemizygous variant in each of *WAS* and *STAT1* genes.

Twenty-three patients had 'Pathogenic' or 'Likely Pathogenic' compound heterozygous variants confirmed through Sanger sequencing in *DCLRE1C*, *NCF2*, *TRIM22* and *NOD2*. One additional patient (harbouring *TRIM22* R317K/R442K) was assumed to be compound heterozygote but segregation analysis was not possible due to lack of parental DNA. Without exception all *potential* compound heterozygote variant pairs within *NOD2, TRIM22* and *DCLRE1C* (representing 16, 3 and 2 patients respectively) was confirmed following segregation analysis. Conversely, potential compound heterozygote variants in *NCF2* correctly segregated in only two out of eight patients where failure to segregate was consistently due to the *NCF2* P454S variant co-occurring on the same parental haplotype as the H389G variant.

4.3.2 Phenotypic Characteristics of Monogenic Variants

Phenotypic characteristics of the 46 patients with a 'Pathogenic' or 'Likely Pathogenic' monogenic IBD variant(s) are detailed in Table 9.

4.3.2.1 *NOD2* variants - a monogenic stricturing disease phenotype

Twenty patients (5% of all IBD) harboured one or more of 11 variants consistent with an AR pattern of inheritance, 19/20 (95%) patients had a diagnosis of CD, representing 7.3% of CD patients. A novel variant (E963G) predicted to be highly deleterious (CADD 27.3) was observed in a single patient

In the 19 CD patients with 'Pathogenic' or 'Likely Pathogenic' *NOD2* variants, 13 had stricturing disease (68.4%). Stricturing disease behaviour was seen in 38/240 (15.8%) of the remaining CD patients translating to an odds ratio (OR) of 11.52 (relative risk, RR 4.32) in patients with monogenic *NOD2* CD (χ^2=30.3, p=2.0 x 10^{-6}). To assess whether this stricturing phenotype was solely a function of disease location, we tested the rate of stricturing disease in monogenic *NOD2*-related disease patients with ileal location compared to those non-*NOD2* patients with ileal location. Where approximately 20% of non-*NOD2* CD patients with ileal disease developed strictures, 70% of patients with 'Pathogenic' or 'Likely Pathogenic' *NOD2* variation and ileal location were subsequently diagnosed with stricturing disease (χ^2=20.4, p=6.0 x 10^{-6}, Supplementary data).

Patients with monogenic *NOD2*-related disease were at significantly increased risk of undergoing intestinal resection (right hemicolectomy). Surgical resection had occurred in 12/19 (63.2%) monogenic *NOD2* CD patients, compared to 33/259 (13.8%) non-*NOD2* CD patients (OR 10.75, RR 4.59, χ^2=29.8, p=4.9 x 10^{-8}, Supplementary data).

4.3.2.2 TRIM22 variants- severe variable disease phenotype

All five patients with 'Pathogenic' or 'Likely Pathogenic' *TRIM22* variants had at least one copy of the R321K variant, with a single patient harbouring this variant in homozygote form. A variable but severe disease phenotype was seen in all five patients. Patient #19 had a moderate-severe disease course requiring treatment with anti-TNF monoclonal therapy but no fistulating or stricturing disease emerged during the follow-up period. Patient #20 was diagnosed with CD aged 12 years, was treated with monoclonals but developed a stricturing phenotype within 2.5 years of follow-up. Patient #21 was diagnosed aged 14 years with UC and had a mild disease phenotype requiring 5-ASA and thiopurine treatment. Patient #22 had very early onset CD, a severe fistulating perianal phenotype requiring multiple surgical procedures and anti-TNF therapy consistent with the phenotypic spectrum previously reported in three *TRIM22* cases[197]. Patient

#23 was diagnosed with CD at 9 years of age and a severe disease course leading to subtotal colectomy for refractory disease aged 11 years.

4.3.2.3 WAS variant (P460S) - severe ulcerative colitis with liver disease

Of the four patients with 'Pathogenic' or 'Likely Pathogenic' *WAS* alleles, two patients (hemizygous for P460S) presented with a markedly distinct phenotype. Patient #6 was diagnosed with severe and extensive UC aged 11 years, requiring an early colectomy within 2 years and a subsequent diagnosis of primary sclerosis cholangitis (PSC). Intermittent thrombocytopenia and recurrent infections was recorded throughout their disease course consistent with the previously described phenotype for this variant in Wiskott-Aldrich Syndrome[198,199]. Patient #7 was diagnosed aged 14 years and also had severe UC refractory to treatment that requiring a colectomy. This patient was also diagnosed with PSC.

4.3.2.4 Additional monogenic variants

Patients with *XIAP*, *MASP2* and *NCF2* 'Pathogenic' or 'Likely Pathogenic' variants had a phenotype largely consistent with previous reports for those genes[200–202], whereas those with *CD40LG, DKC1, DCLRE1C, NCF1 and STAT1* variants were more heterogenous in their clinical profile.

4.3.3 Monogenic Genes Harbour Significantly Higher Mutation Burden in IBD Patients

Forty-four patients were of non-Caucasian ethnicity and excluded from association analyses. Patients of different ethnicities have the potential to confound GenePy scores as variants having differing minor allele frequencies in different ethnic groups. GenePy scores were successfully generated for 67 of the 68 monogenic IBD genes where at least one high-quality missense or insertion/deletion variant was annotated in exonic regions.

When comparing the top 10% of GenePy scores between IBD and controls, eight genes accrued significantly higher scores in IBD cases. Following FDR correction *ADA*, *FERMT1*, *LRBA* and *NOD2* remained significant (Table 10). Patients identified as having extreme GenePy scores within *NOD2*

were significantly enriched for CD patients (0.0046). No other genes were enriched for either IBD subtype. Of the 20 patients with 'Pathogenic' or 'Likely Pathogenic' *NOD2* variant(s), 15 were present in the top 10% *NOD2* GenePy scores. We excluded all patients with monogenic *NOD2* variants and recalculated the Mann-Whitney U statistics. This confirmed a persistent significant difference (p=0.0035) in cases compared to controls and confirms that those patients failing the threshold to have 'Pathogenic' or 'Likely Pathogenic' *NOD2* variation harbour a significant excess of pathogenic *NOD2* mutations.

Table 10- GenePy score comparison between top 10% of IBD patients (n= 36) versus top 10% of controls (n= 18).

Gene	***Cases vs Controls significance***	***FDR correction***	***CD***	***UC***	***IBDU***	***CD/UC enrichment***
ADA	0.0023	**0.0364**	20	14	2	0.3504
COL7A1	0.0339	0.2671	23	12	1	1
FERMT1	0.0023	**0.0364**	23	11	2	1
LRBA	0.0014	**0.0364**	28	6	2	0.0812
NOD2	0.0015	**0.0364**	32	4	0	**0.0046**
STXBP2	0.0279	0.2515	22	13	1	0.7081
TRIM22	0.0141	0.1538	28	8	0	0.1946
TTC37	0.0146	0.1538	26	10	0	0.5785

False discovery rate (FDR) correction for multiple testing was applied. Enrichment for CD / UC subtypes was investigated using Fishers exact test.

4.3.3.1 Phenotypic assessment of patients with extreme GenePy scores

Following correction for multiple testing, four genes maintained evidence for a significant burden of gene pathogenicity scores (*ADA*, *FERMT1*, *LRBA* and *NOD2*). Evidence for distinctive phenotypic characteristics conferred by each of these genes was tested using linear regression. GenePy scores for all patients (regardless of IBD subtype) were regressed against clinical features and significant associations are detailed in table 11. Patients with inflated GenePy scores in the *ADA* gene were enriched for males (p=0.021) and presented with isolated colonic disease (p=0.033).

No clinical characteristics were significantly associated with *FERMT1* scores. Patients with higher *LRBA* gene pathogenicity scores more often underwent any IBD-related surgery (p=0.006).

Table 11- Clinical phenotype characteristics associated with genes overburdened with pathogenic mutations in IBD patients.

Gene	Effect	B	Std. Error	p
ADA	Gender (Male)	0.277	0.119	0.021
	Isolated colonic disease	0.253	0.118	0.033
LRBA	Surgery	0.011	0.004	0.006
NOD2	5-ASA	-0.056	0.018	0.002
	Stricturing disease	0.134	0.026	$6.6x10^{-7}$
NOD2 excluding reportable monogenic patients	5-ASA	-0.047	0.017	0.006
	Stricturing disease	0.094	0.028	$7.3x10^{-4}$

Across all European ancestry patients, higher mutational burden within *NOD2* was associated with lower use of 5-ASA medication (p=0.002), consistent with this drug being of primary use in UC. These patients also had a significantly higher rate of stricturing disease ($p=6.6 \times 10^{-7}$) confirming the association observed at the monogenic variant level. We hypothesised that GenePy may identify an association between stricturing disease driven by patients that carry a high *NOD2* mutational burden but did not fulfil the criteria for 'Pathogenic' or 'Likely Pathogenic' monogenic *NOD2*-related disease. Therefore, we excluded patients with 'Pathogenic' or 'Likely Pathogenic' monogenic *NOD2* variants and for the remaining 338 patients repeated the regression analysis of *NOD2* GenePy scores. Despite limiting this analysis to patients who do not fulfil the criteria for 'Pathogenic' or 'Likely Pathogenic' NOD2 variation, there endures a striking negative correlation between a high *NOD2* gene pathogenicity score and use of 5-ASA (p=0.006) and a strong positive correlation with stricturing disease ($p=7.3x10^{-4}$).

4.3.4 *TTC7A* variants were observed on the same haplotype and appear non-pathogenic

Five patients were identified harbouring the p.K606R and p.S672P variants (Table 12). Sanger sequencing confirmed all patients had these variants on the same chromosome (inherited maternally or paternally). Neither variant was observed in isolation (without the other) in the remaining 396 patients. In all cases, the parent who transmitted these variants did not have IBD. The brother of patient 1, who also had IBD (Crohn's disease, presenting in childhood), was found to have both the p.K606R and p.S672P variants. The phenotype of these TTC7A patients was varied, 4/5 probands had Crohn's disease. One had required surgery during follow-up (mean follow-up time 4.0 years) due to stricturing disease but none of the remaining patients had developed complications and disease was controlled with 5-ASA, thiopurine or anti-TNF therapy.

Table 12- Genotype and phenotype characteristics of five patients identified harbouring the p.K606R and p.S672P TTC7A variants

Patient	*TTC7A* variants	gNomad frequency	CADD score	Inherited from	Second TTC7A variant?	Second TTC7A variant CADD score	Second TTC7A variant frequency	Sex	Age at diagnosis	Follow-up duration	Disease	Paris Classification	Complications including evidence of immunodeficiency	Medications used	Additional information
1	K606R	0.0022	25.6	Mother	Synonymous G696A:p.E232E	9.522	0.1778	Male	12.9 years	4.4 years	Ulcerative colitis	E4	None evident	5-ASA Thiopurine	No FH
	S672P	0.0022	27.5												
2	K606R	0.0022	25.6	Father	None	-	-	Female	12.6 years	8.9 years	Crohn's disease	L3+L4a	None evident	5-ASA Thiopurine	Brother with IBD also heterozygote for variants
	S672P	0.0022	27.5												
3	K606R	0.0022	25.6	Mother	Synonymous G696A:p.E232E	9.522	0.1778	Male	15.2 years	2.7 years	Crohn's disease	L1	None evident	Anti-TNF	No FH
	S672P	0.0022	27.5												
4	K606R	0.0022	25.6	Father	Synonymous G696A:p.E232E	9.522	0.1778	Male	13.4 years	2.8 years	Crohn's disease	L2	None evident	5-ASA Thiopurine	No FH
	S672P	0.0022	27.5												
5	K606R	0.0022	25.6	Mother	Synonymous C198T:p.S66S	17.63	0.08	Male	15.7 years	1.4 years	Crohn's disease	L3+L4a	Stricturing disease	Anti-TNF	No FH Has *NOD2* variant
	S672P	0.0022	27.5												

The p.K606R and p.S672P variants are rare, both having a gNomad (all genomes) frequency of 0.0022 and an EXAC (non-Finnish European) frequency of 0.0035. Identical frequency suggests inheritance on a common haplotype. Both are predicted to be highly deleterious, possessing CADD scores of 25.6 and 27.5 respectively. The percentage of our patients, as a proportion of IBD patients sequenced, possessing these alleles is 1.2% (5 patients), contrasting to the one patient (0.22-0.35%) expected from curated frequency databases. To determine whether an additional variant could be impacting on the function of the wild-type allele patient VCFs were reviewed. Three patients (patient 1, 3 and 4) harboured the common synonymous p.E232E variant (CADD 9.5) and one (patient 5) harboured the synonymous p.S66S variant (CADD 17.63). Both of these are common, minor allele frequency 0.178 and 0.08 respectively, and have modest deleteriousness scores. Segregation analysis was not performed for these variants. Sanger sequencing to confirm segregation unambiguously confirmed heterozygote status and excluded a deletion in the five patients at the site of the p.K606R and p.S672P variants. To further assess the evidence for large deletions and copy number variants (CNV), WES data from patients harbouring TTC7A variants were examined in the integrative genome viewer (IGV) and compared to patients not harbouring any TTC7A variants, sequencing using the same capture kits. Read depth was calculated across the entire gene and compared to patients without the variants (figure 10). No deletions or CNVs were identified through either method although we acknowledge the limitations of WES in detection of deletions.

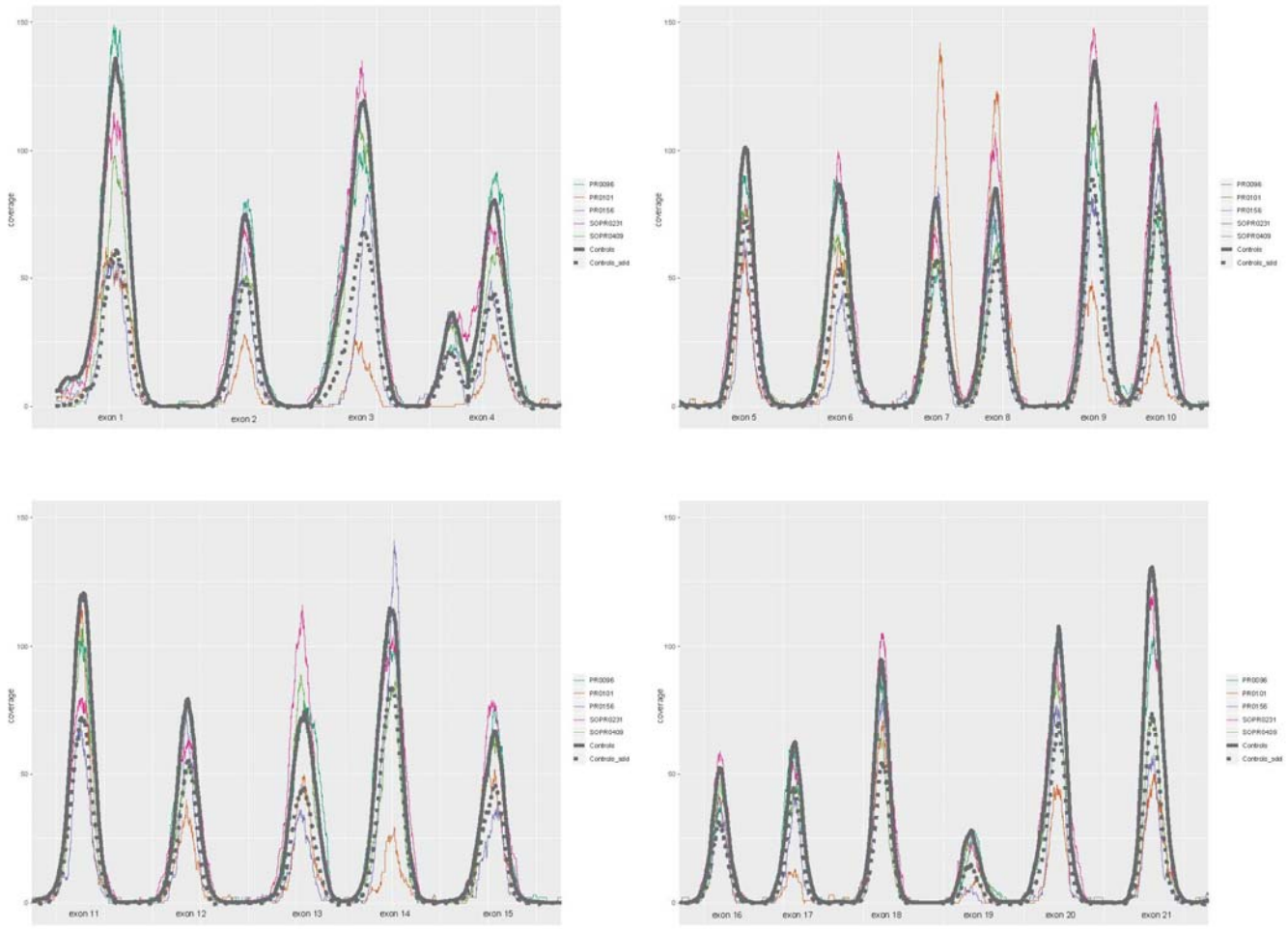

Figure 10- Coverage plot for TTC7A. Per-base read coverage for 5 patients harbouring TTC7A variants. Coverage is shown for each exon with a 500 base pair padding to each side. The bold line represents the mean coverage observed in 15 individuals not harbouring TTC7A variants. The dashed line indicates 1 standard deviation below the control mean coverage. Exons from patients PR0096, SOPR0231 and SOPR0409 were captured using Agilent SureSelect V6. Exons from patients PR0101 and PR0156 were captured using Agilent SureSelect V4 and V5 respectively reflecting a chemistry with worse coverage of this gene. All patients lie within 2 standard deviations of the mean.

4.3.4.1 Role of *TTC7A* variants in polygenic IBD

Based on the key role TTC7A has in regulation of normal intestinal epithelial cell development we hypothesised that high mutation burden in this gene may contribute to IBD through mutations not classified as casual of a monogenic disorder. We utilised a novel per gene, per individual deleteriousness score, GenePy, to assess whether variation in TTC7A was contributing to the risk of developing inflammatory bowel disease but not in a monogenic inheritance pattern. GenePy provides a single score to a gene, for each individual, based on all variants harboured within that gene. Even in genes known to play a role in the development of polygenic disease high scores are uncommon in individual risk genes. Therefore, comparison of extreme scores (top 10%, Mann Whitney U-test) in TTC7A between all cases and a set of non-IBD controls allows determination of whether variation in TTC7A was more deleterious in the IBD cohort. Following application of GenePy, mutation burden was identical in cases and controls, with the median of the top 10% of scores being 0.72 in both, p=>0.05. This provides evidence that TTC7A variation observed in our cohort was not contributing to disease development.

4.4 Discussion

In our unselected cohort of paediatric patients, we identified a 'Pathogenic' or 'Likely Pathogenic' variant, in a known monogenic IBD gene, in 11.5% of patients. When considering VEOIBD only, this rate was 16.7%. Over recent years NGS has identified Mendelian causes of patients presenting with particularly severe IBD-like phenotypes [78,79]. It is generally thought that patients with monogenic IBD are exceptional cases masquerading as IBD and contemporary molecular diagnostics are unlikely to yield clinically relevant diagnostic rates within the general IBD population[190]. In this study we applied extremely stringent filtering criteria, insisting on validated functional evidence for variants and our observations are likely to underestimate the true prevalence of monogenic gene variants. Whether this non-trivial rate is maintained in adult cohorts remains to be seen.

NOD2 was the first genetic locus identified in IBD and was designated the IBD1 locus through linkage studies in the 1990s[74]. These analyses were focused on pedigrees with early-onset and severe disease and suggested an AR inheritance pattern[75,76]. More recently *NOD2* has been the most consistent hit in GWAS of IBD, and fuelled the argument for common variation predisposing to disease, with many studies focusing on R702W, G908R and 1007fs only[203,204]. Our findings are consistent with monogenic *NOD2*-related disease representing the molecular basis of 5% of all paediatric IBD cases, increasing to 7.3% for CD. However, the majority of cases harboured compound heterozygous variants, having one low frequency variant and a second different very rare/novel mutation. This supports previous data from the 2000s where studies independently identified a RR of 9.8-44 in individuals with compound heterozygous or homozygous *NOD2* variants. Although, these studies were limited by identifying only commonly reported *NOD2* variants and did not account for additional rare or novel variants[73,85,183]. Our results confirm recent data from Horowitz *et al* who suggested up to 7.8% of paediatric patients had monogenic *NOD2*-related disease and the modest differences in diagnostic rates between both studies are likely due to our application of conservative and stringent filtering criteria[184]. Our data additionally report the distinct clinical characteristics that segregate with monogenic *NOD2*-related disease. This chapter describes the E963G variant as 'Likely Pathogenic' based on *in silico* evidence, segregation with a known pathogenic variant and presence with a distinct stricturing phenotype, however classification of novel variants without functional validation remains challenging and confirmation of the impact of this variant is important.

A relationship between *NOD2* variation and stricturing disease phenotype was first discussed in 2002. Abreu *et al* reported ORs of 2.4 and 7.4 for heterozygous and pooled compound heterozygous/homozygote variants respectively in an analysis limited to R702W, G908R and 1007fs mutations[205]. Subsequent studies, summarised elsewhere, have questioned whether this association is due to *NOD2* predisposing to ileal disease, rather than fibrostenotic disease *per se,* with conflicting results[206]. Many of these studies examined a limited number of variants (R702W, G908R or 1007fs), did not correct for ileal disease location or did not differentiate

between heterozygous and compound heterozygous/homozygote variants[206]. To our knowledge this study is the first to analyse stricturing phenotype whilst considering *NOD2* as a monogenic cause of disease, including all rare and novel variants passing stringent filtering criteria. Following correction for disease location we observe a striking increase in stricturing disease and surgical resection risk in monogenic *NOD2* patients, with the highest reported RR to date (4.32 and 4.59 respectively).

Additional monogenic causes for disease were identified in this cohort, with phenotype-genotype correlation observed for some variants. We describe the second report of monogenic IBD associated with variants in *TRIM22*[197]. One of our patients has a phenotype consistent with the recent description of severe early onset perianal CD, however we describe four patients with a severe but variable phenotype. This suggests a spectrum of *TRIM22*-related disease presenting throughout childhood. We identify a novel relationship between severe extensive UC, PSC and the *WAS* P460S variant not previously reported in IBD, which may have treatment implications for patients presenting with this genotype[198]. This variant allele has a frequency of 0.0023 and may impart variable penetrance similar to other X-linked IBD genes[190].

Application of a whole gene pathogenicity scoring tool enabled us to assess the burden for any given gene in individuals, rather than assessing single variants only. Despite no individual patient fulling strict criteria for monogenic disease, *ADA*, *FERMT1* and *LRBA* accumulated significantly higher mutation burden in cases compared to controls. Either the observed variants in these genes play a role in polygenic IBD risk or there are additional non-coding mutations undetected by exome sequencing that lead to AR disease in patients with higher mutation burden. Nevertheless, we discern clinically informative significant associations between *LRBA* pathogenic burden in children with increased rates of IBD-related surgery.

Utilising GenePy we confirm a significant role for *NOD2* in stricturing disease, excluding monogenic *NOD2* diagnoses. Our results indicate that within the set of patients not achieving a *NOD2* monogenic disease diagnosis, there remain patients whose disease is underpinned by

deficient *NOD2* signalling. These data provide further evidence that *NOD2* heterozygosity has either some penetrance, that cumulative burden of mild *cis* or *trans* variants impact on disease, or more likely, an undetected mutation in non-coding regulatory region(s) constitute the 'second hit' under a recessive model. Whilst GenePy provides a contemporary method for assessing pathogenicity across a gene it is limited by the imperfection of deleteriousness metrics, as evidenced by modest CADD score assigned to the *NOD2* V955I variant, despite this variant's known role in disease. It is possible that variants such as V955I are in linkage disequilibrium with additional intronic or promotor variants, not detected through WES, which is the true 'second hit' in a recessive model in these patients. There is a clear necessity for more functional work-up of variant impact, both on a *per variant* basis as well as in combination (both *cis* and *trans*). Multiple variants within a single haplotype have the potential to behave diversely – acting in synergy to reduce or increase functionality or may mutually compensate in ways that cannot be predicted by single variant functional analyses.

Missing heritability of IBD remains. Twin studies estimate heritability at 0.75 in CD and 0.67 in UC, compared to 0.37 and 0.27 from GWAS data[72]. As GWAS are only powered to detect features attributable to common variation some of this missing heritability is likely due to very rare or private mutation, as observed through identification of rare variants and increased mutation burden in patients in this study.

The p.K606R and p.S672P variants appear to be most commonly inherited on the same haplotype. The number of patients harbouring these heterozygote variants exceeds that expected from both gNomad and EXAC allele frequency databases. However, there is no evidence a single affected allele would translate into impact on protein expression or function, based on all previous reports of disease associated with TTC7A variants it appears to be a Mendelian disorder with autosomal recessive inheritance, although previous data has indicated a role for heterozygote variants in genes associated with autosomal recessive monogenic IBD for other genes[200]. Where these variants have been inherited in trans, previous reports have detailed a partial loss-of-function

associated with mild CVID and enteropathy phenotype. Interestingly, neither of parents of the patient described by Lawless et al possessed the p.K606R and p.S672P haplotype, with the mother having the wild-type allele at amino acid position 606 and the father being wild-type at position 672.

In this project we identify five paediatric IBD patients harbouring variants on the same chromosome, previously described as disease-causing when inherited in compound heterozygosity. These variants, p.K606R and p.S672P, appear to exist on a relatively rare ancestral haplotype and are seen in 1.2% of our cohort. In silico analysis indicates that exonic variation in TTC7A was not contributing to disease in this cohort. We find no evidence of increased TTC7A mutation burden in the coding regions for IBD patients compared to controls. These data imply that TTC7A is only associated with an autosomal recessive Mendelian disorder, with disease occurring in patients who harbour rare, deleterious variants. We were unable to assess the impact of variation in intronic/promotor regions, and alteration of TTC7A expression or alternative splicing could still contribute to disease pathogenesis. Functional analysis is required to assess the role of rare heterozygote variants, not fulfilling the criteria for causing monogenic disease, in the regulation of intestinal development and inflammation.

4.4.1.1 Conclusion

Accumulating data provide evidence to advocate for *NOD2* screening in newly diagnosed IBD to inform predictive algorithms for treatment including need for surgical resection[184]. Enabling personalised therapy becomes more important with the development of new drugs, such as *RIPK2* inhibitors, that modulate the *NOD2* signalling pathway[207,208]. However, focussing only on *NOD2* would limit the potential benefits of precision diagnostics by excluding analysis of other genes, as evidenced by clinically relevant variation in *TRIM22* and *WAS* in this modest cohort. Any monogenic IBD gene panel would need to adapt flexibly alongside gene discovery. National programmes are committing to providing NGS for any child admitted to intensive care with an unknown diagnosis as this approach has achieved a diagnostic rate of 25%. The value of a

molecular diagnosis, especially one that provides certain prognosis and bespoke management to a child with a serious chronic disease, is invaluable. Our data deliver persuasive evidence for NGS diagnostics as a standard of care in paediatric-onset IBD, providing a precise diagnosis and personalised therapy for a substantial number of patients.

Chapter 5 Combined digenic *NCF4* and *NOD2* variation is associated with a fistulating Crohn's disease phenotype

Chapter summary- *This chapter investigates the combined effects of genetic defects across the NADPH oxidase complex and the NOD2 gene. We utilise in silico burden scoring, through integration of GenePy to whole exome sequencing data, to determine the relationship of monogenic, digenic, and oligogenic variation with disease subtype and Crohn's disease phenotype. Small subsets of patients are identified with digenic variation appearing to contribute to phenotype.*

Chapter contributions- *Whole exome sequencing data were processed by Imogen Stafford and Guo Cheng. GenePy scores were generated by Imogen Stafford. NADPH oxidase and related genes were identified by James Ashton. Clinical data were collected by James Ashton. GenePy score and statistical analysis were performed by James Ashton.*

Supplementary data can be found at https://doi.org/10.5258/SOTON/D1657

5.1 Background

Monogenic forms of IBD are now well established, and provide clinical translation to patients diagnosed with these conditions[190]. Despite this, only a small fraction of patients with IBD appear to harbour a true monogenic defect, with recent estimates ranging from 3-8%[209,210].

The role of variation in genes within pathways or complexes are of increasing interest. Previous studies utilising a GWAS approach are unable to the impact of rare variation on disease pathogenesis, rather than utilising common 'marker' variants, WES data allows for the analysis of all exonic pathogenic variants within a gene, facilitating the analysis of epistasis. It should be noted that WES data may fail to capture splicing variants and will not cover intronic or non-exonic parts of the gene, such as promotor regions.

CYBB/NOD2 knockout murine models demonstrate digenic forms of IBD and interaction between *TRIM22* and *NOD2* appears to lead to IBD in specific patients[211–213]. It appears increasingly likely that in some patients who do not fulfil diagnostic criteria for monogenic disease, the role of secondary variants in interacting genes, or genes within the same pathway, may lead to disease, so called digenic (variation in two genes), or oligogenic (variation in a defined number of interacting genes), inflammatory bowel disease (IBD).

Traditionally, IBD has been considered a complex disease arising from variation in multiple genes interacting with environmental stimuli. A multitude of pathways may harbour genetic risk variation, with the individual cause of disease increasingly felt to be specific to an individual, or to a family[91]. These personal genetic risk profiles will result in similar, broad, phenotypes- Crohn's disease and ulcerative colitis[214]. Despite this, there is huge intra-disease heterogeneity and specific subgroups of disease behaviour are clearly present, including development of specific complications and presence of additional autoimmune co-morbidities, such as primary sclerosing cholangitis. Within Crohn's disease, identification of patients at increased risk of stricturing or penetrating (fistulating) disease has huge treatment implications. Furthermore, these disease behaviours may have their own risk genes, potentially allowing identification of these distinct patients[206,215]. Within Chapter 4 we have identified *NOD2* variation as a risk factor for stricturing disease development, but not penetrating (fistulating) disease. Penetrating disease appears to be a more genetically complex phenotype with no monogenic genes proving significant in chapter 4's regression analysis. Environmental influence through microbial profiles/infection

appearing to play a large role[216]. Despite this, variation within the NADPH oxidase complex and within *IL10*-pathway genes have been associated with the occurrence of fistulae in IBD patients, although in both cases these were considered as part of a monogenic disease[123,188].

The NADPH oxidase complexes are a number of highly conserved protein complexes that produce reactive oxygen species (superoxides), key to the killing of bacteria both within blood and specific tissues[217]. Neutrophil-derived NAPDH is a key component in innate immunity, providing host defence, but additional complexes are also involved in cellular signalling and regulation of gene expression[217]. There are six NADPH complexes, alongside related proteins (figure 11). It is well established that highly deleterious genetic defects in NADPH oxidase genes, such as *NCF1, CYBA* and *CYBB* present with chronic granulomatous disease (CGD). This condition frequently has co-existing features of inflammatory bowel disease, including colitis, granulomata or recurrent perianal abscesses[218]. In paediatric-onset Crohn's disease there appears to be role for variants not fulfilling ACMG criteria to be either causative CGD, or for monogenic forms of IBD[219].

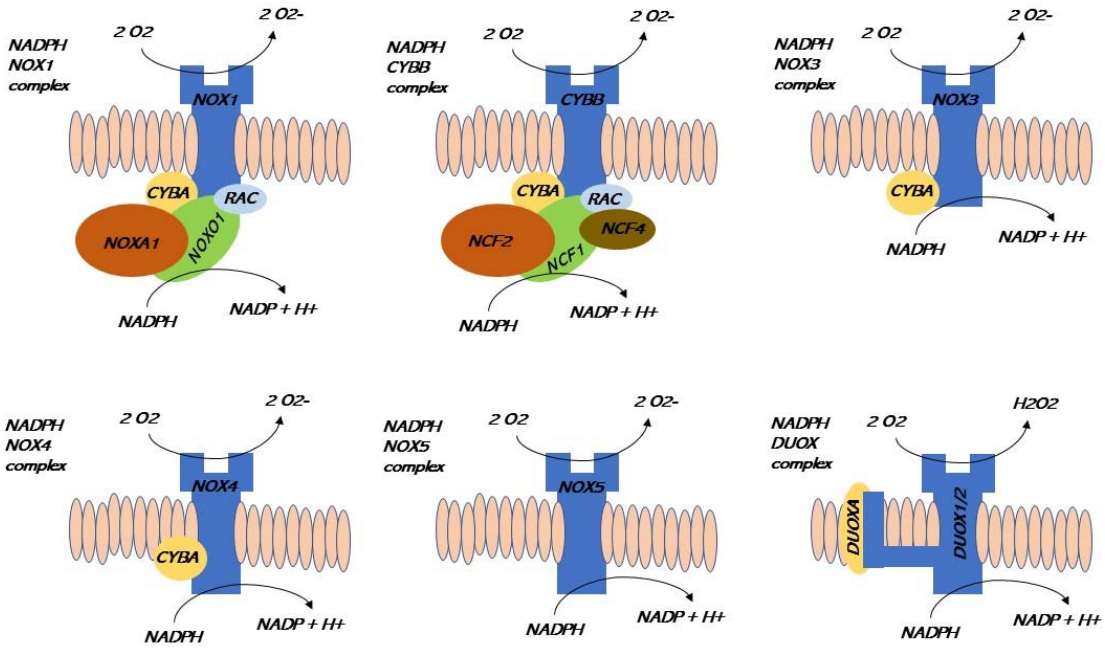

Figure 11- Six NADPH complexes and constituent genes. All complexes are transmembrane and produce reactive oxygen species

Variation across a pathway, or protein complex, may lead to a specific manifestation of disease, such as a Crohn's phenotype or fistulating behaviour. Both NAPDH and *NOD2* pathways are

integral to the immune response to bacteria, enabling adequate killing and clearance of invasive species. Based on previous data we hypothesised that patients harbouring variation in multiple NADPH oxidase complex genes, and/or in *NOD2,* but not fulfilling diagnostic monogenic criteria would be at increased risk of a Crohn's disease phenotype, beyond the risk conveyed by variation in one of these genes. Specific patients may have true oligogenic disease. As fistulating (and specifically perianal) disease appears to be associated with inability to clear bacteria resulting in abnormal connections between epithelialised surfaces, we hypothesised that patients with oligogenic disease ('significant' variations in both NADPH genes and *NOD2*) would be at increased risk of fistulating disease, whereas patients with variation in only the NADPH oxidase complex or in *NOD2,* would be at lower risk.

5.2 Methods

Patients were recruited from the Wessex regional paediatric IBD service at Southampton Children's Hospital to the genetics of paediatric IBD study (2010 to present), as described in previous chapters. The eligibility criteria for recruitment was a confirmed histological diagnosis of either Crohn's disease (CD), ulcerative colitis (UC) or IBD unclassified (IBDU), in line with the Porto criteria, and age less than 18 years[2,220]. All patients with IBD were included in order to determine if any genetic variation in NADPH oxidase genes were associated with disease subtype.

5.2.1 DNA Extraction

Patient DNA was extracted from peripheral venous blood samples collected in EDTA using the salting-out method, or from saliva, as previously described[138].

5.2.2 Whole Exome Sequencing (WES) Data Processing

Whole exome sequencing data were processed as previously described in Chapter 4. For these analyses all patients had updated sequencing data using Agilent SureSelect V5 or V6.

5.2.3 Application of GenePy *in-silico* Score

We utilised GenePy to integrate into downstream statistical analysis[153]. In this analysis we were not comparing cases to controls so all patients with WES data were included, rather than those of the same ethnicity. GenePy scores were generated from the most up to data sequencing available (all patients Agilent SureSelect V5 or V6). Patients sequenced on V3 or V4 had sequencing data regenerated during the final year of the PhD, providing updated sequencing data for this chapter and chapter 7. Only genetic variants called in genomic regions common to both exon enrichment capture kits (intersection of versions 5 and 6 bed manifest file) were included to avoid unbalanced calls. GenePy scores were based on CADD, version 1.6, which integrates spliceAI to improve annotation of splicing variants[221].

5.2.4 NADPH Oxidase Gene List

A a literature review was performed to collate a list of all genes involved in NADPH oxidase complexes, or production of reactive oxygen species. As these genes are well established it was not necessary to conduct a systematic review[217]. Each gene was annotated with information including tissue expression, function and chromosome. There are six NADPH complexes (table 13) and each gene was also annotated with which complex it is present in.

5.2.5 Statistical Analysis

All GenePy scores were scaled to between 0 and 1 to facilitate comparison, summing and multiplication of genes. Using the maximum and minimum scores in the entire cohort as reference points the following formula was applied to each score, for each patient, for each gene-

Scaled GenePy score = (Gene X GenePy score-MIN(All gene X GenePy scores))/(MAX(All gene X GenePy scores)-MIN(All gene X GenePy scores))

Following scaling, GenePy scores for each NADPH oxidase complex (as detailed in table 11) were summed for each patient. This resulted in a GenePy score for each NADPH oxidase complex, in

addition to each individual gene. In order to assess for interaction of NADPH genes/complexes with *NOD2* we performed multiplicative transformation of NADPH gene or complex GenePy score, with *NOD2* GenePy score. The analysis was performed for each patient; every NADPH gene's GenePy score, and each complex's GenePy score, was multiplied by the scaled *NOD2* GenePy score for that patient, resulting in a score combining the NADPH gene or complex, and *NOD2*, into a multiplicative model. Supplementary data.

Downstream analysis was with forward binary logistic regression (Wald) and survival analysis (Cox proportional hazard modelling). Both were performed using SPSS (v25, IBM) software.

5.2.6 Phenotypic Characterisation

We extracted whether patients had developed fistulating disease, alongside their maximal follow-up duration. Fistulating disease was defined as presence of perianal fistulae, entero-entero fistulae, enterocutanous fistulae or recto-vaginal fistulae[29]. Patients with recurrent perianal abscesses were included within this definition. Disease subtype (Crohn's disease or non-Crohn's disease), and presence (or absence) of fistulating disease was transformed to binary data for use in logistic regression analyses. Cox proportional hazard modelling was performed, with patients censored at maximal follow-up duration, if an event had not occurred for that patient. Follow-up duration was calculated for each patient based on date of diagnosis and last recorded clinical contact.

5.3 Results

Five-hundred-and-one patients were included in the analysis. Of these 330 had a diagnosis of Crohn's disease, and of these 57 had manifested a fistulating phenotype at the time of analysis. All patients had GenePy scores calculated from whole exome sequencing data. Clinical follow-up data were available for all patients. The mean age at diagnosis was 12.09 years (range 1.3-17.4

years), with a mean follow-up time of 4.6 years (range 0.1-18.4 years). No patients had an underlying clinical diagnosis of chronic granulomatous disease.

5.3.1 Genes in NADPH oxidase complexes

Eighteen genes were identified in NAPDH oxidase complexes, or as directly interacting genes. Genes included are seen in table 13. We were able to calculate GenePy scores on 16 of these genes: *RAC1* was invariant across the cohort, and therefore no score was calculated. *NCF1* had very poor coverage and quality across the gene, stringent quality filtering preventing accurate calculation of a GenePy score.

Table 13- Summary of NADPH oxidase complex and related genes. Coverage of capture kits used to generate whole exome sequencing in out cohort are included.

Gene	Synonyms	Chromosome	Agilent SureSelect coverage		Gene function	Intestinal expression?	Immune cell expression?	Protein complex					
			V5	V6				NADPH complex-NOX1	NADPH complex-CYBB	NADPH complex-NOX3	NADPH complex-NOX4	NADPH complex-NOX5	NADPH complex-DUOX
NCF2	gp67phox	1	0.511083	0.445479	Forms NOX2 enzyme complex, with NCF1, NCF4, RAC2(?) and RAP1A		Yes		**X**				
NOX3		6	0.682673	0.535344	Similar enzyme complex to NOX1/NOX2, found in different tissues/cells.	Yes				**X**	**X**		
NCF1	gp47phox	7	0.49892	0.45463	Forms NOX2 enzyme complex, with NCF2, NCF4, RAC2 and RAP1A		Yes		**X**				
RAC1		7	0.556938	0.541244	Present in NOX1 enzyme complex		Yes	**X**	**X**				
NOXA1		9	0.545622	0.475487	Oxidase activity of complex supporting ROS production (NOX1 enzyme complex) potentiated by this gene and NOXO1.Homolog of NCF2.	Yes		**X**					
NOX4		11	0.502646	0.443894	Similar enzyme complex to NOX1/NOX2 found in different tissues/cells.						**X**		
DUOX1		15	0.595996	0.468955	Similar enzyme complex to DUOX2, found in different tissues/cells.	Yes							**X**
DUOX2		15	0.613191	0.548507	Protein encoded forms similar enzyme complex to NOX1, NOX2 etc., but end product is hydrogen peroxide	Yes							**X**
DUOXA1		15	0.7805	0.636455	Gene encodes maturation factor for DUOX1 function	Yes							**X**
DUOXA2		15	0.718356	0.65908	Gene encodes maturation factor for DUOX2 function	Yes							**X**
NOX5		15	0.719965	0.592605	Similar enzyme complex to NOX1/NOX2, found in different tissues/cells.							**X**	

CYBA	p22phox	16	0.283206	0.255535	Contributes to NOX1, NOX2 and NOX3 enzyme complexes		Yes	**X**	**X**	**X**			
NOXO1		16	0.671166	0.674393	Oxidase activity of complex supporting ROS production (NOX1 enzyme complex) potentiated by this gene and NOXA1. Homolog of NCF1.	Yes		**X**					
CYBC1		17			Membrane-spanning protein controlling phagocyte respiratory burst		Yes						
NCF4	gp40phox	22	0.565237	0.483092	Forms NOX2 enzyme complex, with NCF1, NCF2, RAC2(?) and RAP1A		Yes		**X**				
RAC2		22	0.419067	0.373059	Present in NOX2 enzyme complex		Yes	**X**	**X**				
NOX1		X	0.620387	0.528157	Associated with oxidative stress. Proteins encoded by NADPH oxidase (NOX1), NOXA1 (homolog of NCF2), NOXO1 (homolog of NCF1), RAC1 and CYBA combine to form complex that supports ROS production. The complex is responsible for one-electron transfer of oxygen to generate superoxide.	Yes		**X**					
CYBB	NOX2 gp91phox	X	0.64316	0.538437	Contributes to NOX4 enzyme complex		Yes		**X**				

5.3.2 Crohn's disease subtype is related to deleterious variation in *HMOX1* and *NOXO1*

Binary logistic regression was performed using Crohn's disease status (diagnosis of CD versus either IBDU or ulcerative colitis) as the dependant variable. Independent variables were scaled GenePy scores for the 16 included NADPH oxidase genes, *NOD2,* the summed GenePy scores for the NADPH complexes and the multiplicative scores combining *NOD2* and each of NADPH genes/complexes, table 14.

Table 14- NADPH genes, complexes and multiplicative combinations used in binary logistic regression models

Genes	**Genes combined with *NOD2* (multiplied GenePy scores)**	**NADPH complexes (summed GenePy scores)**	**Complexes combined with *NOD2* (multiplied GenePy scores)**
CYBA	*CYBA_NOD2*	NADPH complex-NOX1	NADPH complex-NOX1_*NOD2*
CYBB	*CYBB_NOD2*	NADPH complex-CYBB	NADPH complex-CYBB_*NOD2*
DUOX1	*DUOX1_NOD2*	NADPH complex-NOX3	NADPH complex-NOX3_*NOD2*
DUOX2	*DUOX2_NOD2*	NADPH complex-NOX4	NADPH complex-NOX4_*NOD2*
DUOXA1	*DUOXA1_NOD2*	NADPH complex-NOX5	NADPH complex-NOX5_*NOD2*
DUOXA2	*DUOXA2_NOD2*	NADPH complex-DUOX	NADPH complex-DUOX_*NOD2*
HMOX1	*HMOX1_NOD2*	All NADPH	All NADPH_*NOD2*
NCF2	*NCF2_NOD2*		
NCF4	*NCF4_NOD2*		
NOX1	*NOX1_NOD2*		
NOX3	*NOX3_NOD2*		
NOX4	*NOX4_NOD2*		
NOX5	*NOX5_NOD2*		
NOXA1	*NOXA1_NOD2*		
NOXO1	*NOXO1_NOD2*		
RAC2	*RAC2_NOD2*		
CYBC1	*CYBC1_NOD2*		
NOD2			

As expected, *NOD2* was strongly correlated with Crohn's disease, Beta 2.523, $p=6x10^{-6}$. Also significantly correlated were *HMOX1* (Beta 4.18, p=0.018 and *NOXO1* (Beta 1.977, p=0.032), after accounting for the effect of *NOD2*. These data imply that deleterious variation in *HMOX1* and *NOXO1* is a driver behind the Crohn's disease phenotype. There were no other significant associations between genes and disease subtypes.

The distribution of the significantly implicated gene's GenePy scores was examined for Crohn's disease and non-Crohn's disease patients. Using violin plots it was demonstrable that a subset of Crohn's disease patients had GenePy scores higher than non-Crohn's disease. Figure 12.

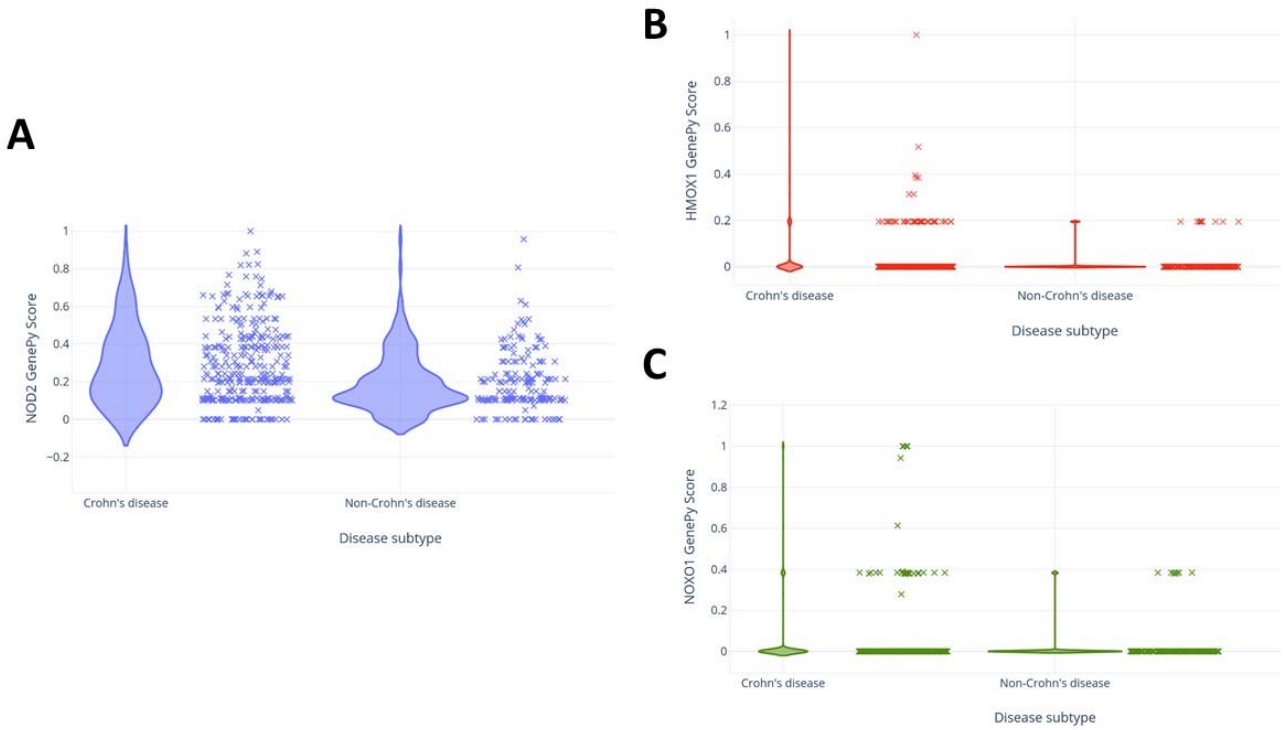

Figure 12- Violin plots for A) NOD2, B) HMOX1, and C) NOXO1. Higher GenePy scores in NOD2 are seen in multiple Crohn's disease patients, whereas in HMOX1 and NOXO1 extreme scores are seen in a small subset of seven independent patients for both genes.

5.3.3 Fistulating Crohn's disease phenotype correlates with digenic variation in *NCF4* and *NOD2*

We hypothesised that fistulating disease would be associated with oligogenic or digenic variation across the NADPH gene complexes, or between *NOD2* and NADPH genes. Using only Crohn's disease patients, a binary logistic regression using fistulating disease status as the dependent variable against the same independent variables detailed in table 14. The strongest relationship with fistulating disease status was observed for the interaction variable that incorporated the *NCF4* and *NOD2* genes ($NCF4_{GenePy}$*$NOD2_{GenePy}$) (Beta 6.614, p=0.041). This indicates patients with significant pathogenic variation in both *NCF4* and *NOD2*, but not in *NCF4* or *NOD2* alone, have increased risk of fistulating disease. The summed GenePy scores for the NOX4 NADPH complex, consisting of the *CYBA* and *NOX4* genes achieved borderline significance and positive correlation with a fistulating phenotype in this model, Beta 1.719, p=0.052. No single genes were significantly identified through this regression analysis.

5.3.4 Patients with extreme digenic variation in *NCF4* and *NOD2* have a two-fold increased risk of fistulating disease

Patients within the cohort have a variable follow-up duration as recruitment has occurred over a 10-year period. The regression analyses above did not account for variation in follow-up duration and patients recruited more recently, or lost to follow-up, may not have had the opportunity to develop a fistulating phenotype yet. In order to account for this, we performed a Cox proportional hazard survival model to determine to what extent patients with high digenic variation in *NCF4* and *NOD2* were at increased risk of fistulating disease, figure 13. As we would not expect all fistulating disease patients to have a fistulating phenotype due to digenic *NCF4*NOD2* variation, with other polygenic and environmental factors accounting for some disease, we chose to assess the top 10% GenePy scores against the remaining 90%, as previously described[153]. If these genes have no interacting effects, we would not expect a difference to be observed between the

top scores in either group. No other statistical tests were performed on different group percentages.

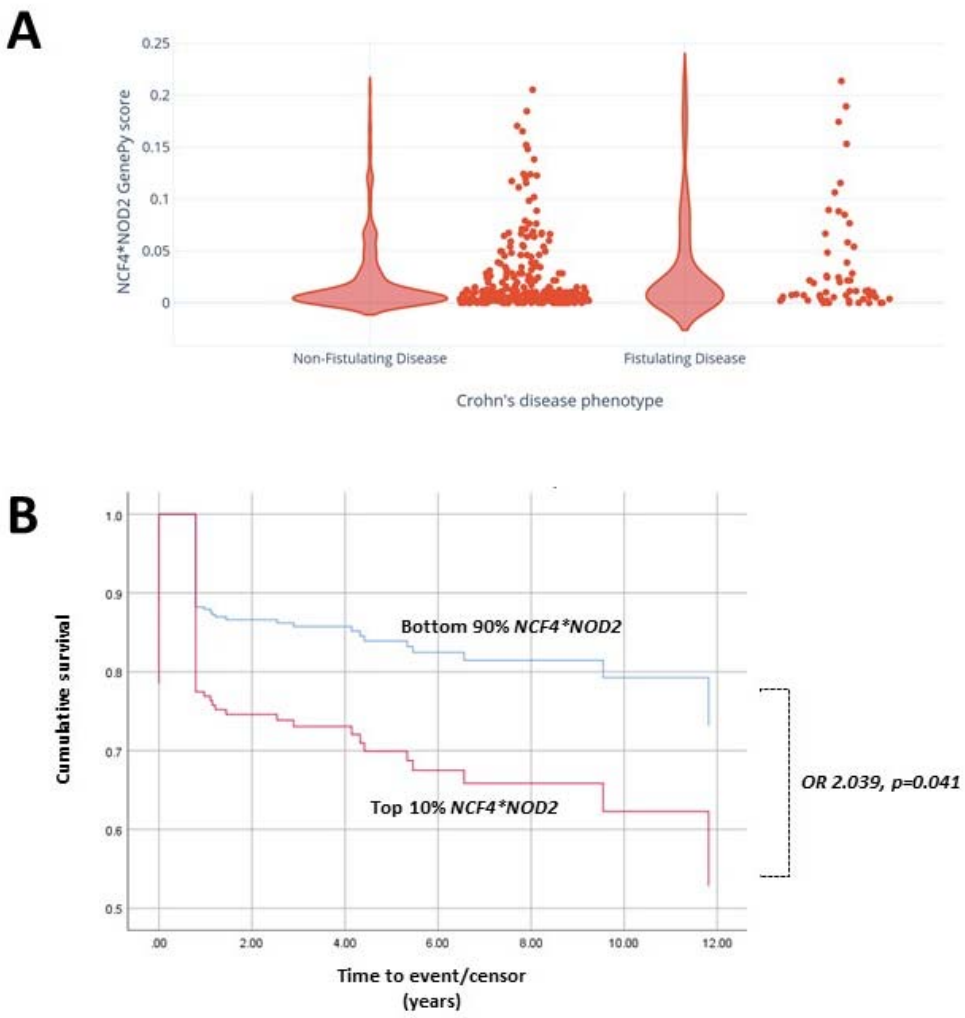

*Figure 13- A) Distribution of NCF4*NOD2 GenePy scores in patients with fistulating disease and non-fistulating disease. B) Cox proportional hazard model demonstrating significantly higher incidence of fistulating disease in the top 10% of NCF4*NOD2 GenePy score group, compared to all other patients.*

5.3.5 Deleterious variation in the NOX4 NADPH complex confers increased risk of fistulating disease

Although not reaching statistical significance in the binary logistic regression model we tested the NOX4 NADPH complex in a CPH analysis. Patients in the top 10% of summed GenePy scores for this complex had significantly increased risk of development of fistulating disease, OR 2.02, p-0.044, figure 14. Interestingly, the distribution of GenePy scores between the fistulating and non-fistulating groups did not demonstrate a stark difference between patient groups, with the differences being driven by a modest increase in the proportion of patients with moderate-high scores.

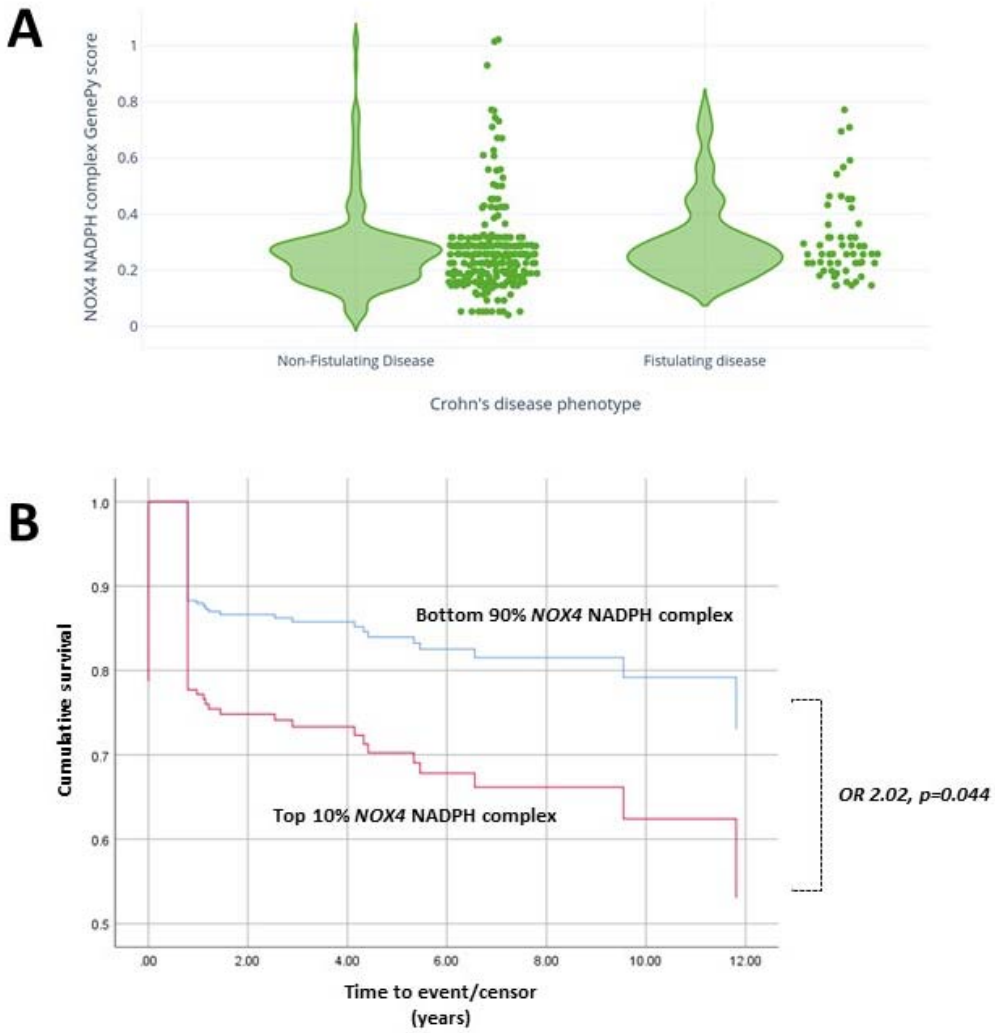

Figure 14- A) Distribution of NOX4 NADPH complex GenePy scores (summed NOX4 and CYBA GenePy scores) in patients with fistulating disease and non-fistulating disease. B) Cox proportional hazard model demonstrating significantly higher incidence of fistulating disease in the top 10% of NOX4 NADPH complex GenePy score group, compared to all other patients.

5.3.6 Patients harbouring higher variant burden across all NADPH genes have increased risk of fistulating disease

Finally, we hypothesised that variation across the NADPH complex would increase risk of fistulating disease. In comparison to variation across a single or several genes we would expect

variation across all NADPH genes to account for more fistulating disease. We therefore chose to compare the bottom 50% of summed GenePy scores against the top 50% of summed GenePy scores. No other statistical tests were performed on different group percentages. We performed a χ^2 test to assess this hypothesis. The 'top 50%' group harboured significantly more patients with fistulating disease compared to the 'bottom 50%' group, 22.4% (37 patients) vs 12.1% (20 patients) respectively, odds ratio 2.1, p=0.013.

5.4 Discussion

Deleterious variation in *HMOX1* and *NOXO1* are correlated with a Crohn's disease phenotype in paediatric IBD patients. The relationship observed in our cohort appears to be driven by a small subset of Crohn's disease patients harbouring significantly higher values than non-Crohn's disease patients. Reassuringly we replicate the strong signal of *NOD2* as the major driver for a Crohn's disease subtype in our paediatric IBD cohort. Digenic variation in *NCF4* and *NOD2,* where patients harboured high mutation burden in both genes, were linked to a fistulating disease phenotype. Patients harbouring the highest 10% of digenic GenePy scores had a two-fold increased risk of fistulating disease, compared to patients with a lower burden of pathogenic variants across both genes. These data demonstrate the impact of digenic variation on disease phenotype, it appears that modifying effect of variants in multiple genes contribute to the risk of developing fistulating disease within Crohn's disease patients.

Previous data has indicated a significant role in monogenic IBD pathogenesis for genetic variation within the NADPH complex. Denson *et al* demonstrated that 26 patients with significant missense mutations in *CYBA, CYBB, NCF1, NCF2*, and *NCF4,* had a 3x increased risk of perianal disease, and a more subtle increased risk of surgery and stricturing complications[219]. Furthermore, evidence of defective *NOX1* and *DUOX2* function, in relation to deleterious variation in these genes, was associated with reduced reactive oxygen species production in the colonic epithelium[222]. Four variants were identified across five very-early onset IBD patients. Dhillon *et al* demonstrated

targeted sequencing of *CYBA, CYBB, NCF1, NCF2, NCF4, RAC1,* and *RAC2* was able to identify eight causative variants in paediatric IBD patients, but these were largely related to a pan-colitis phenotype. In our analysis we demonstrate a modest impact of digenic *NCF4* and *NOD2* variation on the development of fistulating Crohn's disease phenotype. Whilst variation in either of these genes alone was insufficient to be statistically associated with fistulae formation, when combined we observe a two-fold increase in fistulating disease for the top 10% of *NCF4*NOD2* GenePy score patients. A similar observation was evident for the top 10% of NOX4-NADPH complex patients, who also had a two-fold increase in risk. The number of fistulating patients for which NADPH genes appear to confer risk for fistulae formation is relatively modest, however this is expected as there are many other implicated genes in fistulating disease pathogenesis[223]. The *IL10* pathway and its associated genes, *IL10*, *IL10RA* and *IL10RB*, are highly implicated, particularly in perianal disease[224]. In addition, studies have shown that polymorphisms altering IL10 serum levels have been linked to both Mendelian causes of IBD (specifically to severe penetrating disease behaviour), but also to polygenic IBD risk[225–229]. Additionally, whilst we focused on the role of genomic variation in predisposing or causing penetrating disease behaviour, there is significant evidence implicating infection and microbiome in the formation of fistulae[230,231]. The aberrant regulation of host inflammatory response to host microbes is the basis of IBD pathogenesis, with invasion of bacteria and active infection frequently seen in fistulae and penetrating disease[91].

Within our cohort we identify *HMOX1* as a driver of a Crohn's disease subtype, *HMXO1* is an enzyme catalysing the degradation of iron (heme) in a number of proteins including haemoglobin[232]. A by-product of this reaction is carbon monoxide, a reactive oxygen species (ROS), which has the ability to downregulate immune response within the intestine, alongside modulation and clearance of intestinal bacteria. *HMOX1* has previously been reported to have increased expression within the intestine during inflammation and it has been mooted as a potential therapeutic target in IBD[232]. It appears that *HMOX1* plays an important role in intestinal bacterial clearance particularly of *Salmonella* and *E. Coli*, with murine *HMOX1* knockout models displaying impaired bacterial killing and clearance[232]. It is highly plausible that

deleterious variants within *HMOX1*, as identified in this chapter, account for a small subset of Crohn's disease patients.

Similarly, we identify variation in *NOXO1*, a protein related to *NOX1* that regulates ROS formation by the NOX1 NAPDH complex, as correlating with a Crohn's disease phenotype[233]. Moll *et al* have previously demonstrated mouse knockouts for *NOXO1* have reduced ROS production within the colon, and *NOXO1* appears to prevent intestinal inflammation through influence of natural killer cells. *NOXO1* is expressed throughout the gastrointestinal tract, but is expressed at higher levels in the ileum and colonic tissue[234]. There is no previous data to implicate *NOXO1* variation in Crohn's disease pathogenesis, however deleterious variation in the closely related *NOX1* gene has proven effect on ROS production in the gastrointestinal tract[234]. In their analysis, Schwerd *et al* identified eight patients with *NOX1* variants impacting on ROS production, presenting with variable phenotypes including four patients with Crohn's disease, three of whom had fistulating disease[234]. Interestingly, a single case report from Germany detailed a hemizygous *NOX1* variant leading to altered host antimicrobial immune function and impaired immune signalling through *NOD2*, illustrating a direct link from *NOX1*, and related variants, to Crohn's disease phenotype, characterised by altered *NOD2* signalling[235].

This study has several strengths and limitations; in an effort to synthesise genetic variation across a number of genes and complexes we utilised a contemporary whole gene deleteriousness score, this facilitated integration with downstream statistical analysis but does not resolve causation to an individual variant basis. Additionally, as we were using strict quality and mapping filters, we may have failed to include some true variation in genes. *NOX1* had only a single patient with a GenePy score of >0, quality filtering removed other potentially correct variants called in the WES data. This strategy is conservative and follows best practice guidelines[153]. In addition, due to mappability issues NCF1 was excluded from all analysis. It has two closely related pseudogenes, *NCF1B* and *NCF1C*, where a premature stop codon is present in both, however they have >99.5% sequence homology with *NCF1*[236]. This creates significant issues in mapping of sequences to

the correct part of the genome, and reliably calling variants in whole exome sequencing data. Finally, the variants included in calculation of the GenePy scores are limited to the intersection of the BED files for that capture kit, this is required to avoid addition calls in patients sequenced on capture kits with better coverage. Despite this we may be forced to exclude good quality deleterious variants due to differences between the parts of the exome captured by SureSelect v5 and v6. Longitudinal follow-up data and a larger number of patients sequenced with recent capture kits are additional strengths of this study, although longer follow-up data would provide more opportunity for a fistulating phenotype to emerge and potentially produce a stronger CPH model. A specific independent cohort to test these models in would provide additional strength to these data.

5.4.1 Conclusion

This chapter presents evidence of digenic variation across NADPH oxidase genes and NOD2 in the fistulating disease phenotype, with patient harbouring high deleteriousness scores being at significantly higher risk of disease. We identify *HMOX1* and *NOXO1* as potential drivers of a Crohn's disease subtype. Further functional work-up of patients with high GenePy scores is desirable to confirm an effect on ROS production or bacterial clearance. These results require replication in independent cohorts with longitudinal phenotyping, however these data provide a framework for personalising care and predicting outcome through genetic sequencing.

Chapter 6 *NOD*- and *IL17*-signalling characterise the ileal transcriptome in paediatric Crohn's disease

Chapter summary- *This chapter utilises contemporary targeted autoimmune RNA sequencing, in parallel to single-cell sequencing, on ileal tissue derived from paediatric Crohn's disease and controls. We establish a 31-gene signature characterising treatment naïve Crohn's disease patients. The CSF3R gene is a hub within this module and is key in neutrophil expansion and differentiation. Antimicrobial genes including S100A12 and the calprotectin subunit S100A9 were significantly upregulated. Gene-enrichment analysis confirmed upregulation of the IL17-, NOD- and Oncostatin-M-signalling. An upregulated gene-signature was enriched for transcripts promoting Th17-cell differentiation and correlated with prolonged time to relapse. Single-cell sequencing of TN-CD patients identified specialised epithelial cells driving differential expression of S100A9. Cell groups, determined by single-cell gene-expression, demonstrated enrichment of IL17-signalling in monocytes and epithelial cells.*

Chapter contributions- *Patients were recruited and biopsies were retrieved by James Ashton and Rachel Haggarty. RNA was extracted from biopsies by Konstantinos Boukas. Histological analysis was conducted by Bhumita Vadgama. Targeted RNA sequencing was performed by Konstantinos Boukas and James Ashton. Fresh biopsies for Drop-Seq single cell sequencing were processed by James Davies. Drop-Seq sequencing and analyses was performed by James Davies. Targeted RNA analyses were performed by James Ashton. Integration of single-cell and targeted RNA sequencing was performed by James Ashton with help with James Davies.*

Supplementary data can be found at https://doi.org/10.5258/SOTON/D1657

6.1 Background

Paediatric-onset Crohn's disease is a heterogenous condition characterised by chronic, relapsing and remitting inflammation, largely of the intestinal tract. Paediatric-onset Crohn's disease has a greater genetic contribution to pathogenesis compared to adult-onset disease, with multiple genes and pathways implicated in the inflammation observed in the condition[59]. These genes are largely centred on innate and adaptive immune pathways, cytokine signalling pathways, and bacterial recognition and response pathways[91]. Recently Mendelian causes of inflammatory bowel disease (IBD) have given additional insights into causes and risk factors for polygenic disease, with variation in a number of genes including *IL10* pathway, *NOD2* and NADPH oxidase complex genes being implicated in both forms of disease[188,209,219]. Non-Mendelian forms of Crohn's disease may present with similar phenotypic appearance. However, it is becoming increasingly clear that individual patients are likely to have a specific molecular diagnosis related to their underlying genetic variation. This may present through either a limited number of genes (oligogenic IBD) or through interaction of many genes (polygenic IBD), often resulting in perturbation of inflammatory pathways common to all genetic causes[91]. The ability to determine this genetic signature within an individual patient will bring new opportunities for predicting disease outcome and personalisation of therapy[214].

RNA sequencing allows identification of abnormal gene expression, and specific gene signatures, which are associated with disease subtypes. This information allows insight into the biological processes underlying pathways driving inflammation. Previous studies have identified differentially expressed genes, including *OSM* and *TREM1,* associated with disease onset, treatment response and have been able to predict disease course[67,237,238]. Whilst there is clear utility in determining markers of disease in blood, insights from extra GI tissues are likely sub-optimal to elucidate drivers of intestinal inflammation[239]. In addition, bulk RNA sequencing

where all cell types resident within a single biopsy sample are assessed concurrently, may to fail to sequence lowly expressed genes, with reads being taken up by housekeeping transcripts that provide no biological insight[240]. Contemporary efforts at targeted sequencing in cancer have revealed novel pathways and genes associated with disease, and provided clinical diagnostic utility[241,242]. Furthermore, integration of targeted RNA sequencing, single cell sequencing and clinical outcome data provides the opportunity for improved molecular profiling of patients, garnering understanding of the specific cells driving inflammatory pathways whilst simultaneously using RNA gene signatures to stratify patients. Recently, single cell analysis in 22 Crohn's disease patients determined a specific transcriptomic module from cells derived from the lamina propria, which was reproducibly found in bulk RNA sequencing and was associated with failure to respond to anti-TNF therapy[243].

In this study we apply cutting-edge autoimmune targeted RNA sequencing of ileal biopsy tissue from a cohort of paediatric Crohn's disease patients. We utilise these data to characterise patients by underlying gene transcription signatures, identify differentially expressed genes impacting on signalling pathways in treatment naïve and established disease patients. We integrate single cell RNA sequencing performed on a limited number of patients to determine cell populations driving specific gene expression.

6.2 Methods

Paediatric IBD patients were recruited through from the Paediatric Gastroenterology service at the Southampton Children's Hospital. All patients were diagnosed under the age of 18 years, according to the modified Porto criteria[2]. Two patient populations were recruited, the first consisted of patients referred to paediatric gastroenteology with a suspected diagnosis of IBD, recruited prior to diagnostic endoscopy. Patients who were diagnosed with Crohn's disease, who had successful ileoscopy, were termed the treatment-naïve Crohn's disease group. Children who had a normal endoscopy and were not diagnosed with IBD or any other gastrointestinal

pathology, were included in a control group. The control group were followed-up for a minimum of 6 months to confirm there was no subsequent diagnosis of IBD. The second population consisted of patients with established (ED) Crohn's disease undergoing routine endoscopy for reassessment, termed the established disease group, figure 15.

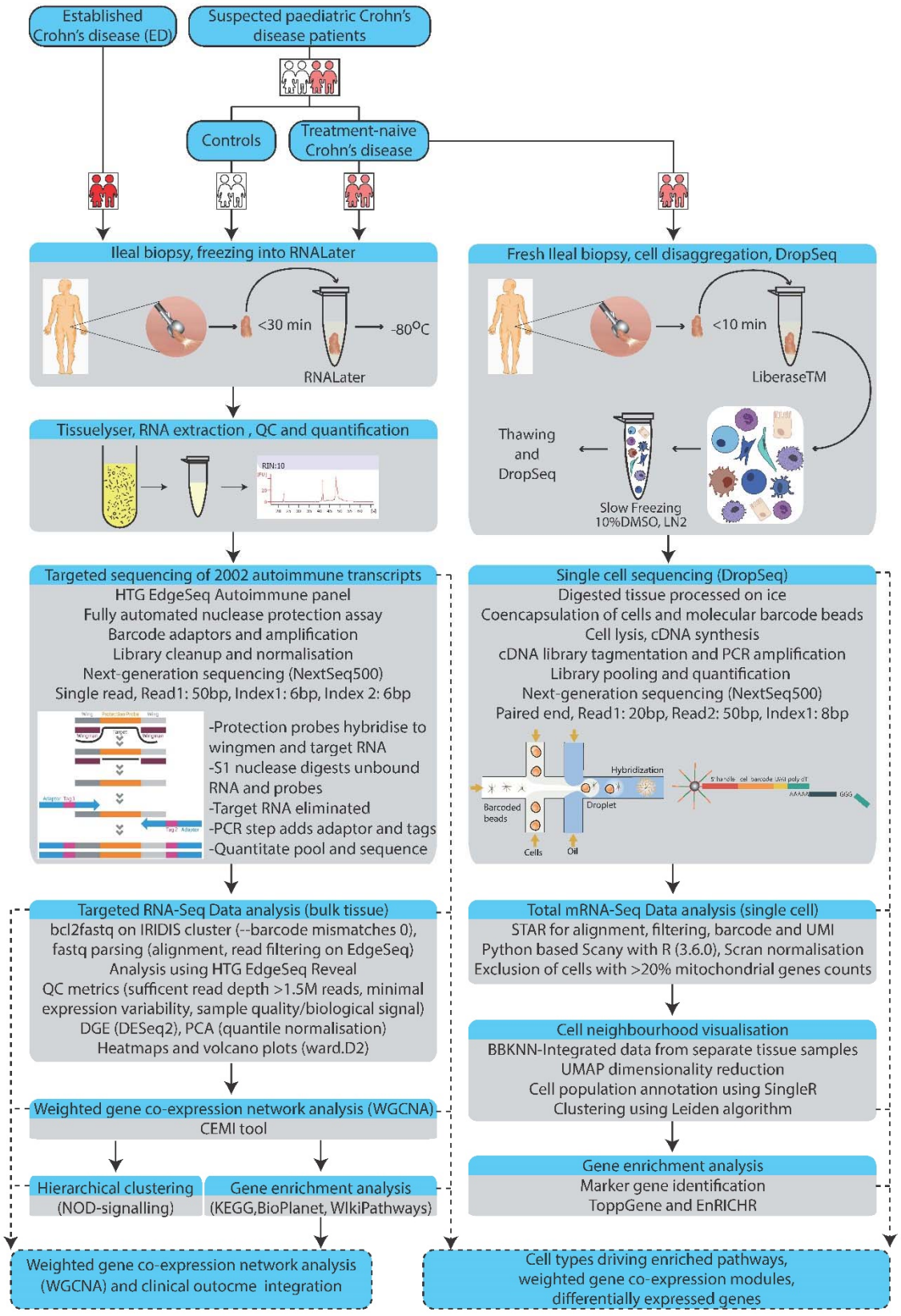

Figure 15- Summary of patient recruitment, sample processing and data analysis pipelines.

Patients were recruited in two groups, established Crohn's disease (ED) and suspected Crohn's disease patients, consisting of treatment-naïve patients (TN) and controls. All

groups underwent endoscopy with ileal biopsy. All patients had ileal biopsies retrieved and stored in RNAlater at -80. These biopsies underwent bulk RNA extraction and subsequent targeted RNA sequencing. A subgroup of TN patients had fresh ileal biopsies processed for single cell sequencing. Data quality control and processing steps can be seen in the figure. Integration of targeted RNA sequencing and single-cell sequencing was conducted following individual pipeline analyses.

6.2.1 RNA sequencing of terminal ileal biopsies

6.2.1.1 Sample acquisition, processing and storage

Terminal ileal biopsies were obtained during endoscopy and immediately placed into a cryovial containing 1ml of RNAlater (Sigma Aldrich), and frozen at -80°C within 30 minutes from collection. The diameter of each biopsy was an average of 2.5mm (range 1-4.5mm) with the mean volume of 27mm^3 (equivalent of 30mg).

6.2.1.2 RNA extraction

Frozen biopsies were transferred to the WISH laboratory to be extracted. Biopsies were transported on dry ice and remained frozen in RNAlater. Biopsies were processed in batches of 12 using the Maxwell RSC simply RNAtissue kit.

Maxwell homogenisation solution was prepared using DX reagent to avoid foaming at 0.5%.

- Maxwell Homogenisation solution: Before use, chill homogenisation solution on ice. In 2,985µl of chilled Maxwell homogenisation solution, add 15 µL of reagent DX. Also add 60µl of 1-Thioglycerol.

6.2.1.3 TissueLyser

Biopsies stored in RNAlater were thawed to room temperature. A sterile filter tip was used to transfer the biopsies into the pre cooled tubes containing TissueLyser beads and incubated at

room temperature for 2 min to avoid freezing of lysis buffer. 200µl of the prepared Maxwell homogenisation solution was added in the appropriate bead containing tubes (tube 1-12). The tubes containing the bead, the biopsy and the homogenisation solution were placed into the insert of the TissueLyser LT adapter. Samples underwent lysing for 5 min. Tubes were then removed and centrifuged for 3min at full speed. Supernatant was then transferred into RNAse-free fresh tubes.

6.2.1.4 Maxwell processing

Solution preparation:

- 1-Thioglycerol/Homogenization Solution: 20µl of 1-Thioglycerol per millilitre of Homogenization Solution. A volume of 200µl of 1-Thioglycerol/ Homogenization Solution for each sample.
- DNase I Solution: 275µl of Nuclease-Free Water to lyophilized DNase I. 5µl of Blue Dye to the reconstituted DNase I as a visual aid for pipetting. Dispense the DNase I Solution into single-use aliquots in nuclease-free tubes.
- Cartridge Preparation (maximum of 16 samples. 1sample per cartridge). A Plunger was placed in well 8 of each cartridge. 0.5ml Elution Tubes were placed in the Deck Tray. 50µl of Nuclease-Free Water was added to the bottom of each Elution Tube.

200µL of Lysis buffer was added to the 200µL of supernatant from the Tissuelyser and vortexed vigorously, 400µl was then transferred to well 1 of each of the 12 Maxwell RSC cartridges and 5µl of DNase I solution was added to well 4. The Maxwell instrument was run on the simplyRNA Tissue method. The resulting solution was transferred into 2 aliquots of 20µl and stored at -80^{o}C.

6.2.1.5 RNA quantification and quality check

The Bioanalyzer was used in conjunction with the Agilent RNA 6000 Nanokit. The samples were loaded onto the Gel-Dye mix RNA chip following priming. The chip was vortexed and run on the

Agilent Bioanalyzer to give RIN values and RNA concentrations for each extracted sample. Traces and values for each batch of samples can be seen in supplementary data.

6.2.2 Targeted RNA sequencing

The contemporary HTG EdgeSeq Autoimmune Panel was used to measure mRNA expression levels in 2002 genes associated with autoimmune disease, including inflammatory bowel disease[146]. Briefly, an RNA sample for each patient was thawed, diluted (to standardise concentration across all samples) and loaded onto the HTG EdgeSeq 96 well-instrument. Utilising custom chemistry for the autoimmune panel, the nuclease protection chemistry run was commenced. A PCR reaction using primers designed for the sequencing adaptors and sample barcodes was conducted, using 12 forward (F1-F12) and 8 reverse primers (RA-RH) by well location. 20 PCR cycles resulted in amplified autoimmune gene targets. Following a standardised PCR clean-up procedure, all samples were quantified using both Qubit fluorometry and qPCR techniques to ensure successful amplification of PCR product and to provide concentrations for sequencing dilutions. Based on quantification results, each sample underwent a dilution-based normalisation process to ensure that each sample was represented by a suitable (>750,000) number of reads. Due to high quantities of PCR product, 1/10 dilution in Tris buffer was performed for all samples. Individual libraries were pool at equimolar quantities into the final library for sequencing. The library was denatured and prepared for sequencing in line with Illumina and HTG practice guidelines. Sequencing was performed on the Illumina NextSeq platform.

6.2.2.1 RNA data processing

Output files from the NextSeq run were converted from BCL format to FASTQ files using the bcl2fastq, using standard parameters. Barcode mismatch filter was set to 0 to ensure reads were linked to the correct patient. FASTQ files were loaded into the HTG EdgeSeq parsing software. Probes sequences were used to identify specific genes. Using barcodes identifying specific

patients a gene expression count matrix was constructed for each gene and each patient. These were merged to form a single output file containing all genes and all counts. Downstream analyses of RNA data were performed using HTG Reveal software.

Gene counts were normalised using quantile normalisation (QN) based on best practice guidelines utilising the HTGEdge sequencing technology, specifically applying the previously developed immune-oncology targeted panel[148,149]. QN assumes the same distribution of gene expression across samples and therefore is there is expected to be significantly different expression between samples, such as for different tissues.

6.2.2.2 Downstream RNA analysis and quality control

Quality control of RNA sequencing data were performed in line with recommendations from HTG. HTG recommends a cut-off of 750,000 reads their data indicates biologically useful information can be gained from much lower sequencing depths.

Differential expression was assessed using DESeq2 package (Python, within Reveal software)[157]. Gene-co-expression networks enable regulatory hubs and gene-gene associations to be determined. CEMiTool was used to assess weighted gene co-expression networks within normalised data, and to determine modules and hub regulatory genes observed in different categorical groups[156]. Gene-Gene interactions within co-expressed genes were determined using the HitPredict database[164]. WGCNA (R package) was used to establish gene co-expression modules and assess correlation between continuous clinical outcome variables (time to relapse)[155].

We assessed for enrichment of genes in specific WGCNA modules, and differentially expressed genes (DEGs), in specific pathways using ToppFun[165], EnRichR[166] and BioPlanet[167]. Statistical analysis was performed using Reveal software and SPSS (v25, IBM). Multiple testing correction was conducted using false discovery rate (FDR) methodology.

6.2.3 Single-cell transcriptomic analysis

Tissue biopsies from IBD patients were digested in LiberaseTM research grade (Roche, UK, 2 h at 37 °C) and cryopreserved in 90% FBS (Gibco, UK), 10% DMSO (Sigma, UK). Prior to Drop-seq, cells were unbanked from cryo-storage and processed on ice. Co-encapsulation of single cells with genetically encoded beads was performed following the Drop-seq pipeline[244,245]. 1nl sized droplets were generated using microfluidic devices created in the Centre for Hybrid Biodevices, University of Southampton. Optimised microfluidics parameters were used, ensuring the generation of single-cell/single barcoded Bead SeqB (Chemgenes, USA) encapsulation events. Following encapsulation, ~4500 STAMPS (beads exposed to a single cell) from 1.2 ml of cell suspension were generated. 1000 STAMPS for each biopsy were taken further for library preparation (High Sensitivity DNA Assay, Agilent Bioanalyser, 12 peaks with the average fragment size 500 bp). Prepared libraries were run on an Illumina NextSeq (1 × 105 reads/cell) at the Wessex Investigational Sciences Hub laboratory, University of Southampton, to obtain single-cell sequencing data.

De-multiplexing of samples was performed using the bcl2fastq tool from Illumina. Alignment, read filtering, barcode and UMI counting were performed using kallisto-bustools[158]. For gene filtering, genes detected in less than 10 cells were excluded. Subsequent data analyses were run using the python-based Scanpy[159] with R (3.6.0) embedded via rpy2. High quality barcodes were selected based on the overall UMI distribution using EmptyDrops[246] to identify the number of true cells amongst empty beads. Cells of low quality with high fraction of counts from mitochondrial genes (20% or more), which indicates stressed or dying cells, were removed. Data was normalised using Scran[160]. Highly variable genes were selected using distribution criteria: min_mean=0, max_mean=4, min_disp=0.1. A single-cell neighbourhood graph, with data integrated from separate tissue samples, was computed using BBKNN[161], using 50 principal components that sufficiently explained the variation within the data. Data was visualised using

Uniform Manifold Approximation and Projection (UMAP), with the Leiden algorithm used to identify cell clusters (r = 0.5, n_pcs=50)[247].

Cell type annotation was performed using SingleR (database: BlueprintEncodeData) and enrichment analysis (ToppGene and EnRICHR) using the top 50 marker genes of each cluster (linear regression FDR <0.05)[162,165,166].

6.3 Results

Ninety-one patients with ileal biopsies were recruited to the study. Confirmation of diagnosis by Porto criteria resulted in 70 patients being included: 27 TN Crohn's disease patients; 17 controls; and 26 ED Crohn's disease. Twenty-one patients that were diagnosed with IBDU or ulcerative colitis following endoscopy were excluded. A single ED sample failed quality control (QC2) leading to exclusion. Patient characteristics can be seen in table 15.

*Table 15- Patient characteristics for those included in thetranscriptomic analysis. *following quality control*

	Treatment naïve patients	**Established disease patients**	**Controls**
Number of patients	27	26	17
Number of biopsies included in analysis	27	25	17
Mean age at diagnosis (range)	13.46 years (9.26-16.75)	12.03 years (5.79-15.3)	N/A
Mean age at ileal biopsy (range)	13.46 years (9.26-16.75)	14.64 years (8.68-17.70)	11.56 years (6.01-15.10)
Percentage female (%)	29.6% (n=8)	42.3% (n=11)	35.3% (n=6)
Percentage with ileitis at biopsy (histologically proven)	74% (n=20)	27% (n=7)	0%

6.3.1 Targeted RNA sequencing of 2002 autoimmune genes

6.3.1.1 A thirty-one gene module characterises treatment-naïve Crohn's disease patients

Gene modules associated with TN patients, controls and ED patients were established. Three gene modules where identified containing 104, 47 and 31 genes respectively. Module 3, containing 31 genes, was significantly upregulated in TN patients (normalised expression score, NES 3.07, p=0.0006) and downregulated in controls (NES -2.73, p=0.0004), table 16 + figure 16A-B. Module 3 did not correlate with ED patients. Module 1 co-expression was significantly increased in controls (NES 1.95, p=0.0004) and significantly decreased in TN (NES -1.7, p=0.0015) and ED (NES -2.35, p=0.0012) patients. Whilst module 2 contained 47 co-expressed genes it was not significantly associated with any patient group.

Table 16- Weighted gene co-expression analysis and association with patient groups.

	Control adjusted p-value	***Control network expression score***	***Crohn's established disease adjusted p-value***	***Crohn's established disease network expression score***	***Crohn's treatment-naïve adjusted p-value***	***Crohn's treatment-naïve network expression score***
Module 1	*0.00044*	*1.95*	*0.0012*	*-2.35*	*0.00153*	*-1.7*
Module 3	*0.00044*	*-2.73*	*0.10693*	*-1.39*	*0.00062*	*3.07*

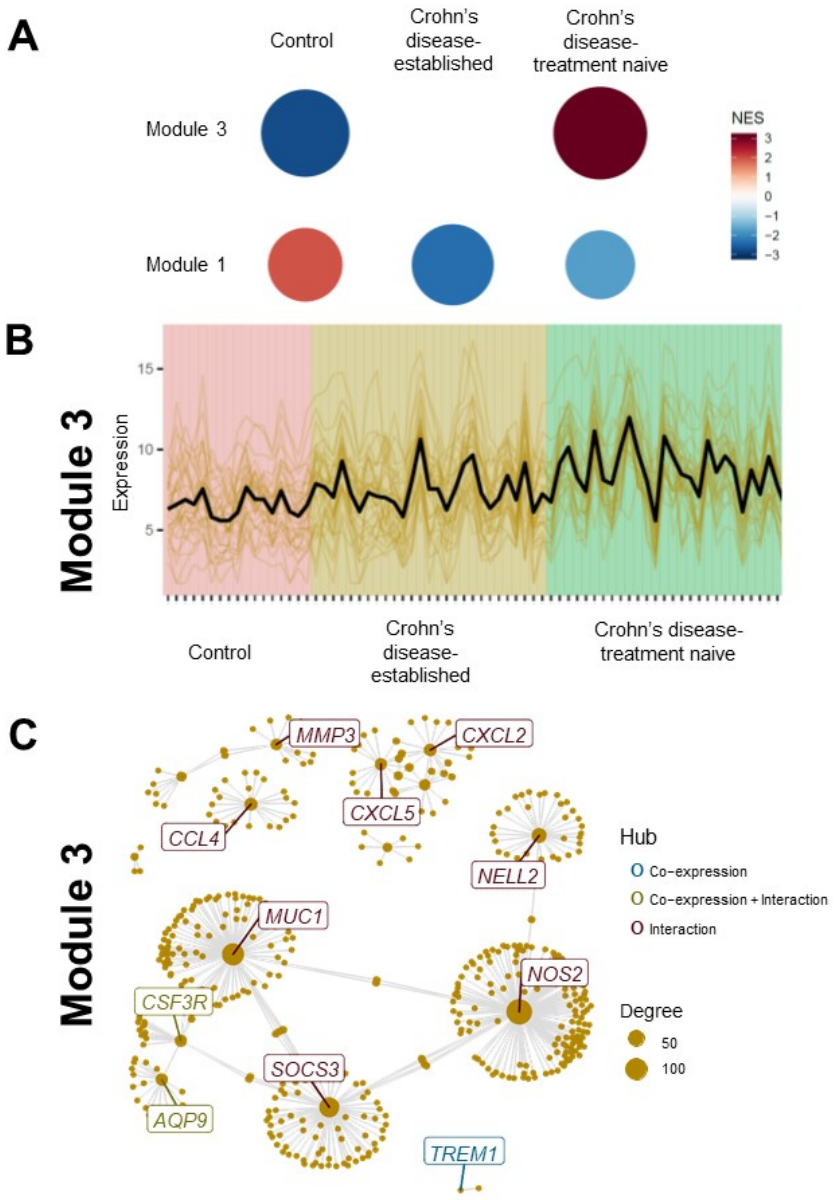

Figure 16- Weighted gene co-expression analysis reveals a 31-gene module specifically upregulated in treatment-naïve Crohn's disease patients. ***A****) Normalised expression score (NES) of modules correlated with patient groups, module three genes have increased expression in treatment-naïve patients only. The NES represents a scaled measure of expression compared to 'normal' within the samples. A 'normal' expression is represented by 0.* ***B)*** *Mean expression of module three genes across individual patients in the three patient groups. Each point on the X axis represents an individual patient.* ***C)*** *Identification of hub co-expression and interacting genes within*

module three, utilising the HitPredict database demonstrates 11 hub genes within the 31-gene module.

6.3.1.2 The treatment-naïve gene co-expression module is associated with upregulation of Oncostatin-M and NOD-signalling pathways

Thirty-three pathways were significantly associated with module 3 genes following multiple testing correction. The most implicated pathway was the Oncostatin-M (OSM) signalling pathway (adj-p=$4.47x10^{-22}$), upregulation of which was seen in treatment-naïve patients. OSM signalling results in activation of proinflammatory pathways including *JAK/STAT3*, *MAPK*, and *PI3K.* Interestingly, activation of the NOD-signalling pathway was also significantly enriched for in module 3 (adj-p=0.0008). Table 17.

		IL17 signalling		**NOD-signalling**	
		WGCA genes	DEGs	WGCA genes	DEGs
Bioplanet*	Odds ratio	43.01	11.59	30.36	10.23
	Adj P value	0.4	0.016	0.0008	$9.99 x-10^{-10}$
KEGG*	Odds ratio	76.31	14.96	18.12	6.19
	P value	$2.88 x-10^{-16}$	$1.23 x-10^{-19}$	0.0003	$4.51 x-10^{-9}$
WikiPathways (Human)*	Odds ratio	-	5.61	15.74	8.48
	P value	-	0.06	0.59	0.0007

*Table 17- DEG and WGCNA pathway enrichment analysis. *Data collated through ToppFun, EnRICHR, and g:Profiler*

6.3.1.3 CSF3R appears to act as a regulatory hub within the treatment-naïve module

In order to assess regulatory hub genes within the module 3 network we performed an interaction network analysis. This revealed six genes with >3 hub interactions within the module, figure 16C. Of these six genes *CSF3R* was also co-expressed within the network with *AQP9. CSF3R* is the receptor for colony stimulating factor 3, it's related pathway functions to control expansion, differentiation and role of neutrophils, with highly deleterious variants in *CSF3R* resulting in congenital neutropenia.

6.3.1.4 S100A9 (Calprotectin subunit) antimicrobial gene are significantly upregulated in treatment-naïve patients

Differential gene expression was assessed between TN patients, controls and ED patients using DESeq2 (supplementary data 3 and 4). Following multiple testing correction, 342 genes were significantly differentially expressed between TN patients and controls, 259 of which were upregulated in TN patients. Figure 17A. The five most significant upregulated DEGs in TN patients were *S100A12* (fold change 32.5, adj-p= $2.6x10^{-15}$), *CXCL8* (IL-8)(fold change 20.2, adj-p= $5.5x10^{-15}$), *S100A9* (fold change 12.6, adj-p= $9.1x10^{-14}$), *FCGR3A/B* (fold change 7.3, adj-p= $5.0x10^{-13}$), and *IL1RN* (fold change 8.1, adj-p= $3.9x10^{-12}$). The difference between TN and ED patients was less marked whereby just 12 genes were significantly upregulated in TN patients compared to ED patients (Figure 17D, specify fold change and FDR cut-offs). The five most significantly upregulated genes were *CSF3R* (fold change 2.5, adj-p= 0.0055), *IL1RN* (fold change 3.0, adj-p= 0.0055), *S100A9* (fold change 3.6, adj-p= 0.0055), *S100A12* (fold change 4.8, adj-p= 0.0055) and *AQP9* (fold change 3.6, adj-p= 0.03).

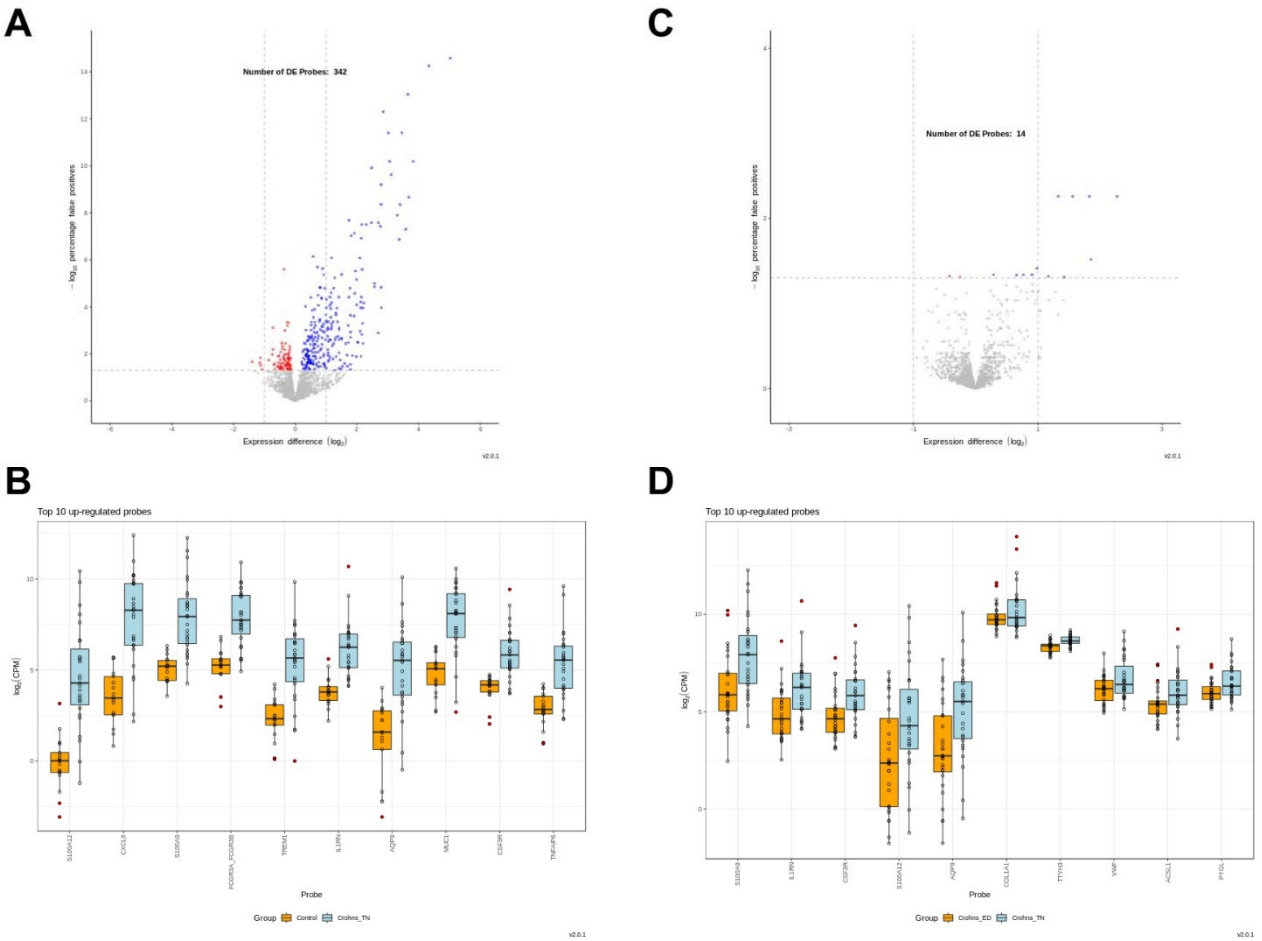

Figure 17- Differentially expressed genes (DEGs) were identified utilising the DESeq2 package. ***A)*** *Volcano plot demonstrating DEGs between treatment-naïve Crohn's disease (TN CD) patients vs controls. A total of 342 genes were differentially expressed between groups.* ***B)*** *The top 10 upregulated DEGs between TN CD patients and controls.* ***C)*** *Volcano plot demonstrating DEGs between TN CD patients vs established Crohn's disease patients. A total of 14 genes were differentially expressed between groups.* ***D)*** *The top 10 upregulated DEGs between TN CD patients and established Crohn's disease patients, highlighting S100A9, S100A12 and CXCL8 (IL8) as remaining significantly upregulated in TN CD patients.*

6.3.1.5 Treatment naive Crohn's disease is characterised by elevated IL17- and NOD-signalling.

We utilised gene enrichment analysis to assess for pathways associated with TN patients using genes implicated by *both* WGCNA and differential gene expression analysis. The IL17- and NOD-signalling pathways as recurrently implicated across multiple gene enrichment databases (Supplementary data).

6.3.2 Gene expression differences are not driven solely by inflamed tissue

To ensure differences between groups were not driven solely by active inflammation we conducted analysis on inflamed vs. non-inflamed tissue. Whilst there were differentially expressed probes between inflamed and non-inflamed biopsies (n=185), none of the top differentially expressed genes were the same as those seen between TN CD and controls, figure 18.

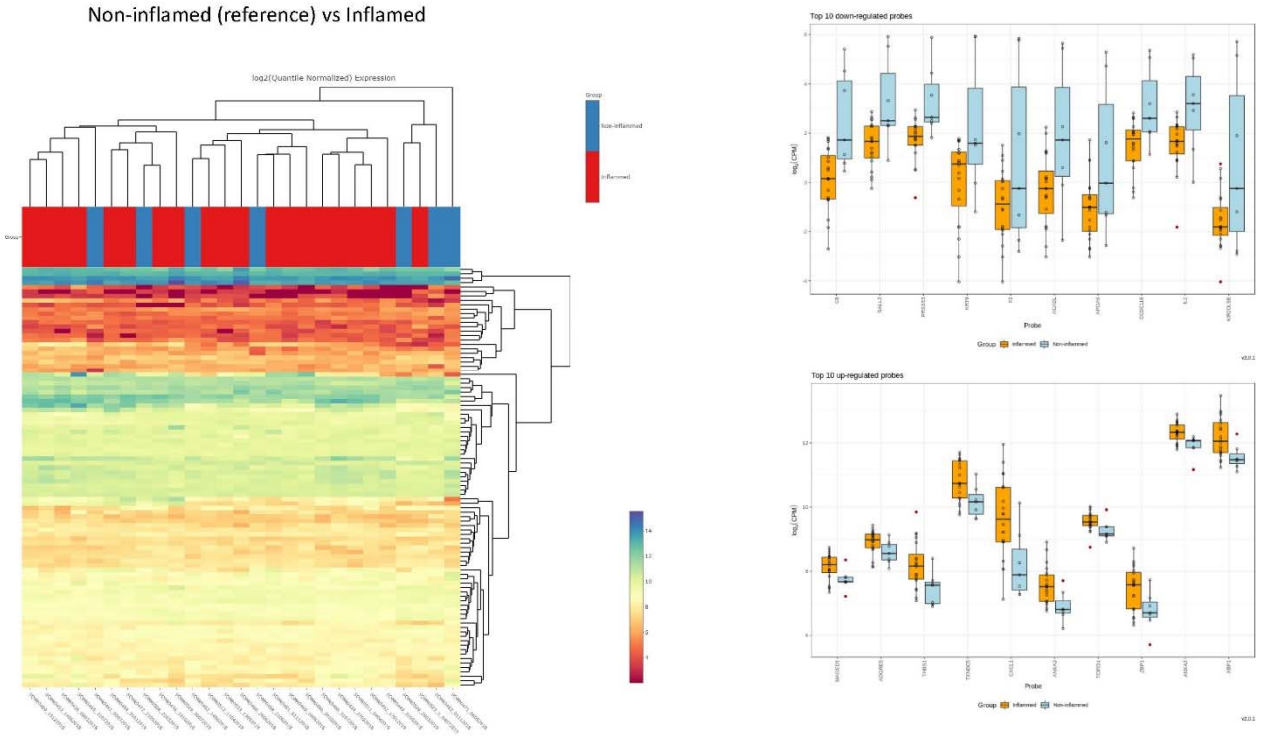

Figure 18- Comparison of inflamed vs non-inflamed tissue in treatment naïve and established Crohn's disease patients. Hierarchical clustering does not demonstrate clustering of inflamed tissue. Top differentially expressed genes between inflamed vs non-inflamed are different to those seen between treatment-naive Crohn's disease and established Crohn's disease.

6.3.2.1 Gene expression in NOD-signalling pathway clusters Crohn's disease patients distinctly from controls

NOD2 is the most heavily implicated gene in Crohn's disease pathogenesis, with variation in interacting genes, including *XIAP, RIPK2* and *ATG16L1,* described as increasing risk of Crohn's disease[91]. The NOD-signalling pathway was one of the most significantly enriched in treatment naïve patients. We hypothesised that aberrant NOD-signalling gene expression could be used to classify patients from controls. Utilising 95 genes curated by the HTG platform as being in the NOD-signalling pathway we performed hierarchical clustering of all patients (quantile normalised data, average distance clustering), figure 19. Three clusters emerged, with 8 patients not falling into any of these clusters. All but three of the controls grouped together in cluster 3, characterised by low *CXCL8 (IL-8), CXCL2 and CASP5* expression. In contrast clusters 1 and 2 had increased expression of pro-inflammatory *CXCL1* and *STAT1. NOD2* expression itself did not appear to differ between groups.

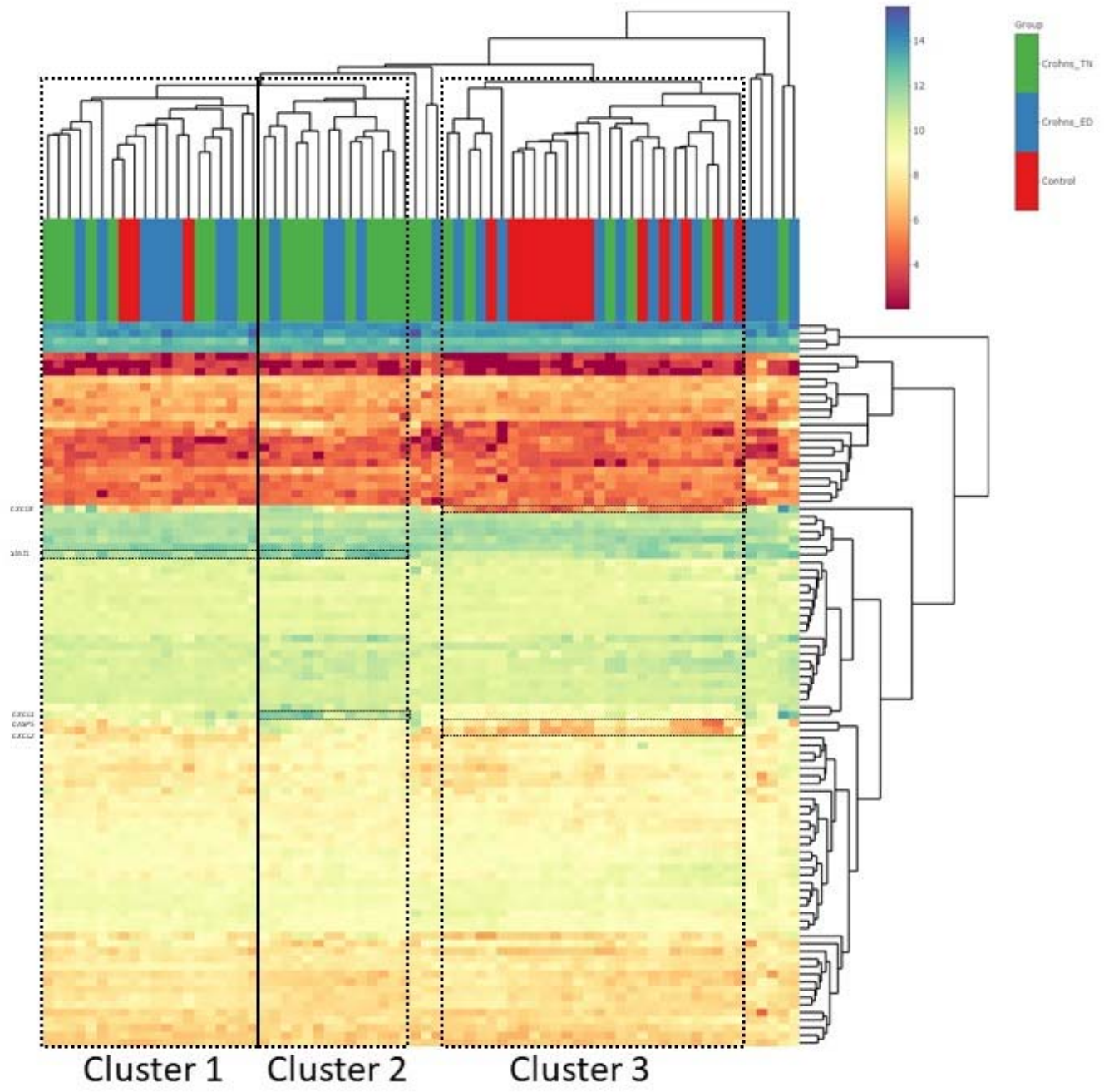

Figure 19- Hierarchical clustering (quantile normalised data, average distance clustering) of all patients using 95 genes in the NOD-signalling pathway. Clustering demonstrated grouping of controls together, characterised by reduced expression of CXCL8 (IL8), CASP5 and CXCL2 (cluster 3). Clusters 1 and 2, mainly consisting of Crohn's disease patients, were characterised by increased CXCL8 (IL8) and STAT1 expression, with cluster 2 also having increased CXCL1 expression.

6.3.3 Clinical data integration

6.3.3.1 Th17-cell differentiation gene module associated with prolonged time to relapse

Using WGCNA, we assessed whether co-expressed gene modules were predictive for patient prognosis. The number of days from diagnosis to relapse were entered as a continuous variable. Six of the 27 treatment naïve Crohn's disease patients had not relapsed at the time of analysis and their time to relapse was set as the number of days from diagnoses to most recent follow up. A co-expression module ('blue', supplementary data 5) characterised by 55 significantly upregulated genes (p= <0.05), was positively correlate with time to relapse (correlation coefficient 0.36, p=0.07) (Figure 20A). The 55 genes comprising this module were enriched for Th17 cell differentiation (KEGG, adjusted p value 9.21×10^{-11}), supplementary data.

In order to determine if any DEGs were associated with early relapse we stratified the 27 treatment naïve CD patients into those that relapsed within 16 weeks (n =14) and those that either relapsed after 16 weeks or did not relapse (n =13). Expression of the *PI3* gene, an antimicrobial peptide expressed in response to lipopolysaccharide and IL17-signalling, was significantly upregulated in patients with early relapse (fold change = 2.13, adjusted p=0.0358, table 18)[248].

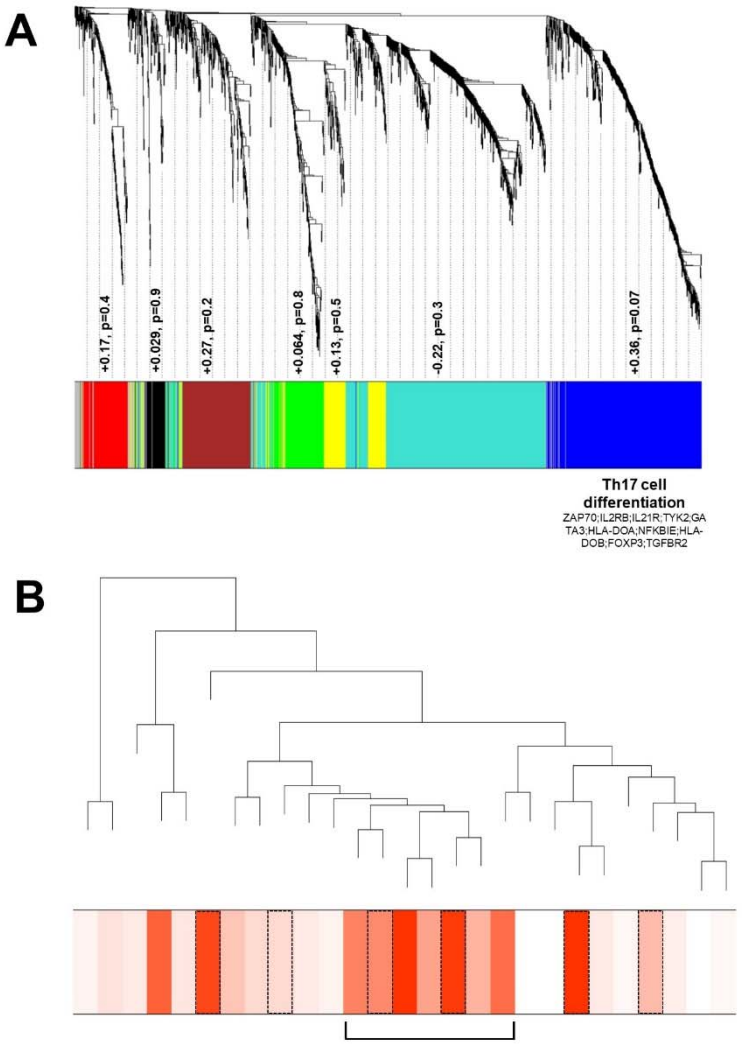

Figure 20- Weight gene co-expression network analysis determines a 'blue' gene module, containing 55 significantly upregulated genes, associated with increased time to relapse. ***A)*** *Cluster dendrogram indicating genes contained within each coloured module. The distance from gene to gene indicates the similarity in expression profile, and determines the module in which that gene lies. Upregulated genes within this cluster (n=55) were associated with a Th17 cell differentiation signature (KEGG, adjusted p value* 9.21×10^{-11}*). Correlation of each module expression in patients, with time to relapse, revealed patients with increased expression of the 'blue' module also had increased time to relapse (correlation co-efficient 0.36, p=0.07).* ***B****) Clustering of treatment naïve Crohn's disease patients by similarity of gene expression across all*

sequenced probes revealed a small cluster with increased time to relapse (brighter red = increased time to relapse).

Table 18-

Gene	Fold Change in gene expression Early relapse vs. No early relapse	Raw p value Early relapse vs. No early relapse	Adjusted p value Early relapse vs. No early relapse
PI3	2.31	1.94E-05	3.58E-02*
CEACAM6	2.08	1.36E-04	1.25E-01
CASP5	1.87	1.10E-03	4.39E-01
CCL23	-1.91	1.20E-03	4.39E-01
IL1R2	1.91	1.10E-03	4.39E-01
G0S2	1.80	1.70E-03	5.09E-01
HLA-DMA	-1.26	2.30E-03	5.53E-01
RCAN1	1.36	3.20E-03	5.53E-01
RNASET2	-1.19	3.30E-03	5.53E-01

Differentially expressed genes in treatment-naïve patients with early relapse vs no early relapse.

**significant following multiple testing correction*

6.3.4 Single-cell sequencing of treatment naïve Crohn's disease patients

6.3.4.1 Specialised gastrointestinal epithelial cells drive differential expression of S100A9, a key molecule in IL17 signalling pathway.

Single cell RNA sequencing of ileal biopsies from two IBD patients (patient IDs SOPR0472 and SOPR0476) were undertaken to identify distinct cell populations driving the enrichment of specific biological pathways. After filtering, a total of 1458 cells (IBD2=710 cells, IBD3=748 cells) and 4,731 genes were used for analyses. ScanPy UMAP dimensionality reduction of the 1458 cells integrated from both patient biopsies (SOPR0472=710, SOPR0476=748 cells) was performed (Figure 21A),

followed by annotation of cell populations as CD8+ effector memory T cells (CD8+ Tem), memory B-cells, monocytes, epithelial cells and plasma cells, contributed by cells from each sample (Figure 21A, 20B). We identified 9 distinct clusters within the SingleR annotated populations (Figure 21C), with each cluster defined by the expression of unique marker genes (Figure 21D). Interestingly, the 2 genes that most strongly defined cluster 7 as epithelial cells included both calprotectin subunits, *S100A8* and *S100A9*. The *S100A9* gene was one of the top DEGs between TN patients and controls. Gene ontology analysis of the 50 genes that define cluster 7 (n=50) revealed associations with inflammatory immune response processes, including a epithelial defence response to bacterium (adj-p=$2.2x10^{-5}$).

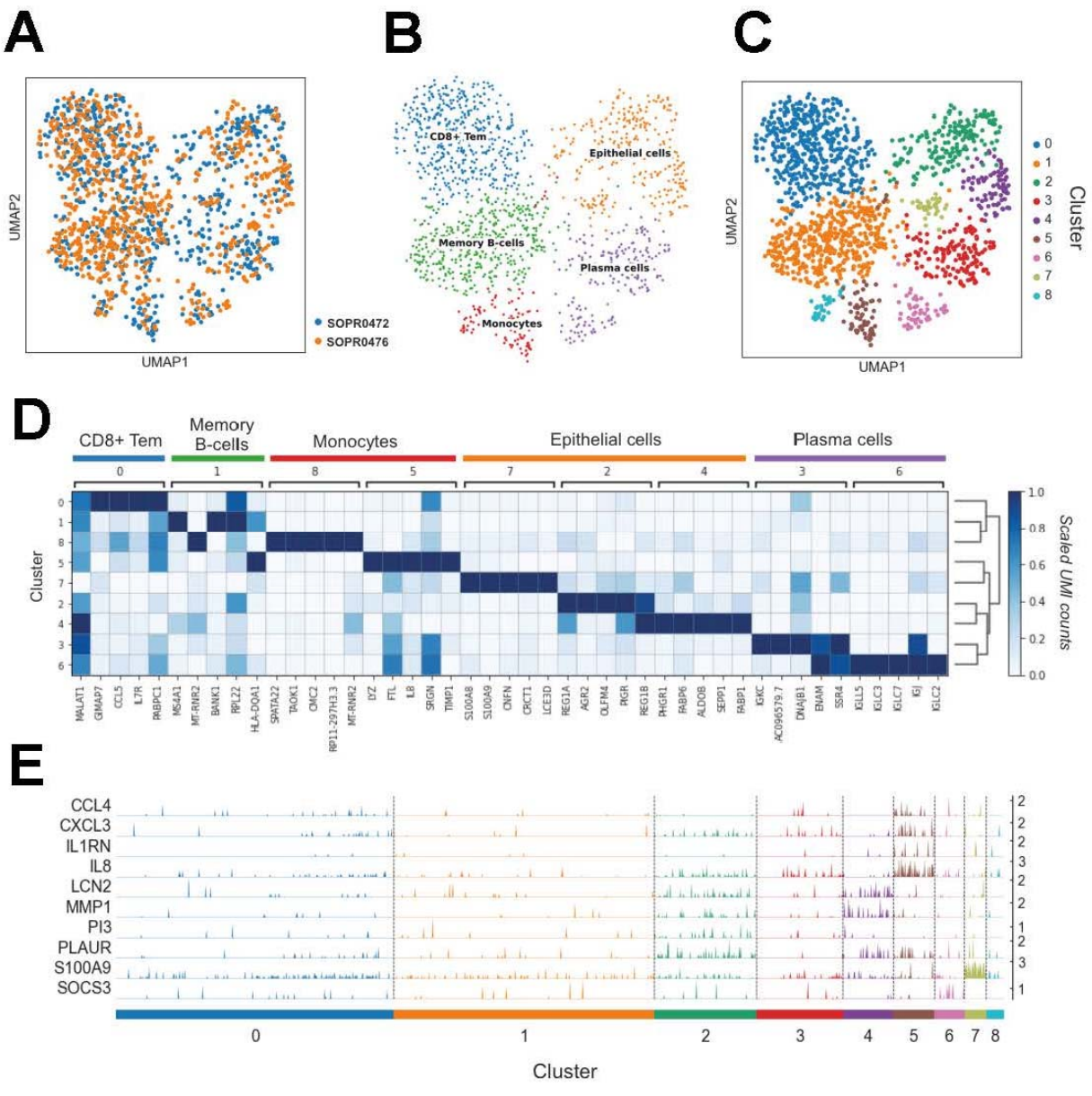

Figure 21- Single cell transcriptomics identifies monocyte and epithelial cells populations which contribute to the IL-17 signature in IBD. ***A)*** *UMAP plot of 1458 cells originating from two independent, digested IBD ilea samples (IBD2=710 cells, IBD3=748 cells), integrated into one neighbourhood graph using BBKNN (ScanPy, n_pcs = 50, 1999 highly variable genes (min_mean=0, max_mean=4, min_disp=0.1)).* ***B)*** *SingleR database annotation (database: BlueprintEncodeData) assigned cells into populations of CD8+ Tem, memory B-cells, monocytes, epithelial cells and plasma cells.* ***C)*** *Leiden clustering (r = 0.5), identified nine clusters (0-8) amongst the cell populations.* ***D)*** *Hierarchical clustering matrix plots with top 5 marker genes (scaled UMI counts) for each Leiden cluster are displayed.* ***E)*** *Barplots displaying frequency and amplitude expression of indicated gene transcripts, that characterise treatment naïve patients in IBD. Bars are colour coded for cells as in panel* ***C)****, identified using Leiden clustering. Each bar shows the Scran normalised expression level of indicated transcript, in a given cell.*

6.3.4.2 Specific cell populations drive gene expression seen in treatment naïve patients

Identified cell populations were interrogated for the expression of genes included in the gene module 3, characterising TN IBD patients. Clusters showing elevated expression of module genes included cluster 5 monocytes (*CCL4, CXCL3, IL1RN, IL8* and *PLAUR*) and cluster 7 (*S100A9*), cluster 4 (*LCN2, MMP1* and *PLAUR*) and cluster 2 (*IL8, LCN2, MMP1, PI3* and *PLAUR*) epithelial cells (Figure 21E).

6.3.4.3 Enrichment of Il17-signalling was observed in monocytes and specialised epithelial cells

We hypothesised that specific cell populations were driving the pathways implicated by targeted bulk RNA sequencing. IL17-signalling genes were enriched for in the specialised epithelial cell cluster 7, and in cluster 5 (monocytes). Significant enrichment for the IL-17 pathway (adj-p=0.016)

in cluster 7 was largely attributed to elevated expression of *S100A7, S100A8* and *S100A9*. Cluster 5 monocytes markers (n=50) were enriched for genes involved in IL-17 signalling pathway (adj-p=0.0012); due to elevated expression of *CXCL8* (IL8), *CXCL3, IL1B, NFKBIA, HSP90B1,* as well as the broad pathways of 'response to cytokines' (adj-p=$1.1x10^{-9}$) and the 'inflammatory response' (adj-p=$2.6x10^{-6}$).

6.4 Discussion

We present data utilising a targeted autoimmune panel for the first time in IBD. These data demonstrate an ileal gene expression signature identified through both WGCNA and differential gene expression analysis, characterised by NOD- and IL17-signalling, and specific to treatment-naïve Crohn's disease patients. Contemporary single cell analysis identified specialised epithelial cells driving differential expression of several genes, including the calprotectin subunits (*S100A8*/*S100A9*). Enrichment of IL17-signalling genes was observed in both this epithelial cell cluster as well as specific monocytes. Finally, we identify a gene module characterised by the Th17 cell differentiation pathway that is observed in patients with increased time to relapse following diagnosis.

Our data confirms and corroborates findings in a number of previous studies, implicating *OSM*, *CXCL8* and *AQP9* as upregulated DEGs in IBD patients compared to controls[67,237]. We also provide tissue-specific evidence for several upregulated genes that have previously been observed in blood, including *TREM1*[238]. Considering the largest previous study detailing paediatric Crohn's disease ileal transcriptomic signatures we identify a large number of overlapping genes from their 'core iCD' expression module, including upregulation of *CXCL5*, *IL8* and *S100A9*[67]

Stratification of patients based on differential gene expression provides the opportunity to predict treatment response and patient prognosis. Recently the RISK cohort from North America has been used to develop prediction algorithms, specifically utilising integration of ileal transcriptome

profiles into a multifactorial risk score for stricturing disease, identifying an extracellular matrix gene signature associated with stricturing disease[120]. Haberman et al also identified a model, including APOA1 gene expression, able to predict 6-month steroid free remission. We identify a gene module, characterised by Th17 cell differentiation, associated with prolonged time to relapse. Results must be replicated in external cohorts and transitioning these models into clinical practice will be key for personalising therapy in patients.

Th17 cells and IL17 signalling are of great interest in IBD pathogenesis. Differentiated Th17 cells produce several effector IL-17 cytokines, promoting inflammation and mucosal pathogen clearance[249]. We implicate ileal activation of IL17 signalling within TN Crohn's disease patients, compared to both controls and ED patients. Several previous studies have not identified increased serum IL17 in Crohn's disease patients, however IL17 levels are increased in affected tissues[250,251]. We replicate these findings in paediatric patients and for the first time identify a specialised epithelial cell cluster and a monocyte cell cluster appearing to drive the IL17-signalling at a tissue level. It appears that the epithelial cells are a target tissue for IL17-signalling, resulting in heighten proinflammatory and anti-microbial response within this specific population of cells, characterised by secretion of calprotectin (*S100A8/S100A9*). This pro-inflammatory effect of IL17-signalling on epithelial cells has been observed in colonic epithelial cell cultures, with upregulation of CXCL8 and CXCL1 promoting neutrophil chemotaxis[252]. Our data identifies distinct cell clusters driving these processes, whilst targeting RNA sequencing replicates similar expression profiles within the ileum of treatment naïve patients.

Previously a single study has identified an expression module in ileal Crohn's disease, characterising IgG plasma cells, mononuclear phagocytes, activated T cells and stromal cells[243]. Through single cell sequencing of ileal tissue derived from two TN patients we were able to identify a novel group of epithelial cells driving differential expression of the calprotectin complex (*S100A8/S100A9*), which was echoed in the targeted sequencing of all patients. Typically, it has been thought that most calprotectin is mainly derived from colonic neutrophils, with small bowel

inflammation less reflected in faecal sampling[253]. These data indicate that ileal epithelial cells, rather than neutrophils, highly express *S100A8/S100A9* in Crohn's disease patients, driving the differential gene expression between patients and controls. Interestingly, previous transcriptomic analysis of only intestinal epithelial cells of paediatric Crohn's disease patients did not identify any differentially expressed genes correlated to active inflammation[254]. In their article, Howell et al do describe several genes, including *DEFA5*, *DEFA6, LYZ*, *PLA2G2A*, *CD40*, and *CD44*, that were differentially expressed between controls and TN Crohn's disease patients, concluding that there was minimal molecular impact of disease on the epithelial cells[254]. Identification of high expression of *S100A8/S100A9* in ileal biopsies, or staining for the calprotectin complex, may aid with histological diagnosis of small bowel Crohn's disease.

Within TN patients we also identify the upregulation of NOD-signalling genes, associated with bacterial recognition, response and proinflammatory downstream signalling[255]. Taken together, activation of these pathways infers increased inflammatory response to pathogenic bacteria, or an aberrant response to normal bacteria. Alternatively ineffectual bacteria clearance, related to a downstream 'hypoimmune' response, has recently been postulated as a cause of IBD[256]. Through WGCNA we identified a module of genes, associated with increased time to relapse, characterised by upregulation of the Th17 cell differentiation pathway. The control of Th17 differentiation is complex, several key cytokines, including IL-1β, IL-6, IL-23 and TGFβ, suppressing FOXP3 expression and inducing RORC-dependant Th17 differentiation[249]. We hypothesise that in patients able to mount a good Th17 response bacteria are cleared quicker, resulting in immediate reduction of chronic inflammation following induction therapy. Conversely at diagnosis in most Crohn's disease patients there is a huge upregulation of IL17- and NOD-signalling as a response to ineffectual bacterial clearance, resulting in chronic inflammation. Whether IL17-signalling is driving chronic inflammation in response to invasive bacteria, or as a primary 'hyperinflammatory' response is uncertain. However, it appears that downstream IL17-signalling within these cell populations may result in production of antimicrobial peptides and proinflammatory cell infiltration as part of the disease process.

This study has several strengths and limitations. Through a targeted sequencing approach we reduce the number of reads lost to 'house-keeping' genes, enabling identification of low expression probes which impact on biological processes, a method previously successfully applied in cancer samples[242]. Applying single-cell transcriptomics gave us the means to identify cell types driving differential gene expression, and gain insight into the biology of specific cells in disease. Despite this, analyses of primary patient biopsies are associated with challenges and our study has several limitations. Whilst analysis was limited to Crohn's disease patients and controls there remains heterogeneity between individuals, exacerbated by the relatively modest numbers in each subgroup. Additionally, at the time of analysis, follow-up duration was insufficient to assess WGCA and DEGs with long-term disease behaviour and response to therapy. Despite these limitations, applying novel targeted and single cell sequencing methods to Crohn's disease biopsy material, we uncover specific disease -associated pathways, identified the driver cell populations, and characterised a gene module associated with disease prognosis.

6.4.1 Conclusion

This study demonstrates the high granularity of targeted RNA sequencing to identify pathways in paediatric Crohn's disease, particularly the IL17- and NOD-signalling pathways. We identify a Th17 cell differentiation gene module associated with increased time to relapse in treatment naïve patients and utilise single-cell RNA sequencing to determine an epithelial cell population driving differentially expressed genes. Personalising therapy based on underlying molecular diagnosis and stratification is an exciting prospect. Replication of these findings is required integration of long-term outcomes may yield improved predictive models.

Chapter 7 Deleterious genetic variation within the NOD-signalling pathway is associated with reduced transcription of promotors of *NFKB*-signalling

Chapter summary- *Focusing on the NOD-signalling pathway, this chapter integrates genomic variation in key genes and protein-complexes through application of GenePy scoring. Using these data, we assess the impact of genetic variation on transcription of 95 genes in the NOD-signalling pathway. An overall hypoimmune response is observed, related to deleterious variation in many key genes in the pathway. Utilising these techniques we are able to stratify patients by perturbation of NOD-signalling in their IBD pathogenesis.*

Chapter contributions- *Patients were recruited and biopsies were retrieved by James Ashton and Rachel Haggarty. RNA was extracted from biopsies by Konstantinos Boukas. Targeted RNA sequencing was performed by Konstantinos Boukas and James Ashton. Targeted RNA analyses were performed by James Ashton. Whole exome sequencing data and GenePy scores were processed by Imogen Stafford. Analysis of genomic data and GenePy scores was by James Ashton. Integration of RNA sequencing and genomic data was by James Ashton.*

Supplementary data can be found at https://doi.org/10.5258/SOTON/D1657

7.1 Background

Integration of data types to assess the impact of genetic variation on gene expression, and immune function remains difficult. Typically, assessment of deleterious genomic variation in an individual is through quantification of downstream protein (cytokine) levels, either in peripheral blood mononuclear cells or constructed organoids from the tissue of interest. Stimulation with cytokines or bacterial stimuli allows assessment of the impact of a mutation on downstream immune function. In addition, analysis may focus on transfection of cell lines with the mutation of interest, although this is difficult to do for numerous or compound heterozygous variants. For protein-truncating mutations the effect may be visualised through immunohistochemistry techniques or knock-out murine models. Whilst these techniques are effective in Mendelian disease, in a polygenic disease, such as IBD, where multiple genetic variants may result in the same phenotype, establishing the cumulative effect of multiple heterogenous mutations across a cohort is more difficult. Typically, functional work-up focuses on a single patient (or a single variant) and on a specific downstream cytokine, preventing identification of additional gene-gene interactions and effects.

One strategy is to consolidate genetic variation across a protein complex or pathway, allowing grouping of patients with deleterious variation in a limited number of interacting genes. One approach is consolidation through a mathematical model based on a whole gene pathogenicity score, such as GenePy, summing the variation within the genes, complexes or pathways of interest[153]. Differences in downstream RNA expression between groups of patients, stratified by genetic variation across a molecular complex, could provide functional evidence to cluster patients based on molecular similarity. An established example of this can be seen in chronic granulomatous disease, an immunodeficiency resulting from inability to produce reactive oxygen species and impaired clearance of bacteria. There are six NADPH complexes, expressed in different tissues and formed of a variable number of key subunits[217]. Chronic granulomatous disease may arise from deleterious mutation in any one of the complex's subunits but is most

commonly results from highly deleterious homozygote, hemizygote or compound heterozygote variant(s) in either *CYBA, CYBB, NCF1, NCF2* or *NCF4.* The phenotype is distinctive and similar amongst patients with mutations across any of these genes and grouping deleterious variation by NADPH complex captures the downstream effect on reactive oxygen species production[257].

Where it is recognised that monogenic disease with highly similar phenotypes can result from deleterious variation in different genes, the occurrence of polygenic conditions, such as IBD, are also likely to have contribution from variation in a number of related genes within an individual. The NOD-signalling pathway is highly implicated in Crohn's disease pathogenesis, both through genomic analysis and transcriptomic analysis[255,258]. *NOD2* is the most implicated gene in Crohn's disease pathogenesis is a central player in the pathway, acting as an intracellular pattern recognition receptor for bacterial components, specifically muramyl dipeptide (MDP)[255], figure 22. Variation within other genes within this pathway, including *XIAP, ATG16L1, CARD9* and *RIPK2,* have all been implicated in both polygenic and monogenic forms of IBD[210]. Data from chapter 6 indicated a strong upregulation of NOD-signalling transcripts in treatment-naïve Crohn's disease and patients clustered distinctively from controls based only on the expression of 95 NOD-signalling genes.

It is therefore plausible that genomic variation across the NOD-signalling pathway, focusing on *NOD2* as a key signal transduction gene, will have direct impact on transcription of related genes within the pathway. Whilst this altered expression may be subtle, the resultant signalling changes may contribute to disease pathogenesis in a subgroup of patients. In this chapter, we assess whether variation in single genes, and across complexes, are associated with altered gene expression within the NOD-signalling pathway through assessment across a cohort, rather than in an individual. Utilising mathematical models of the variation in genes and protein complexes we integrate whole exome sequencing data through GenePy scores. We apply these data to targeted transcriptomic data from treatment naïve IBD patients to determine the impact of genetic variation on gene transcription through regression modelling.

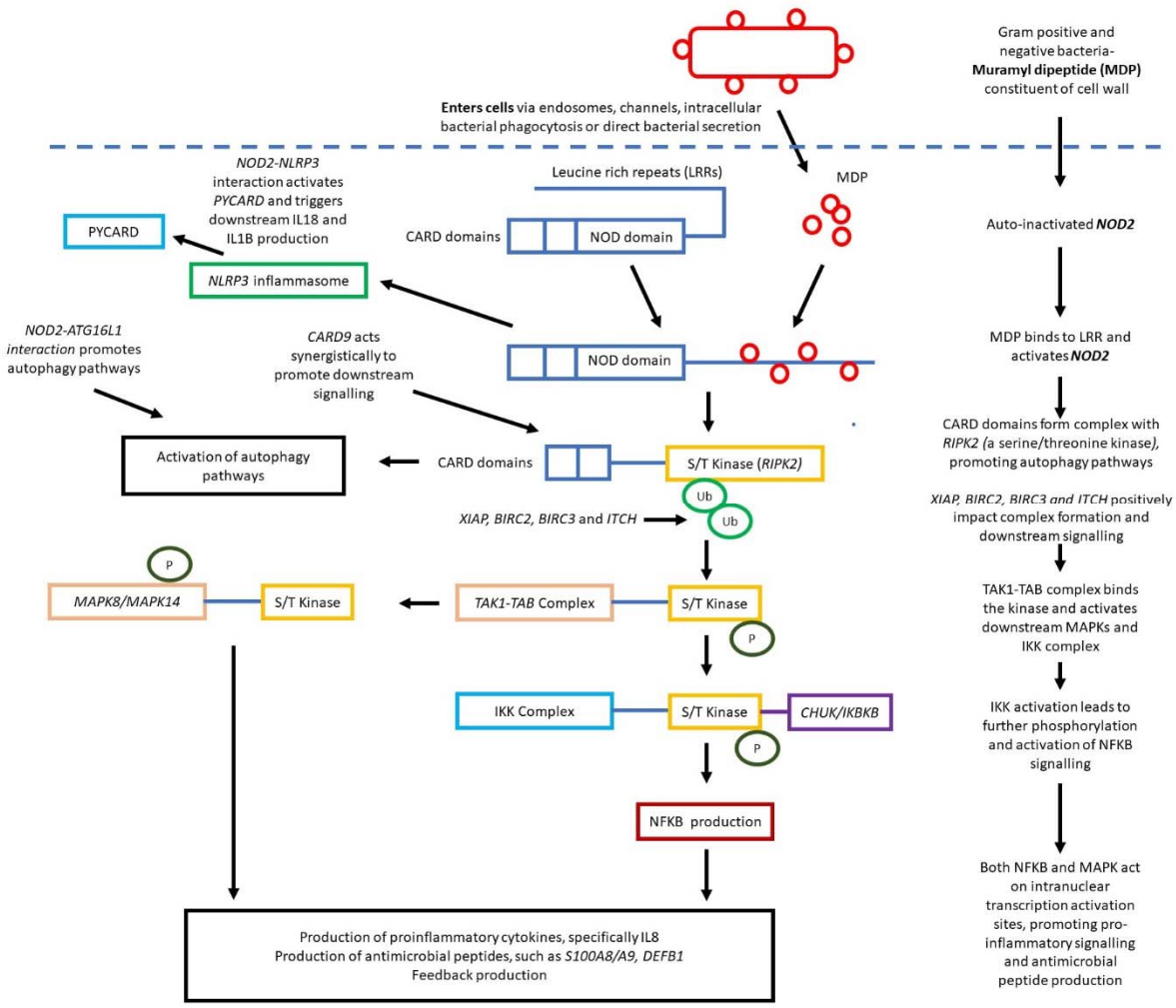

Figure 22- NOD2-signalling cascade and directly related inflammatory signalling pathways. Kinases, demonstrated in yellow, are S/T kinase domains formed as part of that protein complex.

7.2 Methods

Paediatric IBD patients were recruited through from the Paediatric Gastroenterology service at the Southampton Children's Hospital. In this chapter we included only treatment-naïve IBD patients for whome we had both ileal transcriptomic and whole exome sequencing data. We included patients with a subsequent diagnosis of Crohn's disease, ulcerative colitis or IBDU. Patient details are described in results.

7.2.1 Genetic analysis

Genomic analysis was performed as previously described in chapter 5.

7.2.2 Normalisation and application of LOEUF score to GenePy

GenePy scores were calculated as previously described in chapter 5. As some genes can accrue very high GenePy scores (due to length or mutability), when summing GenePy scores across a complex we first needed to normalise the values in order to make each gene comparable. We normalised all GenePy scores between 0 and 1, where 0 represented the lowest GenePy score for that gene across all patients in the analysis, and 1 represented the highest score for that gene across all patients in the analysis.

The LOEUF score is a metric developed as part of the Genome Aggregation Database (gnomAD) which assigns a score to each gene based on the gene's intolerance to inactivation[151]. This score can then be applied to determine which genes are able to accrue variation whilst maintaining activity, compared to those in which variation will be highly damaging. Higher LOEUF scores are associated with increased tolerance to variation. Genes within the NOD-pathway which are highly conserved and key to multiple inflammatory processes, such as *RIPK2* or *TAK1*, have very low LOEUF scores, whereas genes in which variation is more commonly seen, such as *NOD2* have higher scores.

As these analyses involved assessment of many genes across a complex, we chose to integrate the LOEUF score into the GenePy score in order to upweight the importance of variation in genes predicted to be intolerant to inactivation. LOEUF scores for genes in this analysis can be seen in supplementary data. The normalised GenePy score was divided by the respective LOEUF score for each gene.

7.2.3 Ileal transcriptomic analysis

Ileal transcriptomic analysis, including quality control, was performed as described in chapter 6. Gene counts were normalised by quantile normalisation.

7.2.4 Key genes and protein complexes within the NOD-signalling pathway

We hypothesised that summing variation across key protein complexes within the NOD-signalling pathway would enable us to discern the impact of underlying genomic variation on gene expression. We utilised a pre-collated list of 95 genes in the NOD-signalling pathway, produced by HTG as part of their autoimmune panel product[259].

The *NOD*-signalling pathway genes were cross-referenced with genes known to be implicated in IBD by either GWAS, or as a monogenic IBD gene, chapter 2. Invariant genes or complexes were not assessed. Single genes within the pathway and implicated in IBD were included in the analysis. Molecular complexes in the NOD-signalling pathway, with 2 or more constituent proteins, containing genes with variation, were included in the analysis. Molecular complexes activated as a result of multiple inflammatory pathways, including the IKK complex and MAPK complexes, were excluded due to the lack of specificity to *NOD*-signalling.

7.2.4.1 Summing GenePy scores for protein complexes

We summed GenePy scores across the genes in key protein complexes. For each gene within a molecular complex the LOEUF corrected, normalised, GenePy scores for each gene were summed to create a GenePy score for the molecular complex (supplementary data). This provides a quantitative score reflecting the cumulative sum of variation within that complex.

7.2.5 Genomic and transcriptomic integration

We utilised a stepwise linear regression model to determine the impact of genomic variation on transcription. The analyses were conducted as follows:

1. To determine whether genomic variation within the NOD-signalling pathway impacted on gene expression within the same pathway, we utilised the GenePy scores for an individual gene, or summed scores for a complex, as the dependant variable. The quantile normalised expression values for the 95 NOD-signalling genes were entered as independent variables.

   GenePy score for gene/complex = intercept + slope(quantile normalised transcript 1, QN transcript 2…… QN transcript 95)

2. To better understand the drivers of differential gene expression (within the NOD-signalling pathway) observed in chapter 6, we applied the converse approach. To determine whether variation in any of the 95 NOD-signalling genes was associated with transcription levels we entered the quantile normalised transcription level for *CXCL8* (IL8) as the dependant variable. The independent variables were the GenePy scores for the 95 NOD-signalling genes.

   Quantile normalised transcript number for CXCL8= intercept + slope(GenePy score for gene 1, GenePy score for gene 2…… GenePy score for gene 95)

7.2.5.1 Weighted gene co-expression network analysis

We hypothesised that patients with deleterious variation in *NOD2* would correlate with specific gene expression modules across all autoimmune transcripts (2002), thus identifying transcriptomic signatures and inflammatory pathways associated with *NOD2* variation. Co-expression modules were identified using WGCNA (R package). Gene co-expression modules were correlated with the *NOD2* GenePy scores for treatment naïve patients to determine whether treatment-naïve patients with high burden of *NOD2* variation led to specific gene expression patterns[155].

7.2.5.2 Statistical analysis

Statistical analysis was performed using SPSS (v25, IBM) and WGCNA (R package).

7.3 Results

Treatment naïve patients with WES and transcriptomic data were included. There were a total of 39 patients, 27 had a diagnosis of Crohn's disease, 9 had a diagnosis of ulcerative colitis and 3 had a diagnosis of IBDU. The mean age at diagnosis was 13.2 years (range 2.9-16.8 years)

7.3.1 Genes and complexes included in the analysis

The core NOD-signalling pathway was interrogated. Three Genes and three protein complexes were selected for analysis based on the criteria described above, table 19. The IRAK-TRAF6 complex acts synergistically with the core NOD2-signalling pathway and was included, in addition to the core NOD2-RIPK2 and TAK1-TAB complexes. The IKK complex, consisting of IKKa, IKKb and NEMO, was excluded as it is the target of multiple activators including TNFa and IL1 signalling, and is not specific for NOD-dependant bacterial recognition and response[260]. Similarly, MAPK complexes were excluded due to the large number of MAPK genes that may be included and the lack of specificity to NOD-signalling, figure 22.

Table 19- Genes and complexes to be entered as dependant variables in regression analysis, and the constituent proteins (genes). All gene's GenePy scores are scaled to between 0-1 and corrected by LOEUF score prior to being summed to form the 'complex's GenePy score'

Gene or complex	**Proteins comprising complex**
NOD2	*NOD2*
ATG16L1	*ATG16L1*
CARD9	*CARD9*
NOD2-RIPK2 complex	*RIPK2, NOD2, XIAP, BIRC2, BIRC3, ITCH*
TAK1-TAB complex	*TAK1, TAB2, TAB3*
IRAK-TRAF6 complex	*IRAK1, IRAK2, IRAK4, TRAF6, MYD88*

7.3.2 Patients harbouring deleterious *NOD2* gene variation have reduced *NOD2* gene expression and increased expression of *NFKB* inhibitor-α

We expected *NOD2* gene variation to impact on downstream gene expression within the signalling pathway. We examined the effect of variation in *NOD2* through stepwise linear regression, with *NOD2* GenePy score as the dependant variable and all 95 gene transcript levels as the independent variables. Increased *NOD2* GenePy scores, reflecting increased deleterious variation, was associated with a decrease in *NOD2* transcripts (figure 23) and increased expression of *NFKBIA*, encoding a key inhibitory protein preventing NFKB signalling (table 20). The effect appears to reflect a decrease in NFKB signalling as a result of deleterious *NOD2* variation. *CCL5*, a T-cell chemokine, was also downregulated in patients with high variant deleteriousness in the *NOD2* gene.

Table 20- Impact of NOD2 variation on NOD-signalling gene expression. Dependant variable is NOD2 GenePy score

Independent variable- Gene expression	**Beta coefficient**	**P value**
NOD2	-0.702	0.000043
NFKBIA	0.486	0.001
CCL5	-0.414	0.008

Figure 23- Relationship between quantile normalised NOD2 transcript levels and NOD2 GenePy score. Four patients harbouring the 1007fs variant are seen in red.

7.3.2.1 1007fs *NOD2* variant appears to impact on transcription but is not sole driver of reduced expression

NOD2 harbours a protein truncating variant, 1007fs, commonly seen in Crohn's disease patients. We assessed whether this specific nonsense variant within *NOD2* was driving the inverse relationship between GenePy score and transcript number. Four of the 39 patients were heterozygous for the 1007fs *NOD2* variant. A T-test demonstrated there was no significant difference in the *NOD2* expression level between those with the 1007fs variant (mean QN transcripts 42.9) and those without (mean QN transcripts 59.5), p= 0.12.

No other protein truncating variants within *NOD2* were identified in the 39 patients.

7.3.3 Deleterious variation in *ATG16L1* increases expression of *IKBKB*

ATG16L1 encodes for a protein key in autophagy pathways. Activated *NOD2* works synergistically with *ATG16L1* to promote autophagy. Variation within *ATG16L1* is an established risk for Crohn's

disease development. Regression analysis, utilising *ATG16L1* GenePy score as the dependant variable demonstrated increased *ATG16L1* GenePy score was associated with an increase in *IKBKB*, an activator of NFKB-signalling. Table 21. Autophagy represents a different pathway to core *NOD2*-signalling and variation in *ATG16L1* may be pro-inflammatory via NKFB signalling, whilst reducing autophagy.

Table 21- Impact of ATG16L1 variation on NOD-signalling gene expression. Dependant variable is the ATG16L1 GenePy score

Independent variable- Gene expression	**Beta coefficient**	**P value**
IKBKB	0.504	0.001
IRAK4	-0.498	0.001
CYBA	-0.288	0.041

7.3.4 Deleterious variation in *CARD9* decreases expression of *IKK-α (CHUK)*

CARD9 interacts directly with *NOD2*, resulting in activation of downstream pro-inflammatory signalling through MAPK activation. It also functions independently as a signal transduction complex alongside *BCL10* and *MALT1,* largely in response to fungal infection, which then activates the IKK complex triggering NFKB activation. Regression analysis, with CARD9 GenePy score as the dependant variable demonstrated increased deleterious variation in *CARD9* was associated with a decrease in *IKK-α* (*CHUK*), an upstream activator of NFKB signalling, figure 24. Table 22.

Table 22- Impact of CARD9 variation on NOD-signalling gene expression. Dependant variable is the CARD9 GenePy score

Independent variable- Gene expression	**Beta coefficient**	**P value**
IKK-α	-0.397	0.007
IFNB1	0.345	0.018

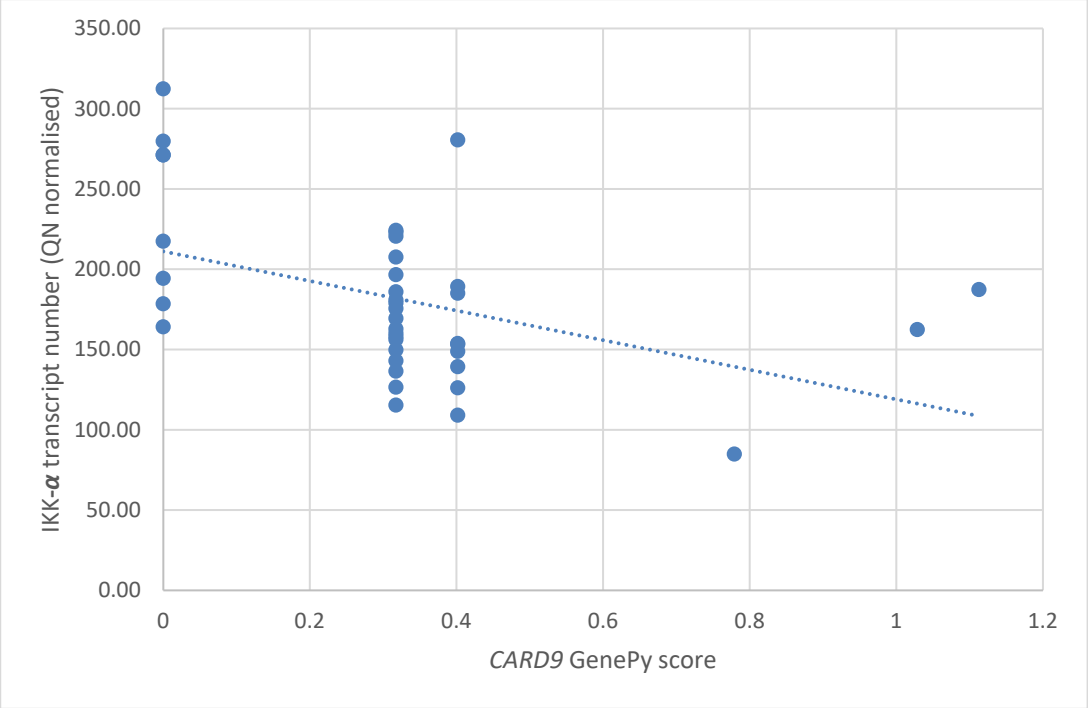

Figure 24- Relationship between quantile normalised IKK-α transcript levels and CARD9 GenePy score

7.3.5 Deleterious variation in the *NOD2-RIPK2* complex is associated with increased expression of *BIRC2, TXN and NLRP3*

Following activation by MDP, *NOD2* forms a complex with *RIPK2*. *XIAP, BIRC2, BIRC3* and *ITCH* all positively associate with the complex promoting downstream *RIPK2* kinase activity. Regression analysis, using the summed *NOD2-RIPK2* complex as the dependant variable (and all 95 NOD-signalling gene transcripts as the independent variables), revealed increased deleterious variation within the complex was related to an increased expression of the *NLRP3* and *TXN*. Table 23. Both of these genes are involved in the pro-inflammatory *NLRP3* inflammasome, however we also observe a decrease in PYCARD expression, encoding a key protein in the inflammasome activation pathway. *BIRC2* was highly significantly upregulated in patients harbouring deleterious variation in this complex, which includes *BIRC2*.

Table 23-Impact of NOD2-RIPK2 complex variation on NOD-signalling gene expression. Dependant variable is the scaled and LOEUF-corrected NOD2-RIPK2 complex (RIPK2, NOD2, XIAP, BIRC2, BIRC3, ITCH) GenePy score

Independent variable- Gene expression	Beta coefficient	P value
BIRC2	.800	3.1475E-8
TXN	.417	0.000084
NLRP3	.245	0.014
PYCARD	-.278	0.014
IRAK4	-.345	0.001
UBA52	-.224	0.033

7.3.6 Genomic variation within the TAK1-TAB complex leads to reduced *MAPK14* expression

The TAK1-TAB complex, including *TAK1*, *TAB2* and *TAB3*, is a key signal transducer, both from the NOD2-RIPK2 complex and toll-like receptors (TLRs). Summed GenePy scores for the TAK1-TAB complex were used as the dependant variable. Expression of *MAPK14* and *BIRC3* was reduced in the presence of deleterious variation within the TAK1-TAB complex. There was also increased expression of *IFNA1*, figure 25 and table 24. *MAPK14* autophosphorylates in the presence of the TAK1-TAB complex, leading to downstream activation of pro-inflammatory and anti-microbial gene expression.

Table 24- Impact of TAK1-TAB complex variation on NOD-signalling gene expression. Dependant variable is the scaled and LOEUF-corrected TAK1-TAB complex (TAK1, TAB2, TAB3) GenePy score

Independent variable- Gene expression	Beta coefficient	P value
MAPK14	-0.677	0.000017
IFNA1	0.479	0.001
BIRC3	-0.375	0.008

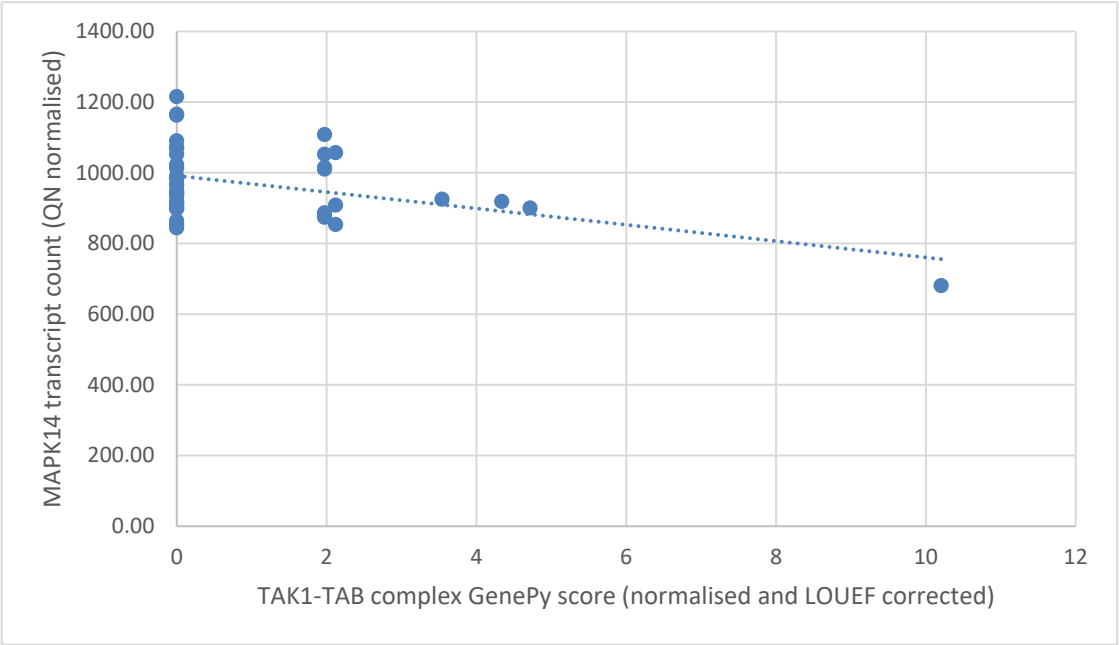

Figure 25- Relationship between quantile normalised MAPK14 transcript levels and TAK1-TAB complex GenePy score

7.3.7 Variation in the IRAK-TRAF6 complex, within the toll-like receptor (TLR) signalling pathway, results in decreased expression of the NFKB activating protein *IKBKG (NEMO)*

The IRAK-TRAF6 complex (*IRAK1, IRAK2, IRAK4, TRAF6, MYD88)* plays a key role in signal transduction from TLRs through interaction with *MYD88*. It acts synergistically with *NOD2* activation and one result of IRAK-TRAF6 activation is downstream activation of the TAK1-TAB and subsequently activation of NFKB signalling. Variation in the IRAK-TRAF6 complex was associated with a decrease in a single gene transcript, *IKBKG*, a component of the NFKB activating complex, figure 26.

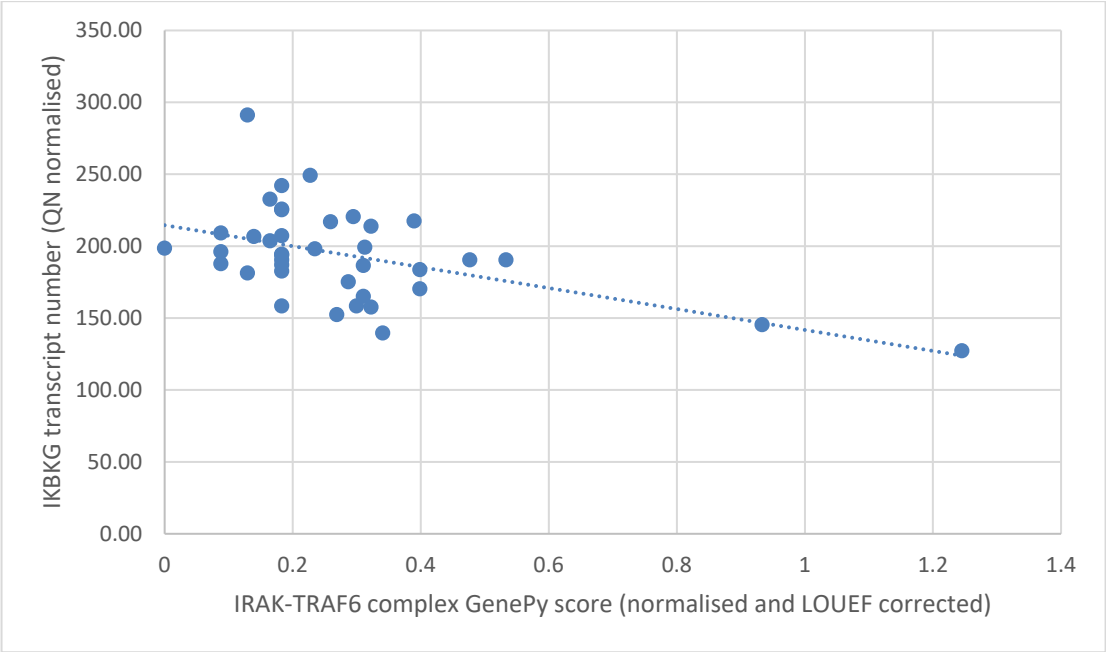

Figure 26- Relationship between quantile normalised IKBKG transcript levels and IRAK-TRAF6 complex GenePy score

7.3.8 Impact of genomic variation in the NOD-signalling pathway on previously identified differentially expressed genes

Utilising differentially expressed genes (between controls and treatment-naïve Crohn's disease) directly downstream of NOD-signalling, identified in chapter 6, we determined if variation within NOD-signalling genes were associated with increased expression of the upregulated probes. IL8 (CXCL8) was one of the top upregulated probes between treatment-naïve Crohn's disease patients and controls and is a downstream transcription product triggered by NFKB signalling. Using IL8 expression levels as the dependant variable, and GenePy scores for all 95 NOD-signalling genes as independent variables we performed a stepwise linear regression model. There was a positive association between GenePy scores in *NLRP3* and *IL18*, and IL8 expression levels. Whereas mutation in *PRKCD* was negatively associated with IL8 expression levels, table 25.

Table 25- NOD-signalling genes significantly impacting on IL8 expression levels. Independent variables are all 95 NOD-signalling gene's GenePy scores. Dependant variable is the IL8 (CXCL8) quantile normalised expression levels

Independent variable- GenePy score for gene	Beta coefficient	P value
NLRP3	0.378	0.011
PRKCD	-0.295	0.045
IL18	0.292	0.046

7.3.9 *NOD2* GenePy score is not associated with specific gene expression modules across all autoimmune genes

Given the established role of *NOD2* in Crohn's disease pathogenesis we hypothesised that subgroup(s) of patients with accumulation of pathogenic *NOD2* variation would be characterised by similar gene expression. To test this hypothesis, we determined gene co-expression modules, identified through WGCNA of all 39 patients, and correlated these modules with *NOD2* GenePy score. Patients were clustered by the similarity of gene expression for all 2002 autoimmune gene transcripts, figure 27. NOD2 GenePy scores were entered as a continuous variable. Hierarchical clustering did not demonstrate patients with similar *NOD2* GenePy scores in the same clusters.

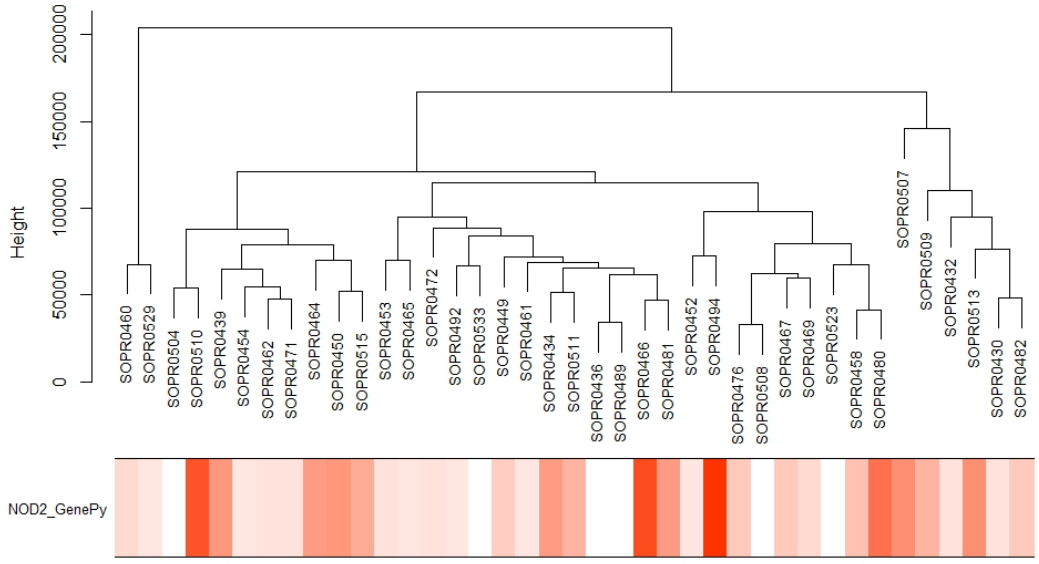

Figure 27- Clustering of patients by WGCNA utilising all 2002 autoimmune gene transcripts. Annotation of patients with NOD2 GenePy score did not reveal clusters of patients with similar gene expression also harbouring similar NOD2 deleteriousness.

7.3.9.1 Gene co-expression modules in treatment naïve patients

Expression modules were identified within the AI transcripts, based on the 39 treatment naïve patients. Two large co-expression signatures emerged, the turquoise and blue modules, alongside several smaller co-expression modules. Figure 28.

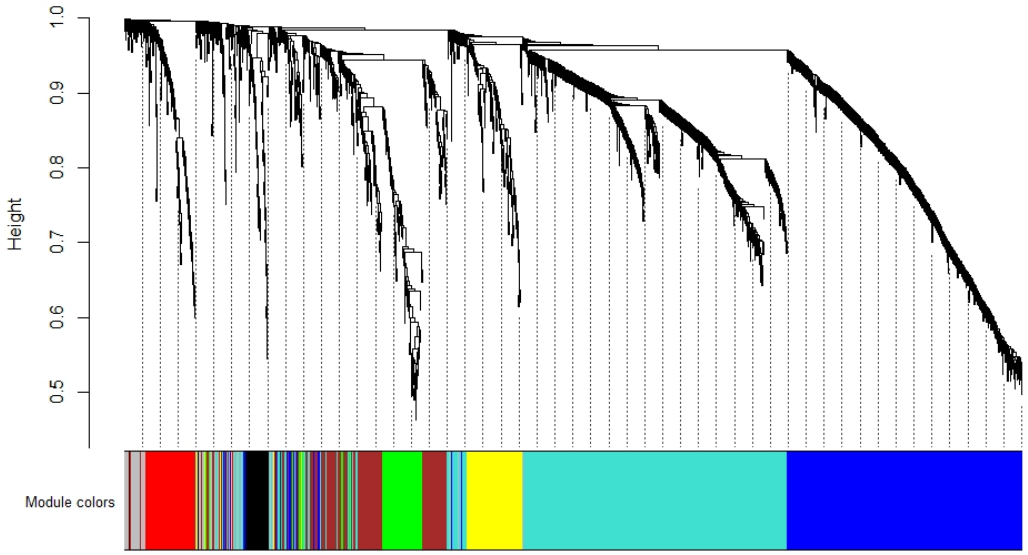

Figure 28- Gene coexpression modules determined using WGCNA performed on 39 treatment naïve IBD patients. Turquoise and blue modules represent large clusters of similarly expressed patterns of genes across the cohort.

7.3.9.2 *NOD2* GenePy scores are not correlated with gene co-expression modules

In order to determine whether specific sets of co-expressed genes were associated genomic variation in *NOD2* we analysed whether the co-expression modules were correlated with *NOD2* GenePy score. None of the 8 modules were significantly correlated with *NOD2*, figure 29.

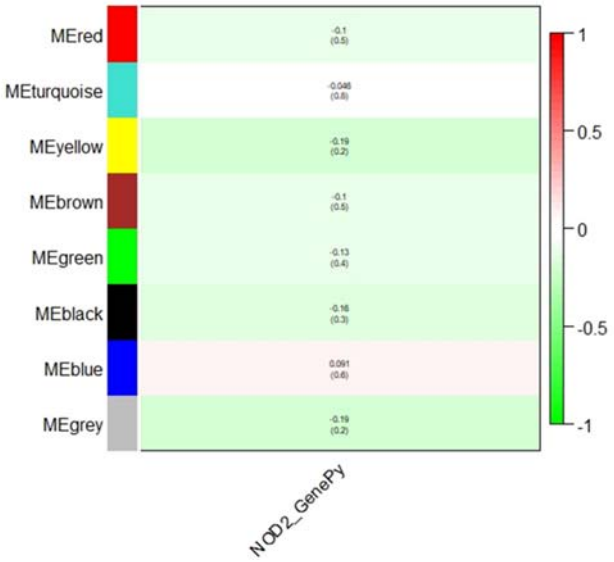

Figure 29- Correlation coefficient values (p values) between gene expression modules and NOD2 GenePy scores

7.4 Discussion

The impact of genomic variation across key genes, and complexes, within the NOD-signalling pathway appears to lead to a hypoinflammatory response, with reduced activation (or increased inhibition) of NFKB-signalling or reduced upstream activation of pro-inflammatory signalling. Variation in *NOD2*, and directly related complexes, appears to act synergistically to reduce *NOD2* transcription, whilst simultaneously increasing transcription of alternative inflammatory pathways including the NLRP3 inflammasome and interferon signalling. We identify variation across the *TAK1-TAB* complex directly resulting in reduced *MAPK14* transcription. Variation in *NOD2*-synergistic activators of autophagy (*ATG16L1*), or NOD-signalling (*CARD9*), impact on downstream transcription, either increasing or decreasing NFKB signalling, respectively.

Previous data has pointed to a hypoimmune response in both Crohn's disease patients and murine models harbouring deleterious *NOD2* variants[255,261]. *NOD2* variants are thought to be loss-of-function, leading to impaired *NOD2* activation[262]. Studies detailing the direct effect of deleterious *NOD2* variants repeatedly identify a decrease in pro-inflammatory cytokine response after MDP stimulation in peripheral blood mononuclear cells, specifically through reduced NFKB production[255]. Overall the mechanism whereby *NOD2* variants increase susceptibility to Crohn's disease appears to be through impaired bacterial recognition/response leading to reduced bacterial clearance and increased chronic inflammation through non-*NOD2* proinflammatory pathways[263]. In addition to *NOD2*, several additional risk susceptibility genes, or monogenic IBD genes, lie within the NOD-signalling pathway including *XIAP, CARD9* and *TAB2*[91]. Loss-of-function variants within these genes are associated with severe monogenic forms of Crohn's-like IBD (*XIAP, CARD9*), or increased risk of 'classical' Crohn's disease (*TAB2*). Several studies have described the function of these genes in downstream NFKB signalling, including variants in XIAP[264–266] and CARD9[267] leading to reduced NFKB production. Whilst the mechanism by which these genes lead to disease may not have the evidence base seen with *NOD2*, it appears that a hypo-inflammatory response is implicated, potentially alongside

activation of additional aberrant pathways (XIAP- loss of anti-apoptotic function and cell death, CARD9- inability to recruit *BCL10* and *MALT1*).

Within this study we hypothesised that variation across the NOD-signalling pathway, with a focus on *NOD2*-signalling, would result in transcription level defects associated with a hypoinflammatory response. Previous data has inferred that disruption at any step on the NOD2-signalling cascade will result in decreased downstream NFKB or MAPK activation, although direct impact of genetic variation at each step has not been studied[268,269]. Through single gene, and whole complex, deleteriousness scoring we identify a consistent pattern of defects associated with decreased transcription of downstream *NFKB* or *MAPK* genes. Importantly, by summing deleteriousness across a complex we were able to observe cohort-level effects that may be missed if assessing a single variant or a single gene within an individual patient. As individual variant effects on gene transcription are likely to be very mild, or private to an individual, the ability to sum the effects of interacting genes allows a statistical association to emerge across a cohort. It is clear that for most patients with IBD, the effect of multiple genomic variants leads to disease, rather than a strong effect from a single gene (monogenic IBD)[210,270].

At an individual gene level, we reveal a striking decrease in *NOD2* transcripts associated with increased *NOD2* deleteriousness. This is not only driven by patients harbouring the nonsense 1007fs variant, with only four of the 39 patients being heterozygote, and no patients being homozygote for this protein truncating variant. It is not possible to determine whether several of the more common variants, harboured by patients with low *NOD2* transcript levels are in linkage disequilibrium with non-coding variants in the promotor region of *NOD2*. Both *NOD2* and *CARD9* variation were significantly associated with downstream gene transcription leading to decreased NFKB signalling. *ATG16L1* synergistically acts with *NOD2* to promote antibacterial autophagy. Additionally, *ATG16L1* has a role in negative regulation of proinflammatory MAPK and NFKB activation cascades, triggered by *NOD2* signalling[271]. Our results demonstrate the direct impact of this additional *ATG16L1* role, with *ATG16L1* variation leading to an increase in the NFKB

activation transcript *IKBKB.* Where deleterious variation in the *NOD2* canonical pathway appears to lead to reduced NFKB/MAPK activation of downstream inflammatory signalling, variation in *ATG16L1* may lead to increased inflammation through impaired autophagy, or directly through the inability to negatively regulate *NOD2* activity.

We identify summed variation in the TAK1-TAB complex directly associated with reduced *MAPK14* transcription. Alongside this we observe an increase in gene transcription associated with alternative inflammatory pathways (*TXN* and *NLRP3*) seen with *NOD2-RIPK2* complex variation. Previous data has indicated a key role for *TAK1* in MDP-stimulated NOD2-signalling, with absence of TAK1 completely admonishing downstream NFKB and MAPK signalling[272]. Variation in the toll-like receptor transduction complex, *IRAK-TRAF6*, was associated with a decrease in IKBKG transcription, a potent activator of NFKB signalling. This membrane receptor-triggered pathway, acts in parallel to intracellular *NOD2* signalling, also leads to NFKB activation in response to bacterial recognition and response. These data imply that variation across this related complex also impairs inflammatory and antimicrobial response.

IL8 is a highly upregulated transcript in Crohn's disease (chapter 6), produced in response to proinflammatory pathway activation, including NOD-signalling. We determine that deleterious variation within alternative NLRP3 inflammasome pathway (specifically NLRP3 and IL18), typically resulting in IL18 and IL1B production, is associated with an increase in IL8 transcription. We hypothesise that poor NLRP3 or IL18 function activates alternative proinflammatory signalling, including NOD-signalling, resulting in increased IL8.

We hypothesise that for the majority of patients, IBD appears to arise due to multiple 'hits' across complexes/genes contributing to impairment of inflammatory pathways. These patients then fail to clear bacteria allowing invasion and chronic inflammation to develop. The precise immune impairment within an individual are likely to lead to commonality between subgroups of patients, with a large number of Crohn's disease patients having disease attributable to impaired *NOD2*-signalling[256]. Other implicated pathways in which loss-of-function variants impair antibacterial

activity include the NADPH oxidase pathways, IL10-signalling and IL23-IL17 signalling[91]. The ability to classify patients into the underlying immune impairment would be a large step forward in the ability to personalise therapy and provide more targeted medicine. The use of summed whole pathway genomic pathogenicity scoring may yield clinically translatable results for patients.

This study has several strengths. To improve accuracy of results we employ the LOUEF score as an addition to GenePy. This allowed our analyses to account for genes intolerant to inactivation, where deleterious variation is more likely to be disease causing[151]. Additionally, we provide a degree of validation of genomic findings without the need for time-consuming functional assays, although caution should be exercised in interpretation of these results. Through use of targeted sequencing we enable identification of lowly expressed transcripts, which are key in many of these analyses. Refinement of GenePy to include LOUEF scores was demonstrated to be a useful addition in these analyses, supplementary data. Weaknesses of this approach include the inability of WES to capture regulatory variation and the dependence of GenePy on *in silico* deleteriousness metrics.

There are several strategies that can be used to analyse these types of data in the future. Whilst we have employed a linear model it is likely several of the relationships between gene variation and transcription are non-linear associations, by use of supervised machine learning modelling, including random forest classifiers, we may be able to determine additional transcripts impacted by variation in key complexes. Providing a functional layer of evidence to these transcriptomic-genomic data would enable application of the techniques within this manuscript to be generalised at a wider level. We also have the opportunity to expand to additional pathways of interest, including NADPH oxidase pathways, IL10-signalling and JAK-STAT signalling. Finally, further integration of data with microbiome sequencing may enable the precise impact of NOD-signalling impairment to be observed at a microbial community level, although unveiling causality will remain difficult.

There is now a clear trajectory to move translation of these analyses to clinical practice. Understanding and detailing the precise immune perturbation in a patient, including the upregulated and downregulated inflammatory signalling may allow targeted therapy for individuals, focused on a specific immune pathway. Novel therapies, targeting different inflammatory cytokines (in addition to anti-TNF, Il12/23 etc.) may prove highly efficacious when used in the correct patient scenarios.

7.4.1 Conclusion

These data demonstrate a pathway-wide effect of genomic variation in NOD-signalling genes, resulting in reduced proinflammatory gene transcription within this pathway. Integration of genomic and transcriptomic data allows for statistical association of genomic variation with downstream transcription. We observe variation at each stage of the NOD2-signalling pathway resulting in broadly reduced NFKB signalling, with frequent upregulation of other inflammatory genes including *NLRP3* and interferons. Expanding these analyses to additional pathways implicated in IBD may allow for precise 'immuno-typing' of patients, identifying defects in specific immune pathways and pacing the way for personalised therapy.

Chapter 8 Summary and future research

Chapter summary- *This chapter summarises the key findings and provides an outline of future work to be conducted on the back of findings, data and samples achieved in this PhD.*

8.1 Summary of findings

The overarching theme of this thesis is the integration of complex 'multi-omic' data, with longitudinal clinical follow-up information, to move towards providing a personalised approach to treatment and management of children with IBD. I have analysed and integrated genomic and transcriptomic data using bioinformatic tools and directly linked these to clinical data and outcomes. These data provide the basis for developing clinically translatable tools, utilising clinical and multi-omic data, to improve prognostication and personalisation in children with IBD.

During this PhD nearly 100 IBD patients were recruited and had intestinal biopsies, 50% of patients were treatment-naïve and recruited prior to diagnosis. In addition, an estimated 20 controls were recruited. Longitudinal clinical data we collected and collated for all patients, including automatic and manual data extraction. Establishing this cohort, nested within the 500 patients in the Southampton genetics of paediatric IBD study, provided samples and data for which all analyses are based. In conjunction with this several literature reviews were conducted, providing data on clinical and scientific background to IBD and facilitating analyses within the PhD[98,273].

The initial project in chapter 3 demonstrated novel clustering of patients based on presenting blood results, and the ability of clinical data to provide stratification of patients. These data identified the utility of longitudinal clinical data, available as part of routine clinical care. We identified an estimated 10% of IBD patients with normal blood tests at diagnosis, and distinct patterns of results characterising patient subgroups. These data provide a clear clinical basis demonstrating that there are hugely heterogenous presentations within paediatric IBD and there are likely to be multiple clinical patient subgroups.

In chapter 4 this thesis demonstrates the utility of whole exome sequencing to identify and diagnose patients with monogenic IBD variants. For a subset of these patients we provide a clinically translatable molecular diagnosis. We also identify clinical traits and phenotypes statistically associated with deleterious variation within monogenic IBD genes. The most important relationship identified was the substantial increased risk (odds ratio >11) of stricturing disease development in patients with monogenic *NOD2* disease. This chapter provides the initial translation of molecular findings to distinct clinical phenotypes.

Within chapter 5 we hypothesised that variation across multiple genes within the NADPH complex and *NOD2* would translate to distinct clinical phenotypes. We utilised GenePy and determined that multiplicative deleterious variation in *NCF4* and *NOD2,* alongside variation in the *NOX4* NADPH complex was associated with increased risk of fistulating disease for a small subset of patients. Patients with increased deleteriousness across all NADPH oxidase genes were significantly more likely to have fistulating disease, providing further molecular characterisation of patients

Chapter 6 introduces contemporary targeted RNA sequencing and single cell Drop-Seq transcriptomics, derived from ileal intestinal biopsies. Utilising these data, we identified perturbation of IL17- and NOD-signalling pathways, specific to treatment naïve Crohn's disease patients. Time to relapse was associated with a specific Th17 differentiation expression signature.

Furthermore, through collaborative analysis we detail a specific epithelial cell subtype driving antimicrobial calprotectin expression.

Finally, through integration of RNA and whole exome sequencing in chapter 7, we have shown genetic variation in specific complexes within the NOD-signalling pathway have a direct effect on gene expression. Grouping of patients by variation across a complex demonstrates the direct, and indirect, effect of variation within the pathway on key downstream gene transcripts. Importantly deleterious variation in *NOD2* was associated directly with decreased *NOD2* transcription. Overall, variation with the pathway resulted in an apparent hypoimmune response. These data provide the framework for identifying and classifying IBD patients by the pathway(s) which are perturbed, providing a molecular diagnosis which may translate to treatment strategies. We aim to build on this work through functional assessment of specific patients with high deleterious burden and altered RNA transcription within the NOD-signalling pathway.

8.2 COVID-19 impact

In March 2020 the overall 'Genetics of PIBD' study was paused, by non-substantial amendment, in line with local, and national, recommendations. The study was reopened in August 2020 and recruitment was recommenced in September 2020. The direct effect of COVID-19 on my PhD are set out below.

8.2.1 Substantial study amendment

We submitted a substantial amendment in November 2019, with the specific aim of contacting patients to feedback monogenic findings, whilst this amendment was approved in February 2020, due to COVID-19 this was not implemented locally. We have not been able to contact patients to date, unless there was a clear clinical priority where withholding this information could cause patient harm. We expect this to be implemented as soon as the study recommences.

8.2.2 Variant confirmation and segregation analysis

We were able to identify patients with possible monogenic variants from WES data from the March 2020 exome batch. However, due to laboratory closures, we were unable to perform Sanger confirmation of any variants identified and were unable to perform segregation analysis for any patients with potential compound heterozygote variants. We have therefore presented data from the initial analysis of 401 patients, with supplementary information to include the latest batch (501 patients).

8.2.3 Microbiome sequencing and analysis

Performing microbiome sequencing on ileal samples from recruited patients, and layering onto RNA and genomic data, is the primary objective of future research. Whilst this element of the project was intended to form part of this thesis, due to the COVID-19 pandemic we have been unable to complete the sequencing within the time scale. We have secured ongoing funding for sequencing and staff costs in order to deliver this within the next 12 months.

8.2.4 Supervisor meetings

All supervisor meetings, alongside all other university work, were moved to Microsoft Teams at two days' notice. This lasted for 6 months from March 2020 until the end of my PhD in September 2020.

8.2.5 Additional clinical work

Due to COVID-19 I was asked to work additional clinical shifts from March-August 2020. In total I worked an additional 38 clinical shifts during this 5-month period.

8.2.6 International collaboration- Oligogenic IBD.

I have an active and ongoing collaboration with Sick Kids in Toronto. I have previously visited the laboratory of Professor Aleixo Muise, and had several visits planned in 2020 to provide joint analysis of Southampton and Toronto data on NADPH/*NOD2* forms of oligogenic IBD. These visits were all cancelled. Analysis was restricted to Southampton data only. We have an ongoing plan to recommence this work after the effects of COVID-19 have been mitigated.

8.2.7 International conference presentations

I had two conference abstracts accepted for presentation at Digestive Diseases Week (May 2020), and an oral presentation accepted at the World Congress of Paediatric Gastroenterology, Hepatology and Nutrition. Both conferences were cancelled due to COVID-19.

8.3 Future work

8.3.1 Rectal biopsy processing and sequencing

Largely due to COVID-19 restrictions we were unable to process, and analyse, rectal biopsies achieved from patients recruited during this PhD. As detailed in Chapter 2 there are 110 individuals for which two rectal biopsies are available. Additionally, we have repeat biopsies on 10 of these patients. This tissue provides a significant resource that can be used in local and collaborative projects. We expect that both RNA and microbial sequencing will be possible from these samples, and funding to achieve this will be sought through future collaborative grant applications. As technology improves we also have the opportunity to utilise metagenomic sequencing of these rectal biopsies, to gain further insight into the microbial communities at a mucosal level[64].

8.3.2 Microbiome

We will perform 16S sequencing of bacterial DNA extracted from ileal biopsies of IBD patients and controls. These data will be utilised to detail community differences between patient groups, map to clinical outcomes and integrate with the genomic and RNA sequencing data presented in this thesis. Specifically, we will assess the hypothesis that perturbation of specific pathways identified through RNA analysis (NOD-signalling, IL17-signalling) will result in differences in bacterial ecology. Furthermore, we will analyse the impact of genomic variation in the *NOD2* bacterial recognition and response pathway, and the NADPH oxidase complexes, through identification of patients with significant variants in *NOD2*, related genes (such as *XIAP*) and within the NADPH oxidase complex genes.

We will test the hypothesis that an individual's genetic risk for developing IBD will result in distinct microbial community profiles. Those clustering with patients known to harbour *NOD2* or NADPH variants, but without exonic evidence of this may harbour intronic or promotor variation leading to a similar phenotype, warranting further genomic analyses. Monogenic forms of IBD may also leads to specific microbial groups.

8.3.3 Disease prediction modelling

Collection of long-term follow-up data on the cohort of treatment-naive patients will be key to providing 12-, 24- and 72-month outcomes. Longer follow-up duration of this cohort will also allow patients to develop complicated disease phenotypes (stricturing/penetrating disease) and need for surgery. These additional outcomes will provide new opportunities for re-analyses of these data and facilitate the development of predictive models for complicated disease phenotypes, response to therapy and growth outcomes. Utilising supervised machine learning models will be key this. We hope integration of these long-term clinical data will allow replication of previous findings (from studies based on the RISK, and similar, cohort), and yield new predictive factors allowing personalisation of therapy and prognosis[120].

There is clear potential for predictive disease modelling for complicated Crohn's disease behaviour, based on genomic data only. The data directly relevant to this thesis can be seen in Chapter 4 and Chapter 5. Further analyses, not presented in this thesis, focused on utilising GenePy scores and random forest classifiers to identify genes in which variation is risk, or protective, for development of stricturing, or penetrating, complications. Further integration of these genes (and GenePy scores) into a Cox proportional hazard model allowed accurate stratification of patients into risk groups for development of complicated disease. These models will form a key part of future grant applications and work to predict disease outcomes and provide clinical translation of basic science data.

8.4 Reflections on personalised medicine

Data presented within the thesis demonstrate the feasibility and practicality of using multi-omic data to stratify and group IBD patients by their underlying molecular diagnosis, and to integrate clinical data into these analyses to move towards clinical application. Within this final chapter I have set out the planned future work to continue this translation. We have a clear plan for building on the findings of this PhD, with further integration of clinical data and utilisation of collected samples.

The ability to predict disease course at diagnosis, alongside tailoring medications based on response gives the potential for a more 'personalised approach'. The move to a pre-emptive strategy to prevent IBD-related complications, whilst simultaneously minimising side effects and long-term toxicity from therapy, particularly in those with relatively indolent disease, has the potential to revolutionise care. This includes patients who are less likely to need an initial intense 'top down' approach. In very early-onset IBD, personalised approaches to diagnosis and management have become the standard of treatment enabling clinicians to significantly alter the outcomes of the few children with monogenic disease. However, the promise of discoveries in genomics, microbiome and transcriptomics in paediatric IBD has not yet translated to clinical

application for the vast majority of patients. Using innovative and cutting-edge machine learning techniques gives the potential to use molecular, and clinical, data to develop personalised clinical care algorithms to treat patients more effectively, reduce toxicity and improve outcome. It will add to clinical trial data on how best to treat this vulnerable paediatric patient group, presenting with a serious chronic pathology. A 'personalised' approach now appears to be increasingly possible; the challenge is to move this routinely into the clinic.

List of References

1 Ashton JJ, Coelho T, Ennis S, *et al.* Presenting phenotype of paediatric inflammatory bowel disease in Wessex, Southern England 2010-2013. *Acta Paediatr Int J Paediatr* 2015;**104**. doi:10.1111/apa.13017

2 Levine A, Koletzko S, Turner D, *et al.* ESPGHAN revised porto criteria for the diagnosis of inflammatory bowel disease in children and adolescents. *J Pediatr Gastroenterol Nutr* 2014;**58**:795–806. doi:10.1097/MPG.0000000000000239

3 Rosen MJ, Dhawan A, Saeed SA. Inflammatory Bowel Disease in Children and Adolescents. *JAMA Pediatr* 2015;**169**:1053–60. doi:10.1001/jamapediatrics.2015.1982

4 Aloi M, Nuti F, Stronati L, *et al.* Advances in the medical management of paediatric IBD. *Nat Rev Gastroenterol Hepatol* 2014;**11**:99–108. doi:10.1038/nrgastro.2013.158

5 Oliveira SB, Monteiro IM. Diagnosis and management of inflammatory bowel disease in children. *BMJ* 2017;**357**:j2083. doi:10.1136/BMJ.J2083

6 Ashton JJ, Ennis S, Beattie RM. Early-onset paediatric inflammatory bowel disease. *Lancet Child Adolesc Heal* 2017;**1**:147–58. doi:10.1016/S2352-4642(17)30017-2

7 Denson LA, Long MD, McGovern DPB, *et al.* Challenges in IBD research: Update on progress and prioritization of the CCFA's research agenda. *Inflamm Bowel Dis* 2013;**19**:677–82.

8 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**:659–64. doi:10.1136/archdischild-2013-305419

9 Su HY, Gupta V, Day AS, *et al.* Rising Incidence of Inflammatory Bowel Disease in Canterbury, New Zealand. *Inflamm Bowel Dis* 2016;**22**:2238–44.

doi:10.1097/MIB.0000000000000829

10 Hope B, R. S, C. D, *et al.* Rapid rise in incidence of Irish paediatric inflammatory bowel disease. *Arch Dis Child* 2012;**97**:590–4. doi:10.1136/archdischild-2011-300651

11 Henderson P, Richard H, L. CF, *et al.* Rising incidence of pediatric inflammatory bowel disease in Scotland. *Inflamm Bowel Dis* 2012;**18**:999–1005. doi:10.1002/ibd.21797

12 Benchimol EI, A. G, M. GA, *et al.* Increasing incidence of paediatric inflammatory bowel disease in Ontario, Canada: evidence from health administrative data. *Gut* 2009;**58**:1490–7. doi:10.1136/gut.2009.188383

13 Sawczenko A, K. SB, A. LRF, *et al.* Prospective survey of childhood inflammatory bowel disease in the British Isles UK collaborative. 2001;**357**:1093–4.

14 Benchimol EI, Bernstein CN, Bitton A, *et al.* Trends in Epidemiology of Pediatric Inflammatory Bowel Disease in Canada: Distributed Network Analysis of Multiple Population-Based Provincial Health Administrative Databases. *Am J Gastroenterol* 2017;**112**:1120–34. doi:10.1038/ajg.2017.97

15 Virta LJ, Saarinen MM, Kolho K-L. Inflammatory Bowel Disease Incidence is on the Continuous Rise Among All Paediatric Patients Except for the Very Young: A Nationwide Registry-based Study on 28-Year Follow-up. *J Crohn's Colitis* 2017;**11**:150–6. doi:10.1093/ecco-jcc/jjw148

16 Uhlig HH, Schwerd T, Koletzko S, *et al.* The diagnostic approach to monogenic very early onset inflammatory bowel disease. *Gastroenterology* 2014;**147**:990-1007.e3. doi:10.1053/j.gastro.2014.07.023

17 Uhlig HH. Monogenic diseases associated with intestinal inflammation: implications for the understanding of inflammatory bowel disease. *Gut* 2013;**62**:1795–805. doi:10.1136/gutjnl-2012-303956

18 Ashton JJ, Coelho T, Ennis S, *et al.* Presenting Phenotype of Paediatric Inflammatory Bowel Disease (PIBD) in Wessex, Southern England 2010-13. *Acta Paediatr* 2015;**104**:831–7. doi:10.1111/apa.13017

19 Sawczenko a. Presenting features of inflammatory bowel disease in Great Britain and Ireland. *Arch Dis Child* 2003;**88**:995–1000. doi:10.1136/adc.88.11.995

20 Ashton JJ, Coelho T, Ennis S, *et al.* Endoscopic Versus Histological Disease Extent at Presentation of Paediatric Inflammatory Bowel Disease. *J Pediatr Gastroenterol Nutr* 2016;**62**:246–51. doi:10.1097/MPG.0000000000001032

21 Ashton JJ, Bonduelle Q, Mossotto E, *et al.* Endoscopic and Histological Assessment of Paediatric Inflammatory Bowel Disease Over a Three Year Follow-up Period. *J Pediatr Gastroenterol Nutr* 2018;**66**:402–9. doi:10.1097/MPG.0000000000001729

22 Van Limbergen J, Russell RK, Drummond HE, *et al.* Definition of phenotypic characteristics of childhood-onset inflammatory bowel disease. *Gastroenterology* 2008;**135**:1114–22. doi:10.1053/j.gastro.2008.06.081

23 Childers RE, Eluri S, Vazquez C, *et al.* Family history of inflammatory bowel disease among patients with ulcerative colitis: A systematic review and meta-analysis. *J Crohn's Colitis* 2014;**8**:1480–97. doi:10.1016/j.crohns.2014.05.008

24 Pascual V, Dieli-Crimi R, López-Palacios N, *et al.* Inflammatory bowel disease and celiac disease: overlaps and differences. *World J Gastroenterol* 2014;**20**:4846–56. doi:10.3748/wjg.v20.i17.4846

25 IBD Working Group of the European Society for Paediatric Gastroenterology H and N. Inflammatory bowel disease in children and adolescents: recommendations for diagnosis--the Porto criteria. *J Pediatr Gastroenterol Nutr* 2005;**41**:1–7.

26 Fell JM, Muhammed R, Spray C, *et al.* Management of ulcerative colitis. *Arch Dis Child*

2016;**101**:469–74. doi:10.1136/archdischild-2014-307218

27 Kammermeier J, Morris MA, Garrick V, *et al.* Management of Crohn's disease. *Arch Dis Child* 2016;**101**:475–80. doi:10.1136/archdischild-2014-307217

28 D'Arcangelo G, Aloi M. Inflammatory Bowel Disease-Unclassified in Children: Diagnosis and Pharmacological Management. *Paediatr Drugs* 2017;**19**:113–20. doi:10.1007/s40272-017-0213-9

29 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**:1314–21. doi:10.1002/ibd.21493

30 Fernandes MA, Verstraete SG, Garnett EA, *et al.* Addition of Histology to the Paris Classification of Pediatric Crohn Disease Alters Classification of Disease Location. *J Pediatr Gastroenterol Nutr* 2016;**62**:242–5. doi:10.1097/MPG.0000000000000967

31 Turner D, Levine A, Escher JC, *et al.* Management of pediatric ulcerative colitis: joint ECCO and ESPGHAN evidence-based consensus guidelines. *J Pediatr Gastroenterol Nutr* 2012;**55**:340–61. doi:10.1097/MPG.0b013e3182662233

32 Ruemmele FM, Veres G, Kolho KL, *et al.* Consensus guidelines of ECCO/ESPGHAN on the medical management of pediatric Crohn's disease. *J Crohns Colitis* 2014;**8**:1179–207. doi:10.1016/j.crohns.2014.04.005

33 Kammermeier J, Morris MA, Garrick V, *et al.* Management of Crohn's disease. *Arch Dis Child* 2016;**101**:475–80. doi:10.1136/archdischild-2014-307217

34 Fell JM, Muhammed R, Spray C, *et al.* Management of ulcerative colitis. *Arch Dis Child* 2016;**101**:469–74. doi:10.1136/archdischild-2014-307218

35 Sigall-Boneh R, Pfeffer-Gik T, Segal I, *et al.* Partial Enteral Nutrition with a Crohn's Disease

Exclusion Diet Is Effective for Induction of Remission in Children and Young Adults with Crohn's Disease. *Inflamm Bowel Dis* 2014;**20**:1353–60. doi:10.1097/MIB.0000000000000110

36 Sigall Boneh R, Sarbagili Shabat C, Yanai H, *et al.* Dietary Therapy With the Crohn's Disease Exclusion Diet is a Successful Strategy for Induction of Remission in Children and Adults Failing Biological Therapy. *J Crohn's Colitis* 2017;**11**:1205–12. doi:10.1093/ecco-jcc/jjx071

37 Svolos V, Hansen R, Nichols B, *et al.* Treatment of Active Crohn's Disease With an Ordinary Food-based Diet That Replicates Exclusive Enteral Nutrition. *Gastroenterology* 2019;**156**:1354–67. doi:10.1053/j.gastro.2018.12.002

38 Levine A, Wine E, Assa A, *et al.* Crohn's Disease Exclusion Diet Plus Partial Enteral Nutrition Induces Sustained Remission in a Randomized Controlled Trial. *Gastroenterology* 2019;**157**:440-450.e8. doi:10.1053/j.gastro.2019.04.021

39 Hyams J, Walters TD, Crandall W, *et al.* Safety and efficacy of maintenance infliximab therapy for moderate-to-severe Crohn's disease in children: REACH open-label extension. *Curr Med Res Opin* 2011;**27**:651–62. doi:10.1185/03007995.2010.547575

40 Hyams JS, Lerer T, Griffiths A, *et al.* Outcome following infliximab therapy in children with ulcerative colitis. *Am J Gastroenterol* 2010;**105**:1430–6. doi:10.1038/ajg.2009.759

41 D'Haens GR. Top-down therapy for IBD: rationale and requisite evidence. *Nat Rev Gastroenterol Hepatol* 2010;**7**:86–92. doi:10.1038/nrgastro.2009.222

42 Van Rheenen PF, Aloi M, Assa A, *et al.* The Medical Management of Paediatric Crohn's Disease: an ECCO-ESPGHAN Guideline Update. *J Crohn's Colitis* 2020;**2020**:1–24. doi:10.1093/ecco-jcc/jjaa161

43 Dave M, Loftus E V. Mucosal healing in inflammatory bowel disease-a true paradigm of success? *Gastroenterol Hepatol (N Y)* 2012;**8**:29–38.

44 Civitelli F, Nuti F, Oliva S, *et al.* Looking Beyond Mucosal Healing: Effect of Biologic Therapy on Transmural Healing in Pediatric Crohn's Disease. *Inflamm Bowel Dis* 2016;**22**:2418–24. doi:10.1097/MIB.0000000000000897

45 Velayos FS, Terdiman JP, Walsh JM. Effect of 5-aminosalicylate use on colorectal cancer and dysplasia risk: a systematic review and metaanalysis of observational studies. *Am J Gastroenterol* 2005;**100**:1345–53. doi:10.1111/j.1572-0241.2005.41442.x

46 Hyams JS, Lerer T, Mack D, *et al.* Outcome following thiopurine use in children with ulcerative colitis: a prospective multicenter registry study. *Am J Gastroenterol* 2011;**106**:981–7. doi:10.1038/ajg.2010.493

47 Riello L, Talbotec C, Garnier-Lengliné H, *et al.* Tolerance and efficacy of azathioprine in pediatric Crohn's disease. *Inflamm Bowel Dis* 2011;**17**:2138–43. doi:10.1002/ibd.21612

48 Hyams JS, Dubinsky MC, Baldassano RN, *et al.* Infliximab not Associated With Increased Risk of Malignancy or Hemophagocytic Lymphohistiocytosis in Pediatric Patients With Inflammatory Bowel Disease. *Gastroenterology* 2017;**152**:1901–14. doi:10.1053/j.gastro.2017.02.004

49 Gordon J, Ramaswami A, Beuttler M, *et al.* EBV Status and Thiopurine Use in Pediatric IBD. *J Pediatr Gastroenterol Nutr* 2016;**62**:711–4. doi:10.1097/MPG.0000000000001077

50 Olén O, Askling J, Sachs MC, *et al.* Childhood onset inflammatory bowel disease and risk of cancer: a Swedish nationwide cohort study 1964-2014. *BMJ* 2017;**358**:j3951. doi:10.1136/BMJ.J3951

51 Hyams J, Damaraju L, Blank M, *et al.* Induction and maintenance therapy with infliximab for children with moderate to severe ulcerative colitis. *Clin Gastroenterol Hepatol* 2012;**10**:391-9.e1. doi:10.1016/j.cgh.2011.11.026

52 Conrad MA, Stein RE, Maxwell EC, *et al.* Vedolizumab Therapy in Severe Pediatric

Inflammatory Bowel Disease. *Inflamm Bowel Dis* 2016;**22**:2425–31. doi:10.1097/MIB.0000000000000918

53 Singh N, Rabizadeh S, Jossen J, *et al.* Multi-Center Experience of Vedolizumab Effectiveness in Pediatric Inflammatory Bowel Disease. *Inflamm Bowel Dis* 2016;**22**:2121–6. doi:10.1097/MIB.0000000000000865

54 Feagan BG, Sandborn WJ, Gasink C, *et al.* Ustekinumab as Induction and Maintenance Therapy for Crohn's Disease. *N Engl J Med* 2016;**375**:1946–60. doi:10.1056/NEJMoa1602773

55 Dayan JR, Dolinger M, Benkov K, *et al.* Real World Experience With Ustekinumab in Children and Young Adults at a Tertiary Care Pediatric Inflammatory Bowel Disease Center. *J Pediatr Gastroenterol Nutr* 2019;**69**:61–7. doi:10.1097/MPG.0000000000002362

56 Ledder O, Assa A, Levine A, *et al.* Vedolizumab in Paediatric Inflammatory Bowel Disease: A Retrospective Multi-Centre Experience From the Paediatric IBD Porto Group of ESPGHAN. *J Crohns Colitis* 2017;**11**:1230–7. doi:10.1093/ecco-jcc/jjx082

57 Olivera P, Danese S, Peyrin-Biroulet L. Next generation of small molecules in inflammatory bowel disease. *Gut* 2017;**66**:199–209. doi:10.1136/gutjnl-2016-312912

58 De Vries LCS, Wildenberg ME, De Jonge WJ, *et al.* The Future of Janus Kinase Inhibitors in Inflammatory Bowel Disease. *J Crohn's Colitis* 2017;**11**:885–93. doi:10.1093/ecco-jcc/jjx003

59 Khor B, Gardet A, Xavier RJ, *et al.* Genetics and pathogenesis of inflammatory bowel disease. *Nature* 2011;**474**:307–17. doi:10.1038/nature10209

60 Franke A, P. MD, C. BJ, *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet* 2010;**42**:1118–25. doi:10.1038/ng.717

61 Kalla R, Adams A, Nimmo E, *et al.* Epigenetic alterations in infl ammatory bowel disease: the complex interplay between genome-wide methylation alterations, germline variation, and gene expression. *Lancet* 2017;**389**:S52. doi:10.1016/S0140-6736(17)30448-8

62 Gevers D, Kugathasan S, Denson LA, *et al.* The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* 2014;**15**:382–92. doi:10.1016/j.chom.2014.02.005

63 Ashton JJ, Colquhoun CM, Cleary DW, *et al.* 16S sequencing and functional analysis of the fecal microbiome during treatment of newly diagnosed pediatric inflammatory bowel disease. *Med* 2017;**96**:e7347. doi:10.1097/MD.0000000000007347

64 Douglas GM, Hansen R, Jones CMA, *et al.* Multi-omics differentially classify disease state and treatment outcome in pediatric Crohn's disease. *Microbiome* 2018;**6**:13. doi:10.1186/s40168-018-0398-3

65 Wright EK, Kamm MA, Teo SM, *et al.* Recent Advances in Characterizing the Gastrointestinal Microbiome in Crohn's Disease: A Systematic Review. *Inflamm Bowel Dis* 2015;**21**:1219–28. doi:10.1097/MIB.0000000000000382

66 Kostic AD, Xavier RJ, Gevers D. The microbiome in inflammatory bowel disease: current status and the future ahead. *Gastroenterology* 2014;**146**:1489–99. doi:10.1053/j.gastro.2014.02.009

67 Haberman Y, Tickle TL, Dexheimer PJ, *et al.* Pediatric Crohn disease patients exhibit specific ileal transcriptome and microbiome signature. *J Clin Invest* 2015;**125**:1363. doi:10.1172/JCI79657

68 Ashton JJ, Gavin J, Beattie RM. Exclusive enteral nutrition in Crohn's disease: Evidence and practicalities. *Clin Nutr* 2018;**38**:80–9. doi:10.1016/j.clnu.2018.01.020

69 Elsamanoudy AZ, Neamat-Allah MAM, Mohammad FAH, *et al.* The role of nutrition related genes and nutrigenetics in understanding the pathogenesis of cancer. *J Microsc Ultrastruct*

2016;**4**:115–22. doi:10.1016/J.JMAU.2016.02.002

70 Liu JZ, Sommeren S van, Huang H, *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet* 2015;**47**:979. doi:10.1038/NG.3359

71 Jostins L, Ripke S, Weersma RK, *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 2012;**491**:119–24. doi:10.1038/nature11582

72 Gordon H, Trier Moller F, Andersen V, *et al.* Heritability in inflammatory bowel disease: from the first twin study to genome-wide association studies. *Inflamm Bowel Dis* 2015;**21**:1428–34. doi:10.1097/MIB.0000000000000393

73 Hugot J-P, Chamaillard M, Zouali H, *et al.* Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* 2001;**411**:599–603. doi:10.1038/35079107

74 Hugot J-P, Laurent-Puig P, Gower-Rousseau C, *et al.* Mapping of a susceptibility locus for Crohn's disease on chromosome 16. *Nature* 1996;**379**:821–3. doi:10.1038/379821a0

75 Lesage S, Zouali H, Cézard J-P, *et al.* CARD15/NOD2 Mutational Analysis and Genotype-Phenotype Correlation in 612 Patients with Inflammatory Bowel Disease. *Am J Hum Genet* 2002;**70**:845–57. doi:10.1086/339432

76 Brant SR, Panhuysen CIM, Bailey–Wilson JE, *et al.* Linkage heterogeneity for the IBD1 locus in Crohn's disease pedigrees by disease onset and severity. *Gastroenterology* 2000;**119**:1483–90. doi:10.1053/GAST.2000.20245

77 Mateos B, Palanca-Ballester C, Saez-Gonzalez E, *et al.* Epigenetics of Inflammatory Bowel Disease: Unraveling Pathogenic Events. *Crohn's Colitis 360* 2019;**1**. doi:10.1093/crocol/otz017

78 Oh SH, Baek J, Liany H, *et al.* A Synonymous Variant in IL10RA Affects RNA Splicing in Paediatric Patients with Refractory Inflammatory Bowel Disease. *J Crohns Colitis* 2016;**10**:1366–71. doi:10.1093/ecco-jcc/jjw102

79 Zeissig Y, Petersen BS, Milutinovic S, *et al.* XIAP variants in male Crohn's disease. *Gut* 2015;**64**:66–76. doi:10.1136/gutjnl-2013-306520

80 Kammermeier J, Drury S, James CT, *et al.* Targeted gene panel sequencing in children with very early onset inflammatory bowel disease--evaluation and prospective analysis. *J Med Genet* 2014;**51**:748–55. doi:10.1136/jmedgenet-2014-102624

81 Ahmad T, Marshall S-E, Jewell D. Genetics of inflammatory bowel disease: the role of the HLA complex. *World J Gastroenterol* 2006;**12**:3628–35. doi:10.3748/WJG.V12.I23.3628

82 van Heel DA, Fisher SA, Kirby A, *et al.* Inflammatory bowel disease susceptibility loci defined by genome scan meta-analysis of 1952 affected relative pairs. *Hum Mol Genet* 2004;**13**:763–70. doi:10.1093/hmg/ddh090

83 Yang H, Plevy SE, Taylor K, *et al.* Linkage of Crohn's disease to the major histocompatibility complex region is detected by multiple non-parametric analyses. *Gut* 1999;**44**:519–26.http://www.ncbi.nlm.nih.gov/pubmed/10075959 (accessed 21 Sep 2018).

84 Stokkers PC, Reitsma PH, Tytgat GN, *et al.* HLA-DR and -DQ phenotypes in inflammatory bowel disease: a meta-analysis. *Gut* 1999;**45**:395–401.http://www.ncbi.nlm.nih.gov/pubmed/10446108 (accessed 1 Mar 2018).

85 Ahmad T, Armuzzi A, Bunce M, *et al.* The molecular classification of the clinical manifestations of Crohn's disease. *Gastroenterology* 2002;**122**:854–66.http://www.ncbi.nlm.nih.gov/pubmed/11910336 (accessed 21 Sep 2018).

86 Goyette P, Boucher G, Mallon D, *et al.* High-density mapping of the MHC identifies a shared role for HLA-DRB1*01:03 in inflammatory bowel diseases and heterozygous

advantage in ulcerative colitis. *Nat Genet* 2015;**47**:172–9. doi:10.1038/ng.3176

87 Ahmad T, Armuzzi A, Neville M, *et al.* The contribution of human leucocyte antigen complex genes to disease phenotype in ulcerative colitis. *Tissue Antigens* 2003;**62**:527–35. doi:10.1046/j.1399-0039.2003.00129.x

88 Matzaraki V, Kumar V, Wijmenga C, *et al.* The MHC locus and genetic susceptibility to autoimmune and infectious diseases. *Genome Biol* 2017;**18**:76. doi:10.1186/s13059-017-1207-1

89 Morgan XC, Kabakchiev B, Waldron L, *et al.* Associations between host gene expression, the mucosal microbiome, and clinical outcome in the pelvic pouch of patients with inflammatory bowel disease. *Genome Biol* 2015;**16**:67. doi:10.1186/s13059-015-0637-x

90 Moustafa A, Li W, Anderson EL, *et al.* Genetic risk, dysbiosis, and treatment stratification using host genome and gut microbiome in inflammatory bowel disease. *Clin Transl Gastroenterol* 2018;**9**:e132. doi:10.1038/ctg.2017.58

91 Graham DB, Xavier RJ. Pathway paradigms revealed from the genetics of inflammatory bowel disease. *Nature* 2020;**578**:527–39. doi:10.1038/s41586-020-2025-2

92 Zeissig Y, Petersen B-S, Milutinovic S, *et al.* XIAP variants in male Crohn's disease. *Gut* 2015;**64**:66–76. doi:10.1136/gutjnl-2013-306520

93 Sartor RB. Pathogenesis and immune mechanisms of chronic inflammatory bowel diseases. *Am J Gastroenterol* 1997;**92**:5S-11S.http://www.ncbi.nlm.nih.gov/pubmed/9395346 (accessed 2 Oct 2018).

94 Choy MC, Visvanathan K, De Cruz P. An overview of the innate and adaptive immune system in inflammatory bowel disease. Inflamm. Bowel Dis. 2017;**23**:2–13. doi:10.1097/MIB.0000000000000955

95 O'Shea JJ, Schwartz DM, Villarino A V., *et al.* The JAK-STAT pathway: Impact on human disease and therapeutic intervention. *Annu Rev Med* 2015;**66**:311–28. doi:10.1146/annurev-med-051113-024537

96 Yang Y, Wang H, Kouadir M, *et al.* Recent advances in the mechanisms of NLRP3 inflammasome activation and its inhibitors. Cell Death Dis. 2019;**10**:1–11. doi:10.1038/s41419-019-1413-8

97 Nguyen GT, Green ER, Mecsas J. Neutrophils to the ROScue: Mechanisms of NADPH oxidase activation and bacterial resistance. Front. Cell. Infect. Microbiol. 2017;**7**:373. doi:10.3389/fcimb.2017.00373

98 Ashton JJ, Latham K, Beattie RM, *et al.* Review article: the genetics of the human leucocyte antigen region in inflammatory bowel disease. *Aliment Pharmacol Ther* 2019;**50**:885–900. doi:10.1111/apt.15485

99 Ahluwalia B, Moraes L, Magnusson MK, *et al.* Immunopathogenesis of inflammatory bowel disease and mechanisms of biological therapies. *Scand J Gastroenterol* 2018;**53**:379–89. doi:10.1080/00365521.2018.1447597

100 Lloyd-Price J, Arze C, Ananthakrishnan AN, *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* 2019;**569**:655–62. doi:10.1038/s41586-019-1237-9

101 Haberman Y, Karns R, Dexheimer PJ, *et al.* Ulcerative colitis mucosal transcriptomes reveal mitochondriopathy and personalized mechanisms underlying disease severity and treatment response. *Nat Commun* 2019;**10**:38. doi:10.1038/s41467-018-07841-3

102 Schreiber S, Rosenstiel P, Hampe J, *et al.* Activation of signal transducer and activator of transcription (STAT) 1 in human chronic inflammatory bowel disease. *Gut* 2002;**51**:379. doi:10.1136/GUT.51.3.379

103 Swain SL. T-Cell Subsets: Who does the polarizing? *Curr Biol* 1995;**5**:849–51. doi:10.1016/S0960-9822(95)00170-9

104 Britton GJ, Contijoch EJ, Mogno I, *et al.* Microbiotas from Humans with Inflammatory Bowel Disease Alter the Balance of Gut Th17 and RORγt+ Regulatory T Cells and Exacerbate Colitis in Mice. *Immunity* 2019;**50**:212-224.e4. doi:10.1016/J.IMMUNI.2018.12.015

105 Sawa S, Lochner M, Satoh-Takayama N, *et al.* RORγt+ innate lymphoid cells regulate intestinal homeostasis by integrating negative signals from the symbiotic microbiota. *Nat Immunol* 2011;**12**:320–6. doi:10.1038/ni.2002

106 Lee P, Yacyshyn BR, Yacyshyn MB. Gut microbiota and obesity: An opportunity to alter obesity through faecal microbiota transplant (FMT). *Diabetes, Obes Metab* 2019;**21**:479–90. doi:10.1111/dom.13561

107 Costello SP, Hughes PA, Waters O, *et al.* Effect of Fecal Microbiota Transplantation on 8-Week Remission in Patients With Ulcerative Colitis. *JAMA* 2019;**321**:156. doi:10.1001/jama.2018.20046

108 Morgan XC, Tickle TL, Sokol H, *et al.* Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol* 2012;**13**:R79. doi:10.1186/gb-2012-13-9-r79

109 Imhann F, Vich Vila A, Bonder MJ, *et al.* Interplay of host genetics and gut microbiota underlying the onset and clinical presentation of inflammatory bowel disease. *Gut* 2018;**67**:108–19. doi:10.1136/gutjnl-2016-312135

110 Shaw KA, Bertha M, Hofmekler T, *et al.* Dysbiosis, inflammation, and response to treatment: a longitudinal study of pediatric subjects with newly diagnosed inflammatory bowel disease. *Genome Med* 2016;**8**:75. doi:10.1186/s13073-016-0331-y

111 Schirmer M, Franzosa EA, Lloyd-Price J, *et al.* Dynamics of metatranscription in the

inflammatory bowel disease gut microbiome. *Nat Microbiol* 2018;**3**:337–46. doi:10.1038/s41564-017-0089-z

112 Lavoie S, Conway KL, Lassen KG, *et al.* The Crohn's disease polymorphism, ATG16L1 T300A, alters the gut microbiota and enhances the local Th1/Th17 response. *Elife* 2019;**8**:e39982. doi:10.7554/eLife.39982

113 Aschard H, Laville V, Tchetgen ET, *et al.* Genetic effects on the commensal microbiota in inflammatory bowel disease patients. *PLoS Genet* 2019;**15**:e1008018. doi:10.1371/JOURNAL.PGEN.1008018

114 Definition of personalized medicine - NCI Dictionary of Cancer Terms - National Cancer Institute. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/personalized-medicine (accessed 2 Aug 2018).

115 Peplow M. The 100,000 Genomes Project. *BMJ* 2016;**353**:i1757.https://www.ncbi.nlm.nih.gov/pubmed/27075170

116 IMPROVING OUTCOMES THROUGH PERSONALISED MEDICINE Working at the cutting edge of science to improve patients' lives. https://www.england.nhs.uk/wp-content/uploads/2016/09/improving-outcomes-personalised-medicine.pdf (accessed 2 Aug 2018).

117 Weiser M, Simon JM, Kochar B, *et al.* Molecular classification of Crohn's disease reveals two clinically relevant subtypes. *Gut* Published Online First: 2016. doi:10.1136/gutjnl-2016-312518

118 Mossotto E, Ashton JJ, Coelho T, *et al.* Classification of Paediatric Inflammatory Bowel Disease using Machine Learning. *Sci Rep* 2017;**7**. doi:10.1038/s41598-017-02606-2

119 Waljee AK, Lipson R, Wiitala WL, *et al.* Predicting Hospitalization and Outpatient Corticosteroid Use in Inflammatory Bowel Disease Patients Using Machine Learning.

Inflamm Bowel Dis 2018;**24**:45–53. doi:10.1093/ibd/izx007

120 Kugathasan S, Denson LA, Walters TD, *et al.* Prediction of complicated disease course for children newly diagnosed with Crohn's disease: a multicentre inception cohort study. *Lancet* 2017;**389**:1710–8. doi:10.1016/S0140-6736(17)30317-3

121 Arijs I, Cleynen I. RISK stratification in paediatric Crohn's disease. *Lancet* 2017;**389**:1672–4. doi:10.1016/S0140-6736(17)30634-7

122 Marigorta UM, Denson LA, Hyams JS, *et al.* Transcriptional risk scores link GWAS to eQTLs and predict complications in Crohn's disease. *Nat Genet* 2017;**49**:1517–21. doi:10.1038/ng.3936

123 Denson LA, Jurickova I, Karns R, *et al.* Clinical and Genomic Correlates of Neutrophil Reactive Oxygen Species Production in Pediatric Patients With Crohn's Disease. *Gastroenterology* 2018;**154**:2097–110. doi:10.1053/j.gastro.2018.02.016

124 Lee JC, Lyons PA, McKinney EF, *et al.* Gene expression profiling of CD8+ T cells predicts prognosis in patients with Crohn disease and ulcerative colitis. *J Clin Invest* 2011;**121**:4170–9. doi:10.1172/JCI59255

125 Gasparetto M, Payne F, Nayak K, *et al.* Transcription and DNA Methylation patterns of blood derived CD8+ T cells are associated with age and Inflammatory Bowel Disease but do not predict prognosis. *Gastroenterology* 2020;**0**. doi:10.1053/j.gastro.2020.08.017

126 Pirmohamed M. Personalized Pharmacogenomics: Predicting Efficacy and Adverse Drug Reactions. *Annu Rev Genomics Hum Genet* 2014;**15**:349–70. doi:10.1146/annurev-genom-090413-025419

127 Arijs I, Quintens R, Van Lommel L, *et al.* Predictive value of epithelial gene expression profiles for response to infliximab in Crohn's disease‡. *Inflamm Bowel Dis* 2010;**16**:2090–8. doi:10.1002/ibd.21301

128 Arijs I, Li K, Toedter G, *et al.* Mucosal gene signatures to predict response to infliximab in patients with ulcerative colitis. *Gut* 2009;**58**:1612–9. doi:10.1136/gut.2009.178665

129 Kolho K-L, Korpela K, Jaakkola T, *et al.* Fecal Microbiota in Pediatric Inflammatory Bowel Disease and Its Relation to Inflammation. *Am J Gastroenterol* 2015;**110**:921–30. doi:10.1038/ajg.2015.149

130 Doherty MK, Ding T, Koumpouras C, *et al.* Fecal Microbiota Signatures Are Associated with Response to Ustekinumab Therapy among Crohn's Disease Patients. *MBio* 2018;**9**. doi:10.1128/mBio.02120-17

131 Ziv-Baran T, Hussey S, Sladek M, *et al.* Response to treatment is more important than disease severity at diagnosis for prediction of early relapse in new-onset paediatric Crohn's disease. *Aliment Pharmacol Ther* 2018;**48**:1242–50. doi:10.1111/apt.15016

132 Deeb SJ, Tyanova S, Hummel M, *et al.* Machine Learning-based Classification of Diffuse Large B-cell Lymphoma Patients by Their Protein Expression Profiles. *Mol Cell Proteomics* 2015;**14**:2947–60. doi:10.1074/mcp.M115.050245

133 Carter H, Douville C, Stenson PD, *et al.* Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* 2013;**14 Suppl 3**:S3. doi:10.1186/1471-2164-14-S3-S3

134 Vural S, Wang X, Guda C. Classification of breast cancer patients using somatic mutation profiles and machine learning approaches. *BMC Syst Biol* 2016;**10**:62. doi:10.1186/s12918-016-0306-z

135 Fabregat A, Jupe S, Matthews L, *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Res* 2018;**46**:D649–55. doi:10.1093/nar/gkx1132

136 Ashton JJ, Versteegh HP, Batra A, *et al.* Colectomy in pediatric ulcerative colitis: A single center experience of indications, outcomes, and complications. *J Pediatr Surg*

2016;**51**:277–81. doi:10.1016/j.jpedsurg.2015.10.077

137 Blackburn SC, Wiskin AE, Barnes C, *et al.* Surgery for children with Crohn's disease: indications, complications and outcome. *Arch Dis Child* 2014;**99**:420–6. doi:10.1136/archdischild-2013-305214

138 Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res* 1988;**16**:1215.http://www.ncbi.nlm.nih.gov/pubmed/3344216

139 Jun G, Flickinger M, Hetrick KN, *et al.* Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data. *Am J Hum Genet* 2012;**91**:839–48. doi:10.1016/j.ajhg.2012.09.004

140 Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *Genomics* 2013;:1303.3997.

141 McKenna A, Hanna M, Banks E, *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;**20**:1297–303. doi:10.1101/gr.107524.110

142 DePristo MA, Banks E, Poplin R, *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 2011;**43**:491–8. doi:10.1038/ng.806

143 Stenson PD, Mort M, Ball E V., *et al.* The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum Genet* 2017;**136**:665–77. doi:10.1007/s00439-017-1779-6

144 Lek M, Karczewski KJ, Minikel E V., *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016;**536**:285–91. doi:10.1038/nature19057

145 Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* 2015;**31**:761–3. doi:10.1093/bioinformatics/btu703

146 HTG. HTG Autoimmune - HTG Autoimmune. https://autoimmune.htgmolecular.com/

147 Peña-Chilet M, Dopazo J, Virgen del Rocío H, *et al.* Pazopanib for treatment of advanced malignant and dedifferentiated solitary fibrous tumour: a multicentre, single-arm, phase 2 trial. *Lancet Oncol* 2020;**21**:134–44. doi:10.1016/S1470-2045(18)30676-4

148 Godoy PM, Barczak AJ, Dehoff P, *et al.* Comparison of Reproducibility, Accuracy, Sensitivity, and Specificity of miRNA Quantification Platforms. *Cell Rep* 2019;**29**:4212–22. doi:10.1016/j.celrep.2019.11.078

149 Dillies M-A, Rau A, Aubert J, *et al.* A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform* 2013;**14**:671–83. doi:10.1093/bib/bbs046

150 Ashton JJ, Colquhoun CM, Cleary DW, *et al.* 16S sequencing and functional analysis of the fecal microbiome during treatment of newly diagnosed pediatric inflammatory bowel disease. *Med (United States)* 2017;**96**. doi:10.1097/MD.0000000000007347

151 Karczewski KJ, Francioli LC, Tiao G, *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv* 2019;:531210. doi:10.1101/531210

152 Shihab HA, Rogers MF, Gough J, *et al.* An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* 2015;**31**:1536–43. doi:10.1093/bioinformatics/btv009

153 Mossotto E, Ashton JJ, O'Gorman L, *et al.* GenePy - a score for estimating gene pathogenicity in individuals using next-generation sequencing data. *BMC Bioinformatics* 2019;**20**:254. doi:10.1186/s12859-019-2877-3

154 Richards S, Aziz N, Bale 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* 2015;**17**:405–24. doi:10.1038/gim.2015.30

155 Langfelder P, Horvath S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* 2008;**9**:559. doi:10.1186/1471-2105-9-559

156 Russo PST, Ferreira GR, Cardozo LE, *et al.* CEMiTool: A Bioconductor package for performing comprehensive modular co-expression analyses. *BMC Bioinformatics* 2018;**19**:56. doi:10.1186/s12859-018-2053-1

157 Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol* 2010;**11**:R106. doi:10.1186/gb-2010-11-10-r106

158 Bray NL, Pimentel H, Melsted P, *et al.* Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* 2016;**34**:525–7. doi:10.1038/nbt.3519

159 Wolf FA, Angerer P, Theis FJ. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol* 2018;**19**. doi:10.1186/s13059-017-1382-0

160 Lun ATL, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol* 2016;**17**. doi:10.1186/s13059-016-0947-7

161 Polański K, Young MD, Miao Z, *et al.* BBKNN: Fast batch alignment of single cell transcriptomes. *Bioinformatics* 2020;**36**:964–5. doi:10.1093/bioinformatics/btz625

162 Aran D, Looney AP, Liu L, *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* 2019;**20**:163–72. doi:10.1038/s41590-018-0276-y

163 Kanehisa M, Susumu G. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids*

Res 2000;**28**:27–30. doi:10.1093/nar/28.1.27

164 Lopez Y, Nakai K, Patil A. HitPredict Version 4: Comprehensive Reliability Scoring of Physical Protein-Protein Interactions From More Than 100 Species - PubMed. *Database* 2015;:bav117.

165 Chen J, Bardes EE, Aronow BJ, *et al.* ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res* 2009;**37**:W305-11. doi:10.1093/nar/gkp427

166 Kuleshov M V., Jones MR, Rouillard AD, *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* 2016;**44**:W90–7. doi:10.1093/nar/gkw377

167 Huang R, Grishagin I, Wang Y, *et al.* The NCATS BioPlanet – An Integrated Platform for Exploring the Universe of Cellular Signaling Pathways for Toxicology, Systems Biology, and Chemical Genomics. *Front Pharmacol* 2019;**10**:445. doi:10.3389/fphar.2019.00445

168 Morpheus. https://software.broadinstitute.org/morpheus/ (accessed 27 Sep 2018).

169 Bolyen E, Rideout JR, Dillon MR, *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat. Biotechnol. 2019;**37**:852–7. doi:10.1038/s41587-019-0209-9

170 DeSantis TZ, Hugenholtz P, Larsen N, *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Env Microbiol* 2006;**72**:5069–72. doi:10.1128/AEM.03006-05

171 iTOL: Interactive Tree Of Life. https://itol.embl.de/ (accessed 17 Jun 2020).

172 Beattie RM, Walker-Smith JA, Murch SH. Indications for investigation of chronic gastrointestinal symptoms. *Arch ofDisease Child* 1995;**73**:354–

5.http://adc.bmj.com/content/archdischild/73/4/354.full.pdf (accessed 5 Apr 2018).

173 Mack DR, Langton C, Markowitz J, *et al.* Laboratory values for children with newly diagnosed inflammatory bowel disease. *Pediatrics* 2007;**119**:1113–9. doi:10.1542/peds.2006-1865

174 Saha A, Tighe MP, Batra A. How to use faecal calprotectin in management of paediatric inflammatory bowel disease. *Arch Dis Child Educ Pr Ed* 2016;**101**:124–8. doi:10.1136/archdischild-2014-307941

175 Henderson P, Casey A, Lawrence SJ, *et al.* The Diagnostic Accuracy of Fecal Calprotectin During the Investigation of Suspected Pediatric Inflammatory Bowel Disease. *Am J Gastroenterol* 2012;**107**:941–9. doi:10.1038/ajg.2012.33

176 Holtman GA, Lisman-van Leeuwen Y, Day AS, *et al.* Use of Laboratory Markers in Addition to Symptoms for Diagnosis of Inflammatory Bowel Disease in Children. *JAMA Pediatr* 2017;**171**:984. doi:10.1001/jamapediatrics.2017.1736

177 Sabery N, Bass D. Use of Serologic Markers as a Screening Tool in Inflammatory Bowel Disease Compared With Elevated Erythrocyte Sedimentation Rate and Anemia. *Pediatrics* 2007;**119**:e193–9. doi:10.1542/peds.2006-1361

178 Day A, Day AS, Hamilton D, *et al.* Inflammatory Markers in Children With Newly Diagnosed Inflammatory Bowel Disease. *J Gastroenterol Hepatol Res* 2017;**6**:2329–32. doi:10.6051/

179 Ricciuto A, Fish JR, Tomalty DE, *et al.* Diagnostic delay in Canadian children with inflammatory bowel disease is more common in Crohn's disease and associated with decreased height. *Arch Dis Child* 2017;**103**:319–26. doi:10.1136/archdischild-2017-313060

180 Weinstein TA, Levine M, Pettei MJ, *et al.* Age and Family History at Presentation of Pediatric Inflammatory Bowel Disease. *J Pediatr Gastroenterol Nutr* 2003;**37**:609–13.https://insights.ovid.com/pubmed?pmid=14581806 (accessed 4 Jun 2018).

181 Ashton JJ, Ennis S, Beattie RM. Early-onset paediatric inflammatory bowel disease. *Lancet Child Adolesc Heal* 2017;**1**. doi:10.1016/S2352-4642(17)30017-2

182 Jostins L, S. R, K. WR, *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 2012;**491**:119–24. doi:10.1038/nature11582

183 Ogura Y, Bonen DK, Inohara N, *et al.* A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* 2001;**411**:603–6. doi:10.1038/35079114

184 Horowitz JE, Warner N, Staples J, *et al.* Mutation spectrum of NOD2 reveals recessive inheritance as a main driver of Early Onset Crohn's Disease. *bioRxiv* 2017;:098574. doi:10.1101/098574

185 Hyams, Jeffrey S. Thomas, Sonia Davis , Gotman, Nathan, Haberman, Yael , Rebekah Karns, Melanie Schirmer, Angela Mo, David R. Mack BB. Clinical and Biological Predictors of Response to Standardised Paediatric Colitis Therapy: A Multicentre Inception Cohort Study. *Lancet* 2018;**393**:1708–1720.

186 Takahashi S, Andreoletti G, Chen R, *et al.* De novo and rare mutations in the HSPA1L heat shock gene associated with inflammatory bowel disease. *Genome Med* 2017;**9**:8. doi:10.1186/s13073-016-0394-9

187 Worthey EA, Mayer AN, Syverson GD, *et al.* Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet Med* 2011;**13**:255–62. doi:10.1097/GIM.0b013e3182088158

188 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**:347–55. doi:10.1053/j.gastro.2012.04.045

189 Katsanis SH, Katsanis N. Molecular genetic testing and the future of clinical genomics. *Nat Publ Gr* 2013;**14**:415. doi:10.1038/nrg3493

190 Uhlig HH, Muise AM. Clinical Genomics in Inflammatory Bowel Disease. *Trends Genet* 2017;**33**:629–41. doi:10.1016/j.tig.2017.06.008

191 Girardelli M, Loganes C, Pin A, *et al.* Novel NOD2 Mutation in Early-Onset Inflammatory Bowel Phenotype. *Inflamm Bowel Dis* 2018;**24**:1204–12. doi:10.1093/ibd/izy061

192 Frade-Proud'Hon-Clerc S, Smol T, Frenois F, *et al.* A Novel Rare Missense Variation of the NOD2 Gene: Evidencesof Implication in Crohn's Disease. *Int J Mol Sci* 2019;**20**:835. doi:10.3390/ijms20040835

193 Landrum MJ, Lee JM, Benson M, *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* 2016;**44**:D862–8. doi:10.1093/nar/gkv1222

194 Pedersen BS, Quinlan AR. Who's Who? Detecting and Resolving Sample Anomalies in Human DNA Sequencing Studies with Peddy. *Am J Hum Genet* 2017;**100**:406–13. doi:10.1016/j.ajhg.2017.01.017

195 Lawless D, Mistry A, Wood PM, *et al.* Bialellic Mutations in Tetratricopeptide Repeat Domain 7A (TTC7A) Cause Common Variable Immunodeficiency-Like Phenotype with Enteropathy. *J Clin Immunol* 2017;**37**:617–22. doi:10.1007/s10875-017-0427-1

196 Jardine S, Dhingani N, Muise AM. TTC7A: Steward of Intestinal Health. *Cell Mol Gastroenterol Hepatol* Published Online First: 2018. doi:10.1016/j.jcmgh.2018.12.001

197 Li Q, Lee CH, Peters LA, *et al.* Variants in TRIM22 that Affect NOD2 Signaling Are Associated With Very Early Onset Inflammatory Bowel Disease HHS Public Access. *Gastroenterology* 2016;**150**:1196–207. doi:10.1053/j.gastro.2016.01.031

198 Zheng Y, Wu S, Yu X, *et al.* The WASP P460S Mutation Causes a New Phenotype of WASP Mutations Related Disorder: X-Linked Pancytopenia. *Blood* 2017;**130**.http://www.bloodjournal.org/content/130/Suppl_1/1044?sso-checked=true (accessed 19 Dec 2018).

199 Ohya T, Yanagimachi M, Iwasawa K, *et al.* Childhood-onset inflammatory bowel diseases associated with mutation of Wiskott-Aldrich syndrome protein gene. *World J Gastroenterol* 2017;**23**:8544–52. doi:10.3748/wjg.v23.i48.8544

200 Ashton JJ, Andreoletti G, Coelho T, *et al.* Identification of Variants in Genes Associated with Single-gene Inflammatory Bowel Disease by Whole-exome Sequencing. *Inflamm Bowel Dis* 2016;**22**:2317–27. doi:10.1097/MIB.0000000000000890

201 Thiel S, Steffensen R, Christensen IJ, *et al.* Deficiency of mannan-binding lectin associated serine protease-2 due to missense polymorphisms. *Genes Immun* 2007;**8**:154–63. doi:10.1038/sj.gene.6364373

202 Dhillon SS, Fattouh R, Elkadri A, *et al.* Variants in nicotinamide adenine dinucleotide phosphate oxidase complex components determine susceptibility to very early onset inflammatory bowel disease. *Gastroenterology* 2014;**147**:680-689.e2. doi:10.1053/j.gastro.2014.06.005

203 Rivas MA, M. B, A. G, *et al.* Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat Genet* 2011;**43**:1066–73. doi:10.1038/ng.952

204 Adler J, Rangwalla SC, Dwamena BA, *et al.* The Prognostic Power of the NOD2 Genotype for Complicated Crohn's Disease: A Meta-Analysis. *Am J Gastroenterol* 2011;**106**:699–712. doi:10.1038/ajg.2011.19

205 Abreu MT, Taylor KD, Lin Y-C, *et al.* Mutations in NOD2 are associated with fibrostenosing disease in patients with Crohn's disease. *Gastroenterology* 2002;**123**:679–88.http://www.ncbi.nlm.nih.gov/pubmed/12198692 (accessed 13 Jun 2019).

206 Verstockt B, Cleynen I. Genetic Influences on the Development of Fibrosis in Crohn's Disease. *Front Med* 2016;**3**:24. doi:10.3389/fmed.2016.00024

230

207 Salla M, Aguayo-Ortiz R, Danmaliki GI, *et al.* Identification and Characterization of Novel Receptor-Interacting Serine/Threonine-Protein Kinase 2 Inhibitors Using Structural Similarity Analysis. *J Pharmacol Exp Ther* 2018;**365**:354–67. doi:10.1124/jpet.117.247163

208 Hrdinka M, Schlicher L, Dai B, *et al.* Small molecule inhibitors reveal an indispensable scaffolding role of RIPK2 in NOD2 signaling. *EMBO J* 2018;**37**:e99372. doi:10.15252/embj.201899372

209 Ashton JJ, Mossotto E, Stafford IS, *et al.* Genetic Sequencing of Pediatric Patients Identifies Mutations in Monogenic Inflammatory Bowel Disease Genes that Translate to Distinct Clinical Phenotypes. *Clin Transl Gastroenterol* 2020;**11**:e00129. doi:10.14309/ctg.0000000000000129

210 Crowley E, Warner N, Pan J, *et al.* Prevalence and Clinical Features of Inflammatory Bowel Diseases Associated with Monogenic Variants, Identified by Whole-exome Sequencing in 1000 Children at a Single Center. *Gastroenterology* Published Online First: 2020. doi:10.1053/j.gastro.2020.02.023

211 Ashton JJ, Mossotto E, Pandey S, *et al.* Sa1121 COMPOUND HETEROZYGOTE TRIM22 VARIANTS ARE A POTENTIAL MODIFIER OF INFLAMMATORY BOWEL DISEASE THROUGH DEFECTIVE MURAMYL DIPEPTIDE-MEDIATED ANTIMICROBIAL ACTIVITY. *Gastroenterology* 2020;**158**:S-282. doi:10.1016/S0016-5085(20)31395-0

212 Li Q, Lee CH, Peters LA, *et al.* Variants in TRIM22 that Affect NOD2 Signaling Are Associated With Very Early Onset Inflammatory Bowel Disease. *Gastroenterology* Published Online First: 2016. doi:10.1053/j.gastro.2016.01.031

213 Caruso R, Mathes T, Martens EC, *et al.* A specific gene-microbe interaction drives the development of Crohn's disease–like colitis in mice. *Sci Immunol* 2019;**4**. doi:10.1126/sciimmunol.aaw4341

214 Ashton JJ, Mossotto E, Ennis S, *et al.* Personalising medicine in inflammatory bowel disease—current and future perspectives. *Transl Pediatr* 2019;**8**:56. doi:10.21037/TP.2018.12.03

215 Cleynen I, González JR, Figueroa C, *et al.* Genetic factors conferring an increased susceptibility to develop Crohn's disease also influence disease phenotype: results from the IBDchip European Project. *Gut* 2013;**62**:1556–65. doi:10.1136/gutjnl-2011-300777

216 Schwartz DA, Loftus E V., Tremaine WJ, *et al.* The natural history of fistulizing Crohn's disease in Olmsted County, Minnesota. *Gastroenterology* 2002;**122**:875–80. doi:10.1053/gast.2002.32362

217 Panday A, Sahoo MK, Osorio D, *et al.* NADPH oxidases: An overview from structure to innate immunity-associated pathologies. Cell. Mol. Immunol. 2015;**12**:5–23. doi:10.1038/cmi.2014.89

218 Kannengiesser C, Gérard B, El Benna J, *et al.* Molecular epidemiology of chronic granulomatous disease in a series of 80 kindreds: identification of 31 novel mutations. *Hum Mutat* 2008;**29**:E132-49. doi:10.1002/humu.20820

219 Denson LA, Jurickova I, Karns R, *et al.* Clinical and Genomic Correlates of Neutrophil Reactive Oxygen Species Production in Pediatric Patients With Crohn's Disease. *Gastroenterology* 2018;**154**:2097–110. doi:10.1053/j.gastro.2018.02.016

220 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.http://www.ncbi.nlm.nih.gov/pubmed/15990620 (accessed 19 Feb 2018).

221 Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* 2019;**176**:535-548.e24.

232

doi:10.1016/j.cell.2018.12.015

222 Hayes P, Dhillon S, O'Neill K, *et al.* Defects in Nicotinamide-adenine Dinucleotide Phosphate Oxidase Genes NOX1 and DUOX2 in Very Early Onset Inflammatory Bowel Disease. *CMGH* 2015;**1**:489–502. doi:10.1016/j.jcmgh.2015.06.005

223 Tozer PJ, Whelan K, Phillips RKS, *et al.* Etiology of perianal Crohn's disease: Role of genetic, microbiological, and immunological factors. Inflamm. Bowel Dis. 2009;**15**:1591–8. doi:10.1002/ibd.21026

224 Beser OF, Conde CD, Serwas NK, *et al.* Clinical features of interleukin 10 receptor gene mutations in children with very early-onset inflammatory bowel disease. *J Pediatr Gastroenterol Nutr* 2015;**60**:332–8. doi:10.1097/MPG.0000000000000621

225 Marlow GJ, Van Gent D, Ferguson LR, *et al.* Why interleukin-10 supplementation does not work in Crohn's disease patients. Published Online First: 2013. doi:10.3748/wjg.v19.i25.3931

226 Engelhardt KR, Grimbacher B. IL-10 in humans: Lessons from the Gut, IL-10/IL-10 receptor deficiencies, and IL-10 polymorphisms. *Curr Top Microbiol Immunol* 2014;**380**:1–18. doi:10.1007/978-3-662-43492-5_1

227 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**:431–5. doi:10.1016/j.humimm.2011.02.014

228 Amre DK, MacK DR, Morgan K, *et al.* Interleukin 10 (IL-10) gene variants and susceptibility for paediatric onset Crohn's disease. *Aliment Pharmacol Ther* 2009;**29**:1025–31. doi:10.1111/j.1365-2036.2009.03953.x

229 Aithal GP, Craggs A, Day CP, *et al.* Role of polymorphisms in the interleukin-10 gene in determining disease susceptibility and phenotype in inflamatory bowel disease. *Dig Dis Sci*

2001;**46**:1520–5. doi:10.1023/A:1010604307776

230 Haac BE, Palmateer NC, Seaton ME, *et al.* A Distinct Gut Microbiota Exists Within Crohn's Disease–Related Perianal Fistulae. *J Surg Res* 2019;**242**:118–28. doi:10.1016/j.jss.2019.04.032

231 Marzo M, Felice C, Pugliese D, *et al.* Management of perianal fistulas in Cohn's disease: An up-to-date review. *World J Gastroenterol* 2015;**21**:1394–403. doi:10.3748/wjg.v21.i5.1394

232 Chieppa M, Specializzato O, Saverio De Bellis G, *et al.* Heme Oxygenase-1 as a Modulator of Intestinal Inflammation Development and Progression. *Front Immunol | www.frontiersin.org* 2018;**9**. doi:10.3389/fimmu.2018.01956

233 Moll F, Walter M, Rezende F, *et al.* NoxO1 controls proliferation of colon epithelial cells. *Front Immunol* 2018;**9**:1. doi:10.3389/fimmu.2018.00973

234 Schwerd T, Bryant R V., Pandey S, *et al.* NOX1 loss-of-function genetic variants in patients with inflammatory bowel disease. *Mucosal Immunol* 2018;**11**:562–74. doi:10.1038/mi.2017.74

235 Lipinski S, Petersen BS, Barann M, *et al.* Missense variants in NOX1 and p22phox in a case of very-early-onset inflammatory bowel disease are functionally linked to NOD2. *Cold Spring Harb Mol Case Stud* 2019;**5**:a002428. doi:10.1101/mcs.a002428

236 Brunson T, Wang Q, Chambers I, *et al.* A copy number variation in human NCF1 and its pseudogenes. *BMC Genet* 2010;**11**:13. doi:10.1186/1471-2156-11-13

237 West NR, Hegazy AN, Owens BMJ, *et al.* Oncostatin M drives intestinal inflammation and predicts response to tumor necrosis factor–neutralizing therapy in patients with inflammatory bowel disease. *Nat Med* 2017;**23**:579–89. doi:10.1038/nm.4307

238 Verstockt B, Verstockt S, Dehairs J, *et al.* Low TREM1 expression in whole blood predicts

anti-TNF response in inflammatory bowel disease. *EBioMedicine* 2019;**40**:733–42. doi:10.1016/j.ebiom.2019.01.027

239 Palmer NP, Silvester JA, Lee JJ, *et al.* Concordance between gene expression in peripheral whole blood and colonic tissue in children with inflammatory bowel disease. *PLoS One* 2019;**14**:e0222952. doi:10.1371/journal.pone.0222952

240 Martin DP, Miya J, Reeser JW, *et al.* Targeted RNA sequencing assay to characterize gene expression and genomic alterations. *J Vis Exp* 2016;**2016**. doi:10.3791/54090

241 Martin-Broto J, Cruz J, Penel N, *et al.* Pazopanib for treatment of typical solitary fibrous tumours: a multicentre, single-arm, phase 2 trial. *Lancet Oncol* 2020;**21**:456–66. doi:10.1016/s1470-2045(19)30826-5

242 Hurtado M, Prokai L, Sankpal UT, *et al.* Next generation sequencing and functional pathway analysis to understand the mechanism of action of copper-tolfenamic acid against pancreatic cancer cells. *Process Biochem* 2020;**89**:155–64. doi:10.1016/j.procbio.2019.10.022

243 Martin JC, Chang C, Boschetti G, *et al.* Single-Cell Analysis of Crohn's Disease Lesions Identifies a Pathogenic Cellular Module Associated with Resistance to Anti-TNF Therapy. *Cell* 2019;**178**:1493-1508.e20. doi:10.1016/j.cell.2019.08.008

244 Vallejo AF, Davies J, Grover A, *et al.* Resolving cellular systems by ultra-sensitive and economical single-cell transcriptome filtering. *bioRxiv* 2019;:800631. doi:10.1101/800631

245 Macosko EZ, Basu A, Satija R, *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* 2015;**161**:1202–14. doi:10.1016/j.cell.2015.05.002

246 Lun ATL, Riesenfeld S, Andrews T, *et al.* EmptyDrops: Distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol* 2019;**20**.

doi:10.1186/s13059-019-1662-y

247 Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* 2019;**9**:5233. doi:10.1038/s41598-019-41695-z

248 Gudjonsson JE, Ding J, Johnston A, *et al.* Assessment of the psoriatic transcriptome in a large sample: Additional regulated genes and comparisons with in vitro models. *J Invest Dermatol* 2010;**130**:1829–40. doi:10.1038/jid.2010.36

249 Hou G, Bishu S. Th17 Cells in Inflammatory Bowel Disease: An Update for the Clinician. *Inflamm Bowel Dis* 2020;**26**:653–61. doi:10.1093/ibd/izz316

250 Fujino S, Andoh A, Bamba S, *et al.* Increased expression of interleukin 17 in inflammatory bowel disease. *Gut* 2003;**52**:65–70. doi:10.1136/gut.52.1.65

251 Sahin A, Calhan T, Cengiz M, *et al.* Serum Interleukin 17 Levels in Patients with Crohn's Disease: Real Life Data. *Dis Markers* 2014;**2014**. doi:10.1155/2014/690853

252 Youssef S, Steinman L, Defea K, *et al.* IL-17 in Colonic Epithelial Cells Differential Regulation of Chemokines by. *J Immunol Ref* 2008;**181**:6536–45. doi:10.4049/jimmunol.181.9.6536

253 Stawczyk-Eder K, Eder P, Lykowska-Szuber L, *et al.* Is faecal calprotectin equally useful in all Crohn's disease locations? A prospective, comparative study. *Arch Med Sci* 2015;**11**:353–61. doi:10.5114/aoms.2014.43672

254 Howell KJ, Kraiczy J, Nayak KM, *et al.* DNA Methylation and Transcription Patterns in Intestinal Epithelial Cells From Pediatric Patients With Inflammatory Bowel Diseases Differentiate Disease Subtypes and Associate With Outcome. *Gastroenterology* 2018;**154**:585–98. doi:10.1053/j.gastro.2017.10.007

255 Caruso R, Warner N, Inohara N, *et al.* NOD1 and NOD2: signaling, host defense, and inflammatory disease. *Immunity* 2014;**41**:898–908. doi:10.1016/j.immuni.2014.12.010

256 Coelho T, Mossotto E, Gao Y, *et al.* Immunological Profiling of Paediatric Inflammatory Bowel Disease Using Unsupervised Machine Learning. *J Pediatr Gastroenterol Nutr* 2020;:1. doi:10.1097/mpg.0000000000002719

257 Chiriaco M, Salfa I, Di Matteo G, *et al.* Chronic granulomatous disease: Clinical, molecular, and therapeutic aspects. *Pediatr Allergy Immunol* 2016;**27**:242–53. doi:10.1111/pai.12527

258 Andreoletti G, Shakhnovich V, Christenson K, *et al.* Exome Analysis of Rare and Common Variants within the NOD Signaling Pathway. *Sci Rep* 2017;**7**:46454. doi:10.1038/srep46454

259 Gene List HTG EdgeSeq Autoimmune Panel.

260 Israël A. The IKK complex, a central regulator of NF-kappaB activation. Cold Spring Harb. Perspect. Biol. 2010;**2**. doi:10.1101/cshperspect.a000158

261 Coelho T, Mossotto E, Gao Y, *et al.* Immunological Profiling of Paediatric Inflammatory Bowel Disease Using Unsupervised Machine Learning. *J Pediatr Gastroenterol Nutr* 2020;**70**:833–40. doi:10.1097/MPG.0000000000002719

262 Bonen DK, Ogura Y, Nicolae DL, *et al.* Crohn's disease-associated NOD2 variants share a signaling defect in response to lipopolysaccharide and peptidoglycan. *Gastroenterology* 2003;**124**:140–6. doi:10.1053/gast.2003.50019

263 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**:1794–6. doi:10.1016/S0140-6736(05)66582-8

264 Lu M, Lin SC, Huang Y, *et al.* XIAP Induces NF-κB Activation via the BIR1/TAB1 Interaction and BIR1 Dimerization. *Mol Cell* 2007;**26**:689–702. doi:10.1016/j.molcel.2007.05.006

265 Parackova Z, Milota T, Vrabcova P, *et al.* Novel XIAP mutation causing enhanced spontaneous apoptosis and disturbed NOD2 signalling in a patient with atypical adult-

onset Crohn's disease. *Cell Death Dis* 2020;**11**:1–11. doi:10.1038/s41419-020-2652-4

266 Damgaard RB, Fiil BK, Speckmann C, *et al.* Disease-causing mutations in the XIAP BIR2 domain impair <scp>NOD</scp> 2-dependent immune signalling. *EMBO Mol Med* 2013;**5**:1278–95. doi:10.1002/emmm.201303090

267 De Bruyne M, Hoste L, Bogaert DJ, *et al.* A CARD9 Founder Mutation Disrupts NF-κB Signaling by Inhibiting BCL10 and MALT1 Recruitment and Signalosome Formation. *Front Immunol* 2018;**9**:2366. doi:10.3389/fimmu.2018.02366

268 Goncharov T, Hedayati S, Mulvihill MM, *et al.* Disruption of XIAP-RIP2 Association Blocks NOD2-Mediated Inflammatory Signaling. *Mol Cell* 2018;**69**:551-565.e7. doi:10.1016/j.molcel.2018.01.016

269 Warner N, Burberry A, Franchi L, *et al.* A genome-wide siRNA screen reveals positive and negative regulators of the NOD2 and NF-κB signaling pathways. *Sci Signal* 2013;**6**:rs3–rs3. doi:10.1126/scisignal.2003305

270 Ashton JJ. Genetic sequencing of paediatric patients identifies mutations in monogenic inflammatory bowel disease genes that translate to distinct clinical phenotypes. *Clin Transl Gastroenterol* 2020.

271 Sorbara MT, Ellison LK, Ramjeet M, *et al.* The protein ATG16L1 suppresses inflammatory cytokines induced by the intracellular sensors Nod1 and Nod2 in an autophagy-independent manner. *Immunity* 2013;**39**:858–73. doi:10.1016/j.immuni.2013.10.013

272 Kim JY, Omori E, Matsumoto K, *et al.* TAK1 is a central mediator of NOD2 signaling in epidermal cells. *J Biol Chem* 2008;**283**:137–44. doi:10.1074/jbc.M704746200

273 Ashton JJ, Gavin J, Beattie RM. Exclusive enteral nutrition in Crohn's disease: Evidence and practicalities. *Clin Nutr* 2018;**38**:80–9. doi:10.1016/j.clnu.2018.01.020

Bibliography

Papers published during this PhD candidature (since September 2017)

1. Ashton JJ, Boukas K, Davies JD, Stafford ISS, Vallejo AF, Haggarty R, et al Ileal transcriptomic analysis in paediatric Crohn's disease reveals IL17- and NOD-signalling expression signatures in treatment-naïve patients and identifies epithelial cells driving differentially expressed genes, Journal of Crohn's and Colitis, Nov 2020 IN PRESS
2. Maclean A, **Ashton JJ,** Garrick V, Beattie RM, Hansen R, The Impact of COVID-19 on the Diagnosis, Assessment and Management of Children with Inflammatory Bowel Disease in the UK: Implications for Practice, BMJ Paed Open, Oct 2020 IN PRESS http://dx.doi.org/10.1136/bmjpo-2020-000786
3. **Ashton** JJ, Kammermeier J, Spray C, Russell RK, Hansen R, Howarth LJ, et al. Impact of COVID-19 on diagnosis and management of paediatric inflammatory bowel disease during lockdown: a UK nationwide study. Arch Dis Child. 2020;0:archdischild-2020-319751. doi:10.1136/archdischild-2020-319751.
4. Beattie RM, **Ashton** JJ, Penman ID. COVID-19 and the gastrointestinal tract: recent data. Frontline Gastroenterol. 2020;0:flgastro-2020-101602. doi:10.1136/flgastro-2020-101602.
5. **Ashton** JJ, Green Z, Young A, Borca F, Coelho T, Batra A, et al. Growth failure is rare in a contemporary cohort of paediatric inflammatory bowel disease patients. Acta Paediatr. 2020;:apa.15383. doi:10.1111/apa.15383.
6. **Ashton** JJ, Smith R, Smith T, Beattie RM. Investigating coeliac disease in adults. BMJ. 2020;369:m2176. doi:10.1136/bmj.m2176.
7. Young A, Andrews ET, **Ashton** JJ, Pearson F, Beattie RM, Johnson MJ. Generating longitudinal growth charts from preterm infants fed to current recommendations. Arch Dis Child Fetal Neonatal Ed. 2020. doi:10.1136/archdischild-2019-318404.

8. **Ashton** JJ, Batra A, Coelho TAF, Afzal NA, Beattie RM. Challenges in chronic paediatric disease during the COVID-19 pandemic: diagnosis and management of inflammatory bowel disease in children. Arch Dis Child. 2020. doi:10.1136/archdischild-2020-319482.
9. Beattie RM, **Ashton** JJ, Penman ID. COVID-19 and the gastrointestinal tract: Emerging clinical data. Frontline Gastroenterology. 2020;11:290–2.
10. **Ashton** JJ, Green Z, Beattie RM. Beyond bedside measures of malnutrition in paediatric Crohn's disease – Should we be thinking of sarcopenia. Clinical Nutrition. 2020;39. doi:10.1016/j.clnu.2020.03.034.
11. Coelho T, Mossotto E, Gao Y, Haggarty R, **Ashton** JJ, Batra A, et al. Immunological Profiling of Paediatric Inflammatory Bowel Disease Using Unsupervised Machine Learning. J Pediatr Gastroenterol Nutr. 2020;:1.
12. **Ashton** JJ, Beattie RM. Can risk stratification help reduce negative appendicectomy rates? The Lancet Child and Adolescent Health. 2020;4:252–3. doi:10.1016/S2352-4642(20)30042-0.
13. **Ashton** JJ, Mossotto E, Stafford IS, Haggarty R, Coelho TAF, Batra A, et al. Genetic Sequencing of Pediatric Patients Identifies Mutations in Monogenic Inflammatory Bowel Disease Genes that Translate to Distinct Clinical Phenotypes. Clin Transl Gastroenterol. 2020;11:e00129. doi:10.14309/ctg.0000000000000129.
14. Barnes C, **Ashton** JJ, Borca F, Cullen M, Walker DM, Beattie RM. Children and young people with inflammatory bowel disease attend less school than their healthy peers. Arch Dis Child. 2019.
15. **Ashton** JJ, Mossotto E, Beattie RM, Ennis S. TTC7A Variants Previously Described to Cause Enteropathy Are Observed on a Single Haplotype and Appear Non-pathogenic in Pediatric Inflammatory Bowel Disease Patients. Journal of Clinical Immunology. 2020;40:245–7. doi:10.1007/s10875-019-00726-0.
16. **Ashton** JJ, Green Z, Kolimarala V, Beattie RM. Inflammatory bowel disease: long-

term therapeutic challenges. Expert Review of Gastroenterology and Hepatology. 2019;13:1049–63. doi:10.1080/17474124.2019.1685872.

17. **Ashton** JJ, Beattie RM. Letter: anti-TNF therapy and intestinal resections in Crohn's disease-are we just delaying the inevitable? Aliment Pharmacol Ther. 2019;50:842–3.
18. **Ashton** JJ, Latham K, Beattie RM, Ennis S. Review article: the genetics of the human leucocyte antigen region in inflammatory bowel disease. Aliment Pharmacol Ther. 2019;50:885–900. doi:10.1111/apt.15485.
19. Young A, Andrews ET, **Ashton** JJ, Pearson F, Beattie RM, Johnson MJ. ' Catch-up' growth of infants with IUGR does not significantly contribute to the whole-cohort weight gain pattern. Archives of Disease in Childhood: Fetal and Neonatal Edition. 2019;104:F663–4. doi:10.1136/archdischild-2019-317566.
20. Andrews ET, **Ashton** JJ, Pearson F, Beattie RM, Johnson MJ. Handheld 3D scanning as a minimally invasive measuring technique for neonatal anthropometry. Clin Nutr ESPEN. 2019;33:279–82. doi:10.1016/j.clnesp.2019.06.012.
21. **Ashton** JJ, Beattie RM. Treatment of Active Crohn's Disease With an Ordinary Food-Based Diet That Replicates Exclusive Enteral Nutrition. Gastroenterology. 2019;157:1160–1.
22. Mossotto E, **Ashton** JJ, O'Gorman L, Pengelly RJ, Beattie RM, MacArthur BD, et al. GenePy - a score for estimating gene pathogenicity in individuals using next-generation sequencing data. BMC Bioinformatics. 2019;20:254. doi:10.1186/s12859-019-2877-3.
23. **Ashton** JJ, Beattie RM. Personalised therapy for inflammatory bowel disease. The Lancet. 2019;393:1672–4.
24. **Ashton** JJ, Beattie RM. Screen time in children and adolescents: is there evidence to guide parents and policy? The Lancet Child and Adolescent Health. 2019;3:292–4. doi:10.1016/S2352-4642(19)30062-8.

25. **Ashton** JJ, Borca F, Mossotto E, Coelho T, Batra A, Afzal NA, et al. Increased prevalence of anti-TNF therapy in paediatric inflammatory bowel disease is associated with a decline in surgical resections during childhood. Aliment Pharmacol Ther. 2019.
26. **Ashton** JJ, Borca F, Mossotto E, Phan HTT, Ennis S, Beattie RM. Analysis and Hierarchical Clustering of Blood Results Before Diagnosis in Pediatric Inflammatory Bowel Disease. Inflamm Bowel Dis •. 2018;XX. doi:10.1093/ibd/izy369.
27. **Ashton** JJ, Mossotto E, Ennis S, Beattie RM. Personalising medicine in inflammatory bowel disease—current and future perspectives. Transl Pediatr. 2019;8:56. doi:10.21037/TP.2018.12.03.
28. **Ashton** JJ, Batra A, Beattie RM. Paediatric inflammatory bowel disease- brief update on current practice. Paediatr Child Health (Oxford). 2018;0. doi:10.1016/j.paed.2018.08.007.
29. Gavin J, Marino L V, **Ashton** JJ, Beattie RM. Patient, parent and professional perception of the use of maintenance enteral nutrition in Paediatric Crohn's Disease. Acta Paediatr. 2018. doi:10.1111/apa.14571.
30. Andrews ET, **Ashton** JJ, Pearson F, Mark Beattie R, Johnson MJ. Early postnatal growth failure in preterm infants is not inevitable. Archives of Disease in Childhood: Fetal and Neonatal Edition. 2018.
31. **Ashton** JJ, Beattie RM. Gastro-oesphageal reflux in infants: what are we treating? Lancet Child Adolesc Heal. 2018.
32. **Ashton** JJ, Cullen M, Afzal NA, Coelho T, Batra A, Beattie RM. Is the incidence of paediatric inflammatory bowel disease still increasing? Arch Dis Child. 2018;:archdischild-2018-315038. doi:10.1136/archdischild-2018-315038.
33. **Ashton** JJ, Gavin J, Beattie RM. Exclusive enteral nutrition in Crohn's disease: Evidence and practicalities. Clin Nutr. 2018. doi:10.1016/j.clnu.2018.01.020.

34. Gavin J, **Ashton** JJ, Heather N, Marino LV, Beattie RM. Nutritional support in paediatric Crohn's disease: outcome at 12 months. Acta Paediatr Int J Paediatr. 2018;107.
35. **Ashton** JJ, Beattie RM. Faecal Calprotectin; What Does this Mean for the Paediatric Inflammatory Bowel Disease Phenotype? J Pediatr Gastroenterol Nutr. 2017;:1. doi:10.1097/MPG.0000000000001847.
36. **Ashton** JJ, Harnden A, Beattie RM. Paediatric inflammatory bowel disease: improving early diagnosis. Arch Dis Child. 2017;:archdischild-2017-313955. doi:10.1136/archdischild-2017-313955.
37. **Ashton** JJ, Beattie RM. Improving remission rates in newly diagnosed paediatric ulcerative colitis. Lancet Gastroenterol Hepatol. 2017;2.
38. **Ashton** JJ, Ennis S, Beattie RM. Early-onset paediatric inflammatory bowel disease. Lancet Child Adolesc Heal. 2017;1.