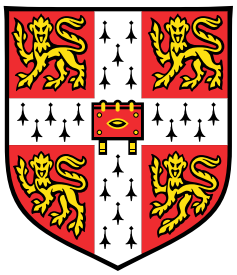
Molecular and clinical profiling of immune disease genetic loci



Omar El Garwany

Wellcome Sanger Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Churchill College

November 2023

I would like to dedicate this thesis to my loving parents and my wife... ...

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 60,000 words excluding tables, footnotes, bibliography, and appendices and has fewer than 150 figures.

Omar El Garwany
November 2023

Acknowledgements

My PhD was far from a solitary journey. I would like to acknowledge the direct and indirect contributions of the wonderful people I met and worked with over the last four years. My heartfelt gratitude goes to Dr. Carl Anderson, my PhD supervisor. His guidance, encouragement and thoughtful feedback over the past four years have made this intellectual journey truly life-changing.

I would also like to extend my gratitude to my friends and colleagues in the Anderson laboratory and in the Department of Human Genetics at the Sanger Institute. In particular, I am grateful to Dr. Nikolaos I Panousis whom I worked with on the MacroMap project and who gave me the opportunity to lead the alternative splicing arm of the project. Special thanks go out to Dr. Laura Fachal who often discussed with me many of the analyses described here and provided invaluable feedback. I had the privilige of working with other members of the laboratory whose expertise on different genetic analyses was extremely helpful: Dr. Aleksejs Sazonovs, Dr. Qian Zhang, Dr Marcus Tutert and every other member of the Anderson laboratory. I am indebted to every participant and patient who donated samples to the HipSci project, the UKB IBD Genetic Cosortium, the IBD BioResource and the UK BioBank.

Finally, I am incredibly grateful to my family, my ultimate motivation throughout this challenging journey. This journey would not have been possible without the love and support of my parents: Mohamed El Gawany and Samia El Saidy. My parents have been, and will always be, my ultimate mentors. Their love, sacrifice and wisdom have guided me through the years. I will always be grateful to my wife Sondos Khedr. Her unceasing support of me has kept me going and her impressive strength and resilience have inspired me at every difficult moment. I owe you so much. I was lucky to be supported by a group of amazing friends who were always there for me: Mohamed Yehia Kamel, Ahmed Abdelbaky, Samuel Gabra, Hani Saleh, Youmna Hussein, Samer Abdelmoeti, John Danial and Abdelhamid Yousef. My heartfelt gratitude goes out to each and every one of you.

Abstract

In the past two decades, genome-wide association studies (GWAS) have identified numerous genetic variants linked to various traits and diseases, including immune-mediated diseases (IMD). However, understanding the downstream effects of these genetic variations, both at the molecular and clinical levels, has proven more challenging than expected in the post-GWAS era. To address these challenges, my thesis focuses on characterizing the effects IMD-associated loci at the transcriptomic and disease sub-phenotype levels.

Many disease-associated variants are found in non-coding regions of the genome, making their functional interpretation elusive. Recent large-scale functional genomic datasets like GTEx and eQTLGen have linked genetic variation to gene expression differences, but most studies have primarily focused on steady-state gene expression at the tissue level, overlooking the impact of environmental factors on gene regulation in different cell types. Additionally, aspects of transcriptomic regulation such as alternative splicing have been understudied. In the first part of the thesis, I used iPSC-derived macrophages to investigate alternative splicing patterns and identify genetic variants regulating alternative splicing in different macrophage environmental contexts. The study found widespread differential splicing between stimulated and unstimulated macrophages, with context-dependent regulation and a link between IMD risk loci and alternative splicing changes.

The second part of the thesis adopts a clinical perspective, focusing on perianal Crohn's disease (pCD) as a case study. A GWAS meta-analysis between CD patients with and without pCD identified a significant genetic locus in the MHC region associated with pCD, previously linked to CD susceptibility. The study also investigated sporadic perianal manifestations in the general population, finding 12 significant genetic loci associated with sporadic perianal disease. An initial assessment shows that none of these loci replicate in the pCD meta-analysis, possibly suggesting distinct mechanisms driving both types of perianal manifestations. In summary, the thesis delves into the functional and clinical aspects of IMD-associated genetic variants, emphasizing the importance of alternative splicing and

exploring a high-burden clinical sub-phenotype, within both disease and sporadic contexts. This research encourages further exploration of these dimensions of IMD genetics.

Table of contents

List of figures

List of tables

Chapter 1

Getting started

1.1 Introduction

On the 30th of August 1958, famous statistician Ronald A. Fisher wrote a letter to the journal *Nature* critiquing the evidence linking smoking to lung cancer. The reasons he cited for his suspicions reflect a wider challenge in biomedical research. In his letter, RA Fisher mentioned that causality is difficult to establish from observational data that show increased rates of lung cancer among smokers. Since then, a huge body of literature established the causal link between smoking and lung cancer (reviewed in [1]), but the problems of causal inference in biology and public health remain alive [2]. Reproducible associations between observed exposures and outcomes have often not withstood more robust experimental designs such as randomised controlled trials. This is often attributed to several limitations of observational data, including confounding, reverse causation, and measurement errors. Confounding manifests as an observed association between an exposure and an outcome that results from a confounding factor that is associated with both. Reverse causation happens when the direction of effect between an exposure and an outcome is not clear.

Over the last 15 years, genome-wide association studies have provided thousands of associations between genetic variants and outcomes of interest. A significant difference between GWAS and epidemiological studies is that genoype-phenotype associations rarely have an ambiguous causal direction. Genetic variants are determined at conception and do not therefore suffer from reverse causation in the same way that other epidemiological exposures do [3]. GWASes have typically used a case-control study design to uncover genetic variants associated with different traits and diseases. As the sample sizes of these studies increased, it became apparent that a large number of genetic loci underpin most complex traits and diseases. These findings naturally posed several questions: which effector genes

are targeted by these risk-modulating genetic variants? Which biological pathways do they implicate? What can these findings tell us about disease pathogenesis? These questions are not uniquely relevant to genetics research. They are important from biological, clinical and drug development perspectives. However, genetics offers a unique angle to answer these questions by minimising the risk of associations driven by reverse causality, something that is difficult to guard against when researchers make conclusions about disease biology in *in vivo* and *in vitro* studies.

1.1.1 Two major gaps in the post-GWAS era

Trait and disease GWASes are often cited as an example of successful population-scale genetics endeavors. The majority of GWASes recruited tens of thousands of disease cases and controls to identify alleles enriched in disease cases compared to healthy controls. However, these associations were often difficult to interpret. First, these efforts have revealed that disease-associated genetic variants are significantly enriched in non-coding sequences such as enhancers and promoters [4–6]. Second, the majority of disease GWASes focussed on disease susceptibility, and the effects of discovered variants on different disease outcomes have remained largely unexplored.

The difficulty of interpreting GWAS results heralded several important "post-GWAS" approaches to understand the effects of genetic variation. The overall theme of these approaches is to bridge the wide gap between genetic variation and the end phenotypes under investigation. At the molecular end of this gap, molecular association studies have focussed on understanding the molecular effects of disease-associated variants. At the phenotype end, GWASes of different disease outcomes have aimed to dissect the genetic underpinnings of various disease sub-phenotypes as follow-up to broad disease susceptibility GWASes.

In this context, the aim of this thesis is to improve our understanding of the effects of genetic variation at these two levels. At the molecular level, a better understanding should improve our ability to understand the biological pathways affected by disease-associated genetic variation, and how these effects manifest in different tissues, cell types and biological contexts. At the disease sub-phenotype level, a better understanding of the genetic determinants of disease sub-phenotypes paves the way for a more nuanced understanding of clinical disease heterogeneity at the molecular level.

1.2 Part I: Understanding the non-coding genome: molecular quantitative trait loci studies

The majority of disease-associated variants are located in the non-coding genome [4–7]. This has made the interpretation of their downstream functional effects difficult. To help bridge the gap between the non-coding genome and function, population-level molecular studies that map genetic variation to variation in molecular traits have been set up (molecular quantitative trait loci or mQTLs). mQTLs reveal how genetic variation regulates different molecular traits, and in doing so can help us link genetically regulated molecular variation to disease risk.

In eukaryotic cells, biological functions are exerted as a complex coordinated program where cells produce effector molecules to exert various functions. These functions aim to sustain cell growth, enable cells to perform their functions or respond to external environmental cues. This process encompasses a wide range of molecular steps that lead to gene expression and end with translation to effector proteins and different post-translation modifications. Moreover, gene expression is regulated via several cis-acting regulatory elements such as promoters, enhancers and silencers. These regulatory elements are the target of epigenetic modifications such as differential chromatin accessibility and histone marker modifications that have been associated with gene expression regulation. The range of genetically regulated molecular traits is therefore wide and includes chromatin accessibility, methylation, gene expression, post-transcriptional modifications, protein levels and post-trasnlational protein modifications. Several studies have investigated the genetic determinants of methylation QTLs [8–13], chromatin accessibility QTLs [14, 15], expression QTLs [16–18], splicing QTLs [16, 19], and protein QTLs [20, 21] in a wide range of cell types and tissues. Although DNA provides a fixed blueprint for cellular function, different molecular aspects of cellular functions are highly dependent on the environmental context of each cell. Moreover, the genetic regulation of molecular traits has also been shown to vary between tissues, cell types and even environmental contexts [22, 23]. Profiling mQTLs in relevant contexts can therefore improve our ability to explain the functional effects of disease-associated variants [24].

Despite the large number of mQTLs, expression QTLs remain the most comprehensively characterised type of mQTLs. The rapid development of experimental and computational RNA-seq methods has accelerated the identification of eQTLs in large numbers of tissues and cell types. eQTLs have been succesfully used to identify effector genes for several complex diseases. For example, using pancreatic islet QTLs, Viñuela et al. robustly linked 22 Type 2 diabetes loci to effector genes [25]. Although eQTLs have been extensively catalogued in

many cell types and tissues, almost 50% of GWAS loci are still unexplained by eQTLs [26]. This gap is at least partly attributed to the lack of diversity of other mQTL types. Relative to eQTLs, fewer studies have comprehensively catalogued the several post-transcriptional steps that follow gene expression such as alternative splicing.

Alternative splicing (AS) is a widespread post-transcriptional modification, whereby intronic sequences are removed from transcribed mRNA and exonic sequences form mature mRNA transcripts. Since its discovery in the 1970s, our appreciation of the role of AS in eukaryotic gene expression has increased. Due to their limited scope, earlier transcriptomic profiling methods showed that 5-35% of human genes are alternatively spliced [27, 28]. However, over the last 15 years, high-throughput RNA-seq methods enabled a less biased and more comprehensive profiling of the human transcriptome. They showed that 90-95% of human genes undergo AS [29].

1.2.1 Alternative splicing in eukaryotes

AS is a complex combinatorial process where different combinations of exons can remarkably increase the coding potential of an otherwise fixed repertoire of genes. Different modes of AS include exon skipping, mutually exclusive exons, intron retention and alternative acceptor or donor splice sites. These modes enable the creation of diverse transcripts from the same DNA sequence. The complex process of splicing starts by the recognition of acceptor and donor splice sites, marked by GU and AG dinucleotides at the 5' and 3' ends of the exon-intron-exon splice junction. Splice site recognition is mediated by the spliceosomal complex, a complex of five small nuclear ribonucleoproteins (snRNPs) and 50-100 small peptides [30]. Two initial snRNPs bind to the acceptor and donor splice site and commit the splice junction to the splicing process (U1 and U2AF, respectively). Bridging interactions then bind these two snRNPs leading to the formation of a pre-spliceosomal complex. Further binding of snRNPs to the pre-spliceosomal complex marks the maturation of the spliceosomal complex, and leads to the release of the spliced intron (U4, U5, and U6).

AS is pervasive in most eukaryotic cells, but its evolutionary origin is subject to debate. The absence of AS in prokaryotes and ancient eukaryotes suggests that AS evolved at a late stage in eukaryogenesis [31]. Whenever its evolutionary origin may have been, AS seems to be a dynamic evolutionary process, where organisms gain novel introns over long evolutionary periods [32]. In support of this, intron gain seems to be a particularly expedient evolutionary process in aquatic species, where horizontal gene transfer is more common [33]. However, AS is still a very relevant layer of complexity in all species. A well-recognised

paradox in modern biology is that the total number of genes does not necessarily reflect organismal complexity. Several plant genomes have more genes than mammalian genomes, which arguably have more complex biology [34]. Conversely, the diversification of the transcriptome via AS seems to correlate with organismal complexity [35], reflecting the importance of AS in shaping complex physiological functions. In line with this, AS is more common in multicellular eukaryotes than unicellular eukaryotes, where genes have fewer and shorter introns [36].

Several physiological functions have been shown to be regulated by AS, including immune response, neuronal development, homeostasis, and sex determination. In most cases, a single gene produces several isoforms which have either distinct or complementary functions. The *Drosophila melanogaster* gene *DSCAM* is perhaps the most striking example of the pivotal role of AS in physiological processes. *DSCAM* is an cell surface immunoglobulin that plays an essential role in establishing neural circuits. By allowing neuronal self-avoidance and axon guidance and targeting [37], *DSCAM* ensures correct neuronal wiring in *Drosophilas*. The complex multi-exonic structure of *DSCAM* results in a total of 38,016 alternatively spliced protein isoforms. These cell surface receptor isoforms have poor self-affinity, which is important for self-avoidance and proper axonal guidance [38]. Sex determination in *Drosophilas* is another example, where sex-specific RNA binding proteins guide the expression of sex-specific transcripts [39]. It is clear that the detailed dissection of different gene isoforms in several model organisms has uncovered a crucial role of AS in core physiological processes.

1.2.2 Genetic regulation of alternative splicing

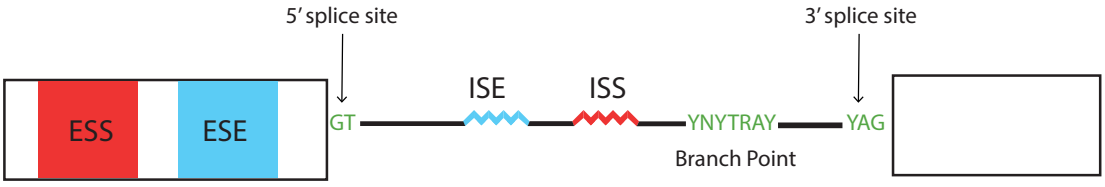


Fig. 1.1 Cis-acting splicing motifs in eukaryotes shown in an exon-intron-exon junction. The acceptor and donor dinucleotides are indicated at the 5' and 3' ends of the intron. ESE and ESS = exonic splicing enhancers and silencers. ISE and ISS = intron splicing enhancers and silencers. Y = pyrimidines.

AS is tightly regulated by several cis- and trans-acting factors. This tight regulation ensures splicing fidelity by correctly guiding the splicing machinery towards the target acceptor and donor splice sites, and by a complex interplay of splicing factors that promote and/or inhibit splicing (Figure 1.1). Despite the apparent complexity of the splicing code [40], direct mutagenesis as well as computational approaches have elucidated several cis-acting sequence elements that guide the choice of splice sites and improve spliceosomal efficiency. These include exonic splicing enhancers (ESE), exonic splicing silencer (ESS), intronic splicing enhancers (ISE) and intronic splicing silencers (ISS). Splicing regulatory elements mostly work by recruiting various classes of trans-acting splicing factors to their target splicing sites. These factors either promote or hinder the recruitment of the spliceosomal complex. Most ESEs are bound by members of the serine/arginine rich proteins (SR proteins), which enhance the recruitment of several snRNPs necessary to initiate the splicing process (reviewed in [41]). The promotion of splicing is often countered by the recruitment of heterogenous nuclear ribonucleoproteins (hnRNPs) to ESS, which often block the recruitment of the splicing machinery [42]. The disruption of this tight regulation underpins several diseases. Spinal muscular atrophy, a debilitating motor neuron disease, is caused by exon 7 skipping in *SMN1*. Exon 7 skipping is caused by a single nucleotide substitution that alters the ESE sequence and results in a non-functional *SMN1* protein isoform [43].

1.2.3 Cataloguing alternative splicing: progress and gaps

Recent efforts to catalogue the human transcriptome have shown remarkable diversity of alternative isoforms. For example, the Reference Sequence (RefSeq) project uses a multi-modal approach to identify a high-confidence set of splice variants for each gene for thousands of organisms including over 770 mammalian transcriptomes [44]. Manual curation by experts in addition to high-quality RNA-seq, proteomics, and histone marker datasets are used to build a bona fida set of gene splice variants. This effort has led to a 100-fold increase in the number of identified transcripts across mammalian species, from approximately 126,000 transcripts in 2003 to over 12 million transcripts in the latest RefSeq release (September 2023; [45]).

Despite these significant advances, our knowledge of the distribution and roles of these splice variants in different tissues and cell types remains heavily underexplored. The evidence supporting the tissue-specificity of AS is contradictory. Wang et al. estimated that between 55-83% of AS events vary between tissues in 15 studied human tissues and cell lines [46]. Others have shown that the majority of genes have a single dominant protein isoform in most

tissues [47, 48]. However, many of these studies suffer from either biased transcriptomic or proteomic profiling methods or a small number of tissues. Fewer studies have attempted to systematically catalogue splice variants in an unbiased manner. In comparison, overall levels of gene expression in diverse tissues are being extensively studied by collaborative initiatives such as the Human Cell Atlas [49]. Similar collaborative efforts that catalogue splice variants in an unbiased manner are warranted given the central role of AS in human health and disease.

1.2.4 Technological limitations

Several factors may explain why AS has received less attention compared to other transcriptional processes. The combinatorial nature of AS means that up to thousands of transcripts can be produced from the same genetic code. This poses several technological and analytical challenges. Most large-scale RNA-seq projects so far have relied on short-read sequencing to study the transcriptome. The complexity of AS patterns therefore makes it difficult to distinguish between distinct isoforms using 50-150 bp reads, as exonic sequences significantly overlap in alternative transcripts [50]. In principle, it is not possible to assign short reads to specific isoforms.

Creative technological and analytical techniques have been developed to assign short reads to their original transcript molecule. For example, Hagemann-Jensen et al. have recently applied a tagmentation strategy to map reads originating from the internal segments of gene bodies to UMI-tagged 5' reads. Using this technique, 30-50% of reconstructed molecules were successfully assigned to a specific isoform [51]. Additionally, computational techniques to reconstruct full isoforms from short reads have been developed. For example, Cufflinks relies on a reference transcriptome to estimate the most likely proportion of each splice variant given the observed RNA-seq reads [52]. Another method called rMATS estimates isoform proportions from the reads that support each type of AS event such as exon skipping and inclusion [53]. What these computational methods have in common is that they provide probabilistic estimates of isoform proportions, which underscores the inherent difficulty of obtaining a complete picture of isoform diversity from short-read RNA-seq experiments [53, 54]. These challenges explain why transcriptomic studies have focussed mostly on overall levels of gene expression, whose experimental and computational analysis workflow are more mature and suffer from less quantification uncertainty.

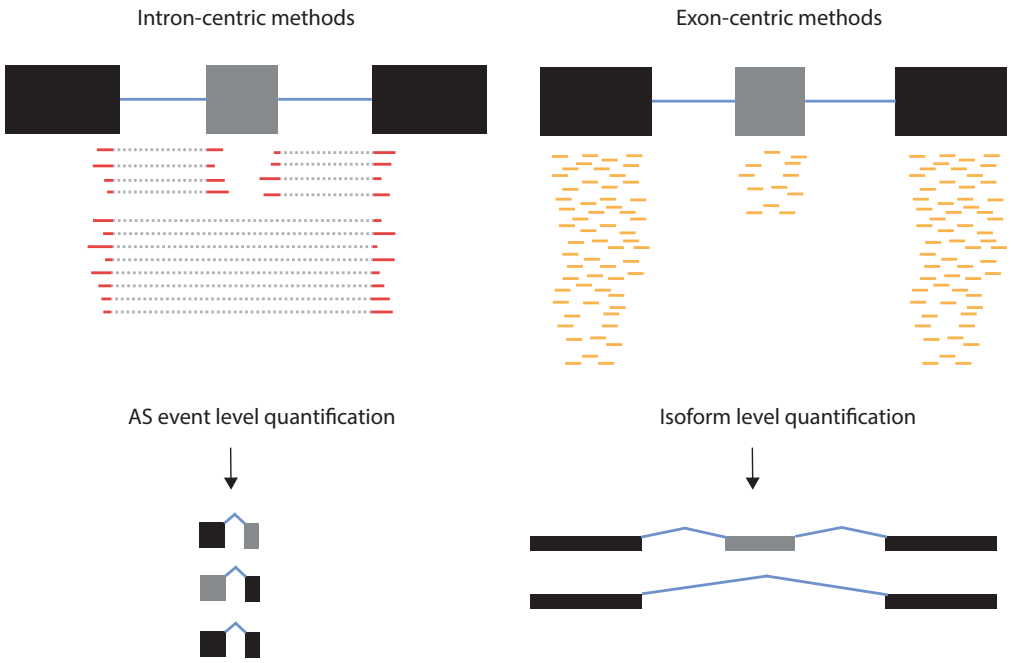


Fig. 1.2 Alternative splicing quantification methods are broadly classified into exon-centric methods and intron-centric methods. Exon-centric methods use exonic RNA-seq reads to estimate isoform abundance. Intron-centric methods such as Leafcutter quantify individual alternative splicing events.

1.2.5 Leafcutter as an intron-centric AS quantification method

AS quantification methods can be broadly divided into exon-centric and intron-centric methods. Exon-centric methods use exonic reads or a combination of exonic reads and reads that span two splice junctions (split reads) to infer isoform-level counts. These methods are heavily dependent on a known reference transcriptome, with some improvements to increase their ability to identify novel splice junctions [55]. Their underlying assumption is that the relative abundance of exonic reads reflects the proportions of the unobserved isoforms. Conversely, intron-centric methods are based on the principle that AS proceeds in a step-wise fashion, where introns are excised from pre-mRNA. Instead of quantifying AS using exonic reads, intron-centric methods use observed split reads at each splice junction to directly quantify local AS events (Figure 1.2). The obvious advantage of intron-centric methods is that they provide less uncertain estimates of AS events as they do not attempt to provide probabilistic isoform-level quantifications. Moreover, they are able to detect novel splice junctions as they do not rely on a reference transcriptome to reconstruct isoforms. However, this comes at the cost of precision and interpretability. By definition, split reads that span exon-intron-exon junctions are less abundant than exonic reads. Consequently, intron-centric quantification methods such as Leafcutter build their AS quantification using many fewer reads than exon-centric methods such as MAJIQ, rMATS, or Cufflinks. Moreover, the interpretation of local AS events is usually less straightforward. Local AS events reflect local intron excision steps at each exon-intron-exon junction, but it is often unclear how different intron excision events relate to one another.

Leafcutter is an example of intron-centric AS quantification methods that use split reads to quantify local intron excision decisions. In its first pass, Leafcutter starts by pooling all observed split reads in all samples to identify a set of high-confidence intron excision events. In a second pass, Leafcutter counts per-sample the number of split reads that map to each intron identified in the first pass. To improve interpretability, Leafcutter then organises individual intron excision events into undirected graph structures called *intron clusters*. Nodes represent local intron excision events which are connected by edges. The Leafcutter algorithm connects two nodes (i.e. introns) if they share a 5' or 3' splice site. The overall Leafcutter procedure results in functionally connected intron cluster where any two connected introns share an acceptor or donor splice site. Within each intron cluster, intron usage is then quantified as the proportion of all split reads that map to each individual intron. This final quantification is performed separately for each RNA-seq sample and the result is a matrix of intron usage ratio for all study samples.

1.2.6 Mapping sQTLs

Given the complex regulatory network that underpins AS regulation, understanding the impact of genetic variation on AS patterns paves the pathway to understand the impact of AS dysregulation on human health. Moreover, understanding how AS patterns are regulated in relevant contexts can help us better understand the impact of disease-associate genetic variant on the transcriptome. Similar to expression QTLs, where genetic variants associated with gene abundance are mapped, AS quantifications can be used as a molecular trait to uncover the genetic determinant of AS (splicing QTLs; Figure 1.3).

General outline of QTL mapping

QTL mapping pipelines are relatively well-established. Typically, a QTL analysis pipeline starts by obtaining an adequate number of samples where a quantitative molecular phenotype of interest is assayed (e.g. gene expression). Initial quality control steps are applied to ensure that experimental issues such as sample mixups are addressed. For transcriptomic studies, the first step after initial QC is to align NGS short reads to a reference genome. To extract quantitative molecular features from aligned reads, a quantification method is applied. The quantification method of choice usually depends on the research question of interest. For example, overall levels of gene expression are quantified using methods that count all short reads that map to each gene, and provide a gene count matrix. Similarly, methods that quantify AS provide an isoform-level or AS-event-level quantification. At this stage, another round of QC is often needed to ensure that low-quality features are removed from subsequent QTL mapping steps. Again, this QC step depends on the molecular QTL of interest. For example, it is important to remove introns detected in a small number of individuals, as tiny individual variations in intron usage can result in spurious sQTL associations.

With a post-QC feature matrix, QTL mapping follows a number of standard steps. The most important step before QTL mapping is to ensure that the molecular feature is properly normalised. Normalisation ensures that features conform to the assumptions of a linear regression model: homoskedasticity and normal distribution. These two assumptions are not only prerequisites of linear regression, but also ensure that effect sizes can be interpreted appropriately. First, heteroskedasticity occurs when the variance of the predicted variable (i.e. feature) is not equal for different values of the independent variable (i.e. different genotypes). Quantile normalisation is one of the most widely used approaches to ensure that a molecular feature has equal variance across all samples in a study, satisfying the

homoskedasticity condition. Second, an inverse normal transformation is applied to each sample to ensure that the feature is normally distributed.

Each molecular QTL can be tested for association with genetic variants in *cis* or in *trans*. Typically, cis-QTL mapping tests the association between a molecular feature and all nearby variants (e.g. within a 1 mbp window), while trans-QTL mapping tests the association between a molecular feature and distant genetic variants (e.g. > 5mbp or on other chromosomes). Cis-QTL mapping is more common as it requires less statistical power to detect an association, owing to the much smaller set of tested variants. For each molecular feature, thousands of genetic variants are usually tested. Compared to GWASes where all variants are tested genome-wide, the number of tests performed in QTL mapping is highly dependent on each individual feature. Setting a significance threshold therefore requires a different approach to a traditional GWAS significance threshold. A common approach to correct for multiple testing is to perform a permutation test between genotypes and features. The genotype-feature mapping is permuted hundreds or even thousands of times and association tests are performed again, resulting in a null distribution of association statistics. The real association statistic is then compared to the null distribution to obtain an adjusted association statistic. This layer of multiple testing correction accounts for the thousands of variants tested for each molecular feature. Another layer of multiple testing correction is applied to account for the thousands of molecular features tested in the QTL study.

Special considerations in sQTL mapping

Although the steps outlined above are standard for all QTL studies, there are a few conceptual and methodological differences between splice and expression QTL mapping. Depending on the AS quantification method, the interpretation of sQTLs can vary. sQTLs discovered using isoform abundance as a molecular trait are the easiest to interpret. A significant isoform-level sQTL would be defined as a genetic variant that increases or decreases a particular transcript abundance. This interpretation is less straightforward when AS is quantified at the AS event level. When intron usage ratios are used as quantitative trait, a significant sQTL can be defined as a variant that changes the proportion of a particular intron within its intron cluster. Therefore, when sQTLs are mapped using intron usage ratios, it is often helpful to examine the effect of the discovered genetic variant on all neighbouring AS events to build a more complete picture of the splicing event under investigation. For example, in Figure 1.2, upon examination of all three AS events in the left-hand panel, it becomes clear that the identified AS events represents an exon inclusion/skipping event. Additionally, it is important to note

that different AS events are often highly correlated. This is because intron usage ratios within an intron cluster always add up to 1. Therefore, a genetic variant that leads to increased usage of one intron also leads to decreased usage of one or more other introns. As a result, multiple significant sQTLs within a single intron cluster do not necessarily represent distinct regulatory effects, but rather highly correlated measurements.

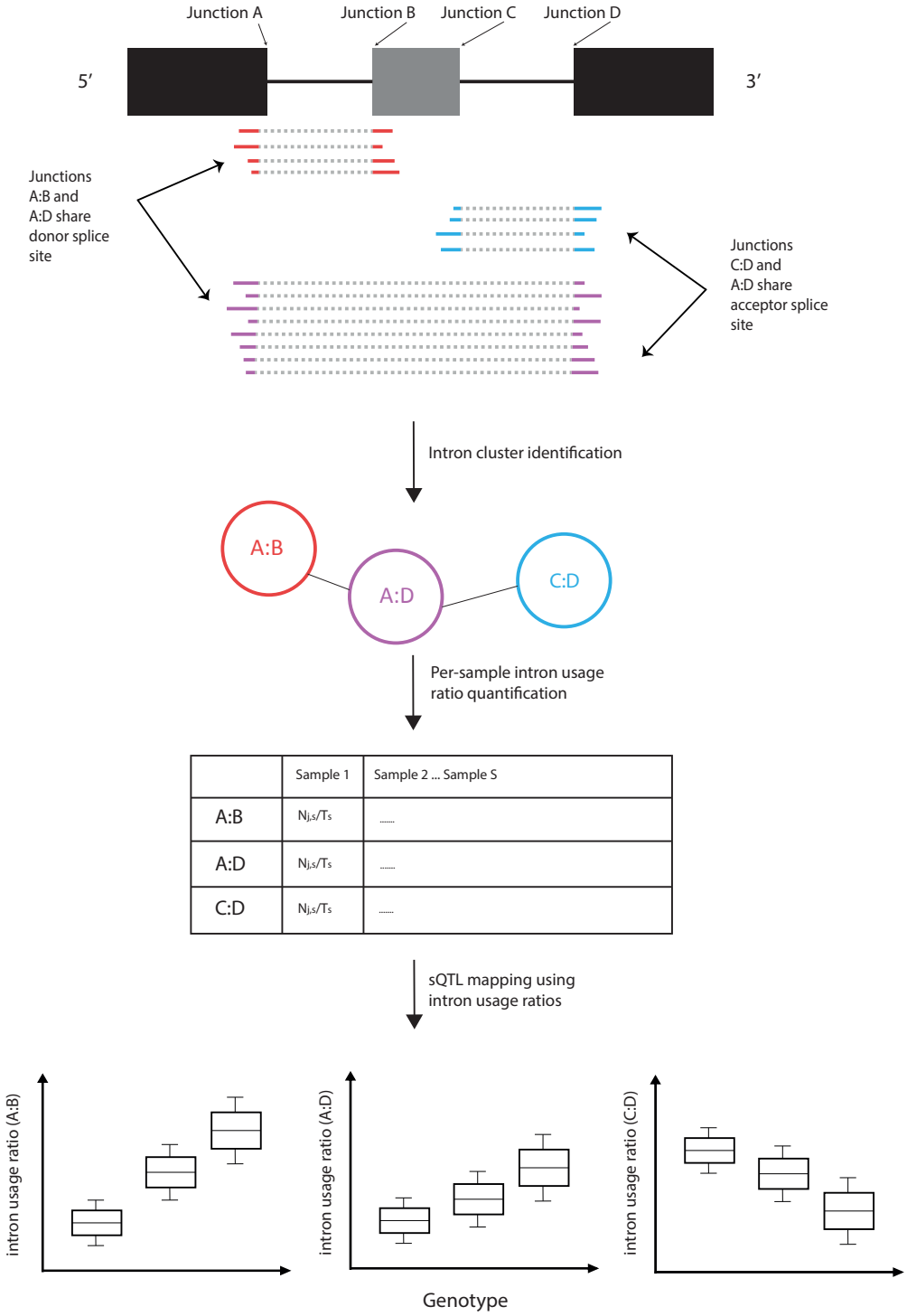


Fig. 1.3 Conceptual outline of sQTL mapping using intron usage ratios as quantitative traits. Intron clusters are identified from all pooled samples in a study. Quantification is then performed for each intron per sample. For a sample S, and an intron cluster with total number of reads T_S, the intron usage ratio for intron j is defined as $\frac{N_{j,S}}{T_S}$. Intron usage ratios are then used as quantitative trait to map sQTLs in *cis* with neighbouring genetic variants.

1.2.7 Comparing QTL effects in multiple contexts

A long-standing question in QTL studies is how gene expression is genetically regulated in different tissues, cell types and environmental contexts. Answering this question is important to understand which transcriptomic effects of genetic variation are shared or distinct in different biological contexts. Context-dependence of QTL effects has often motivated multi-tissue QTL studies, with the assumption that profiling QTL effects in different contexts can draw a more complete picture of gene expression regulation. For example, in a comparison of eQTL effects between CD4+ T-cells and monocytes, Raj et al. [56] found that at least 42 genes had opposing eQTL effects in the two cell types. In line with this, Peters et al. [57] found that 87 genes had discordant eQTL effects in five different immune cell types. Although these dramatically discordant examples of genetic regulation represent a minority of QTL effects, the question of which QTL effects are modulated in a more subtle manner in different contexts remains relevant [58–62].

Assessing the sharing of QTL effects in different treatment groups is non-trivial. In most QTL studies, QTL discovery is carried out separately for different treatment groups. This means that incomplete power may cause truly shared QTL effects to appear non-significant in some groups simply by chance. Direct comparison of statistical significance between different groups is therefore likely to overestimate the number of distinct QTL effects. To address this issue, several methods that probabilistically model effect sizes were developed [63–66]. Earlier methods were inspired by fixed-effects meta-analysis methods, and started from the assumption that any given eQTL effect is shared across all conditions and sought to find statistical evidence to the contrary (i.e. context-specificity).

Later, methods that learn the data-driven correlation structure were developed. For example, multivariate adaptive shrinkage (mash) empirically learns the patterns of effect sharing in the dataset under study, and allows for arbitrary patterns of sharing between different groups. For example, QTLs derived from different brain regions are expected to have highly correlated effect sizes. Usually, this correlation structure is learned from a random unbiased set of QTL effects (i.e. non-significant QTLs). A Bayesian approach is then applied to re-estimate effect sizes for a desired set of QTL effects (e.g. significant QTL effects). The posterior effect sizes are then tested for evidence of effect size heterogeneity between different groups, taking the underlying data-driven correlation structure into account. The obvious advantage of mash is that the re-estimated effect sizes take into account the empirical correlation structure in the dataset. However, this also means that significant QTL effects' sharing may be overestimated when the null QTL effects are highly correlated among the treatment groups. As a result, this

may hide truly context-dependent QTL effects simply because there is not enough statistical evidence to suggest heterogeneity of effect sizes. Additionally, when significant QTL effects are tested for condition-specificity, only the lead QTL SNP is used. In many cases, sharing of the lead QTL SNP does not necessarily mean that the underlying causal variant is shared between different conditions. It has been previously shown that comparing the lead SNP between different association signals can lead to the false conclusion that the effects under comparison are shared [67]. A better approach should leverage the linkage disequilibrium structure to assess if two association signals under comparison are likely to be shared or distinct. Nonetheless, mash can still be useful if the degree of QTL sharing is interpreted as an upper bound, rather than an accurate estimate of QTL sharing.

1.2.8 Linking disease-associated GWAS loci to QTLs

In addition to understanding gene regulation, a major objective of QTL studies is to dissect the effects of disease-associated loci in relevant tissues and cell types. The simplest approach is to test the replication of the lead GWAS SNP in the QTL dataset. It is often compelling to assume that a replicated SNP may indicate that both gene expression and disease risk are driven by the same variant. In fact, lead SNP comparison was commonly used to implicate effector genes at many disease-associated loci. However, this direct SNP comparison was found to result in many false positives [67]. Therefore, more robust methods to compare pairs of association signals were developed to fill this gap. Particularly, statistical colocalisation methods take into account the association signal of all variants in a region to make a conclusion about a pair of association signals. Although the true causal variant may not be genotyped or imputed in either of the association studies, its effect is tagged by other variants in linkage disequilibrium with the true causal variant. Colocalisation methods leverages the linkage disequilibrium in a given locus to make an inference about two association signals. The underlying assumption is that if the two association signals are consistent across the region, it is likely that the same variant is driving both signals. Therefore, colocalisation results are only valid when the LD pattern is similar between the two association signals under comparison. This assumption only holds if the two association signals being compared are derived from populations with matching LD structures, an important consideration when colocalisng signals from GWAS and QTL studies. Additionally, standard colocalisation approaches only test the hypothesis that a *single* shared variant underpins the two association signals. Many QTL studies have shown that secondary and even tertiary association signals have been discovered for several genes [68, 69], and the same observation applies to GWAS signals. Violations of the single causal variant assumption at loci with multiple causal

variants will result in decreased power to detect true colocalisations. Therefore, extensions to standard colocalisation identify independent association signal in each of the two cohorts, before proceeding to perform colocalisation analysis for each of the identified signals. This approach has been shown to increase the number of colocalised signals detected [70].

1.2.9 Current status of sQTL studies

Several studies have investigated the genetic regulation of alternative splicing. These studies include relatively fewer multi-tissue or multi-cell-type studies as well as single-tissue or single-cell-type studies. Notably, virtually all of these studies are bulk RNA-seq studies because the most widely used single cell RNA-seq methods capture only a few bps at the start or end of each RNA molecule. The imminent development of single-cell long-read RNA-seq methods is set to change this by allowing full-length RNA isoforms to be directly mapped at a single-cell level [71]. Still, bulk sQTL studies have taught us a few interesting insights. In this section, I will focus on two relevant aspects of the genetic regulation of AS: context-specificity and role of AS in complex disease risk.

The most diverse sQTL study to date is the GTEx consortium, where eQTLs and sQTLs were mapped across 49 human tissues from 70-700 individuals. GTEx tissues represent a heterogenous mix of different cell types, which makes an investigation of cell-type-specificity of alternative splicing difficult. However, GTEx was still useful understand the patterns of alternative splicing sharing across tissues. It was found that tissue sharing follow a U-shaped distribution, meaning that the majority of sQTL effects were either found in a small number of tissues (1-5) or shared across almost all tissues (45-49 tissues). This pattern was less evident for eQTL effects, which suggested that sQTLs are generally more tissue-specific than eQTLs [16]. However, it is difficult to disentangle the effect of different statistical power in different tissues from true tissue specificity. Indeed, a subsequent re-analysis of the GTEx splicing data (Garrido-Martín et al. 2021 [72]) found that sQTL effects with large effect sizes, which are more readily discovered with small sample sizes, tend to be shared between tissues, and vice versa. When Garrido-Martín et corrected for sample size differences between various tissues, they found that brain, testis, liver and skeletal muscles had large numbers of tissue-specific sQTLs. These observations show that sQTL effects, especially one with smaller effect sizes, show patterns of tissue-specificity.

sQTLs were also shown to contribute significantly to complex disease risk in two major meta-analysis of brain and immune cell sQTL studies. Qi et al. 2022 [19] mapped sQTLs

from over 2800 RNA-seq samples, and discovered thousands of significant sQTLs and eQTLs. The availability of both eQTLs and sQTLs allowed them to compare their contribution to complex disease risk. To this end, they quantified the contribution of both sQTLs and eQTLs to the SNP heritability of 12 brain-related complex diseases, and showed that both molecular QTLs contributed roughly equally to disease heritability. Moreover, when they performed colocalisation with disease-associated loci, they found that both molecular QTLs implicated distinct gene sets. This finding was also mirrored in another eQTL and sQTL meta-analysis of 18 different immune cell types (Mu et al. 2021 [23]). Mu et al. compared the contribution of eQTLs sQTLs in different cell types to immune disease risk. Similar Qi et al., they also found a significant contribution of sQTLs, as measured by the number of loci that colocalised with high confidence with disease-associated loci. For many complex traits, this contribution often exceeded the contributionof eQTLs.

1.2.10 Contribution of this thesis

In the first part of my thesis, I will use iPSC-derived macrophages (MacroMap) to address the previously mentioned two aspects of alternative splicing regulation: the context specificity of AS regulation, and the contribution of AS to immune-mediated disease risk. Although the GTEx sQTL results indicate that AS regulation may be tissue specific, it does not provide an account of which genes are alternatively spliced in a cell-type or context-specific manner. This is mainly because GTEx tissues are a mixture of several cell types. Moreover, even fewer studies addressed the question of whether cells respond to environmental stimuli at the AS level. For example, do cells respond to pathogens by altering isoforms in genes relevant for microbial clearance? What role does the genetic regulation of AS play in such a response?

As part of the MacroMap project, iPSC-derived macrophages from over 200 individuals were exposed to a wide range of stimuli. I will describe how I used genotype and RNA-seq data to study the patterns of alternative splicing in different stimulation conditions, and to identify genetic variants that regulate alternative splicing in different macrophage environmental contexts. I will describe examples of the widespread differential splicing between stimulated and unstimulated macrophages, often implicating core innate immune response pathways. Additionally, I will show that alternative splicing regulation is often context-dependent, and that a considerable proportion of sQTL effects are modulated upon exposure to environmental stimuli. Finally, I will describe how I linked hundreds of immune disease loci to alternative splicing changes in macrophages and give examples of alternative splicing dysregylation by disease-associated risk loci. An important insight of this work

is that in the context of IMD risk, the dysregulation of alternative splicing is at least as important as the dysregulation of overall levels of gene expression. Finally, I will highlight the potentially important role of lowly-used splice junctions in immune disease risk, a role that has been, so far, under-appreciated.

1.3 Part II: sub-phenotype GWAS

1.3.1 Genome-wide association studies

Complex disease risk is determined by a multitude of genetic and environmental factors. Over the last 18 years, genome-wide association studies (GWAS) have revolutionised our understanding of the genetic component of complex disease risk. The Wellcome Trust Case Control Consortium (WTCCC) ushered the era of GWAS studies by designing large-scale case-control cohorts for several common disorders. Since then, the case-control experimental design has been exploited in thousands of GWAS studies to uncover the genetic determinants of cardiovascular, metabolic, immune-mediated, musculoskeletal, neurological, and gastrointestinal diseases. In most cases, these cohorts are built through collaborative efforts between recruitment centres, hospitals and other healthcare facilities and research centers that identify disease cases and controls, and provide biological samples needed to conduct array-based or exome-based genotyping. The continuous growth of sample sizes has increased our ability to detect genome-wide significant loci associated with disease risk. These efforts have also revealed the extensively polygenic nature of most complex diseases, whereby hundreds of genetic loci increase or decrease disease risk with small effect sizes. The complexity of the genetic architecture of most common disease has initially made it more challenging to draw biological insights. Although most GWAS results were initially puzzling, over the last few years massive GWASes have revealed biological insights about common diseases [73, 74]. This increased understanding was facilitated by the availability of functional genomic datasets as well as methodological advances in linking genetic variants to biological functions.

1.3.2 Inflammatory bowel disease

Epidemiology and classification

Inflammatory bowel disease (IBD) encompasses a group of immune-mediated disorders of the gut. IBD poses a considerable burden for healthcare systems globally. In 2017, IBD affected over 6.8 million individuals worldwide, with a rising global burden since at least

the 1990s. IBD incidence shows notable geographical variation, with the highest incidence reported in North America, the UK and northern Europe. Moreover, IBD incidence has notably risen in countries that are becoming increasingly "westernised" in terms of their environmental risk factors, such as China and South Korea, consistent with a significant environmental contribution to IBD [75].

IBD is broadly classified into two broad categories based on radiological, clinical and endoscopic features: ulcerative colitis (UC) and Crohn's disease (CD), although 6-13% are classified as IBD unclassified (IBDU) [76]. The two classes show differences in terms of disease behaviour and location, clinical manifestations and prognosis. UC inflammation occurs in a continuous manner and is often characterised by chronic mucosal inflammation and leukocyte infiltration [77]. UC usually starts near the rectum and diffuses proximally to different parts of the colon. On the other hand, CD can affect any part of the GI tract from mouth to anus and is characterised by patches of inflammation (skip areas). Inflammation often extends beyond the gut mucosa involving the submucosa. CD most frequently occurs in the ileo-coecal region followed by isolated terminal ileal inflammation. Clinically, CD is a heterogenous disease characterised by a remitting-relapsing clinical picture. Most patients experience abdominal pain, rectal bleeding, and altered bowel habits. However, other clinical aspects of CD vary between patients and can often make the difference between favourable or unfavourable disease course and prognosis. Some CD patients experience relatively infrequent CD flares, with milder symptoms that respond well to treatment. Others experience more frequent episodes of severe GI symptoms. Severe CD patients also often develop transmural manifestations such as penetrating disease, fistulas and abscess as well as extraintestinal manifestation involving the eye, joints and/or other systemic manifestations. Although the majority of CD patients undergo surgery at least once over their lifetime, patients who have non-penetrating non-fistulising CD manifestations are less likely to require surgery [78].

Risk factors of IBD

IBD is a complex disease, which is likely caused by an interaction of genetic, environmental, and lifestyle factors. IBD has often been described as an "industrialised nations" diseases, with higher prevalence in developed countries. Epidemiological studies have shown increasing prevalence of IBD in nations that are becoming increasingly industrialised. Interestingly, second-generation immigrants from low-prevalence countries have experienced increasing incidence of IBD [79]. These observations have linked IBD risk to "industrialised" lifestyle factors, whereby environmental and lifestyle factors common in industrialised countries are

thought to contribute to IBD risk. These changes have led to reduced exposure to infectious agents, improved hygiene and sanitation, an increasingly sedentary lifestyle and increased consumption of processed foods, and foods rich with sugar and saturated fats.

Smoking is the best described lifestyle factor linked to IBD risk. Smoking has been shown to increase risk of CD and decreasing risk of UC. However, the mechanism of this paradoxical association between smoking and IBD is not completely understood [80]. Other non-dietary factors include oral contraceptive pill intake, which was shown to increase both CD and UC risk [81], and appendectomy which was associated with reduced UC risk [82].

The effect of lifestyle choices and diet on IBD have been extensively studied, but the results are often difficult to assess. Exercise is known to boost immunity and decrease proinflammatory cytokines. However, the severity of IBD symptoms often impacts patients' physical activity, and studies linking exercise to IBD progression have been therefore confounded by IBD severity [83]. Similarly, alcohol and coffee consumption were not conclusively linked to IBD development or progression [84]. However, obesity has been shown to independently worsen IBD behaviour and increase likelihood of relapse [85]. Diet composition also plays an important role in IBD risk. Its role has been attributed to the effect of diet on the gut microbiota composition and bahaviour. For example, a Japanese study has shown a significant association between IBD risk and total fat and unsaturated fat intake, fish and shellfish consumption, and ω-3 and ω-6 fatty acids [86].

1.3.3 Inflammatory bowel disease genetics

The genetic component of IBD has been recognised for over 70 years via family studies on twins. Family studies have shown that monozygotic twins are more likely to co-inherit IBD compared to dizygotic twins, often with similar disease behaviour and location [87]. Over the last decade, several GWASes have identified over 240 loci associated with IBD susceptibility [88–91]. The largest IBD GWAS studies have focussed on discovering both common and rare genetic variants associated with IBD susceptibility. These studies have revealed several key mechanistic insights regarding the pathogenesis of IBD including autophagy, host-microbe interactions, intestinal innate immune response, and impaired epithelial barrier function [92, 88]. These pathways seem to converge on a IBD pathogenesis model whereby impaired intestinal permeability, leads to microbial infiltration into the gut mucosa. This microbial incursion activates intra-epithelial cells to initiate a cascade of innate and adaptive immune responses aiming to limit microbial spreading and restore normal barrier function.

The integration of hundreds of genetic loci with functional genomic datasets have clearly improved our understanding of IBD susceptibility. However, given the clinical heterogeneity of IBD sub-types, understanding the genetic determinants of their different clinical aspects is crucial for a more nuanced biological insight into what drives disease course.

Growing interest in the genetic determinantsof disease sub-phenotypes

The evident success of GWAS studies in improving our understanding of CD biology has sparked the interest in using them to dissect complex disease sub-phenotypes. However, disease sub-phenotype GWASes have generally lagged behind susceptibility GWASes, due to the difficulty of obtaining deep phenotypic or longitudinal data for similarly large cohorts. It has been previously suggested that the same genetic variants drive both disease susceptibility and disease sub-phenotypes. However, evidence in relatively smaller sub-phenotype GWASes suggests that the genetic variants that underpin disease susceptibility and disease sub-phenotypes may also be distinct [93, 94]. Both paradigms raise interesting questions about the genetic architecture of disease susceptibility and sub-phenotypes. Under the former paradigm, it will be particularly important to understand the relationship between the effect of each variant on susceptibility and sub-phenotype risks. For example, for a given variant associated with both disease sub-phenotype and susceptibility, is the susceptibility risk truly driven by sub-phenotype risk? As sub-phenotype GWASes become more commonplace, it will be particularly interesting to compare the effects sizes of each variant on both susceptibility and disease sub-phenotypes. This may lead to better stratification of disease risk based on distinct sub-phenotype risk profiles. Under the latter paradigm, it is important to understand which distinct biological pathways are involved in disease sub-phenotypes risk and how they interact with disease susceptibility pathways. For example, fistulising CD has been hypothesised to originate as an epithelial-to-mesenchymal transformation, whereby stationary epithelial cells gain migratory features. Is this transformation accelerated by the impaired intestinal barrier and subsequent immune activation that likely underpins CD susceptibility?

The genetic architecture of CD sub-phenotypes has been explored in a number of studies. Longitudinal and deep phenotypic data from CD cohorts of thousands of patients were used alongside genetic data to map genetic variants associated with disease location, behaviour, surgery and prognosis. Because these studies are considerably less powered than susceptibility GWASes, they have only given an initial glimpse into the genetic architecture of CD sub-phenotypes. Therefore, it is still not possible to draw a conclusive answer to the question

whether or not CD susceptibility and sub-phenotypes share genetic underpinnings. In the rest of this section, I will focus on two studies by Cleynen et al. 2016 [95] and Lee et al. 2017 [96] as notable examples of CD sub-phenotype GWASes.

Cleynen et al. [95] used 11 years of longitudinal data from approximately 17,000 CD patients to study the genetic determinants of CD sub-phenotypes. They performed several within-case GWASes of CD age-at-diagnosis, location, behaviour and surgery. Across all their CD analyses, they found three genome-wide significant loci at *MST1* and *NOD2* as well as an MHC association. Both the MHC and *NOD2* loci were associated with ileal versus colonic disease and with a penetrating (B3) versus inflammatory or stricturing disease (B1 or B2). Despite the small number of genome-wide significant loci, the authors built IBD polygenic risk score to compare individuals with different disease locations and behaviours. The main insight from their work was the rejection of the binary classification of IBD into CD and UC. Based on a polygenic risk score (PRS) comparison between individuals with different disease locations, they proposed a novel classification which places ileo-colonic CD as an intermediate form between ileal CD and UC.

Based on the discovered loci, the genetic relationship between CD susceptibility and CD sub-phenotypes is difficult to assess. For example, the *NOD2* lead variant (rs2066847) is a frameshift variant that has been previously associated with CD susceptibility in several studies [97, 88] (P-value=6×10^{-209} in Jostins et al. data). However, the lead MHC variant (rs6930777) was not genome-wide significant in any CD susceptibility GWAS (source: GWAS catalogue accessed in November 2023).

Comparatively, Cleynen et al. had a much smaller sample size than many CD susceptibility GWASes. The growth of sub-phenotype GWASes' sample sizes will enable a more systematic comparison of genome-wide significant associations of CD susceptibility and sub-phenotype variants. However, their PRS comparison may give a clue that may contribute to answering this question. The PRSs they built to compare different IBD forms were based on lead variants from all known IBD susceptibility loci. The ability of these PRSs to distinguish "macroscopic" features of IBD (e.g. CD versus UC or ileal versus colonic CD) lends support to the hypothesis that the genetic architectures of susceptibility and location are not entirely distinct. It is worth noting that the differences in PRS distribution between these groupings were quite small, but this could also be attributed to the low predictive power of PRSs in general.

A later study by Lee et al. [96] showed that the genetic determinants of CD prognosis were largely distinct from CD susceptibility variants. Lee et al. performed a within-case GWAS between two CD subpopulations at opposite ends of the prognosis spectrum. Poor prognosis individuals were defined as patients who experienced frequent CD flares that did not respond to treatment with immunomodulators and/or surgery. Good prognosis patients were defined as CD patients who showed good long-term response to treatment. Their analysis found four genome-wide significant loci in *FOXO3* and *XACT*, an intergenic locus in *IGFBP1-IGFBP3*, as well as an MHC association. These association were not found to be associated with CD susceptibility in any previous CD GWAS studies. Additionally, the authors found no evidence of genome-wide genetic correlation between CD susceptibility and CD prognosis, although it has to be noted that CD prognosis GWAS may have been underpowered to detect a significant genetic correlation ($r_g = -0.51$; P-value=0.12). Unlike Cleynen et al., these lines of evidence support the alternate hypothesis that the genetic architectures of CD susceptibility and CD prognosis, an example of a CD sub-phenotype, are distinct. Interestingly, a recent GWAS of multiple sclerosis progression reported a similar observation. The authors compared the long-term cognitive outcomes of MS patients with the highest and lowest MS PRSs, and found that PRSs did not differentiate between individuals with poor versus good MS prognosis [98]. Similarly, a GWAS of epilepsy prognosis concluded that it is unlikely that epilepsy susceptibility variants affect epilepsy prognosis [99].

At this stage of CD sub-phenotype GWASes, it is perhaps too early to answer the question conclusively, since very little is known about the biological mechanisms behind each CD sub-phenotype. This is set to change as large-scale IBD resources grow and enable well-powered sub-phenotype GWASes that give us a better understanding of the biology behind each CD sub-phenotype.

1.3.4 Contribution of this thesis to sub-phenotype understanding

In my thesis, I attempt to identify the genetic variants associated with perianal CD, as an example of a CD sub-phenotype. pCD is a severe CD sub-phenotype characterised by abscess and fistula formation around the anal region. Collaborative efforts have enabled detailed clinical phenotyping of thousands of CD patients, including information on perianal symptoms. I describe a GWAS meta-analysis between CD patients with and without pCD in two IBD patient cohorts (UK IBD Genetics Consortium and IBD BioResource). As a follow-up, I also studied the genetics of sporadic perianal manifestations in the general population. Although perianal symptoms are highly enriched among CD patients, perianal fistulising

disease also occurs sporadically, often not accompanied by CD. To understand the genetics of the two types of perianal manifestations, I performed an additional GWAS meta-analysis between individuals who report sporadic perianal disease and healthy controls in the UK Biobank and FinnGen. This meta-analysis revealed several genome-wide significant loci associated with sporadic perianal manifestations. My initial assessment shows that none of these loci replicate in the pCD meta-analysis, possibly suggesting distinct mechanisms driving both types of perianal manifestations. Overall, the aim of chapters 3 and 4 of my thesis is to study perianal manifestations as an example of a CD sub-phenotype, both in the context of CD and in its sporadic forms in the general population.

Chapter 2

Genetic regulation of alternative splicing in iPSC-derived macrophages

2.1 Introduction

Genome wide association studies (GWAS) have uncovered thousands of genetic loci associated with susceptibility to immune-mediated diseases (IMD). Over 90% of these loci are located in non-coding regions of the genome [**?**], making it difficult to gain insights into causal disease biology. These non-coding disease-associated loci are enriched in gene regulatory regions, and are therefore thought to modulate gene expression [**?**]. Expression quantitative trait loci (eQTL) mapping has been widely used to characterise the downstream effects of genetic variants on gene expression [17, 16]. Despite the increasing number of available eQTL datasets [18], IMD-associated loci have remained largely unexplained by existing QTL maps. For example, Chun et al. 2017 (ref: [**?**]) found that only 25% of IMD-associated loci colocalised with eQTLs from three immune cell types.

Multiple explanations have been put forward to justify the incomplete overlap between GWAS loci and existing eQTL maps, including a need for more diverse molecular QTL maps across disease-relevant cell types and environmental conditions. Moreover, it has recently been suggested that common variants driving complex diseases and gene expression are systematically different, and that alternative molecular QTLs (such as those affecting splicing, chromatin accessibility and chromatin interactions) may be more likely to colocalise with disease associated loci [**?**]. Unfortunately, most QTL mapping studies have focussed on associating genetic variation with overall levels of gene expression without considering, for example, variation in transcript isoforms. The few studies that have mapped genetic variants

associated with alternative splicing (splicing quantitative trait loci or sQTLs) have shown their promise for understanding disease [**? ?**]. There is thus an urgent need for sQTL maps to be constructed across a broad range of environmental contexts for disease relevant cell types.

In this chapter, I mapped sQTLs in iPSC-derived macrophages in 12 different cellular conditions obtained from 209 individuals at two timepoints after stimulation. I quantify the extent to which alternative splicing responds to macrophage stimulation, and how sQTLs and response sQTLs (re-sQTLs) are shared across environmental contexts. I also explore the contribution of alternative splicing and sQTLs to IMD risk. Finally, I contextualise these findings within an ongoing scientific debate about the functional and evolutionary relevance of low-usage splicing events and discuss the implications of this work on the design of future transcriptomics studies.

2.2 Methods

2.2.1 Differentiation of Induced Pluripotent Stem Cell Lines (iPSC) Into Macrophages

iPSCs obtained from healthy donors of European descent were selected from the HipSci Consortium [**?**], and were differentiated to macrophages using a previously published protocol [14]. Of 315 lines initially selected, 227 (71.6%) were successfully differentiated. RNA-seq libraries were produced for 217, which represented 209 lines after quality control. Differentiated macrophages have been shown to be transcriptionally similar to monocyte-derived macrophages [14]. The differentiated naive macrophages were then stimulated with a diverse panel of adjuvants, resulting in 10 stimulation conditions, as well as naive differentiated (Ctrl_6 and Ctrl_24) and undifferentiated controls (Prec_D0 and Prec_D2). mRNA was harvested 6 and 24 hours after stimulation. Multiple cell lines were derived from a varying number of individuals in each stimulation condition (ranging between 177-202 individuals), resulting in a total of 4,698 unique RNA-seq libraries across all conditions. Detailed experimental protocol for this study is provided by Panousis et al. 2023 [68].

2.2.2 Genotype Imputation and Quality Control

Individuals who donated cell lines were previously genotyped through the HipSci project [**?**]. Genotypes were obtained for each cell line from the HipSci Consortium (hipsci.org) and genotype calling and imputation to UK10K and 1000 Genomes Phase I is described in [**?**]. CrossMap [**?**] was used to lift over from GRCh37 to GRCh38. Imputed variants with a low imputation score (INFO < 0.4), Hardy-Weinberg equilibrium P-value < 10^{-6}, a minor allele frequency (MAF) < 0.05 or a missingness rate > 0.001 were filtered out. For the remaining variants, genotype principal components (PCs) were calculated using EIGENSTRAT [**?**] to correct for population stratification.

2.2.3 RNA-seq Quality Control And Read Mapping

RNA-seq reads (FASTQ files) were mapped to the reference genome build GRCh38 using splice-aware aligner STAR v2.6.1 [**?**] using the following parameters:

```
--twopassMode Basic --outSAMstrandField intronMotif
--outSAMtype BAM SortedByCoordinate --outSAMunmapped Within
--outSAMattributes All --outFilterMismatchNoverReadLmax 0.04
--outSAMmultNmax 1 --limitBAMsortRAM 40000000000
--sjdbOverhang 74 --waspOutputMode SAMtag
```

The parameter `sjdbOverhang` is particularly relevant as it sets the minimum number of basepairs each split read has to map to on either side of the splice junction. I used a a value of 74 as it is recommended to be set to $readlength - 1$. Finally, this step outputs BAM files that I used for the identification of split reads using LeafCutter. `outSAMstrandField` outputs the strand to which each read was mapped to. This is an important check when splice junctions are linked to genes to ensure that the identified splice junction and gene are transcribed from the same strand.

2.2.4 WASP filtering of ambiguously-mapped reads

It has been shown that ambiguously-mapped reads (RNA-seq reads that map to multiple genomic positions) can bias the number of split reads mapped across splice junctions [**?**]. Genetic variants that overlap with splice junctions can potentially introduce unnoticed mapping ambiguity when an alternative allele increases mapping ambiguity. It is particularly important to guard against this when mapping sQTLs, since decreased splice junction can indicate a mapping ambiguity effect of an alternative allele rather than a true genetic effect.

WASP [**?**] identifies an ambiguously-mapped read by replacing the reference allele in each genomic location that harbours a genetic variant with the alternative alleles and then repeats read mapping. Reads whose mapping ambiguity increases after allele replacement are flagged (`vW`=2-7). In the previous step, I used STAR with the parameter `waspOutputMode`, which flags reads that pass the WASP filter with `[vW]=1`. I therefore only removed read alignments which did not have tag `[vW]=1` using this command in SAMtools [**?**]:

```
samtools view -S -b -e
"[vW]!=2 && [vW]!=3 && [vW]!=4 && [vW]!=5 && [vW]!=6 && [vW]!=7"
```

2.2.5 Identification of split reads

I identified split reads from BAM files from the previous step using `regtools` [**?**] which outputs the number of split reads supporting each intron as `.junc` files. I used the following command and parameters:

```
regtools junctions extract -s 0 -a 8 -m 50 -M 500000
```

These parameters specify the minimum and maximum intron length (50 bp and 500,000 bps in length, respectively), and a minimum splice junction anchor length of 8 bps. The last parameter means that there must be reads supporting at least 8 basepairs at either side of an intron.

2.2.6 Intron clustering using LeafCutter

Mapped split reads (as `.junc` files) were used to perform intron clustering using the LeafCutter script `leafcutter_cluster.py`:

```
python leafcutter_cluster_regtools_py3.py -j .junc -m 50 -l 500000
```

A minimum number of 50 split reads across samples was required to support an intron cluster. Maximum intron length used was 500 kbp. For each identified intron cluster with introns $1 \ldots j$, LeafCutter quantifies intron usage ratio R for intron k as:

$$R_k = \frac{X_k}{\sum_{i=1}^{j} X_i}$$

where X_k is the number of split reads that belong to an identified intron k. A conceptual outline of the intron clustering process is provided in the thesis Introduction and in Figure 1.3.

2.2.7 Differential splicing analysis

I performed differential splicing analysis between each of the 10 stimulated conditions and their corresponding timepoint controls (e.g. sLPS_6 vs Ctrl_6 and sLPS_24 vs Ctrl_24). LeafCutter differential splicing analysis tool (script `leafcutter_ds.R`) [**?**] was used with eight experimental covariates: run ID, donor, library preparation method, sex, differentiation media, purity, estimated cell diameter, and differentiation time, and with the following parameters:

```
-min-samples-per-intron 50 -min-samples-per-group 50 -min-coverage 30
```

This ensures that for a splice junction to be considered differentially spliced, it must be supported by at least 50 RNA-seq samples per condition, and 30 split reads in total. This ensures the robustness of the differential splicing analysis and avoids false positives due to technical differences in splice junction calls between conditions.

2.2.8 UMAP clustering

I performed the UMAP clustering on normalised intron usage ratios for introns that that were detected across all 24 conditions. This ensures that clustering is not based on introns that are not detected in a one or a small number of condition. I performed UMAP clustering after 3 pre-processing steps. First, I performed quantile normalisation on raw intron usage ratios using the R function `normalize.quantiles()`. Second, I performed rank-based inverse normal transformation using the R function `rbint()`. Third, I regressed out eight technical covariates similar to the ones included in the differential splicing analysis: run ID, Donor, library preparation method, sex, differentiation media, purity, estimated cell diameter, and differentiation time. I used the `umap` R package to perform UMAP clustering [**?**] (github.com/tkonopka/umap).

2.2.9 Intron usage ratio quality control and normalization

Because LeafCutter performs annotation-free quantification of intron usage, the output introns are not assigned to specific genes. One way to assign introns to genes is to map the 5' and 3' ends of each identified intron to known exon-intron junctions in an annotation database such as GENCODE. However, this will result in novel introns not being assigned to any gene since their 5' and/or 3' ends will not map to any known exon-intron junction. To overcome this issue, I mapped intron clusters, rather than individual introns, to known exon-intron junctions in GENCODE v27. Specifically, an intron cluster will be assigned to a specific gene if *any* intron within the intron cluster maps to a known exon-intron junction

within the gene. After this procedure, intron clusters with no gene assignment were removed from the sQTL analysis.

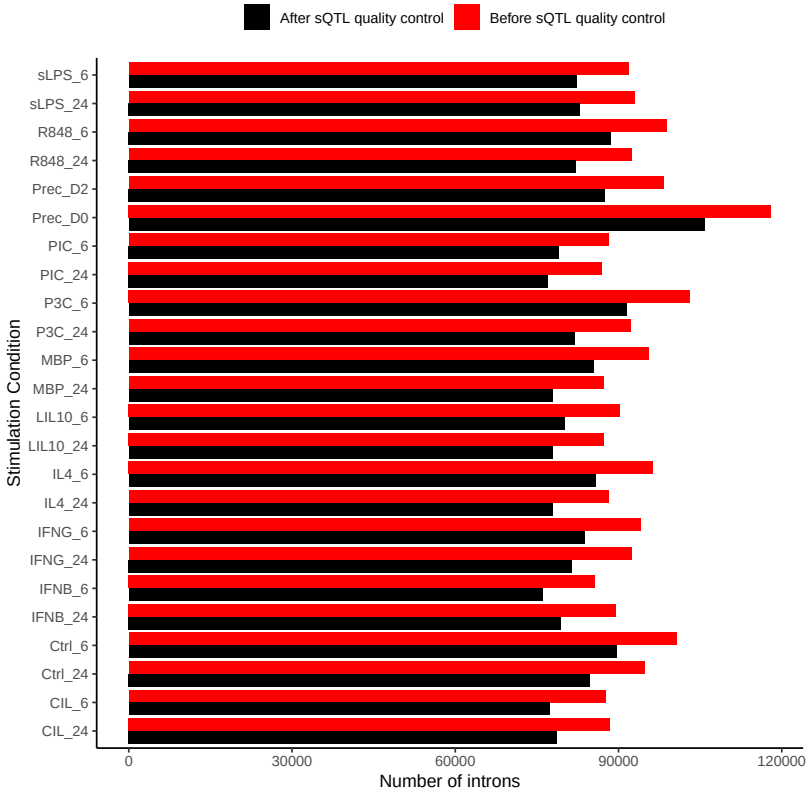


Fig. 2.1 Number of introns identified in the intron clustering procedure of LeafCutter (before intron sQTL quality control; in red) and the number of introns that were used as input to map sQTLs (after intron sQTL quality control; in black).

To minimise the risk of false positive sQTLs, I removed quantified splice junctions that may cause spurious associations when sQTLs are mapped across individuals. These spurious associations can be driven by splice junctions supported by a small number of samples. Therefore, I removed intron clusters that have non-zero usage ratios in more than 40% of samples. I also removed introns with low variance across samples (standard deviation < 0.005). Additionally, in order to make all introns follow the same distribution and to ensure the interpretability of resulting effect sizes, I applied quantile normalisation to intron usage ratios.

2.2.10 Mapping sQTLs using intron usage ratios

I mapped sQTLs using normalised intron usage ratios and genotype data from samples within each condition separately. Variants in a 1 mega base pair (mbp) window around the

transcription start site (TSS) were tested for association with intron usage ratios. Genotype-intron association was modelled using a linear regression model implemented in QTLtools [**?**] with the parameters:

```
--permute 1000 --window 1000000 --grp-best --normal
```

The `--normal` option ensures that intron usage ratios follow a normal distribution by applying a rank-based inverse normal transformation. The option `-grep-best` allows phenotypes (intron usage ratios) to be grouped. Within each phenotype group, the genotype-phenotype sample labels are permuted in exactly the same way. This allows P-values for phenotypes within the same group to be compared with each other and for QTLtools to report the best associated phenotype per group. I grouped introns by the gene they belong to, and report the best associated variant-intron per gene.

In all of my subsequent analyses, I included the first three genotypic principal components (PC) as covariates. In order to remove unwanted variability in intron usage ratios, I mapped sQTLs separately in different conditions using 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20 and 50 intron usage ratio PCs as additional covariates. I then counted the number of genes with significant sQTL effects (sGene) at a false discovery rate (FDR) ≤ 0.05 using the R package `qvalue` v2.16.0 [**?**]. In all my downstream analyses, I use the number of intron usage ratio PCs that maximises the number of sGenes discovered (Figure 2.2)

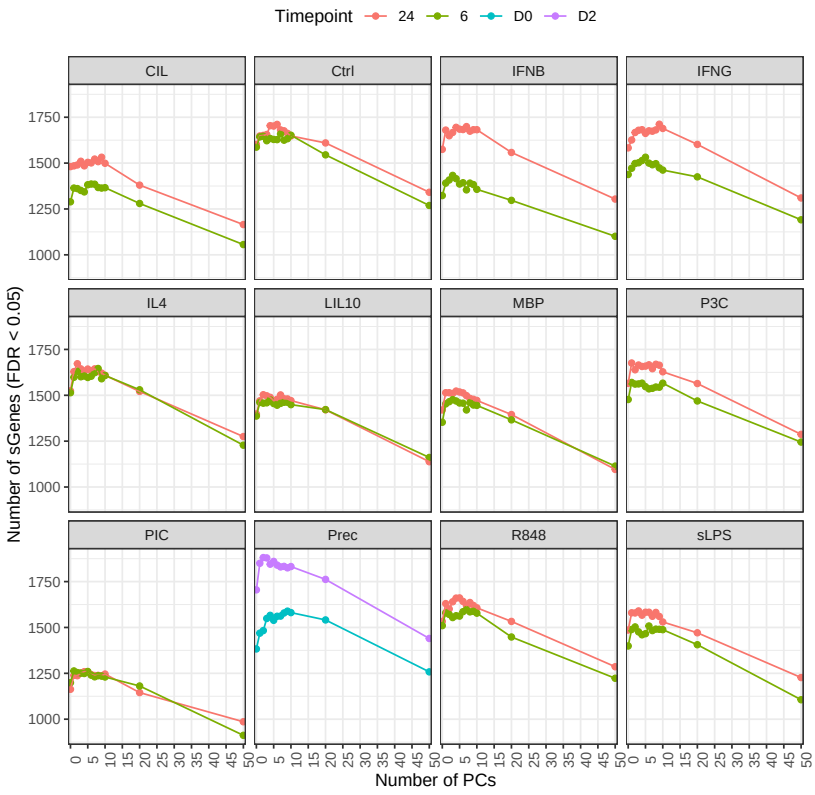


Fig. 2.2 Number of intron usage ratio principal components used as covariates (x-axis) versus the number of genes for which a significant sQTL effect was found (y-axis) coloured by time point (6=6 hours, 24=24 hours, D0=Day 0, D2=Day 2).

2.2.11 Condition-specificity analysis using mash

I tested the condition-specificity of sQTLs using the R package `mashR` v0.2.57 [66]. mashR is an adaptive shrinkage framework that can be used to compare effect size estimates in a multi-tissue or multi-condition association study. It outputs re-estimated effect size estimates in addition to a statistic indicating if effect sizes between two conditions are significantly different from each other (local false sign rate; LFSR). I used this framework to test if a given sQTL effect is a "response sQTL", meaning that its effect size is significantly different in a stimulated macrophage condition from unstimulated macrophages. mashR requires training data to learn the correlation structure in the data from a set of canonical and data-driven covariance matrices (i.e. adaptive). To achieve this, mashR requires a randomly sampled set of QTL associations to learn the mixture components of a diverse set of covariance matrices. I used 10^6 randomly sampled sQTL associations as the random training subset (effect sizes and standard errors). Mash can also be trained using a baseline condition, against which re-estimated effect sizes are compared (I used Ctrl_24 as a baseline condition). The learned

model can be used to re-estimate "posterior" summary statistics for any desired set of sQTL associations by providing their effect sizes and standard errors. In my analysis, I recomputed effect sizes and LFSR for two sets of sQTL effects:

1. Lead SNP for significant sQTLs (FDR < 0.05) per sGene within each stimulated condition. I used this to estimate the number of sGenes with at least one response sQTL (Figure 2.11b).

2. Lead SNP for colocalised genes (PP4 $\geq$ 0.75) in the condition in which the colocalisation was detected. I used this to estimate the proportion of colocalised genes with a response sQTL within each stimulated condition (Figure 2.13).

2.2.12 Genome-wide summary statistics preprocessing

GWAS summary statistics used for the colocalisation analysis were downloaded from either the GWAS catalogue [**?**] or from UK BioBank GWAS summary statistics ([**?**]; Table 2.1). Summary statistics were formatted using a custom script so that each file contains at least: the chromosome and position (GRCh38) of each associated variant, effect size, and standard error.

2.2.13 Identification of genome-wide significant loci from IMD GWAS summary statistics and colocalisation analysis

To define genome-wide significant loci, I identified all variants that passed the genome-wide significance P-value threshold $< 5 \times 10^{-8}$ in each downloaded GWAS summary statistics file. I ordered these SNPs from the lowest to the highest P-value and then iterated over each of them. At each iteration, I defined a 1 Mbp window centred around the SNP. If a SNP falls within a locus that has been defined in a previous iteration, it is not used to define a new locus.

2.2.14 Colocalisation analysis

In order to link the IMD genome-wide significant loci identified in the previous step to effector genes, I performed statistical colocalisation between IMD-associated loci and sQTLs from all MacroMap conditions. Colocalisation analysis is a statistical approach that uses summary statistics from two association studies in order to make an inference about whether the two association signals are likely to be driven by a shared causal variant. In this regard,

Table 2.1 GWAS studies used in the colocalisation analysis. Study accession for GWAS summary statistics downloaded from the GWAS catalogue are shown. Phenotype IDs for the UK Biobank summary statistics are also shown (obtained from https://pheweb.org/UKB-SAIGE/pheno/). PMID = PubMed ID.

Trait	Journal	Publication Date	PMID	Study Accession
Atopic dermatitis	Nat Genet	19/10/2015	26482879	GCST003184
Allergic asthma	Nat Genet	21/05/2018	29785011	GCST007563
Allergy	Nat Genet	30/10/2017	29083406	GCST005038
Ankylosing spondylitis	Nat Genet	01/07/2013	23749187	GCST005529
Asthma	Nat Commun	15/04/2020	32296059	GCST010043
Childhood onset asthma	Am J Hum Genet	28/03/2019	30929738	GCST007800
Crohn's disease	Nat Genet	09/01/2017	28067908	GCST004132
Celiac disease	Nat Genet	06/11/2011	22057235	GCST005523
Diverticulosis and diverticulitis	UK BIOBANK	24/10/2018	30104761	562
Fasciitis	UK BIOBANK	24/10/2018	30104761	728.7
Multiple sclerosis	Nat Genet	01/11/2013	24076602	GCST005531
Osteoarthritis	Nat Genet	21/01/2019	30664745	GCST007093
Osteoarthrosis	UK BIOBANK	24/10/2018		740
Contracture of palmar fascia	UK BIOBANK	24/10/2018	30104761	728.71
Primary biliary cirrhosis	Nat Genet	01/10/2012	22961000	GCST005581
Psoriasis	Nat Genet	01/12/2012	23143594	GCST005527
Rheumatoid arthritis	Nature	25/12/2013	24390342	GCST002318
Primary sclerosing cholangitis	Nat Genet	19/12/2016	27992413	GCST004030
Systemic lupus erythematosus	Nat Commun	17/07/2017	28714469	GCST007400
Systemic sclerosis	Nat Commun	31/10/2019	31672989	GCST009131
Ulcerative colitis	Nat Genet	09/01/2017	28067908	GCST004133

five different hypothesis regarding the relationship between the two association signals are tested:

- H_0: none of the two signals are associated with their corresponding traits
- H_1: only the first signal is associated with its corresponding trait
- H_2: only the second signal is associated with it corresponding trait
- H_3: the two signals are associated with their corresponding traits, with different underlying genetic variants
- H_4: the two signals are associated with their corresponding traits, and share a single underlying genetic variant.

Certainty about each of these hypotheses is quantified as a posterior probability. Therefore, colocalisation analysis outputs five different posterior probabilities: PP_0, PP_1, PP_2, PP_3, and PP_4. Statistical colocalisation is implemented in the R package `coloc v5.1.2`.

Within each genome-wide significant locus, I identified a list of splice junctions to be tested. To achieve this, for each locus I defined a 1 mbp window around the index variant at each locus, and created a set of splice junctions whose respective gene TSS is located within this window. Next, I performed colocalisation analysis between the IMD summary statistics and each sQTLs summary statistics for each gene in the window. I used the `coloc.abf()` function, which takes as input effect sizes and standard errors of each variant from the IMD and sQTL being tested. Importantly, `coloc.abf()` does not require the effect sizes to be aligned to the same effect allele, as the Bayes Factor calculation implemented in `coloc` relies on the Z^2 statistic to compute the posterior probabilities. Finally, I used the default priors implemented in `coloc.abf()`: prior probability a SNP is associated with IMD=10^{-4}, prior probability a SNP is associated with sQTL=10^{-4}.

2.2.15 Colocalisation between significant sQTLs and eQTLs

I performed statistical colocalisation between all significant sQTLs (FDR < 0.05) and eQTLs from the same genes to investigate whether the sQTL and eQTL signals within sGenes are likely to be driven by the same causal variant. I only colocalised the intron with the highest association P-value per sGene using the same approach described in the previous section.

2.3 Results

2.3.1 MacroMap: a resource for studying macrophage transcriptome

Induced pluripotent stem cell lines (iPSC) from 209 healthy unrelated individuals, generated as part of the HipSci project [**?**], were differentiated into iPSC-derived macrophages (experimental protocol described in Panousis et al. 2023 [68]). RNA was harvested from macrophage precursors at day 0 (Prec_D0) and day 2 (Prec_D2). Naïve macrophages were exposed to a panel of 10 stimuli, and RNA was obtained and processed 6 and 24 hours after stimulation (in addition to unstimulated controls; Ctrl_6 and Ctrl_24), resulting in a total of 24 different conditions. I quantified alternative splicing from split reads (reads mapping across two splice junctions) using LeafCutter, which quantifies alternative splicing as intron usage ratios [**?**]. I derived an intron usage ratio matrix for each of the 24 conditions and used this for differential splicing analysis and sQTL mapping.

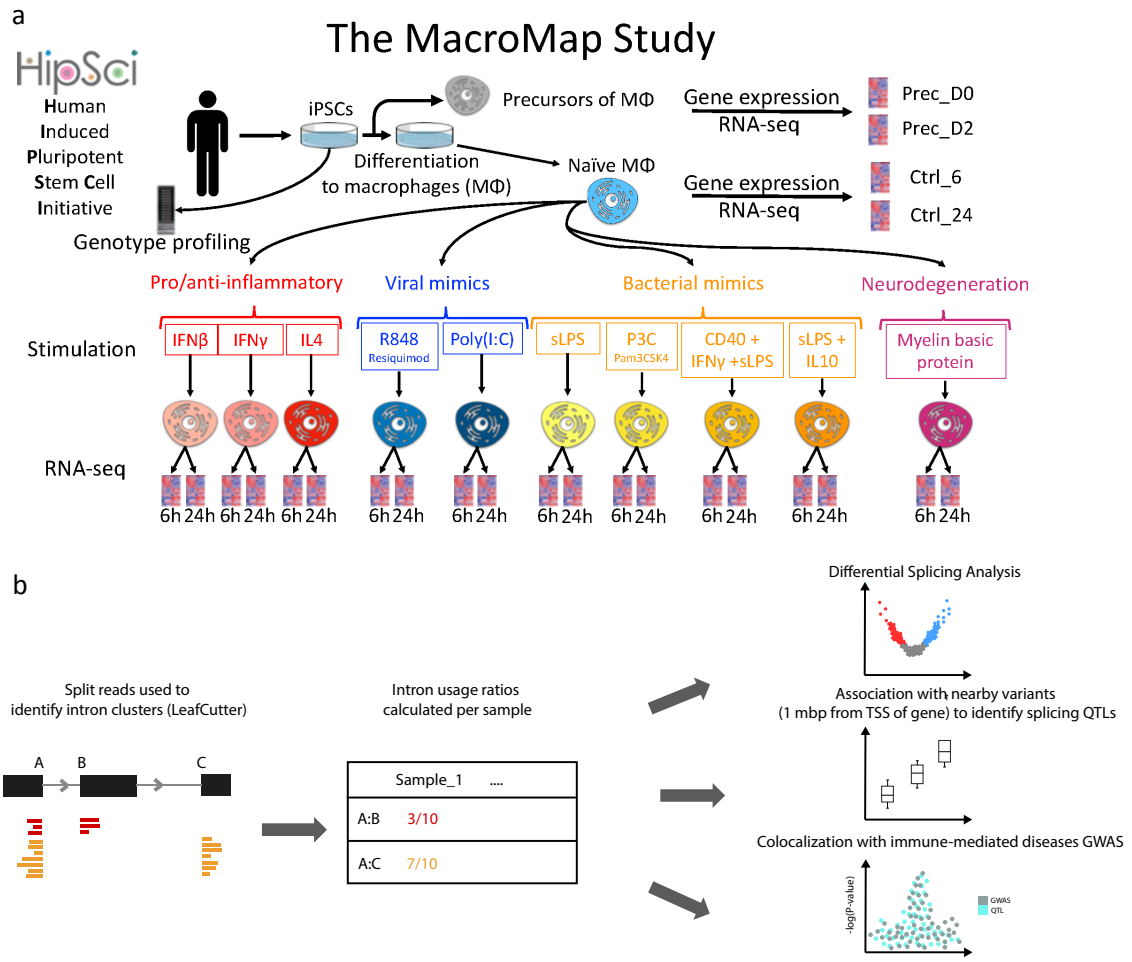


Fig. 2.3 Overview of study: (a) Genotyped iPSC cell lines were differentiated into macrophages, and RNA was harvested before differentiation (Prec_D0) and 2 days after starting differentiation (Prec_D2). RNA was also harvested from differentiated macrophages at 6 and 24 hours (Ctrl_6 and Ctrl_24). Naive macrophages were then exposed to a panel of 10 stimuli and RNA was harvested at 6 and 24 hours after stimulation. (b) Split reads were used to quantify intron usage ratios on an individual level using LeafCutter. Split reads were then used for differential splicing analysis between naive and stimulated conditions, and as a quantitative trait to map splicing quantitative trait loci (sQTLs). sQTLs were then colocalised with 21 immune-mediated disease GWAS summary statistics.

2.3.2 Alternative splicing patterns during the macrophage differentiation process

In order to visualise general patterns of intron usage across conditions, I projected intron usage ratios in all samples on a UMAP. Since intron usage QC was performed separately for

different conditions, not all introns were shared across conditions. Therefore, I created the UMAP projection using only the introns that passed QC in all conditions (N=40,044).

Overall, I found that macrophage precursors (Prec_D0) clustered separately from all other conditions. Moreover, precursors at day 2 (Prec_D2) clustered together with naïve macrophages, suggesting that splicing changes start early during the seven-day macrophage differentiation process (Figure 2.6a). I also observed a clear separation between stimulated cells harvested after 6 and 24 hours, a separation that I did not observe between unstimulated macrophages (Ctrl_6 and Ctrl_24; Figure 2.6b). These observations show that splicing changes are observed both during iPSC differentiation into macrophages and following macrophage stimulation.

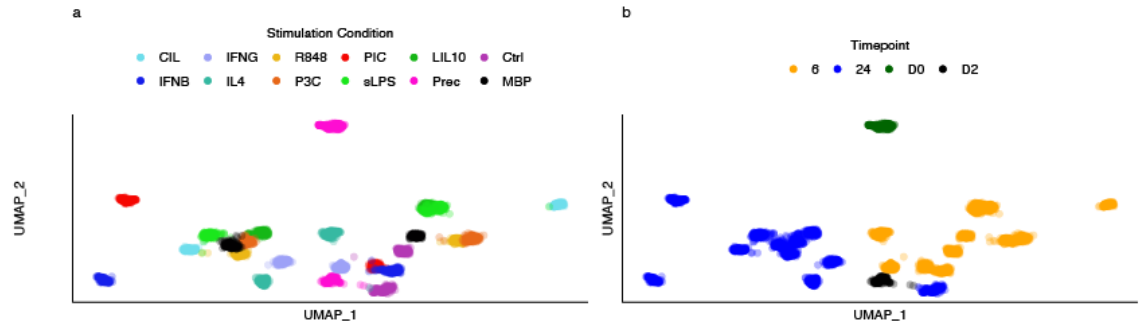


Fig. 2.4 UMAP of intron usage ratios in different stimulation conditions, coloured both by (a) different stimulation conditions and (b) by time point.

Although UMAPs are useful to visualise general patterns in the data, they are less useful to make inferences about the correlation structure due to the non-linear nature of a UMAP transformation. Therefore, I confirmed these patterns by measuring the correlation between different conditions. Since a large proportion of introns show low variability across samples, which may inflate correlation estimates, I only used the top 20,000 variable introns. In confirmation of the UMAP patterns, Prec_D0 showed relatively weaker correlation with both differentiated macrophages at day 2 (Pearson correlation coefficient ρ with Prec_D2=0.84) and the two differentiated control conditions (ρ=0.85 and 0.84 with Ctrl_6 and Ctrl_24). Conversely, correlation between Prec_D2 and both Ctrl_6 and Ctrl_24 was stronger (ρ=0.94 and 0.95, respectively), confirming the trends observed in the UMAP. This difference in correlation strength was even more striking when I used the top 10,000 and top 5,000 variable introns, showing that it is not an artefact of the number of variable introns used to measure

correlations between these conditions.

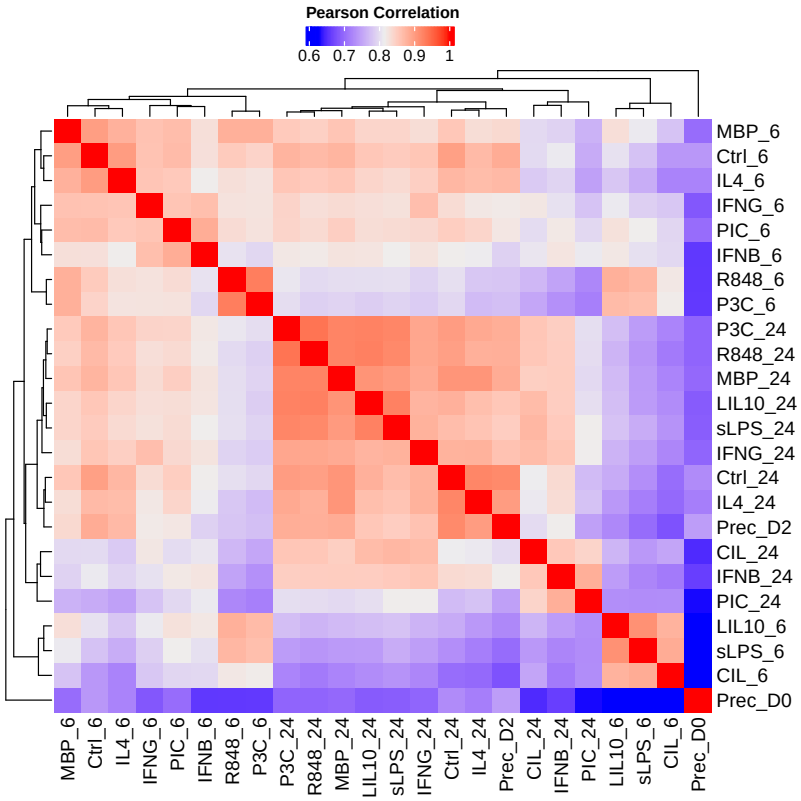


Fig. 2.5 Heatmap showing Pearson correlation coefficient values between all 24 conditions. Usage ratios of the top 20,000 variable introns (highest variance across all conditions) were log_{10} transformed and the mean usage was calculated across samples within each condition. Pairwise correlation coefficients were then calculated between the means of log transformed ratios.

2.3.3 Macrophage response genes are differentially spliced upon stimulation

In order to formally quantify how many genes are alternatively spliced upon stimulation, I performed differential splicing analysis (DSA) between naïve and stimulated macrophages. The DSA method implemented in LeafCutter jointly quantifies the overall changes at the level of intron clusters, which are groupings of introns that share an acceptor and/or donor splice site. In total, I found that 3,464 genes were alternatively spliced upon stimulation (adjusted P-value < 0.05 and |log effect size| > 0.5). Notably, stimulation with IL4 had the least effect on splicing (110 and 94 genes in macrophages harvested after 6 and 24 hours,

respectively; Figure 2.6), corroborating previous reports where stimulation with IL4 did not cause dramatic changes to the splicing patterns of macrophages [**?**].

The large number of differentially spliced genes motivated me to understand which biological pathways are subjected to differential splicing upon macrophage stimulation. In particular, I wanted to investigate whether macrophages respond to pathogens by activating splicing programmes in genes that are import to eliminate pathogens and initiate an immune response. In this section, I will summarise the enriched REACTOME pathways in two representative stimulation conditions. These two conditions were chosen to represent two broad classes of stimuli used in the MacroMap experimental design: Lipopolysaccharides as an example of bacterial stimulation and PolyI:C as a representative of viral stimulation.

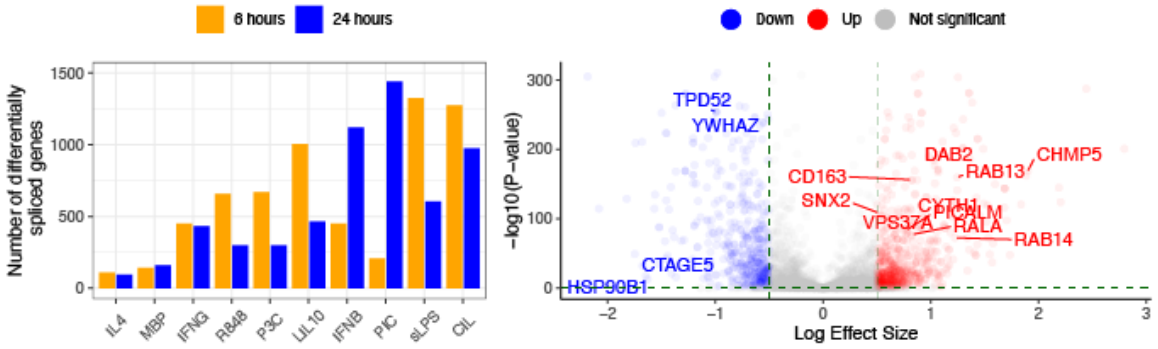


Fig. 2.6 (Left) Number of differentially spliced genes between naive and stimulated macrophages after 6 hours (yellow) and 24 hours (blue) (Right) Volcano plot showing differentially spliced genes 6 hours after sLPS stimulation, with log effect size on the x-axis (for each gene the intron with the largest absolute effect size is shown) and -log10 of adjusted P-value on the y-axis. Colours indicate the direction of intron usage change (blue indicating reduced usage and red indicating greater usage in stimulated cells versus naive cells). Some genes that belong to the "vesicle-mediated transport" REACTOME pathway are indicated.

Stimulation with LPS led to the differential splicing of genes in the cytokine signalling pathway (84 genes; P-value=9.1×10^{-8}), the vesicle-mediated transport pathway (68 genes; P-value=10^{-6}), and the class I MHC mediated antigen processing and presentation pathway (47 genes; P-value=7.95×10^{-6}; Figure 2.7). Within genes that belong to the vesicle-mediated transport pathway, several genes that code for members of the RAB protein family were differentially spliced upon stimulation (including *RAB1B, RAB4A, RAB9A, RAB11A, RAB13, RAB14* and *RABEPK*; Figure 2.6). RAB proteins are GTPases that coordinate membrane trafficking by regulating the formation, movement and fusion of vesicles with

their destination membrane [**?**]. For example, I observed greater usage of the first exon of a non-canonical transcript of *RAB13* (*RAB13-205*), and lower usage of the canonical first exon (*RAB13-201*; difference in percentage spliced in (ΔPSI)=0.046 and -0.05, respectively). The alternative usage of the first exon may also be reflected at the level of protein products as the two annotated transcripts that start with these two alternative first exons have different amino acids sequences (203 and 122 amino acids respectively). Although this finding suggests that stimulation with LPS leads to an alternative *RAB13* isoforms being expressed by macrophages, little is known about the particular functions of these isoforms. *RAB13* has been previously shown to coordinate the formation of actin filaments through which vesicles are transported across the cell [**?**], but the functional characterisation of the different isoforms of *RAB13* are still lacking.

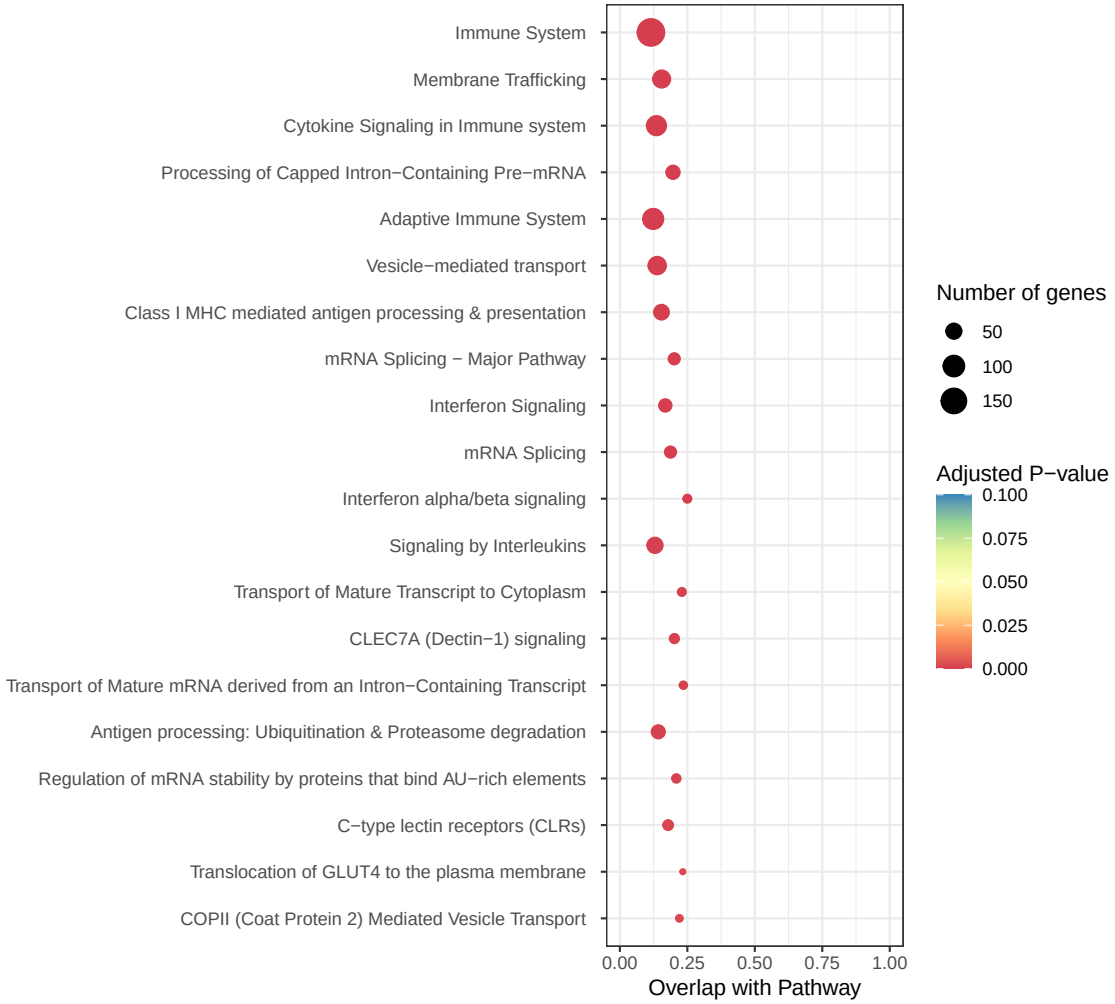


Fig. 2.7 Top 20 REACTOME pathways enriched in differentially spliced genes between Ctrl_6 and sLPS_6. Enriched pathways are shown on the y-axis and the overlap between differentially spliced genes and all genes in the pathway are shown on the x-axis. Colors indicate the enrichment adjusted P-values and the size of the points indicate the number of genes in each enriched pathway.

Stimulation with dsRNA viral mimic PolyI:C (PIC_6) also led to dramatic splicing changes (Figure 2.8). For example, genes involved in viral sensing and RIG-I/MDA5-mediated activation of the antiviral cytokine Interferon β (IFNβ) were differentially spliced (P-value=1.7×10^{-5}; 7 genes). RIG-I/MDA5 receptors belong to the RIG-I-like family of receptors which sense dsRNA and activate type I interferons in response (e.g. IFNβ) [**?** **?**]. In addition to the genes coding for the RIG-I and MDA5 receptors (*DDX58* and *IFIH1* respectively), genes that regulate the expression of cytokine IFNβ, such as *TRIM25* [**?**], *TANK* [**?**], and *RIPK1* [**?**] were also differentially spliced. These observations reinforce previous functional work that showed the modulated antiviral response of RIG-I/MDA5-mediated

IFNβ activation pathway by the different isoforms of its constituent components [**? ? ?**].

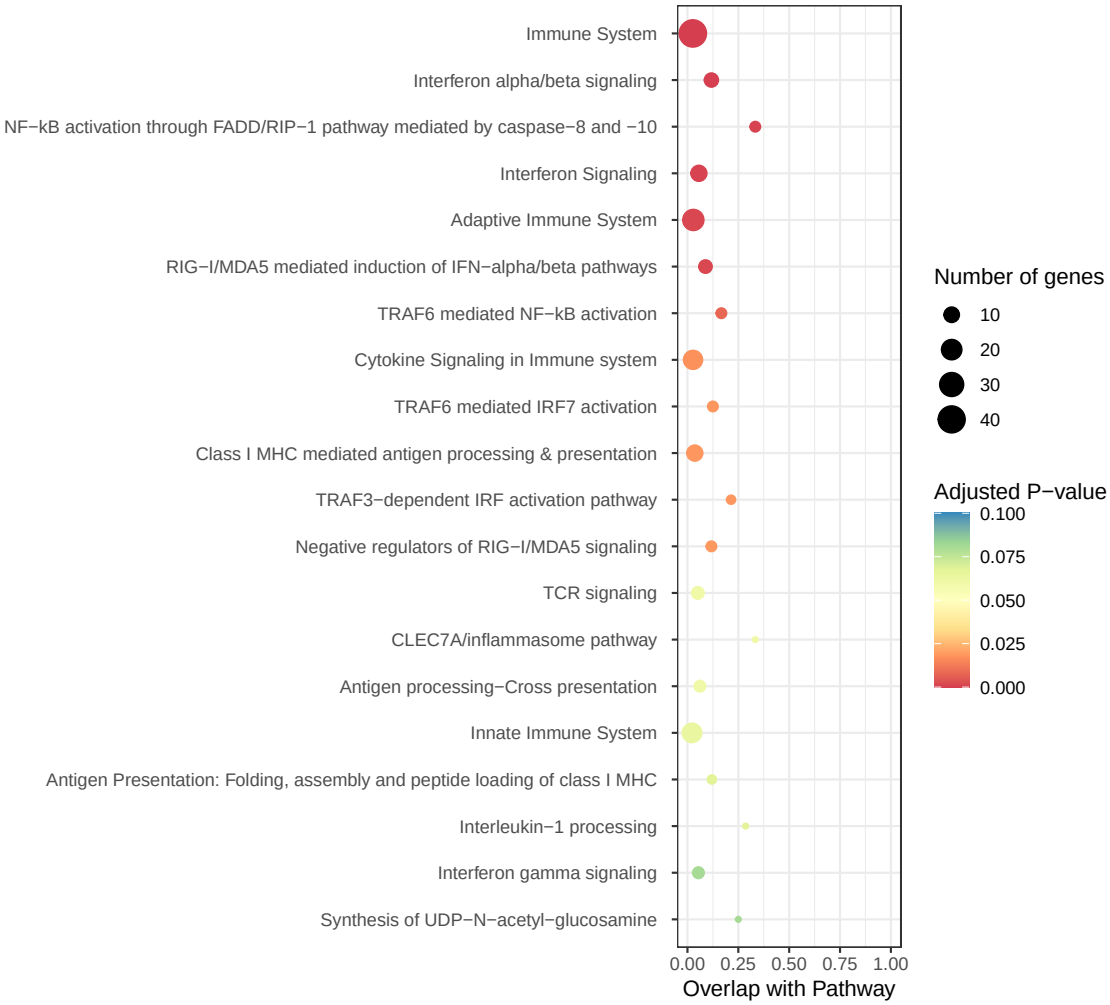


Fig. 2.8 Top 20 REACTOME pathways enriched in differentially spliced genes between Ctrl_6 and PIC_6. Enriched pathways are shown on the y-axis and the overlap between differentially spliced genes and all genes in the pathway are shown on the x-axis. Colors indicate the enrichment adjusted P-values and the size of the points indicate the number of genes in each enriched pathway.

This study is not the first to demonstrate the important but underappreciated role of alternative splicing in innate immune response [**? ?**]. For example, Kalam et al. [**?**] showed that a shorter isoform of *RAB8B* is expressed in macrophages upon stimulation with Mycobacterium tuberculosis. Their work showed a subtle mechanism whereby the survival of mycobacteria inside macrophages was controlled by expressing long or short isoforms of *RAB8B*, which affected the ability of lysosomes to target mycobacteria. Although

a detailed account of the different alternative splicing patterns in response to pathogens is not the primary focus of this chapter, it reinforces its relevance to innate immune response. It also shows that iPSC-derived macrophages are a suitable model that captures relevant transcriptomic changes upon macrophage stimulation. This analysis served as a motivation to understand the genetic regulation of alternative splicing by mapping splicing QTLs.

2.3.4 Macrophage stimulation increases the number of genes with significant sQTL effects

After intron usage quality control, I identified a median of 82,058 introns per condition, with a total of 160,748 unique introns across all conditions (75,987-105,841 introns per conditions with the greatest number of introns seen in Prec_D0; Figure 2.1 in Methods). Approximately 81% of identified introns were independently identified in at least 2 conditions (Figure 2.9). Interestingly, 53% of single-condition introns were identified in Prec_D0. This finding is in line with the earlier correlation analysis that showed that intron usage correlation between undifferentiated iPSCs and differentiated macrophages is generally weaker than intron usage ratio correlation among differentiated macrophages (Figure 2.5). Furthermore, it shows that a largely distinct set of splice junctions are used in undifferentiated iPSCs that become undetectable in fully differentiated macrophages.

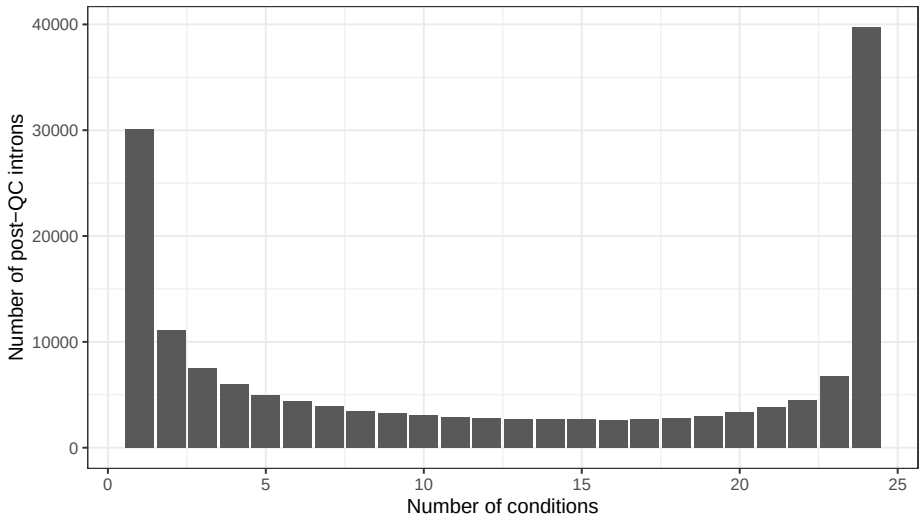


Fig. 2.9 Distribution of the number of introns that were detected and passed QC in different numbers of conditions.

After mapping the quantified introns to genes (see section 2.2.9 in Methods for details), I found that each gene had a median of seven introns and up to 10,851 genes were quantified per condition. In total, introns were mapped to 12,792 genes across all conditions, and over 94% of genes were detected in at least 2 conditions (Figure 2.10). Similar to introns, over half of the 6% single-condition genes were unique to Prec_D0.

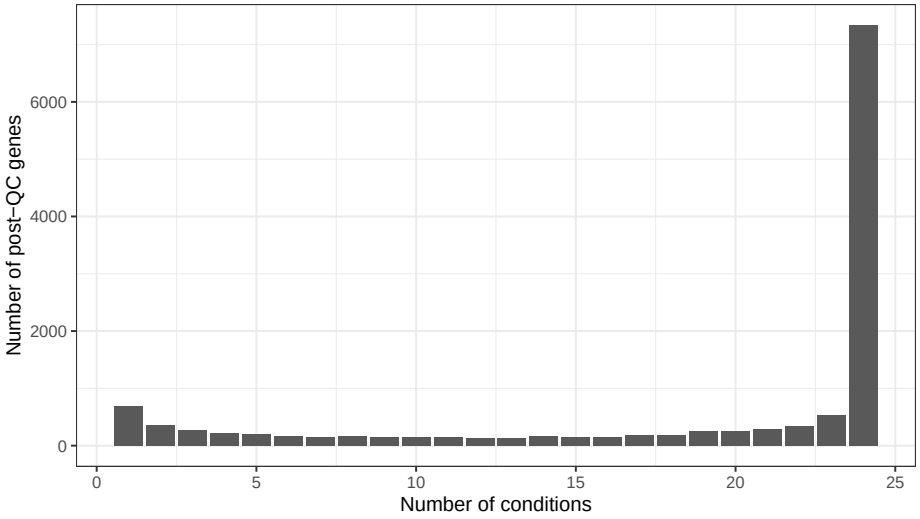


Fig. 2.10 Distribution of the number of genes that were detected and passed QC in different numbers of conditions.

Using normalised intron usage ratios as normally distributed quantitative traits, sQTLs were mapped within a $\pm$ 1 Mbp window centred around the transcription start site (TSS) of each gene. I also used multivariate adaptive shrinkage (mash [66]) to compare sQTL effect sizes between naive and stimulated conditions (I used Ctrl_24 as a baseline condition). Mash reports a significance measure known as local false sign rate (LFSR; section 2.2.11 in Methods), which I used to identify sQTLs whose effect sizes change significantly upon stimulation (response sQTLs).

I called significant sQTLs at a false discovery rate (FDR) < 0.05. Across all conditions, I detected a total of 5,734 sGenes (median number of sGenes per condition=1580 and Prec_D2 had the most sGenes=1,881) (Figure 2.11a,b). Of these, 878 sGenes (15.3%) had at least one response sQTL (LFSR < 0.05). As expected, Prec_D2, Ctrl_6, IL4_6, and IL4_24 had the smallest proportion of response sQTLs (< 9%). This is expected because Prec_D2 and Ctrl_6 represent naïve macrophages and it is therefore unlikely that any significant transcriptomic changes will have occurred when their RNA was harvested. Similarly, stimulation with the anti-inflammatory cytokine IL4 is unlikely to result in any significant

response sQTLs, consistent with results from the differential splicing analysis where stimulation with IL4 led to the smallest number of differentially spliced genes across all conditions.

I then asked which of the two stimulation timepoints (6 and 24 hours) was more likely to have response sQTL effects across stimulation conditions. Up to 29% of response sGenes per condition had a response sQTL after 24 hours that was not detected after 6 hours. Conversely, up to 72% of response sGenes per condition had a response sQTL after 6 hours that was undetectable after 24 hours. This suggests that response sQTLs are more likely to be detected 6 hours after stimulation (Figure 2.11c), in agreement with previous work that suggested 4-6 hours for optimal detection of transcriptomic changes following macrophage stimulation [**?** **?**].

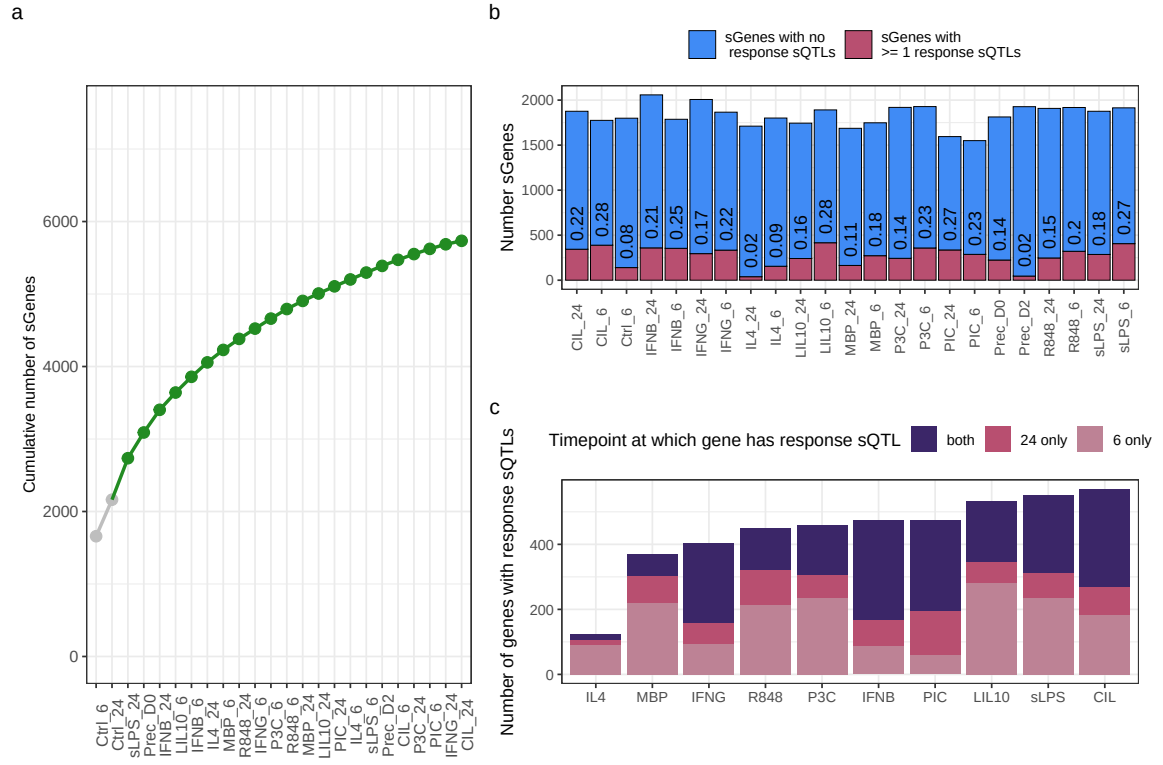
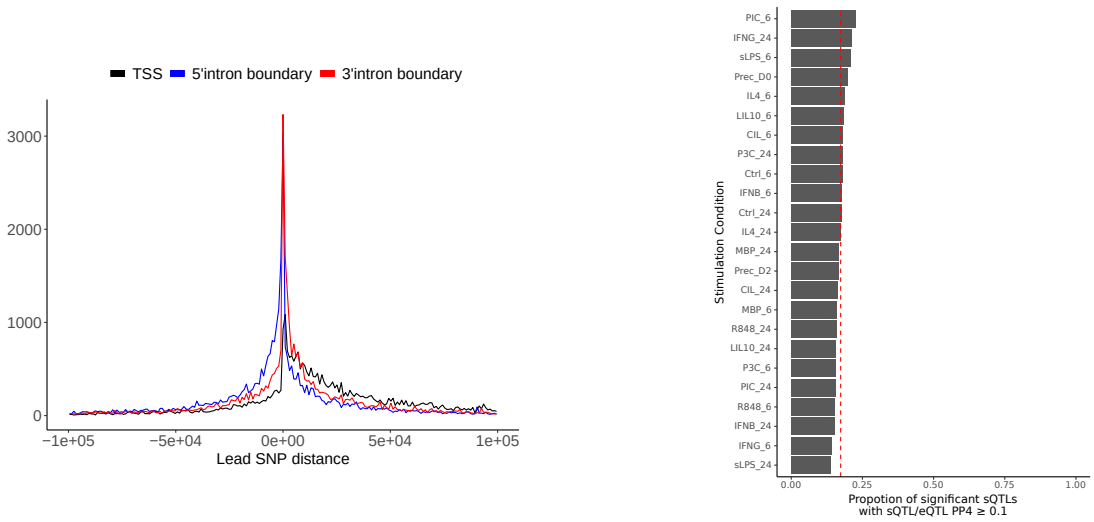


Fig. 2.11 (a) Cumulative number of genes with significant splicing QTL effects, with unstimulated conditions indicated in grey (b) Total number of significant sQTLs per condition and proportion of response sQTLs within each condition (sQTLs with LFSR < 0.05; Methods). (c) Number of genes, per condition, with at least one response sQTL at 6 hours, 24 hours or both.

Similar to previous reports [72], I found that lead sQTL SNPs are located closer to intron boundaries than to the TSS of their genes. On average per condition, 24.5% of sGenes had a lead SNP within 10 kbp of their TSS, compared to 46.8% within 10 kbp of either the 5' or 3' intron boundaries (Figure 2.12a). However, the lead SNP was located within the boundaries of its associated intron in less than 4% of significant sQTLs. Although it is compelling to think that distal enhancer or silencer sequences may explain this observation, it is also likely that the lead SNP is an LD proxy for the truly causal SNP. The significant sQTL introns had a median length of 2.5 kbp, which is much shorter than the European-ancestry LD ranges of many loci.



(a) (b)

Fig. 2.12 (a) Distribution of the distance between the lead SNPs of significant sQTL effects (across all conditions) and transcription start site (TSS; in black) of the sQTL gene, 5'intron boundary (in blue) and 3' intron boundary (in red). (b) Proportion of significant sQTL effects that share a single causal variant (PP4 $\geq$ 0.1) with the same eQTL gene in the same condition. Red line indicates the average proportion of sQTLs with sQTL/eQTL colocalisation $\geq$ 0.1 across conditions.

I next sought to understand the relationship between the sQTL and eQTL signals at the discovered sGenes. This is an important question in the context of understanding the transcriptomic effects of disease-associated loci. If it is the case that a large proportion of sQTL signals colocalised with eQTL signals, then the added value of sQTLs in terms of explaining disease-associated loci would be limited. Existing evidence on the relationship between eQTL and sQTL signals is contradictory. While some sQTL mapping efforts have shown that sQTL signals are largely independent from eQTL signals, others maintain that there is a large

overlap between eQTLs and sQTLs [72**? ? ?**]. Therefore, it is unclear whether overall levels of gene expression and alternative splicing are genetically co-regulated or are regulated via distinct mechanisms [**? ? ?**]. To verify this, I performed statistical colocalisation (Methods; ref [**?**]) between sQTLs and eQTLs derived from the same data (Panousis et. al. 2023 [68]). Specifically, for all sGenes, I colocalised the sQTL and eQTL signals in the exact same window (1 mbp around TSS). I found that on average across conditions, 75% of sGenes were extremely unlikely to share a causal variant with eQTLs from the same genes (PP4 < 0.1; Figure 2.12b; Methods), indicating that the majority of sQTL signals are largely independent from eQTL signals. Building on this evidence, I therefore hypothesised that colocalisation of sQTLs with disease-associated loci may explain additional disease-associated loci distinct from those already implicated via eQTLs.

2.3.5 Splicing QTLs identify GWAS effector genes undetected by expression QTLs

I then aimed to identify IMD-associated loci that were likely linked to alternative splicing changes in macrophages. To this end, I used statistical colocalisation analysis between sQTL and GWAS association signals (using R package coloc; more details in Methods). Since macrophages play a major role in innate immune defence, I focussed on 21 IMD GWAS summary statistics to quantify the probability of a sQTL sharing a causal variant with genetic association signals. IMD GWAS summary statistics were downloaded from the GWAS catalogue [**?**] (See methods for how GWAS loci were defined and Table 2.1 for a complete list of downloaded GWASes). In order to compare the colocalisation yield between eQTLs and sQTLs, I also obtained colocalisation results from eQTLs mapped from the same dataset and colocalised with the same GWAS loci.

Across all 21 IMDs, 1,528 GWAS loci were tested against 4490 genes across all conditions (34,128 introns). I identified 707 unique genes (1,337 introns) with an sQTL signal in at least one condition that likely shares a causal variant with an IMD risk locus (PP4 $\geq$ 0.75; Figure 2.13). Approximately, 60% of these genes passed the colocalisation threshold in a single condition only. However, these should not be interpreted as condition-specific colocalisations as the PP4 values could be close to the colocalisation threshold in other conditions. Therefore, I compared the effect sizes of the colocalised sQTL across all condition using mash. I observed that 68 (9.6%) of the colocalisaed genes implicated a response sQTL, indicating that the genetic effects of these variants on alternative splicing could not

be detected in unstimulated macrophages. Based on hierarchical clustering of LFSR values, Prec_D2, IL4_6 and IL4_24 had the fewest response sQTLs that colocalised with GWAS signals, while sLPS_6, CIL_6 and LIL10_6 (all stimulated with LPS) yielded the most, recapitulating results from the differential splicing analysis (Figure 2.13).

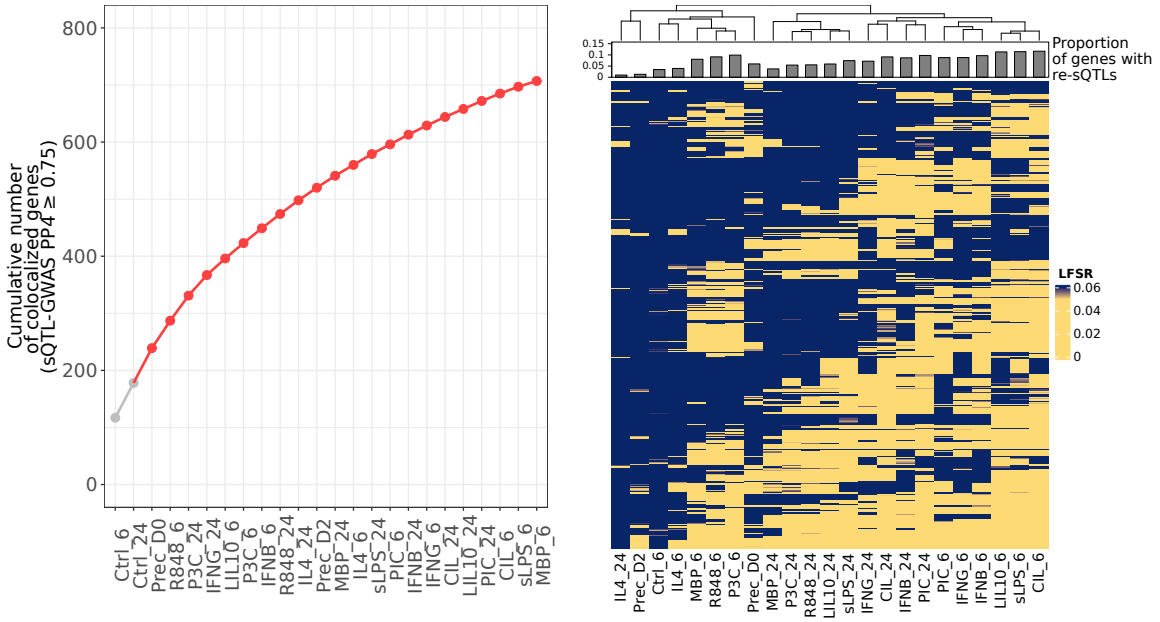


Fig. 2.13 (left) Cumulative number of genes with a GWAS-sQTL (PP4 $\geq$ 0.75) across different conditions, with unstimulated conditions shown in grey on the left. (right) Heatmap and hierarchical clustering of LFSR values for all the colocalised sQTL effects (PP4 $\geq$ 0.75) across all 21 IMDs. On top of the heatmap is a barplot showing the proportion of colocalised sQTL effects that are response sQTLs (LFSR < 0.05)

I then compared how many tested loci colocalised with each type of molecular QTL and found that 50.4% (771/1,528) of tested loci were likely to share a single causal variant with either an eQTL, sQTL or both. Recently, Mountjoy et al. 2021 (ref: [26]) colocalised 50.7% of tested GWAS loci with protein and expression QTLs from 92 tissues and cell types. In comparison, this high colocalisation yield from a single cell type shows the promise of profiling the transcriptome of a relevant cell type in understanding the effects of GWAS loci. Moreover, unlike other tissue-level QTL maps, interpreting these colocalisations in the context of macrophages potentially aids the interpretation and functional follow-up of these loci.

Approximately half of the colocalised loci (385 loci or 25.2% of tested loci) colocalised solely with an sQTL (sQTL PP4 $\geq$ 0.75 > eQTL PP4), clearly demonstrating both the value

of sQTLs for identifying GWAS effector genes and the important role that alternative splicing plays in complex disease risk (Figure 2.14). However, this percentage may also be affected by eQTL colocalisations that are close to the colocalisation threshold (e.g. eQTL PP4 = 0.74). When I relaxed the eQTL PP4 threshold (sQTL PP4 $\geq$ 0.75 and eQTL PP4 $\geq$ 0.5), I found that 16.8% of loci colocalised solely with an sQTL signal. On the other hand, at the same thresholds applied in reverse (i.e. eQTL PP4 $\geq$ 0.75 and sQTL PP4 $\geq$ 0.5), only 2% of loci colocalised solely with an eQTL signal.

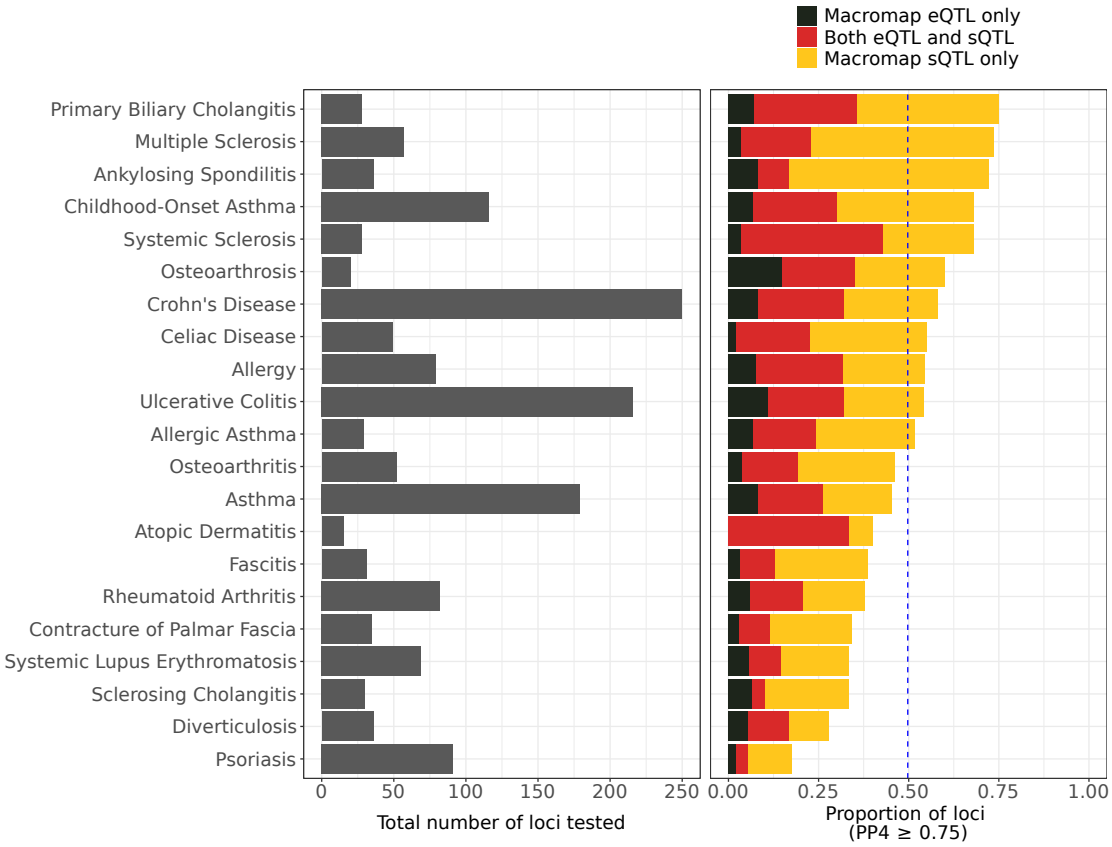


Fig. 2.14 Total number of loci tested for colocalisation (left) and proportion of genome-wide significant loci that share a single causal variant (PP4 $\geq$ 0.75) with an eQTL only, an sQTL only or both (right).

2.3.6 Lowly-used alternative splicing events underlie complex disease risk

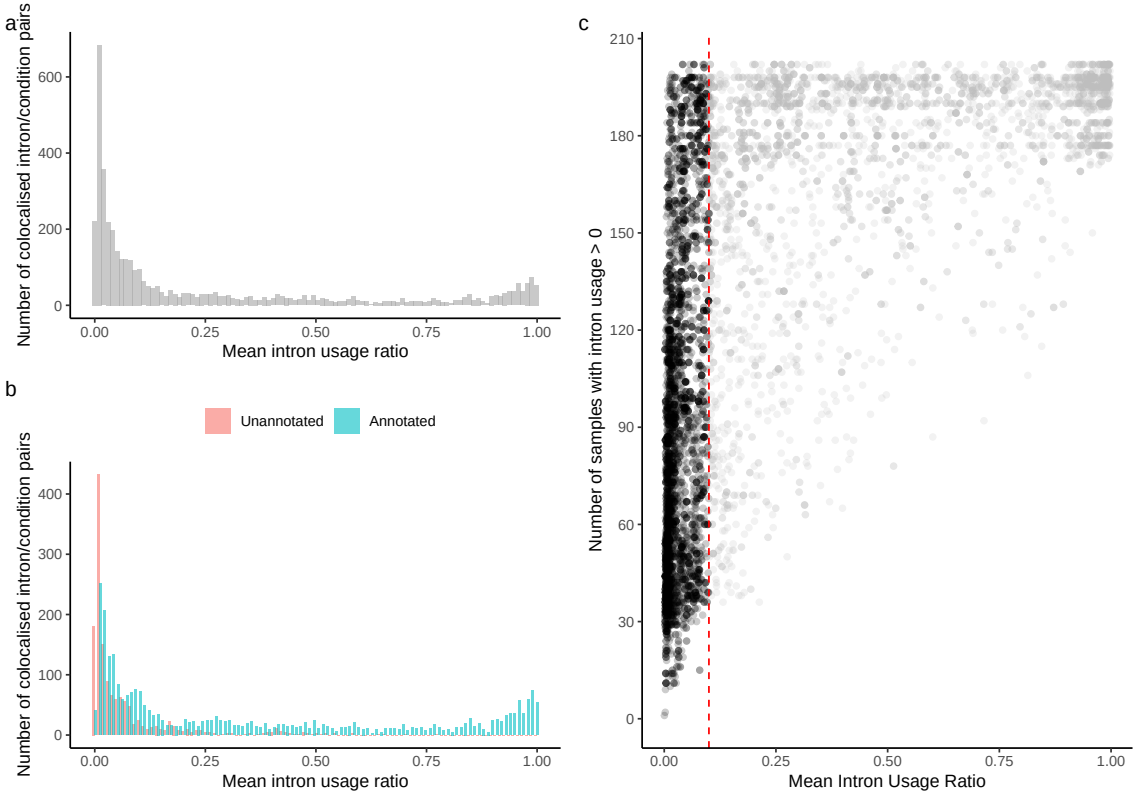


Fig. 2.15 (a) Distribution of mean intron usage ratio for colocalised introns, showing a peak close to 0, and (b) coloured by annotation in GENCODE v27, showing an enrichment of unannotated introns among introns with low mean usage ratio (c) Number of samples in which the intron is supported by split reads (y-axis) shown against the mean intron usage ratio of each sample (x-axis). Red vertical line at mean intron usage ratio = 0.1.

I next sought to characterise the colocalised sQTL introns, by asking how often colocalised sQTL splicing events are used in observed transcripts. There is ample evidence that aberrant splicing underpins several inherited diseases such as Spinal Muscular Atrophy and Duchenne Muscular Dystrophy [**? ? ?**], but it is unclear to what extent lowly-observed splicing events contribute to complex diseases. To evaluate this role, I assessed the usage of introns that were implicated in IMD risk via colocalisation analysis.

I observed that 53.4% of colocalised sQTL introns have a mean intron usage ratio (IUR) < 0.1 across samples (Figure 2.15a). Over 96% of these introns had non-zero usage in at least 30 samples (Figure 2.15c), indicating that these splicing events can be reliably observed in

multiple RNA-seq samples and individuals, but with relatively low IUR. Moreover, 50.6% of these introns are not found in any annotated transcripts in GENCODE v27, whereas only 12% of introns with mean IUR ≥ 0.1 are absent from GENCODE v27 (Figure 2.15b), in line with previous reports showing that lowly-observed splicing events tend to be unannotated in transcript databases [**? ?**].

The observation that over half of colocalised sQTL introns are lowly used and that they tend to be unannotated strongly emphasises the need to investigate their functions and role in the context of complex disease. For example: are low-abundance transcripts translated into protein products or do they exert gene regulatory functions? What is the effect of up- or down-regulating these isoforms on different cellular phenotypes? Sampling these rare transcripts will thus shed light on the transcriptomic effects of IMD-associated risk loci, an avenue that has remained largely unexplored in most large-scale transcriptomic cohorts.

2.3.7 A rare alternative splicing event likely underpins IBD risk at the *PTPN2* locus

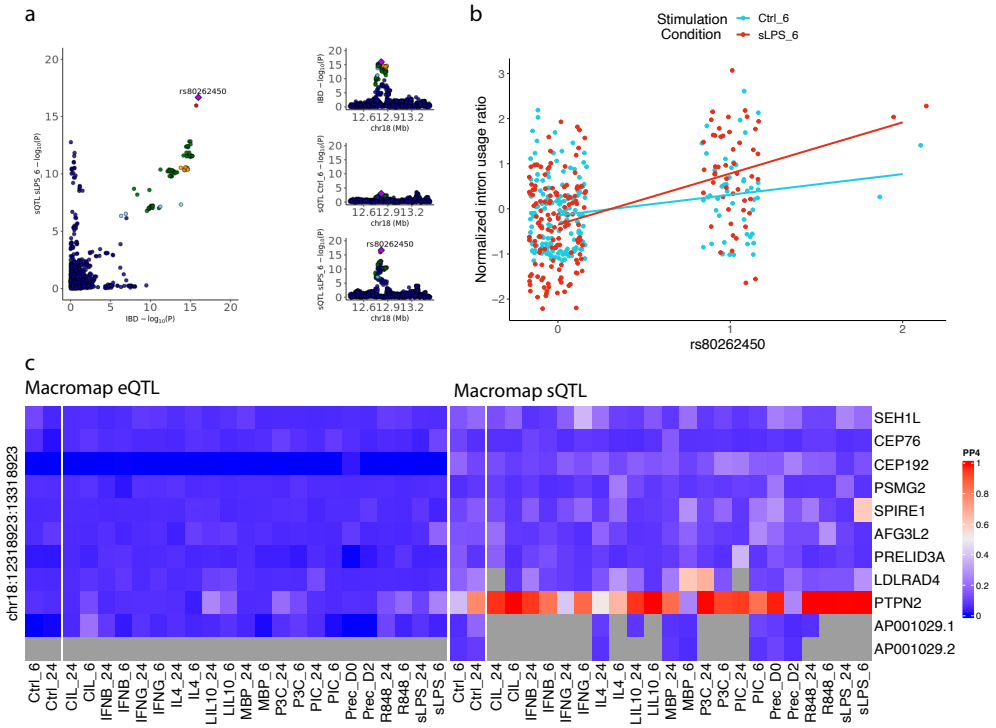


Fig. 2.16 Example of colocalisation between an IBD risk locus at 18p11.21 and an sQTL for *PTPN2*. (a) Regional association plots of the IBD association signal, and sQTL association signal in unstimulated macrophages (Ctrl_6) and macrophages stimulated with sLPS after 6 hours (sLPS_6), (b) Normalised intron usage ratios of different genotypes of the lead IBD SNP rs80262450 in Ctrl_6 and sLPS_6, (c) Heatmap showing evidence of colocalisation (PP4) between the IBD association signal at 18p11.21 and all macrophage eQTLs/sQTLs in the locus (in all conditions).

To demonstrate how sQTLs for lowly-used introns can dysregulate alternative splicing and predispose to IMDs, I further investigated a *PTPN2* sQTL that implicates a lowly-used intron that colocalised with an inflammatory bowel disease (IBD) associated risk locus at 18p11.21. Multiple lines of evidence, including coding variants associated with monogenic IBD [**? ?**] and mouse knock-out models [**? ?**], have suggested *PTPN2* is the effector gene at 18p11.21, though this remains to be established. It is not yet known if and how common IBD-associated SNPs affect the expression of *PTPN2*.

I observed that the lead IBD SNP at 18p11.21 (rs80626450; 18:12818923_G_A) is associated with higher risk of IBD and with increased usage of intron 1-2 of the non-canonical transcript *PTPN2-205* (chr18:12,817,365-12,818,944; Figure 2.16b and Figure 2.17). rs80626450 is located 21 base pairs downstream of the donor splice site of exon 1 of *PTPN2-205*, and is the lead SNP for both the sQTL and IBD association signals, strongly suggesting its involvement in the aberrant splicing event at this locus (Figure 2.16a and Figure 2.17). The *PTPN2* sQTL signal colocalised (PP4 $\geq$ 0.9) with the IBD signal in 13 conditions, but did not colocalise with any eQTLs mapped from the same data (eQTLs mapped in Panousis et. al. 2023; Figure 2.16c). Despite the relative rarity of this transcript, I detected this colocalisation using GTEx sQTL summary statistics [16], where this sQTL signal was also colocalised with the IBD locus (PP4 $\geq$ 0.9) in 14 tissues, including whole blood (Figure 2.18), indicating that this rare splicing event can be reliably detected in a large number of tissues from an independent dataset.

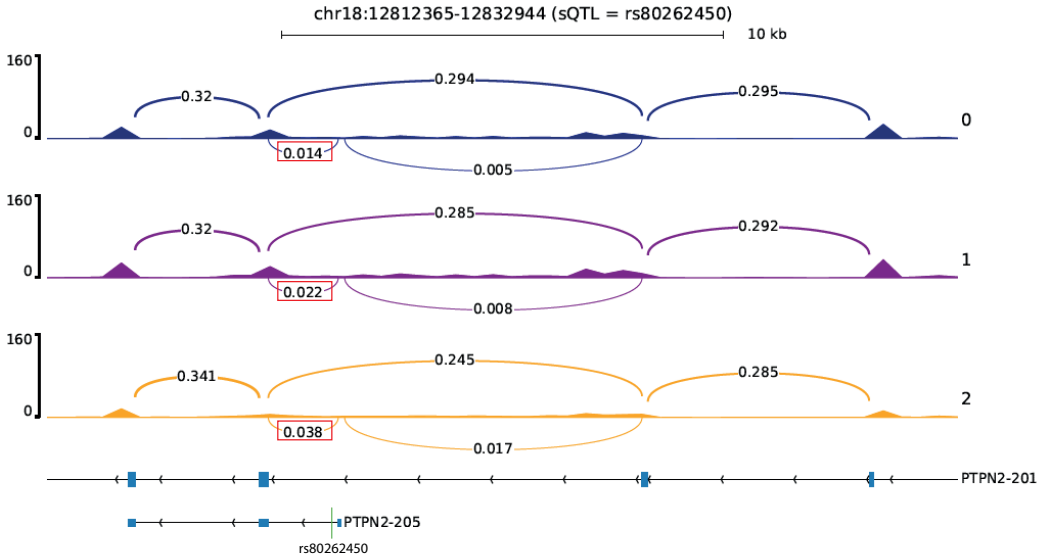


Fig. 2.17 RNA-seq coverage of the intron cluster where the *PTPN2* sQTL effect is detected in sLPS_6. Bars represent the number of reads and arcs represent the usage of different introns (only five splice junctions are shown for clarity and the colocalised sQTL splice junction is indicated in a red box). Canonical transcript *PTPN2-201* and non-canonical transcript *PTPN2-205* are shown underneath, with blue boxes representing introns and the position of rs80262450 on *PTPN2-205* is shown as a green line.

The directions of effects of rs80626450 on intron usage and IBD risk suggest that an increase in the relative abundance of *PTPN2-205* is associated with increased risk of IBD

(for example the effect size in sLPS_6=1.22 and IBD odds ratio=1.17, respectively). Given that mouse knock-out studies of *PTPN2* suggest the gene plays an anti-inflammatory role in macrophages, I hypothesise that increased usage of *PTPN2-205* attenuates the role of *PTPN2* as a negative regulator of inflammation, which in turn increases the risk of IBD.

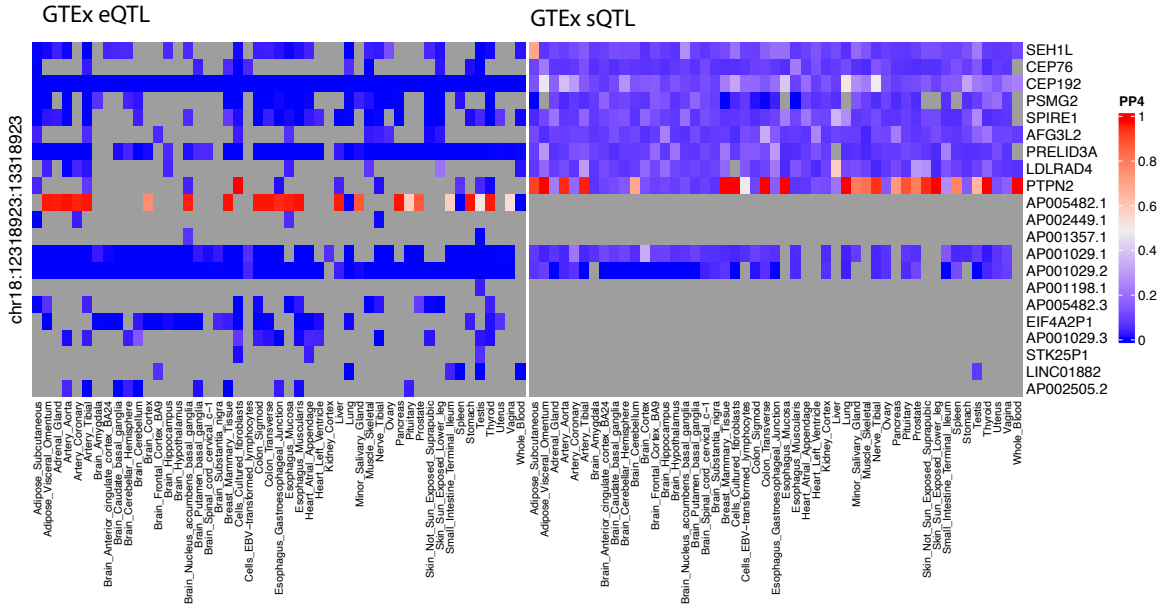


Fig. 2.18 Heatmap of PP4 values for all genes in 18p21.11 (rows) and GTEx tissues (columns) split by type of QTL (eQTL on the left and sQTL on the right). For sQTLs, the intron with the highest PP4 value is shown. Although *AP005482.1* shows strong colocalisation with GTEx eQTLs, it was removed from the current ENSEMBL release (November 2023).

2.3.8 sQTL colocalisations converge on dysregulated pathways in IMDs

IMD-sQTL colocalisations in disease-relevant cell types can reveal how genetic variation dysregulates biological pathways that enable the cell to perform its normal functions. For example, I found that two IBD-associated loci colocalised with sQTL signals for genes that interact with the RAB GTPase Rab35. Rab GTPases are a diverse set of molecules that regulate different aspects of vesicle-mediated transport, with over 70 RAB GTPAses discovered so far [**?**]. RAB GTPases are activated by Guanine Exchange Factors (GEF), and subsequently recruit effector proteins, including proteins required for vesicle uncoating, movement, tethering and fusion, which enables cargo trafficking across cell compartments and membranes [**?**].

IBD-associated loci at 12q12 and 1q31.3 colocalise with *LRRK2* and *DENND1B* sQTLs in 5 and 21 conditions, respectively (PP4 $\geq$ 0.75; conditions with highest PP_4 are shown in

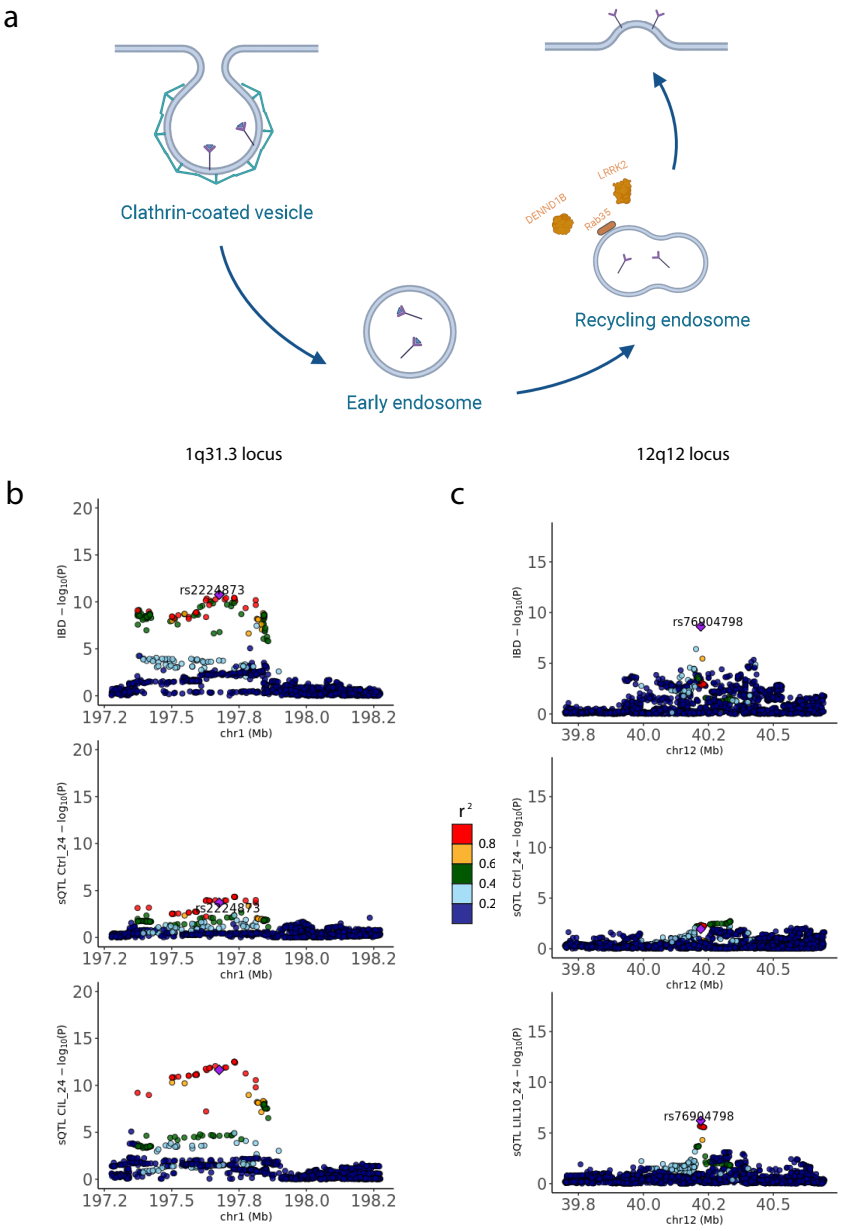


Fig. 2.19 (a) Diagram showing different stages of endocytosis and endosomal recycling. Rab35 and two of its interactors *LRRK2* and *DENND1B* are shown on a recycling endosome (created with BioRender.com). Regional Manhattan plots of the two IBD-associated loci are also shown (b) IBD-associated locus 1q31.1 and *DENND1B* sQTL in naïve and stimulated macrophages (LIL10_24) and (c) IBD-associated locus 12q12 and *LRRK2* sQTL in naive and stimulated macrophages (CIL_24). Panels b and c show high colocalisation evidence with the IBD-associated loci in the stimulated conditions, but not in the naïve conditions.

Figure 2.19). *DENND1B* is a GEF that activates Rab35, while Rab35 was shown to be a bona fide substrate for *LRRK2* [**? ?**], although the exact effect *LRRK2* has on Rab35 is still debated [**? ? ?**]. Rab35 regulates endosomal recycling of both integrins and cadherins, effectively maintaining a balance between cell adhesion and cell migration. Moreover, evidence suggests that endosomal recycling in epithelial cells helps maintain apical and basolateral polarity, and that Rab35 plays a role in maintaining this polarity [**? ?**]. The fact that both *LRRK2* and *DENND1B* interact with Rab35 suggests that endosomal recycling is frequently dysregulated by common IBD-associated variation.

Two *DENND1B* splice junctions with different donor splice sites colocalise with the IBD-associated signal 1q31.3 (chr1:197,715,074-197,772,868 and chr1:197,715,074-197,735,586). The first splice junction is located in the coding sequence of the canonical transcript of *DENND1B* (*DENND1B-211*), while the second is part of the 5' untranslated region of another protein-coding transcript (*DENND1B-201*). Lead sQTL SNPs for both transcripts have opposite directions of effect in different stimulation conditions where the colocalisations are detected (effect size and PP4 for *DENND1B* splice junction chr1:197,715,074-197,772,868 is shown in Figure 2.20) . This suggests that the effect of the risk-increasing allele may have opposite effects on relative transcript abundances is different environmental contexts.

Unlike *DENND1B*, all *LRRK2* sQTLs with high colocalisation probabilities (PP4 $\geq$ 0.75 in 5 conditions) had consistent directions of effects in the conditions where the colocalisation was detected. For example, the *LRRK2* sQTL with the highest colocalisation probability was observed in LIL10_24 (PP4=0.98), where the risk-increasing allele of the lead GWAS SNP (rs76904798) decreases the usage of an *LRRK2* splice junction (chr12:40,316,403-40,319,988; beta=-0.75), and also increases the risk of IBD (beta=0.105). This indicates that decreased usage of this splice junction increases risk of IBD, and conversely that inclusion of this splice junction may have a protective effect against IBD (Figure 2.20).

Although I demonstrate how *LRRK2* and *DENND1B* contribute to a dysregulated endosomal recycling in IBD in different environmental contexts, more research is still needed to understand the exact mechanisms by which the isoforms of these two genes affect such an important biological process in macrophages.

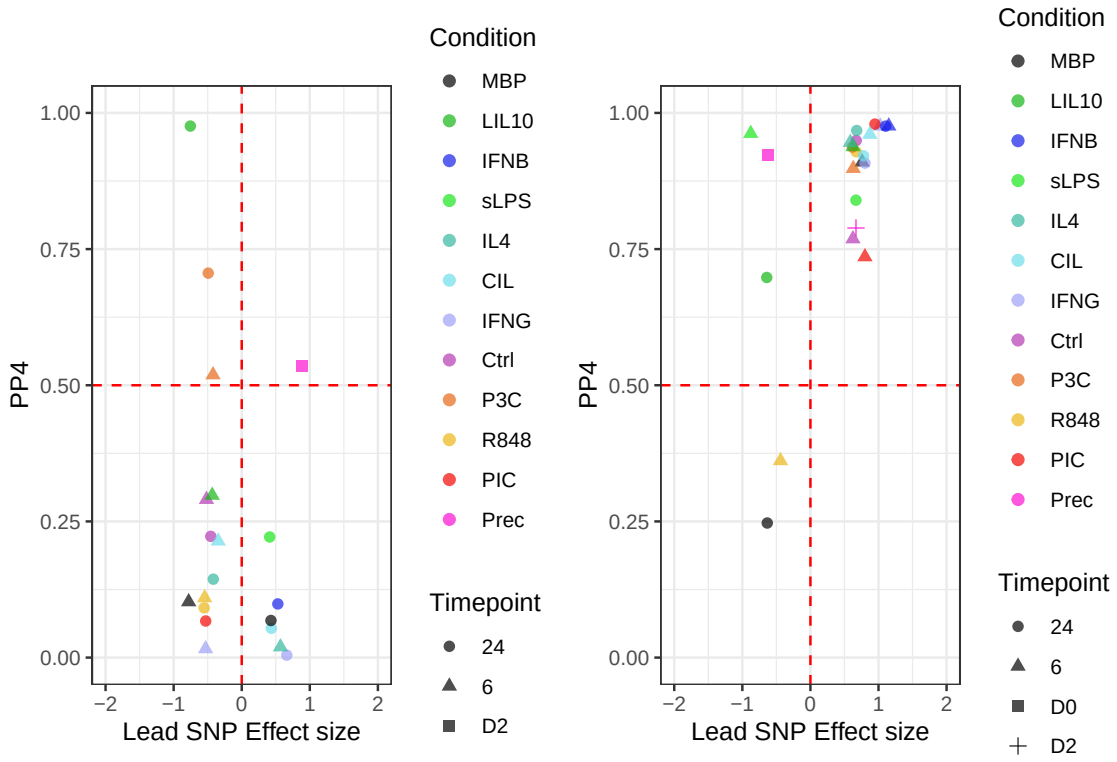


Fig. 2.20 Effect sizes of the most significant SNP and colocalisation PP4 for the colocalised splice junctions of LRRK2 (left) and DENND1B (right).

2.4 Discussion

In this work, I mapped splicing QTLs in iPSC-derived macrophages to understand how splicing is genetically controlled in macrophages in 24 conditions. I found thousands of genes with significant sQTL effects across our array of conditions, and that a considerable proportion of these genes have different effect sizes upon stimulation and were thus defined as response sQTLs.

The primary motivation behind several QTL mapping efforts is to understand the transcriptomic consequences of disease-associated genetic variation. Recently, Mostafavi et al. 2022 (ref 8) suggested that eQTL and GWAS studies are powered to discover systematically different types of variants, and that this may partially explain the limited colocalisation between eQTLs and GWAS risk loci. More than half of the IMD-associated risk loci that I tested likely share a single causal variant with either an eQTL or sQTL, with sQTLs solely contributing half of these loci, which clearly demonstrates the added value of sQTLs. This echoes previous work that showed the promise of sQTLs in closing the colocalisation gap [23**?** **?**]. Mostafavi et al. (ref [**?**]) proposed that systematic differences between GWAS and eQTL association signals are behind this colocalisation gap. Along the same lines, I attribute the large number of sQTL colocalisations to three important features of the discovered sQTLs that make them likely to colocalise with GWAS association signals. First, unlike eQTL variants, sQTL variants tend to be located closer to intron splice sites than gene TSS (Figure 2.12a), and are thus closer to intronic GWAS association signals that do not colocalise with eQTL signals. Indeed, 63.4% of the lead SNPs in the GWAS loci that I tested were intronic variants. Second, sQTL signals largely do not colocalise with eQTL signals, suggesting that they are driven by distinct genetic associations (Figure 2.12b). Third, I show that a considerable number of colocalisation events implicate unannotated introns that are lowly used across transcripts. These subtle changes in intron usage are unlikely to be reflected in overall levels of gene expression levels and may therefore remain undetected in eQTL studies.

Currently, it is unclear whether lowly-used splicing events are simply splicing errors [**?** **?**] or if they have functional consequences[**?** **?**]. For example, Pickrell et al. 2010 (ref [**?**]) interpreted the lack of evolutionary conservation around lowly-used splice sites as evidence that they are functionally irrelevant. This interpretation has been debated over the past decade [**?**]. Here, I show that many of these lowly-used splice junctions may underpin disease-associated genetic effects on alternative splicing, and should not be considered noisy and functionally irrelevant. The *PTPN2* example is particularly intriguing as the implicated splice junction falls within an Alu element, a class of Short Interspersed Nuclear Elements

(SINE; Dfam E-value=4.7×10^{-98}) that constitute 11% of the human genome [**?**]. RNA binding proteins normally repress the expression of newly-incorporated Alu elements, but decreased repression over long evolutionary periods provides a substrate for new functions [**? ?**]. The risk-increasing effect of the lowly-used *PTPN2* splice junction could therefore represent a harmful evolutionary byproduct that attenuates the anti-inflammatory effect of *PTPN2*. This remains to be validated by functional studies that profile the functional consequences of *PTPN2* isoforms that contain this splice junction. With the recent success of RNA therapeutics such as splice-switching antisense oligonucleotides (ASO), it may be possible to "contain" these evolutionary side effects via therapeutic interventions that decrease the proportion of these lowly-used splice junctions [**? ?**]. This should provide incentive for the complex disease and transcriptomic community to understand and interrogate the functional consequences of the splicing events that underpin complex disease risk.

Although this dataset represents the largest resource for studying how alternative splicing is genetically controlled in iPSC-derived macrophages, I acknowledge two main limitations. First, although macrophages differentiated using MacroMap's experimental protocol have been shown to be transcriptionally similar to monocyte-derived macrophages [14], they still do not capture the local environment of tissue-resident macrophages. This may limit the ability to understand the effect of local micro-environments on the transcriptome of macrophage [**?**]. Second, intron usage ratios do not provide information about the relative abundance of full transcripts. For example, I demonstrated the potential effects of splicing events on the canonical transcripts of *PTPN2*, but it is still unclear which transcripts these splicing events may have originated from. Fortunately, long-read sequencing and its algorithms for isoform quantification are becoming increasingly mature [**? ? ?**]. I therefore expect that a more direct quantification of transcript usage will be attainable, and that it will make alternative splicing a more routine part of transcriptomic profiling studies.

In summary, these findings highlight an important role for alternative splicing in macrophage response to environmental cues, and that its dysregulation explains a large proportion of IMD-associated susceptibility loci. I anticipate that improved long-read sequencing technologies will facilitate whole isoform quantification in different cellular contexts, which will open the door for a better understanding of the role of different isoforms in innate immune response. Finally, I recommend that alternative splicing quantification should be an integral part of future QTL studies (including single cell studies), and that it will be increasingly relevant to understand the transcriptomic effects of disease-associated risk loci.

Chapter 3

Epidemiological and genetic characterisation of perianal Crohn's Disease

3.1 Contributions

Genotype and imputation quality control was performed by Dr. Laura Fachal and kinship analysis was performed by Dr. Marcus Tutert as part of the ongoing International IBD Genetics Consortium GWAS project that is being undertaken in the Anderson laboratory. HLA allele imputation was performed by Dr. Qian Zhang as part of his IBD-BR drug response and disease progression study, and UKIBDGC HLA allele imputation was performed by Dr. Loukas Moutsianas. I performed all the GWAS analyses, meta-analyses and all downstream analyses described in this chapter.

3.2 Introduction

Perianal Crohn's disease (pCD) is a sub-phenotype of Crohn's disease, a chronic inflammatory disease of the gut that affects 1% of the population worldwide. pCD represents a major burden on both patients and healthcare providers, and is estimated to affect 20-40% of CD patients worldwide, with a higher prevalence in Asia than in Western countries [**?**]. As their disease progresses, CD patients become more likely to develop perianal symptoms. Twenty years after their CD diagnosis, CD patients have a 32% cumulative probability of developing pCD [**?**]. Timing of pCD diagnosis, however, varies significantly between healthcare systems. Previous studies from countries including France, Sweden and Japan

have reported that between 4%-68% of pCD patients present with perianal symptoms before or at the time of CD diagnosis [**? ? ?**].

Clinical picture of pCD

pCD patients present with a variety of perianal symptoms. These include perianal skin tags, fissures, ulcers, faecal incontinence, rectal discharge and bleeding, perianal abscess, and fistulas. Perianal fistulas are the most common form of pCD, followed by perianal abscess [**?**]. Likewise, the impact of pCD on patients is multi-faceted. In addition to physical manifestations, patients report impaired quality of life as well as social and emotional complications of pCD. Furthermore, surgical interventions that aim to treat pCD and restore normal ano-rectal functions, such as seton insertion and fistula drainage, often impact essential functions such as walking and sitting [**?**]. Moreover, pCD patients, who often require multiple surgical interventions, suffer from high recurrence and relapse rates of pCD. In fact, only one third of pCD patients are estimated to achieve remission [**? ?**].

Despite the diversity in presentation and course, pCD patients share a number of Crohn's disease characteristics. pCD is more common among patients with more distal than proximal disease, and patients with colonic or rectal CD are more likely to develop or initially present with pCD. Moreover, pCD tends to drive Crohn's disease towards a more invasive behaviour. Initially, two thirds of patients have inflammatory manifestations, but over time, the majority of pCD patients display increasingly stricturing and penetrative characteristics of CD [**? ?**], which are characterised by a narrowing of the lumen, and development of abdominal fistulas, inflammatory masses and abscess [**?**]. However, invasive distal CD does not always precede pCD, which can sometimes present a diagnostic challenge in clinical settings. Although 95% of patients will eventually develop luminal disease, an estimated 17.2% of patients initially present with pCD only [**?**].

Pathogenesis of pCD

At a more fundamental level, our biological understanding of pCD fistula formation and progression mechanisms is still markedly lacking. One proposed pathophysiological mechanism is epithelial-to-mesenchymal transformation (EMT). EMT is a well-studied biological process, whereby polarised epithelial cells gain mesenchymal functions, such as enhanced cell invasion, and migration (reviewed in [**?**]). The EMT hypothesis is supported by the

presence of transitional cells which express both epithelial and mesenchymal cell markers in fistula tracts. These include epithelial markers cytokeratin 8 and 20, and mesenchymal markers vimentin and actin. Transforming growth factor β and interleukin-13, which have been associated with the initiation of EMT, have also been identified in transitional cells lining pCD fistula tracts [**?** **?** **?**]. Despite these observations, little is understood about the causal biology of pCD. What are the drivers of this transformation? What causes variation in fistulising disease severity? Which biological pathways give rise to fistulas and what facilitates their development into complex branching structures? What is the role of genetic variation in pCD predisposition? In this regard, genome-wide association studies have improved our understanding of the pathophysiology of several complex disease [**?**]. In the case of pCD, a well-powered GWAS between CD patients who develop pCD and CD patients who do not can help us understand which effector genes and biological pathways are causally linked to pCD risk. Unfortunately, none of the pCD GWAS conducted so far were able to identify genome-wide significant variants associated with pCD risk. Some studies have investigated nominally significant association to better understand pCD biology, but the hypotheses about causal biology remain difficult to reconcile. For example, based on a GWAS of 1,720 CD patients with and without pCD, Kaur et al. [**?**] report an enrichment of nominally-associated variants in genes implicated in the JAK/STAT pathway, a proinflammatory signalling cascade that has previously been implicated in several autoimmune diseases [**?** **?**]. More recently, Akhlaghpour et al. [**?**] found a nominally-associated coding variant that impaired macrophage phagocytosis, and hypothesised that it may contribute to the pathogenesis of fistulising pCD. But overall, there is no clear consensus on the genetic underpinnings of pCD risk.

Available pCD cohorts

The NIHR IBD-BR is a UK-wide collaborative project that is part of the NIHR Bioresource, with the aim of recruiting 50,000 patients with Crohn's disease, ulcerative colitis or unclassified IBD. The IBD-BR collects phenotypic and epidemiological information (both clinical and self-reported) as well as DNA samples for both array genotyping and whole-genome and whole-exome sequencing. The aims of the IBD-BR are wide-ranging. These aims include understanding the genetics of IBD response to therapy, disease mechanism as well as determinants of disease course [**?** **?** **?**]. So far, the IBD-BR has recruited over 31,000 patients, with epidemiological characteristics, clinical phenotypes, extra-intestinal manifestations, prescribed medications and treatment history, surgical history and disease behaviour and complications. The recruitment process starts by an expression of interest by

volunteers who visit participating recruitment centres. Interested volunteers are then provided with an invitation letter and a patient information sheet that provides information on study requirements. Patients who agree to take part are then provided with an informed consent form, and subsequently asked to complete a health and lifestyle questionnaire. After these initial steps, the clinical team then proceeds to collect clinical data from hospital records. A clinician or research nurse extracts core information including IBD type, location and behaviour, complications, comorbidities, family history, smoking history, surgical data and drug therapy outcomes [**?**]. Disease location data include details of perianal manifestations, which can be used to define a pCD case-control cohort.

Another pCD case control cohort can be defined using data from The UK IBD Genetics Consortium (UKIBDGC), a large collaborative consortium that studies the genetics of IBD susceptibility, progression and drug response. Patients are recruited from multiple UK centres in Cambridge, Edinburgh, Manchester, Newcastle, Exeter, Oxford, London, Dundee and Nottingham, and other sites across the UKB [**?**]. In addition to basic epidemiological data such as sex, age, smoking and family history, data on type of IBD, location, surgery, and extraintestinal manifestations is recorded. Disease location data also include whether the disease is located in the perianal region.

Although data on perianal disease are recorded for both cohorts, the depth of clinical phenotyping is different. For example, the IBD-BR, contains information about specific manifestations of pCD. Clinicians and clinical nurses who complete the IBD-BR questionnaire perform an automated search of hospital records for clinical IBD information, including perianal manifestations [**?**]. If the search is unsuccessful, they ask patients about perianal involvement: *"Ever had perianal involvement? 1) Yes 2) No 3) Unknown"* and record the answer in the clinical questionnaire [**?**]. A follow-up question about the type of perianal involvement is then asked: *"If Yes - What type of perianal lesion has the patient had? (Select all that apply): 1) Tags/fissures/ulcers 2) Perianal abscess 3) Simple fistula 4) Complex fistula 5) Other"*. Clinicians may report one or more perianal involvement manifestations. Unlike IBD-BR, the specific manifestations of pCD, such as fissures, ulcers or fistulas are not recorded for UKIBGC participants and only a binary phenotype is recorded (pCD+ or pCD-).

In this chapter, I describe several analyses I conducted to characterise pCD. Using the rich clinical phenotyping in the IBD-BR, I first explored the clinical characteristics of pCD+ patients, which largely conformed with what is known about the disease characteristics of

pCD. Moreover, given our limited understanding of the genetic underpinnings of pCD [**? ? ? ? ?**], I also performed a pCD GWAS meta-analysis leveraging the pCD cohorts of IBD-BR and UKIBDGC to identify pCD-associated variants. I conclude with an analysis that may partly explain the discovered pCD-associated hit and outline future steps for a more comprehensive understanding of the genetic underpinnings of pCD.

3.3 Methods

3.3.1 pCD prevalence estimates

IBD-BR patients were diagnosed with CD over several decades, mostly from 1980 till 2018. To investigate the temporal trends of pCD prevalence, I divided participants by year of CD diagnosis into 39 windows, and calculated pCD point prevalence in each two-year period. Additionally, to calculate 95% confidence intervals around each point estimate, I randomly subsampled CD patients 1,000 times using a bootstrap procedure implemented in the `boot()` functions. Finally, I observed that pCD prevalence starts decreasing in patients diagnosed with CD after the year 2010. Therefore, I compared overall trends in pCD prevalence estimate by pooling CD participants diagnosed with CD before and after 2010 and calculated point estimates and confidence intervals as mentioned before. The difference between these overall estimates was then tested using a t-test to determine if pCD prevalence has significantly decreased before and after 2010.

3.3.2 UK IBD Genetics Consortium Genotype Quality Control

UKIBDGC samples were genotyped with two genotyping arrays: Affymetrix Human Mapping 500K Array (I will refer to this as GWAS1; number of variants before QC=469,281), and Illumina Human Core Exome-12v1.0 or its newer version Illumina Infinium Core Exome-24v1.1 (I will refer to this as GWAS2; number of variants before QC=535,434 and 557,662 respectively). Quality control for UKIBDGC genotype data was performed as part of the International IBD Genetics Consortium cases-control meta-analysis. QC was performed using a combination of Plink (v1.9 and v2), bcftools (v1.16), and KING (v2.2.4).

3.3.3 Variant-level QC

Variants that met the following criteria were excluded:

- Low call rate (< 0.95 for variants with minor allele frequency (MAF) > 0.01 or < 0.98 for variants with MAF ≤ 0.01).

- Significant difference in genotype call rate missingness (P-value $< 10^{-4}$) between IBD cases and controls.
- Large allele frequency (AF) differences between UKIBDGC and Gnomad (Non-Finnish Europeans), or TOPMed (global) using the following formula:

$$\frac{(P_1 - P_0)^2}{(P_1 + P_0)(2 - P_1 - P_0)} > \varepsilon$$

  where $\varepsilon = 0.025$ or 0.125, for Gnomad and TOPMed respectively, P_0 is the minor allele frequency (MAF) in Gnomad or TOPMed and P_1 is UKIBDGC MAF. This formula accounts for larger AF differences between UKIBDGC and population references in common than in low-frequency variants. The TOPMed global AF difference cutoff is higher to account for AF computed across diverse populations
- Hardy Weinberg Equilibrium (HWE) P-value $< 10^{-5}$ in IBD controls or $< 10^{-12}$ in IBD cases. or
- Monomorphic variants.

3.3.4 Sample-level QC

Samples that meet the following criteria were excluded:

- Missing genotyping rate > 0.05
- Heterozygosity estimate $\pm$ 4 standard deviations from the European-ancestry mean, or
- mismatch between recorded gender and genotypically-inferred sex.

3.3.5 Imputation to TOPMed

The TOPMed imputation server (imputationserver at 1.5.7) was used for UKIBDGC genotype imputation. Alleles at directly genotyped variants with an empirical imputation $R < -0.5$ were flipped, and variants with empirical $R^2 \leq 0.5$ were excluded after imputation. After their exclusion, imputation was repeated, and another HWE filtering step was performed.

3.3.6 Continental ancestry principal components

Genotypic principal components (PC) were estimated for all participants, using a set of genotyped variants that were also available in the 1000 Genomes Project (100GP; excluding

variants associated with IBD susceptibility (P-value < 10^{-4}), and variants in long LD regions (as defined in [**?**]). This final list was pruned with the following parameters: window size = 50 kbp; step size = 5; R^2 = 0.2. This set of variants was used to compute PCs using 1000GP samples, and PCs that define European ancestry were retained. IBD-BR and UKIBDGC samples were then projected on continental ancestry PCs, and samples within the European ancestry group were retained for the subsequent analyses. Within European samples, European-ancestry PCs were then calculated and were used as covariates to account for cryptic population stratification in the GWAS analysis described in subsection 3.3.9.

3.3.7 IBD-BR Genotype QC and Imputation

The cohort was genotyped with two different versions of the UK BioBank Thermo Fisher genotyping array. The same genotype QC steps as UKIBDGC were applied to IBD-BR, except for 1) The AF difference check, where 1000 Genomes Panel (1000GP) was used as a reference panel 2) Imputation, where the Sanger Imputation Server was used [**?**], with two imputation reference panels: UK10K+1000GP and HRC. Imputed genotypes from both panels were combined. For variants that existed in both panels, HRC imputed genotypes were retained.

3.3.8 Identification of overlapping samples between UKIBDGC and IBD-BR

Identification of duplicate individuals between UKIBDGC and IBD-BR genotyping data was performed with KING [**?**]. Duplicates were defined as sample pairs with a kinship coefficient > 0.354 as recommended in KING documentations [**?**]. Estimation of kinship coefficient was performed using post-QC genotyped SNPs (number of variants used for kinship inference between IBD-BR and GWAS1=42,292; and between IBD-BR and GWAS2=53,431).

3.3.9 Genome-wide association analysis

All genome-wide association analyses were performed using REGENIE v3.2.5 [**?**] following a 2-step approach. This approach is more computationally efficient than other approaches that account for cryptic relatedness between individuals, such as linear mixed models. Briefly, in step 1, a whole-genome regression model is fitted using a subset of high-quality genome-wide variants in order to estimate a set of genome-wide predictors that capture a large fraction of phenotypic variance. These predictors are then included as covariates in the single-variant

association models tested in step 2, where a larger set of variants of interest are tested for association. I used post-QC genotyped variants in step 1 as recommended by REGENIE documentation (N_{IBD-BR}=338,697; $N_{UKIBDGC(GWAS1)}$=436,931; $N_{UKIBDGC(GWAS2)}$=359,209;), and both genotyped and imputed variants in step 2, testing all autosomal chromosomes (N_{IBD-BR}=9,777,139; $N_{UKIBDGC(GWAS1)}$=8,897,554; $N_{UKIBDGC(GWAS2)}$=,916,200;). The step 2 model was specified as following: pCD $\sim$ variant + sex + genotypic PCs, using the first four European-ancestry population PCs. Step 2 reports single-variant association summary statistics.

3.3.10 Meta-analysis of IBD-BR and UKIBDGC cohorts

I used METAL to perform a fixed-effects meta-analysis between IBD-BR and UKIBDGC summary statistics. METAL can perform fixed-effects meta-analysis using one of two different well-established schemes: P-values and effective sample size, or effect sizes and standard errors. The P-value scheme is implemented to enable meta-analysis of GWAS summary statistics that do not report the effect allele, while the effect sizes scheme can be used when each variant's effect size and effect allele are reported. All my pCD GWAS analyses report the effect allele, so I used the effect size scheme of METAL (`SCHEME STDERR`).

There was a total of 8,473,930 overlapping variants across the meta-analysed summary statistics, and an additional 1,645,123 variants that were unique to one of the studies, 42.7% of which were indels. Given that 16% of variants were unique to one of cohorts, I did not remove them from their respective summary statistics file. It is important to note, however, that this choice may favour variants that are available in all studies.

Table 3.1 Number of SNPs and indels in each of the three GWAS summary statistics.

Study	**SNP**	**Indel**	**Total**
IBD-BR	8,626,072	1,150,933	9,777,005
UKIBDGC (GWAS1)	8,307,857	589,198	8,897,055
UKIBDGC (GWAS2)	8,325,721	589,997	8,915,718

Moreover, METAL automatically aligns any variants that may be flipped between the meta-analysed summary statistics. METAL also enables filtering of variants to be meta-analysed based on their allele frequencies, which was not necessary since I previously filtered out variants with MAF < 0.01 in each summary statistics file. Finally, given the potential subpopulation stratification in the IBD-BR GWAS (λ_{GC}=1.08), I enabled a METAL option

to correct genomic inflation before performing the meta-analysis (`GENOMICCONTROL ON`) as recommended in METAL's documentation website. There was no evidence of genomic inflation in the meta-analysed summary statistics (λ_{GC}=1.03).

For each variant, METAL outputs the effect allele, meta-analysed effect size, standard error, and P-values. After performing meta-analyses, it is important to compare the effect sizes between the meta-analysed cohorts. Comparison of both the direction and magnitude of effect sizes gives an indication on how similar the estimated effects of meta-analysed genetic variants are. To formally test this, I used Cochran's Q test of effect size heterogeneity implemented in METAL. Cochran's Q test assesses two or more effect size estimates and their corresponding standard errors and reports a χ^2 statistic that quantifies the deviation from the null hypothesis that the meta-analysed effect sizes are similar. Depending on the number of meta-analysed studies (in this case 3), a P-value is derived from a theoretical χ^2 distribution with $N-1$ degrees of freedom, where N is the number of meta-analysed studies (heterogeneity of effect P-value P_{het}). I used P_{het} to test if the genome-wide significant variants demonstrate heterogeneity of effect size between the meta-analysed cohorts. To account for multiple variants being tested, I set a Bonferroni-corrected P-value threshold for rejecting the null hypothesis (P-value < $\frac{0.05}{k}$, where k is the number of variants tested).

3.3.11 LD calculation from 1000GP

Reference LD panels obtained from the 1000 Genomes High Coverage project [**?**] were used in the post-GWAS check to study the relationship between LD and association strength at the genome-wide significant locus. R^2 values were calculated between the index variant and all variants in the locus. I downloaded VCFs from 1000GP and used PLINK v1.9 to compute LD between all variants and the index variant within the locus boundaries. I used unrelated individuals with non-Finnish European ancestry (NFE; N=426). Relevant samples were included in the LD calculation using the following PLINK command:

```
plink --r2 --keep EUR.samples --ld-window-r2 0
```

3.3.12 χ^2 comparison between different pCD definitions

In order to compare association statistics from the pCD meta-analysis to meta-analyses performed using more severe pCD+ case criteria (all perianal manifestations, abscess and fistula only, fistula only and complex fistula only), I adjusted the broad-definition χ^2 values using this formula:

$$\chi^2_{Broad,n} = \frac{n}{N}\chi^2_{Broad}$$

where n is the sample size of the meta-analysis being assessed, χ^2_{Broad} is the broad-definition observed association statistic and $\chi^2_{Broad,n}$ is the broad-definition association statistic adjusted for sample size. This adjustment ensure that comparison of association statistics from meta-analyses with different sample sizes is valid.

3.3.13 HLA allele imputation

HLA genes located in the major histocompatibility complex region (MHC) are known to contribute to immune disease susceptibility [**?**]. Although sequence-based typing (SBT) is the gold standard to identify HLA alleles, its relatively higher cost and the complexity of HLA sequencing has prevented scaling up of SBT methods to large cohorts [**?**]. HLA allele imputation methods based on SNP arrays are a reliable alternative to SBT methods of HLA typing, and can be performed using a small number of genotyped SNPs in each HLA gene [**? ? ? ? ? ?**] (reviewed in [**?**]).

HLA alleles were imputed for all IBD-BR and UKIBDGC individuals using HIBAG, a computationally efficient prediction algorithm that was pre-trained on a diverse set of haplotypes from different ancestries and is used to impute HLA alleles [**?**]. Model parameters that were pre-trained on individuals from European ancestry with SNPs measured on UK Biobank Affymetrix Axiom array were used in HLA imputation (downloaded from [**?**]). After downloading the pre-trained model parameters, HLA imputation was performed at the HLA allele and HLA allele group levels for a total of 7 HLA genes (4-digit and 2-digit resolutions, respectively; Table 3.2).

Table 3.2 Number of genotyped variants used to perform HLA imputation.

Gene	2 digits	4 digits
HLA*A	723	717
HLA*B	763	752
HLA*C	819	770
HLA*DPB1	478	478
HLA*DQA1	753	655
HLA*DQB1	754	702
HLA*DRB1	675	653

For each HLA allele, I performed the association test using the logistic regression model: pcd status $\sim$ HLA allele copies + covariates, with the R function `glm(family=binomial())` and using the same covariates as used in the GWAS described in section 3.3.9. The association analysis were performed separately for each of IBD-BR, UKIBDGC (GWAS1) and UKIBDGC (GWAS2). The association effect sizes and standard errors were subsequently meta-analysed using the `rma(method="EE")` function from the R package `metafor`. Additionally, I performed conditional association analysis to investigate if any HLA alleles can account for genome-wide significant SNPs. Conditional association analyses were performed by including SNP dosages as covariates in the same model.

3.4 Results

Among 30,894 IBD-BR participants, 15,152 were diagnosed with Crohn's disease, 14,819 of which had perianal involvement data: 4,448 answered "*Yes*" to "*Ever had perianal involvement?*" (pCD+; 30%), 9,751 answered "*No*" (pCD-; 65.8%), and 620 answered "*Unknown*" (4.1%), matching previous pCD prevalence estimates. Perianal simple or complex fistula was the most common manifestation (2327; 52.3% pCD+ participants), followed by perianal abscess (1806; 40.5% of pCD+ participants).

From 26,327 UKIBDGC patients, a total of 8,977 were diagnosed with CD. 7106 CD patients had perianal involvement information: 1,631 of CD patients reported perianal disease location, 5,475 reported a different disease location, and 1,871 answered "*Unknown*". UKIBDGC does not report specific manifestations of pCD.

3.4.1 Epidemiological characteristics

Epidemiological characteristics of pCD+ and pCD- patients were largely similar in both cohorts (Table 3.3). Males were more likely than females to report perianal involvement in both cohorts (P-value=7×10^{-4} and 8×10^{-6} in IBD-BR and UKIBDGC respectively). pCD+ was not associated with a family history of CD, while smoking was slightly less common in pCD+ patients (P-value=0.006 and 0.003).

Table 3.3 Epidemiological characteristics of pCD+ and pCD- patients in IBD-BR and UKIBDGC

	IBD-BR		UKIBDGC	
	pCD+	pCD-	pCD+	pCD-
Male	2115 (47.5)	4339 (44.5)	807 (49.5)	2363 (43.2)
Female	2333 (52.5)	5412 (55.5)	824 (50.5)	3112 (56.8)
Family History	1325 (34.7)	2795 (34.2)	290 (27.1)	598 (24.6)
Surgery	2971 (68.8)	3636 (38.3)	896 (63.1)	1935 (42.6)
Smoking	656 (16.4)	1572 (18.2)	363 (30.1)	913 (29.6)

3.4.2 Clinical characteristics

pCD is associated with lower age-of-CD-diagnosis and rectal CD

Previous pCD studies have reported an association between pCD and distal penetrating CD, as well as pCD and an earlier age of CD diagnosis [**?** **?**]. Compared to pCD- patients, pCD+ patients were significantly younger at diagnosis (t-test P-value $< 2 \times 10^{-16}$; median age of CD diagnosis for pCD+ patients was 24 versus 29 for pCD- patients in IBD-BR; Figure 3.1). Additionally, pCD+ patients were at least twice as likely to have penetrating disease behaviour. In IBD-BR, 19.1% of pCD+ patients had disease behaviour classified as B3 versus 8.1% in pCD-. This enrichment was stronger in UKIBDGC (28.5% versus 10.6%, respectively).

In IBD-BR, more CD patients reported ileal than colo-rectal CD (68.7% versus 56.3% in IBD-BR), while in UKIBDGC both were equally represented (60.4% versus 58%). Ileal and colo-rectal CD were either isolated, or extended to other parts of the gut. In IBD-BR, patients with an isolated colo-rectal CD were 2.4 times as likely to report pCD, compared to patients with an isolated ileal CD (59.3% versus 24.8%; χ^2 test P-value $< 2 \times 10^{-16}$). Despite the lower pCD prevalence in UKIBDGC, patients with isolated colo-rectal CD were similarly enriched with pCD+ patients compared to patients with an isolated ileal disease (26.6% versus 10.9%; χ^2 test P-value $< 2 \times 10^{-16}$).

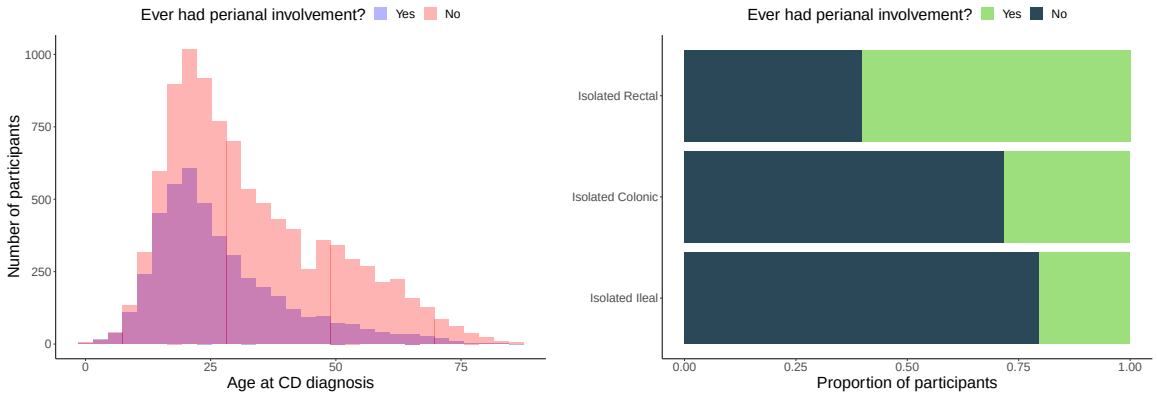


Fig. 3.1 Age at diagnosis (left) and macroscopic extent (right) of Crohn's disease in pCD+ and pCD- patients in the IBD-BR.

Lower rates of drug intake in pCD+ patients in IBD-BR

pCD+ patients were less likely to be actively prescribed a number of CD medications: oral steroids, Infliximab, Adalimumab, Vedolizumab, and Mesalazine (χ^2 test P-value < 0.006 for nine tests; Table 3.4; odds ratio=0.83, 0.77, 0.62 and 0.55 respectively). Anti-TNF therapies, including Infliximab and Adalimumab, are among the first-line drugs for perianal fistulas, and lead to fistula healing in 50% of pCD patients (in combination with other surgical procedures) [**?** **?** **?** **?**]. Additionally, the ENTERPRISE clinical trial found that Vedolizumab achieved remarkable fistula closure and healing [**?**]. On the other hand, oral steroids are known to be ineffective for fistula closure and may even exacerbate perianal abscess [**?**]. It is important to note that the drug intake data report if patients are *currently* on a given drug, and not whether patients were *ever* prescribed a given drug. Patients may discontinue a given drug after being non-responsive or because they lost response after a period of time, and this lack of response may or may not be related to the patients' perianal manifestations. Therefore, it is difficult to establish whether or not the lower drug intake among pCD+ patients is related to their perianal manifestations, but it suggests higher non-response among pCD+ patients.

pCD+ patients are enriched for six extraintestinal manifestations

pCD+ patients were enriched for extra-intestinal manifestation compared to pCD- patients (26.3% versus 19.6%; P-value < 0.05). Enteropathic arthritis was the most most prevalent extraintestinal manifestation among pCD+ patients, followed by serious infections and psoriasis. In total, six extraintestinal manifestations showed significant enrichment in pCD+ versus pCD- patients (P-value < 0.005; Table 3.4).

Table 3.4 Drug intake and extraintestinal manifestations in pCD+ and pCD- patients in the IBD-BR. Percentage of patients are shown between parentheses. Significant differences between pCD+ and pCD- were assesed using a χ^2 test and P-values are shown in the last column.

	pCD+(%)	**pCD-(%)**	**P-value**
Extraintestinal Manifestations			
Primary Sclerosing Cholangitis	25 (0.6)	72 (0.8)	0.29
Enteropathic Arthritis	413 (9.7)	635 (6.7)	2.1×10^{-9}
Erythema Nodosum	199 (4.6)	222 (2.3)	7.7×10^{-13}
Iritis	183 (4.2)	242 (2.5)	1.1×10^{-7}
Orofacial Granulomatosis	153 (3.6)	162 (1.7)	2.4×10^{-11}
Psoriasis	311 (7.2)	518 (5.4)	6×10^{-5}
Ankylosing Spndylitis	110 (2.6)	238 (2.5)	0.89
Multiple Sclerosis	9 (0.2)	27 (0.3)	0.53
Lymphoma	18 (0.4)	36 (0.4)	0.85
Serious Infections	320 (7.4)	465 (4.9)	3.2×10^{-9}
Drugs			
Azathioprine	1373 (41.4)	2649 (42.1)	0.49
Mercaptopurine	271 (35.6)	564 (36.1)	0.83
Methotrexate	192 (28.6)	404 (35.4)	4×10^{-3}
Infliximab	1207 (50.9)	1622 (55.7)	6×10^{-4}
Adalimumab	740 (47.8)	1368 (54.2)	8×10^{5}
Vedolimumab	236 (67.8)	424 (77.4)	2×10^{-3}
Ustekinumab	171 (69.2)	209 (72.6)	0.45
Mesalazine	528 (30)	1649 (43.7)	$< 2.2 \times 10^{-16}$
Oral Steroids	304 (11.3)	823 (14.2)	3×10^{-4}

Surgical burden of pCD

Combined surgical and medical interventions represent some of the few effective interventions available to pCD patients. Different surgical options are available to perianal disease patients depending on its anatomical features, complications and disease severity. Exploration under anaesthesia and seton insertion are the typical first-line management options, and further medical or surgical interventions are based on initial exploration [**?**]. As expected, 2,971 pCD+ patients (66.8%) had undergone any type of surgical intervention compared to 3,637 (37.2%) of pCD- participants in IBD-BR. In total, almost half the pCD+ patients with operative history had undergone one of three pCD-related surgical procedures (1431 patients; 48.2%): drainge of perianal abscess, insertion of seton, or drainage of fistula. Perianal abscess drainage was the most common: 808 pCD+ patients (27.2% of surgically-operated patients) underwent at least one perianal abscess drainage operation, followed by insertion of a seton suture (744 pCD+ patients; 25%), followed by perianal fistula repair operation (438 pCD+ patients; 14.7%).

pCD prevalence decreased over time

Understanding pCD prevalence over time is important to understand how the burden of pCD on patients and healthcare providers has changed. Previous work has shown reduced pCD incidence over the last decade [**?**], which was partly attributed to improved treatment options. In this regard, IBD-BR offers a unique opportunity to assess this trend. Although the precise time of pCD development is not available, the time between CD diagnosis and the last clinical review can be used to compare how pCD prevalence among CD patients has changed in different eras. Over 93% of IBD-BR patients with pCD information were clinically re-assessed in or later than 2016, which reduces the bias introduced by potentially outdated clinical data.

To investigate this trend in the IBD-BR, I partitioned participants according to their year of CD diagnosis into two-year groups (e.g. 2006-2008), and calculated prevalence estimates in each period. As expected, confidence intervals around the point estimates were larger in the years between 1980-2010 since fewer IBD-BR patients were diagnosed with CD in those years (100-230 participants per two years). This rose to 497-734 per two years in the years from 2010 to 2020. Notably, point prevalence estimates decreased starting from the year 2010 onwards. The mean point prevalence between 1980 to 2010 decreased significantly from 35.9% to 25.1% between 2010 to 2020 (t-test P-value $< 2 \times 10^{-16}$; Figure 3.2). The decrease in prevalence remained significant when mean point prevalence was calculated

between 2010 to 2016 only (mean prevalence=26.8%), between 2010 to 2014 only (mean prevalence=28.1%), or between 2010 to 2012 (mean prevalence=29.4%).

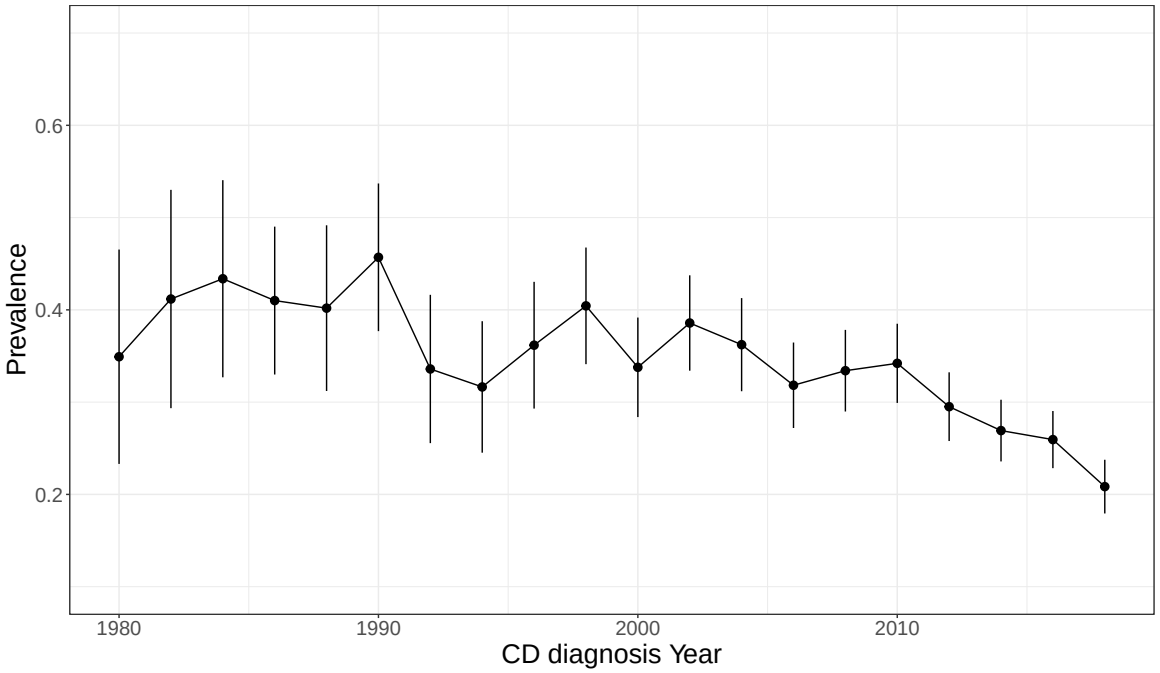


Fig. 3.2 pCD prevalence per year of CD diagnosis, partitioned into two-year groups. 95% confidence intervals around point estimates are calculated using a bootstrap procedure, whereby participants where resampled 1,000 times within each two-year group.

A decrease in pCD prevalence has been observed previously [**?**]. However, such a decrease has not been precisely quantified, partly due to the relatively smaller sample sizes of most studies [**? ? ? ?**]. A potential limitation of this analysis is that censored data may contribute to the observed decrease in pCD prevalence in later years. pCD does not always present at the time of diagnosis, which may bias prevalence estimates downwards in the years after 2010. An important consideration when investigating the effect of data censorship is that pCD prevalence estimates should be up-to-date. For example, a patient who is diagnosed with CD in 2012 for example may be incorrectly considered pCD-free if their clinical information were not updated afterwards. In IBD-BR, this bias is mitigated by the the fact that the majority of patients were clinically assessed between 2016 and 2021. Additionally, the consistent decrease in pCD prevalence even when I excluded patients diagnosed after 2016, 2014, or 2012 indicates that the contribution of censored observation is likely minimal. Although some patients may develop perianal symptoms up to 20 years after diagnosis, the cumulative probability of developing pCD does not increase significantly

after 5 years [**?**]. It is therefore unlikely that the decrease in pCD prevalence is affected by censored observations.

3.4.3 UKIBDGC and IBD-BR definitions of pCD are similar

Similar to IBD-BR, the UKIBDGC can be used to define a pCD+/pCD- cohort. The UK-IBDGC reports a number of clinical and phenotypic characteristics of IBD patients. For each IBD participant, disease subtype diagnosis, and location (including perianal disease) are recorded. However, it is unclear whether the criteria for assigning pCD status is consistent between the different centers, and more importantly if it matches the criteria used to assign pCD status to IBD-BR patients. Heterogeneity in pCD status definition can often arise from different diagnostic criteria being applied, or different times of phenotype update between patients. Ensuring the consistency of pCD status between UKIBDGC and IBD-BR is crucial to minimise the heterogeneity of phenotype definition between the two cohorts and maximise the statistical power gained from a meta-analysis. To assess this, participants who may have taken part in both the IBD-BR and UKIBDGC can be leveraged to understand the level of agreement in pCD status assignment. Since participant identifiers are not mapped across studies, genetic similarity of individuals across cohorts can instead be leveraged to identify overlapping participants (Methods).

Out of 971 overlapping CD participants, only one participant exhibited discordant pCD status between IBD-BR and UKIBDGC. A total of 432 participants had missing or "Unknown" perianal involvement, 406 of which had "Unknown" pCD status in both cohorts (Table 3.5). This strong agreement in indicates that both cohorts assign pCD status in a similar fashion.

I then asked if the UKIBDGC cohort was enriched for particular perianal manifestations. Since this information is not available in the UKIBDGC phenotype data, it can be obtained from the IBD-BR clinical data for the subset of overlapping pCD patients. I found that these patients were not enriched in any particular type of perianal involvement (e.g. 51.9% of overlapping individuals reported either simple or complex fistula versus 52% in IBD-BR). Moreover, 27% of the overlapping patients reported only skin tags, fissures or ulcer, indicating that milder forms of pCD were also included in the UKIBDGC assignment of pCD+ status. Overall, this shows that the definition of pCD status is likely consistent between the cohorts.

Table 3.5 Number of overlapping individuals between UKIBDGC and IBD-BR who answered Yes, No or Unknown to *Ever had perianal involvement?*

	UKIBDGC		
IBD-BR	Yes	No	Unknown
Yes	201	0	1
No	1	337	6
Unknown	6	13	406

3.4.4 Genome-wide association analysis of pCD

Defining pCD+ cases

Genome-wide association studies of disease subphenotypes pose unique challenges compared to traditional case-control GWASes. Unlike GWASes of CD, for example, where robust diagnostic criteria are applied to clearly demarcate cases and controls, in GWAS of disease subphenotype such as pCD it is not obvious which specific manifestations should be considered cases. The IBD-BR questionnaire reports several types of pCD manifestations, including skin tags, fissures or ulcers, perianal abscess, and simple and complex fistulas. In this chapter, my aim is to perform a pCD meta-analysis between IBD-BR and UKIBDGC, and therefore similarity in pCD+ case definition across the cohorts is an important consideration to ensure the robustness of genome-wide significant hits. In the previous section, I showed that leveraging individuals who registered for both studies can give an insight into the composition of the UKIBDGC pCD+ cases. This showed that UKIBDGC pCD+ cases were not particularly enriched in any particular type of perianal manifestations. Additionally, when I inspected the UKIBDGC questionnaire used to collect perianal manifestations data, I found that the relevant question appeared to include all types of perianal manifestations: *"Ever had perianal fistula (incl recto-vaginal), abscess, anal ulcer or significant anal stenosis?"*. I therefore defined pCD+ cases in both cohorts as CD patients that report any type of perianal involvement,

IBD-BR

Although clinical and phenotypic data are available for all participants, not all participants have been genotyped in the current release (04/04/2022). From a total of 15,152 participants with CD diagnosis, 9,458 European ancestry participants with perianal involvement data were genotyped. To ensure that pCD- controls do not include recently diagnosed CD patients who

may develop perianal disease in the near future, I excluded pCD- controls diagnosed with CD less than 5 years before the last clinical review. This choice was informed by previous studies that showed that the cumulative risk of developing perianal disease 5 years and 10 years after diagnosis are similar [**?**]. This resulted in a total of 6,833 participants (2,664 pCD+ cases and 4,169 pCD- controls). After these filters were applied, the composition of genotyped pCD+ cases cohort matched the overall composition of all participants with perianal involvement information reported earlier: 53.6% (1,480) of genotyped pCD+ individuals had either a simple or complex perianal fistula, and 41.2% (1098) had perianal abscess. Together, patients with perianal fistula or abscess account for 74.9% (1995) of genotyped pCD+ cases.

With the pCD case-control cohort defined above, I performed GWAS between pCD+ cases and pCD- controls using REGENIE and used four European-ancestry genotypic principal components and sex as covariates. I removed variants with imputation INFO score < 0.4 and minor allele frequency (MAF) < 0.01, leaving 9,777,139 variants for association analysis (see Methods for detailed genotype and imputation QC). None of the tested variants achieved genome-wide significant association (P-value $< 5 \times 10^{-8}$). There was moderate evidence of genomic inflation (median χ^2=0.49; λ_{GC}=1.08).

UKIBDGC

As mentioned earlier, UKIBDGC only reports whether or not participants report perianal involvement and does not provide specific perianal manifestations. A total of 8,078 genotyped patients of European ancestry were diagnosed with CD, of which 6550 had perianal involvement information. To minimise sample overlap with the IBD-BR, I removed UKIBDGC individuals who showed genetic similarity with individuals from the IBD-BR (see Methods for more details on how genetic similarity was assessed), and performed GWAS with the remaining individuals (1303 pCD+ and 4761 pCD-).

I performed GWAS similar to the IBD-BR analysis, with the difference that UKIBDGC samples were genotyped with two different genotyping arrays and were therefore analysed separately (I will refer to these two cohorts as GWAS1 and GWAS2; Methods). A total of 8,897,554 and 8,916,200 variants were tested in GWAS1 and GWAS2, respectively. No variants achieved genome-wide significant association (P-value $< 5 \times 10^{-8}$). There was no evidence of genomic inflation (GWAS1: median χ^2=0.46; λ_{GC}=1.01; GWAS2: median χ^2=0.47; λ_{GC}=1.04).

3.4.5 Meta-analysis between UKIBGC and IBD-BR: a genome-wide significant locus at 6p21.32

I used METAL to perform a fixed-effects meta-analysis between summary statistics from IBD-BR, and the two UKIBDGC summary statistics GWAS1 and GWAS2, with a total of 3,967 pCD+ cases and 8,930 pCD- controls. Four variants in the MHC region at the 6p21.32 locus showed genome-wide significant association (index variant rs115378818 or 6_32333650_C_T; P-value=8.6×10^{-12}; Table 4.4 and Figure 3.3). None of the variants showed significant heterogeneity of effect size between the constituent cohorts ($P_{het} < 0.01$ for four variants). All four variants were well-imputed across the constituent cohorts (INFO score > 0.7).

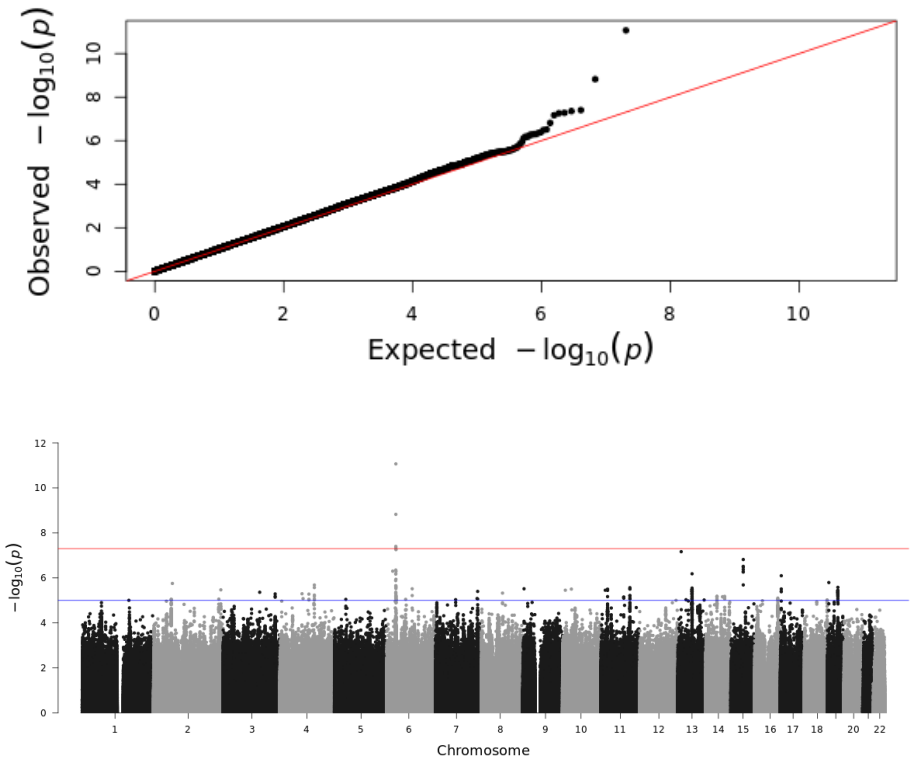


Fig. 3.3 (Top) Quantile-quantile plot for the meta-analysis between UKIBDGC and IBD-BR cohorts, suggesting a good fit to the uniform distribution, and showing no evidence of genomic inflation (median χ^2=0.47; λ_{GC}=1.03; median χ^2 was calculated by converting P-values to χ^2 values using the function `qchisq(P, df=1,lower.tail=F)` in R v4.1.0). (Bottom) Manhattan plot of meta-analysis between IBD-BR and UKIBDGC. pCD+ cases are defined as CD patients with any type of perianal involvement and pCD- controls are defined as CD patients with no perianal involvement.

Table 3.6 Genome-wide significant variants in the 6p21.32 locus. Odds ratio and their 95% confidence intervals are shown. MAF=minor allele frequency.

Chromosome	Position (b38)	Effect Allele	OR	P-value	MAF
6	32,205,822	C	1.45 (1.27 - 1.66)	4×10^{-8}	0.05
6	32,243,461	C	1.38 (1.23 - 1.55)	4.4×10^{-8}	0.08
6	32,279,268	G	1.57 (1.36 - 1.82)	1.5×10^{-9}	0.05
6	32,333,650	T	1.78 (1.51 - 2.1)	8.6×10^{-12}	0.04

Table 3.7 Case and control minor allele frequencies of the four genome-wide significant variants by cohort.

	IBD-BR		UKIBDGC (GWAS2)		UKIBDGC (GWAS1)	
SNP	Cases	Controls	Cases	Controls	Cases	Controls
6:32333650_C_T	0.042	0.027	0.059	0.037	0.055	0.035
6:32279268_T_G	0.054	0.038	0.067	0.046	0.062	0.044
6:32205822_T_C	0.063	0.046	0.070	0.051	0.065	0.047
6:32243461_G_C	0.084	0.066	0.092	0.073	0.099	0.072

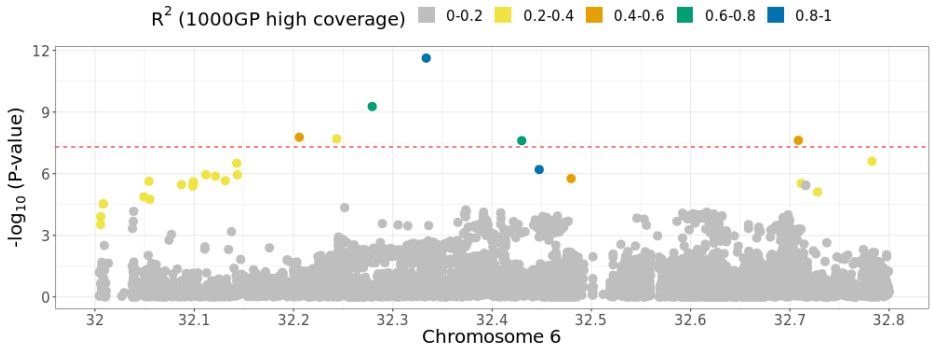


Fig. 3.4 Regional association plot showing meta-analysis association P-values in 6p21.32. Variants are coloured by R^2 with the index variant, derived from 1000 Genomes Project High Coverage study (Non-Finnish Europeans)

Association signal at 6p21.32 matches expected linkage disequilibrium pattern in non-Finnish Europeans

Although the four variants spanned a 130 kbp region, they displayed high LD with the index variant, as the 6p21 region is known to exhibit long-range LD (Figure 3.4). To better understand if the association strength matches the expected LD pattern, I measured the correlation between the association P-value and R^2 with the index variant for all the LD friends of rs115378818. The underlying expectation is that, for a given variant in the locus, the lower its LD with the index variant, the weaker its association is expected to be. P-values that do not match this expectation may suggest a spurious association due to a genotyping or imputation error or cryptic population structure. At 6p21.32, I found that LD friends P-values were correlated with the expected R^2 with the index variant (ρ=0.45; LD friends defined as variants with $R^2 > 0.5$ with the index variant and P-value < 0.01; Figure 3.5). The relatively weak correlation is likely driven by the small number of LD friends that the index variant has (N=4). When I relaxed the R^2 cutoff of LD friends this correlation became increasingly stronger (ρ=0.65 and 0.72 at $R^2 > 0.4$ and 0.2, respectively), showing that overall the variants at this locus follow their expected association strength given the LD structure between variants.

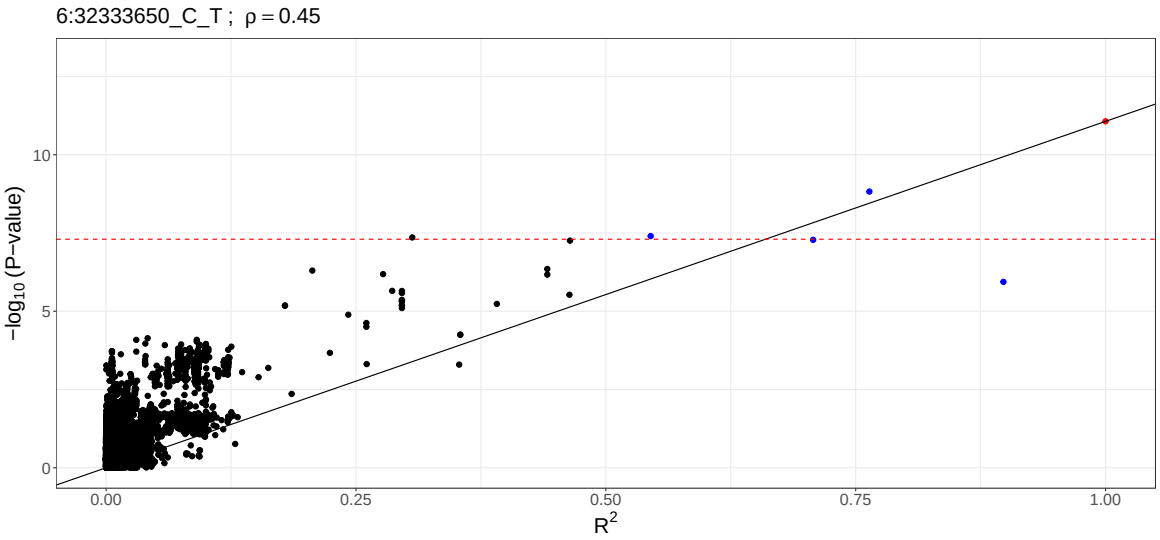


Fig. 3.5 Association P-value for all variants in the genome-wide significant locus at 6p21.32 (P-value < 5×10^{-4}) on the x-axis, and R^2 of variants with the index variant on the y-axis (derived from 1000GP). Blue dots indicate the index variant's LD friends, and the red horizontal line indicates genome-wide significance level (P-value < 5×10^{-8}). The black line is fitted to the origin (0,0), and to the point (1,$-log_{10}(P_{index\_variant})$), and shows the expected association strength given the LD with the index variant.

3.4.6 Association at 6p21.32 is robust to more severe pCD+ definitions

The IBD-BR provides information about the type of perianal involvement each patient presents with. To understand the effect of different definition criteria, I investigated how the meta-analysed association signal at 6p21.32 is sensitive to different definitions of pCD+ cases in the IBD-BR. In addition to the broad-definition meta-analysis described in the previous section ($META_{broad}$), I performed a meta-analysis between UKIBDGC and IBD-BR using three additional pCD+ definitions that have an increasingly severe impact on patients: pCD+ as abscess or simple or complex fistula only ($META_{abscfist}$), as simple or complex fistula only ($META_{fist}$), and as complex fistula only ($META_{complexfist}$).

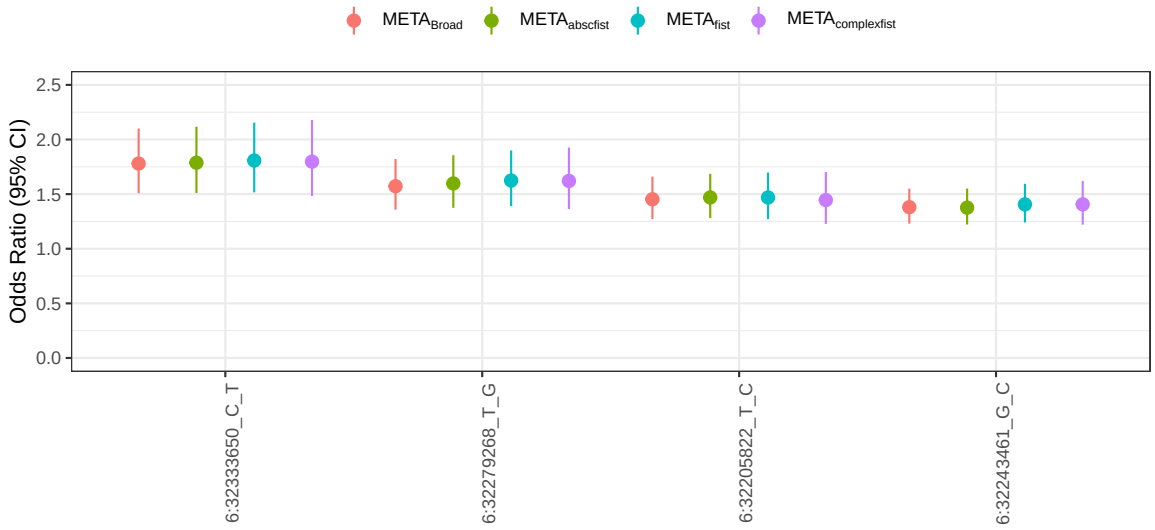


Fig. 3.6 Odds ratios of all four genome-wide significant SNPs in pCD cohorts where IBD-BR pCD+ cases defined with different case inclusion criteria.

I compared both the effect sizes and association statistics of the four-genome-wide significant SNPs between the four meta-analyses outlined above. First, I found that none of the variants exhibited heterogeneity of effect sizes ($P_{het} > 0.01$ for four variants) suggesting that different definitions of pCD are unlikely to bias the effect size estimates of these SNPs. Second, since the stricter definitions resulted in a reduction in the number of pCD+ cases, a proportional decrease in association test statistic (χ^2) may also be expected. Under the hypothesis that the stricter-definition meta-analyses are simply random subsets of $META_{broad}$, the χ^2 observed in any definition meta-analysis should match χ^2 from $META_{broad}$ adjusted for the reduction in sample size (I will refer to this as $\chi^2_{Broad,n}$; see section 3.3.12 in Methods for how this adjustment was performed).

Across all variants in the locus, I compared the χ^2 statistics observed in each of the three stricter-definition meta-analyses to $\chi^2_{Broad,N}$. All four genome-wide significant variants achieved the expected association in the stricter definition meta-analyses. For example, rs115378818 remained genome-wide significant in $META_{fist}$ despite the decrease in sample size (1,234 fewer pCD+ cases; observed P-value=2.2×10^{-11}; broad-definition P-value adjusted for sample size=3×10^{-11}; Figure 3.7). More broadly, across all variants in 6p21.32, I observed strong correlation between observed χ^2 in the stricter-definition meta-analyses and $\chi^2_{Broad,n}$, which shows the robustness of the association signal against different definitions of pCD+ cases in IBD-BR ($META_{abscfist}$=0.95; $META_{fist}$=0.92,$META_{complexfist}$=0.84).

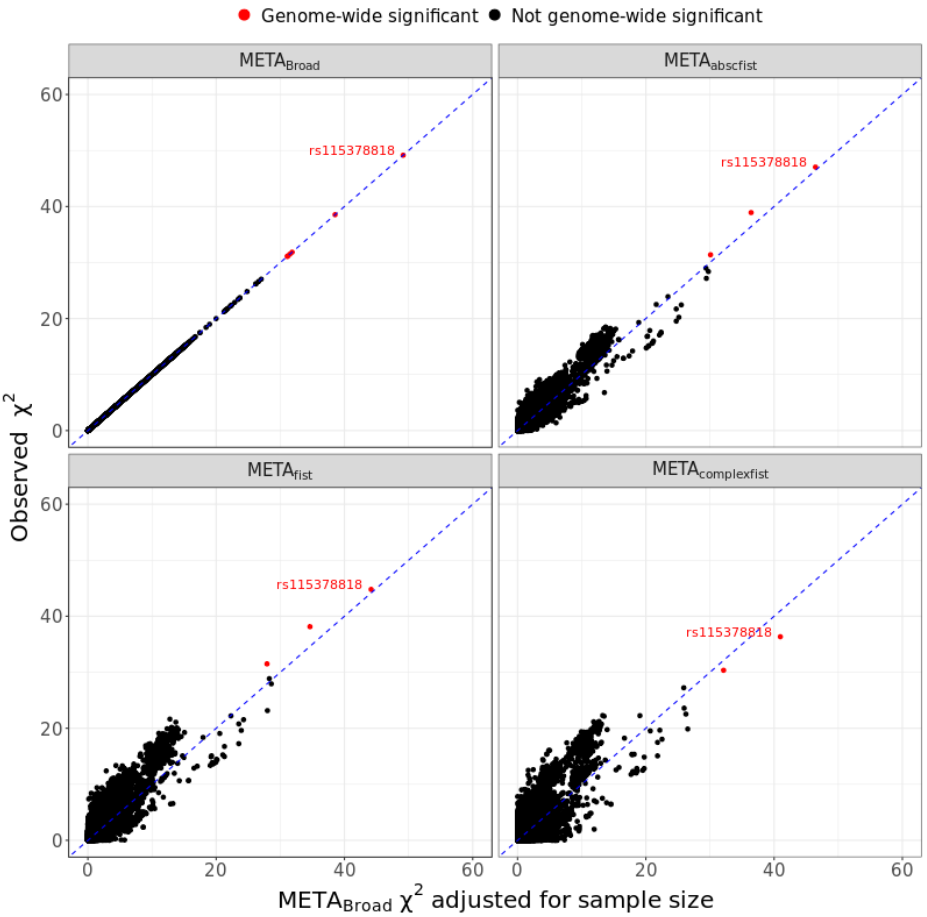


Fig. 3.7 Association P-value for all variants in the genome-wide signficant locus around rs115378818 (P-value $< 5 \times 10^{-4}$) on the x-axis, and R^2 of variants with the index variant rs115378818 on the y-axis (derived from 1000GP). The blue line indicates the line that passess through the points (0,0) and ($\chi^2_{Broad,n}$,$\chi^2_{observed}$).

3.4.7 pCD is nominally associated with HLA allele DRB1*01:03

The MHC region is known to be highly polymorphic and to exhibit long-range LD patterns, which often span several hundred kbps. This makes the mapping of MHC associations to effector genes challenging [**?**]. To this end, HLA imputation based on genotyped variants can aid the interpretation of genome-wide significant hits in the MHC locus. HLA genes are broadly divided into class I and class II genes [**?**]. Both classes of genes are responsible for presenting antigens to T lymphocytes and natural killer cells via antigen-presenting cells in order to initiate an innate and adaptive immune response [**?**]. Due to the extensive polymorphism of the MHC regions, groups of HLA alleles are categorised in HLA groups which are referenced using a 2-digit naming system [**?**] (2-digit resolution; which I will refer to as *allele group*). Two additional digits may be used to reference the specific HLA allele [**?**] (4-digit resolution; which I will refer to as *specific allele*).

To identify which HLA alleles may be associated with pCD, I performed association analyses between pCD status and class I and II HLA alleles, both at the allele group and specific allele levels (2-digit and 4-digit resolutions; see Methods for how HLA imputation was performed). Similar to the genome-wide association analysis, I performed the HLA association analyses separately for IBD-BR and UKIBDGC and subsequently meta-analysed the summary statistics (effect sizes and standard errors).

None of the tested HLA alleles achieved genome-wide significance (P-value $< 5 \times 10^{-8}$) within the cohorts or in the meta-analysis. At the allele group level, DRB1*01 had the strongest association. At a specific allele level, HLA-DRB1*01:03 had the most significant association, and had a stronger association compared to its allele group ($P_{DRB1*01:03} = 1.8 \times 10^{-6}$; $P_{DRB1*01} = 1.4 \times 10^{-3}$). I tested both dominant and additive modes of inheritance and found that the dominant model achieved better model fit at both the allele group and specific allele levels ($AIC_{dominant} < AIC_{additive}$; Table 3.8).

Table 3.8 Top HLA allele associations with pCD status. Both allele groups (2-digit resolution; first two rows) and specific alleles (4-digit resolution; third and fourth rows) are shown. Meta-analysed P-values and odds ratios between UKIBDGC and IBD-BR cohorts are shown (with their 95% confidence intervals). Both dominant and additive modes of inheritance for DRB1*01 and DRB1*01:03 were tested. Akaike Information Content (AIC), a measure of model fit, is shown in the last three columns for each of the three constituent cohorts, and shows a better fit for the dominant model (lower AIC).

HLA Allele	Inheritance	Odds Ratio	P-value	AIC (IBDBR)	AIC (GWAS2)	AIC (GWAS1)
DRB1*01	Dominant	1.2 (1.1 - 1.3)	9.4e-04	9127.887	3901.092	1783.907
DRB1*01	Additive	1.1 (1.1 - 1.2)	1.5e-03	9128.485	3901.413	1783.907
DRB1*01:03	Dominant	1.6 (1.3 - 1.9)	5.3e-07	9122.007	3651.987	1615.647
DRB1*01:03	Additive	1.5 (1.3 - 1.8)	1.8e-06	9122.825	3653.500	1615.647

Conditioning association signal on DRB1*01:03

I then asked if the DRB1*01:03 association with pCD accounted for the genome-wide significant locus at 6p21.32. Linking the two associations can explain which HLA allele the genome-wide significant hit may map to. To this end, I repeated the association tests for all variants in the locus, including DR1*01:03 as a covariate. Additionally, I repeated the HLA allele association test conditioning on the index variant in the locus to understand if the DRB*01:03 association is completely accounted for by the index variant rs115378818. Similar to the GWAS and the HLA allele association tests, I also analysed the different cohorts separately and then meta-analysed the effect sizes and standard errors.

After conditioning on the index variant rs115378818, I did not observe an association with DRB1*01:03, indicating that the DRB1*01:03 association is completely accounted for by the index variant ($P_{DRB1*01:03|rs115378818}$=0.61). Conversely, DRB1*01:03 did not completely account for the rs115378818 association. When I conditioned the rs115378818 association on DRB1*01:03, the index variant remained nominally associated with pCD ($P_{rs115378818} = 8.6 \times 10^{-12}$ and $P_{rs115378818|DRB1*01:03} = 1.1 \times 10^{-5}$; Figure 3.8). Taken together, this evidence suggests that DRB1*01:03 is only nominally associated with pCD and that it only partly explains the observed genome-wide association signal.

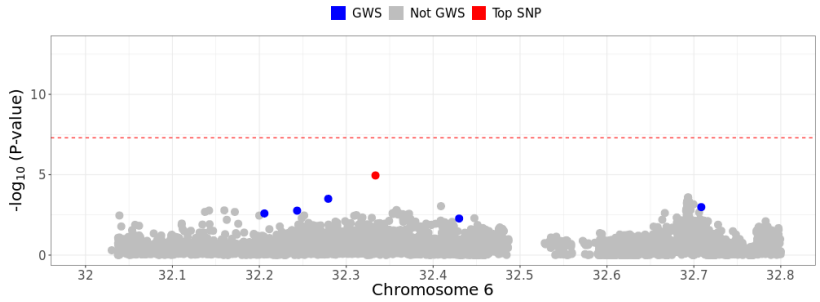


Fig. 3.8 Residual association signal after including DRB1*01:03 as a covariate. Blue points represent variants with a genome-wide significant association in the model that does not include DRB1:01*03. The red point indicates rs115378818 (index variant).

3.5 Discussion

In this chapter, I used two IBD cohorts (UKIBDGC and IBD-BR) with rich clinical data to describe the clinical characteristics of pCD and identify genetic variants associated with pCD risk. With a total of 12,897 individuals, this meta-analysis represents one of the largest pCD GWAS studies to date [**?**]. Although others have attempted to identify pCD-associated loci, none of the studies have found genome-wide significant hits [**?** **?**]. The most recent pCD study by Akhlaghpour et al. identified a Complement Factor B (*CFB*) missense variant that was nominally associated with pCD risk (rs4151651; P-value=9.35×10^{-6}). *CFB* is part of the alternative pathway responsible for the activation of the complement system, an innate immune subsystem that improves the ability of phagocytic cells to clear pathogens [**?** **?**]. Functional follow-up work showed that the *CFB* missense variant impairs the phagocytic function of macrophages. This impaired phagocytic capability was hypothesised to impact the ability of the immune system to fight bacterial strains found in the fistulas of CD patients. Although this study established a plausible factor that may contribute to pCD risk, a more complete picture of its pathogenesis is needed.

In an attempt to fill the gap in our understanding of pCD pathogenesis, I performed a meta-analysis of two pCD cohorts, with a total of 3,967 pCD+ cases and 8,930 pCD- controls. I identified a genome-wide significant locus in the highly polymorphic Major Histocompatibility Locus (MHC) at 6p21.32 that was associated with pCD risk (index variant rs115378818). Additionally, my post-GWAS checks have reasonably confirmed the veracity of this association. Furthermore, the association signal revealed by our meta-analysis is likely distinct from

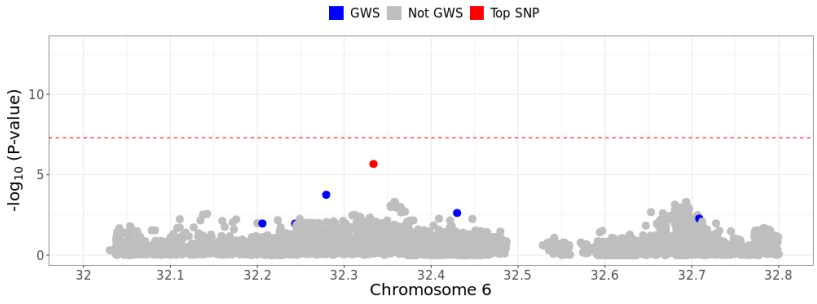


Fig. 3.9 Residual association signal after including rs114969413 (a *CFB* locus variant) as a covariate. Blue points represent variants with a genome-wide significant association in the model that does not include rs114969413. The red point indicates rs115378818 (index variant).

the *CFB* association. First, rs115378818 and rs4151651 are in weak LD (R^2=0.24). Second, upon conditioning on the *CFB* signal (rs114969413; R^2 with index variant from Akhlaghpour et al. $= 0.99$), rs115378818 remained nominally associated with pCD, suggesting that the *CFB* does not completely account for it ($P_{rs115378818|rs114969413}$=$2.1 \times 10^{-6}$; Figure 3.9), and showing that our genome-wide significant variant represents a novel pCD-associated locus.

Despite our ability to identify a novel association, case heterogeneity remains an important limitation of this study. Akhlaghpour et al. have also noted that a potentially heterogeneous composition of their cohorts may have decreased their power to detect genome-wide associations with pCD. Their cohorts, similar to ours, may have had patients with milder forms of pCD such as skin tags and ulcer. Therefore, I tested the association of our locus with pCD+ cases redefined with increasingly severe criteria and found that the genome-wide significant association was robust to different case definitions. However, the heterogeneity of case definition may still have decreased our power to detect *additional* pCD-associated loci. In this study, a more refined definition of pCD+ as fistulising pCD was hampered by the unavailability of more granular data on the specific pCD manifestations in the UKIBDGC cohort. It remains to be seen if larger sample sizes can overcome the heterogeneity in case definition and identify more pCD-associated loci. Increasing power will not only identify more pCD-associated loci, but it will also give us a more complete understanding of the pathogenesis of pCD, and may implicate biological pathways that are targeted by pCD-associated genetic variation.

Nonetheless, the finding that HLA-DRB1*01:03 may explain the genome-wide association signal is a promising starting point. HLA-DRB1*01:03 has been previously shown to be associated with colonic Crohn's disease, ulcerative colitis risk [**?** 95] and rheumatoid arthritis (RA) [**?**]. However, several avenues should be explored to identify which HLA alleles completely account for the genome-wide significant association. Several reasons may explain the incomplete association of HLA-DRB*1:03 with pCD. First, the association signal at 6p21.32 may be explained by multiple HLA alleles, and therefore conditioning on a single HLA allele cannot completely explain the 6p21.32 signal. The highly polymorphic nature of most HLA genes often makes it difficult to completely attribute disease risk to a single HLA allele. This has been previously observed for rheumatoid arthritis (RA), where several *HLA-DRB1* alleles confer risk to RA [**?**]. Second, multiple *HLA-DRB1* alleles with slight differences in their amino acid sequences, especially in the peptide binding regions, have different affinities to antigens being presented to T-cells [**?**]. Therefore, testing the pCD association with multiple HLA-DRB1 alleles that share the same amino acid sequences might better account for the pCD signal [**?**]. Interestingly, Raychaudhuri et al. found that only three amino acid positions in a predictive model of RA risk provided identical prediction to a model that included all *HLA-DRB1* alleles [**?**]. Therefore, an important follow-up to my work should explore the association of HLA-DRB1 alleles with pCD at the amino acid level.

Finally, it is important that future studies do not consider pCD as an isolated disease. The broad clinical phenotyping of IBD-BR showed lower drug intake, higher prevalence of extraintestinal manifestations, and higher prevalence of rectal CD in pCD+ versus pCD- patients. These observations naturally pose a question about how these disease characteristics relate to pCD. A plausible explanation is that these clinical characteristics (including pCD) are simply manifestations of what is collectively termed "severe CD". It is still important however to understand how severe CD drives these seemingly unrelated manifestations. For example, is there common pathogenesis, or genetic predisposition that drives all these manifestations, including pCD? In this regard, my univariate pCD meta-analysis is limited. Future studies should quantify the genetic correlation between pCD and disease severity and location, extraintestinal manifestations, and drug response. Furthermore, multivariate GWAS methods should be employed to identify shared and distinct genetic factors driving each of these manifestations. The observation that clusters of disorders have common as well as distinct genetic factors has been previously shown to further our understanding of multiple groups of disorders such as psychiatric disorders [**?**].

Literature about pCD pathogenesis largely attributes the development of fistlising CD to the theory of epithelial-to-mesenchymal transformation of epithelial cells. With the ongoing growth of IBD-BR, and with more genotyped CD participants, a more nuanced understanding of which dysregulated pathways give rise to pCD risk, and whether or not these pathways corroborate the EMT theory of pCD pathogenesis will become increasingly attainable.

Chapter 4

Genome-wide Meta-analysis of All-cause Perianal Disease

4.1 Contributions

UK Biobank phenotype and genotype data were obtained by Dr. Laura Fachal. Genotype quality control and principal component analysis were also performed by Dr. Laura Fachal. All the other analyses described here were performed by me.

4.2 Introduction

In the previous chapter, I described the characteristics and genetic underpinning of perianal Crohn's disease. However, perianal disease (pAD) is not only associated with Crohn's disease. pAD encompasses a broader set of perianal manifestations such as perianal abscess, fissures, and fistulas.

Anal fissures are tears and/or ulcers in the perianal skin that cause sharp pain associated with defecation and rectal bleeding. They are classified as acute, lasting less than 6 weeks, or chronic. The etiology of anal fissures is still debated [**?**], but they are thought to be caused by trauma resulting from high anal canal pressure associated with constipation or diarrhea [**?**]. Although recent population studies on the etiology of anal fissures are markedly lacking [**?**], a number of studies conducted between 1977 and 1983 found that chronic constipation accompanied fissures in only 25% of patients while diarrhea accounted for less than 7% [**?** **?** **?**]. Anal fissures are conservatively treated with dietary and lifestyle modifications, analgesics, and fiber supplements [**?**]. Anal fistula is a more serious perianal manifestation

of pAD. An anal fistula is defined as an abnormal communication between the anal canal and the perianal skin [**?**]. It is characterised by pain, rectal discharge and bleeding, and causes significant lifestyle difficulties for patients [**?**]. The incidence of perianal fistulas varies per country, and a recent study in four European countries found that it ranges from 1.2 to 2.8 per 10,000 per year [**?**], but its incidence is likely underestimated [**?**]. Moreover, little is known about the pathophysiology of anal fistulas. Cryptoglandular fistulisation, a theory proposed by Parks in 1961 [**? ?**], is the most accepted pathophysiological account for the origin of anal fistulas. According to the cryptoglandulartheory, anal fistulas start as an inflammatory process in the proctodeal glands whose ducts extend to connect the perianal skin to the anal canal. Crohn's disease, tuberculosis [**?**], radiotherapy [**?**], sexually transmitted diseases [**?**] and malignancy can contribute to these initial inflammatory processes that result in anal fistula formation. However, over 90% of cases remain idiopathic [**?**]. Management of anal fistula depends on its type, extension, and underlying cause, but the end goal of all treatments is the complete drainage of abscess, and fistula healing while preserving anal sphincter function and continence. The management plan is typically decided based on a combined clinical, radiological and/or endoscopic assessment. Surgical options of fistula management include fistulectomy, the complete excision of the fistula tract, and fistuolotomy, in which the fistula is laid open and allowed to heal. Surgical procedures that preserve anal function such as fistulotomy with sphincter reconstruction have emerged in the last 30 years as a particularly effective method of treating anal fistulas, with lower incontinence and recurrence rates compared to fistulectomy [**? ?**].

Despite recent advances in the management of pAD, little is known about the biological and pathophysiological processes that lead to anal fissure and fistula formation. Despite the overall consensus on the cryptoglandular theory, it is unclear if some individuals are at higher risk of developing pAD due to their genetic predisposition. Genome-wide association studies (GWAS) have significantly improved our understanding of the genetic factors underlying several complex diseases as well as the biological pathways implicated by disease-associated genetic variants. In the last 15 years, a typical GWAS required extensive coordination between researchers, clinicians and recruitment centres to construct case-control cohorts to study a particular disease or trait of interest. In recent years, efforts to build various national biobanks where genetic, and phenotypic data are available for hundreds of thousands of participants has significantly improved our ability to conduct GWAS of previously understudied diseases and traits. Their large sample sizes have made it possible to carry out GWAS of thousands of binary and continuous traits and diseases, including relatively uncommon

diseases.

Recruitment for the UK Biobank (UKBB) started in 2006, and has so far collected genotypic, biomarker and clinical data from electronic health records as well as blood and urine samples from over 500,000 participants [**?**]. The UKBB queries electronic health records for various types of data including deaths, cancer registrations, and hospital inpatient episodes. Electronic health records data are provided by external sources. Upon receipt of the data, a multi-step approach is followed, where received data is subjected to further pre-processing and quality checks to ensure its alignment with the UKBB data dictionary, and that it does not contain any ambiguities. The data is then consolidated into a central UKBB database and made available to researchers [**?**].

FinnGen is another example of national biobanks that has made GWAS for various traits and diseases more feasible. FinnGen started in 2017 as a public-private partnership between several public Finnish institutions and thirteen international pharmaceutical companies [**?**]. Importantly, FinnGen restricts access to individual-level data to approved researchers only. However, summary statistics data from GWAS analyses are made publicly available. In its latest release (data freeze 9), GWAS are available for over 2,200 binary endpoints from 377,277 individuals [**?**].

The UKBB and FinnGen use slightly different clinical coding systems to register clinical data. The UKBB uses the International Classification of Disease (ICD), a hierarchical clinical framework that organises clinical diagnoses in a tree-like structure. Different groups of diseases are organised in alphabetical chapters and particular diseases within each chapter are given a numeric value (e.g. chapter K contains digestive system disorders and K60 indicates anal fissure and fistula). Further detailed diagnosis subtypes are nested within each alphanumeric code up to four levels of resolution [**?**]. However, the UKBB only records up to two levels of resolution (e.g. K50.0 indicates Crohn's disease of the small intestine). FinnGen, on the other hand uses an expert-curated set of endpoints that are largely parallel to ICD-10 codes. These endpoints are designed to accommodate inclusion and exclusion criteria relevant for GWAS analyses [**?**]. Despite these differences, both resources are valuable for studying pAD.

In this chapter, I will describe a pAD GWAS I performed in the UKBB to map pAD-associated genetic variants and I will describe the post-GWAS quality checks I carried out to ensure the validity of genome-wide significant loci. Additionally, I will outline how I used

summary statistics from the FinnGen GWAS both as a replication dataset and as a constituent cohort in a pAD meta-analysis I performed between the UKBB and FinnGen cohorts. Finally, I will describe two follow-up analyses that I performed to better understand the effects of pAD-associated variants on haemorrhoids, a closely-related disease, and identify effector genes at these loci.

4.3 Methods

4.3.1 UKBB sample preparation and data access

Hospital inpatient episode data and genotyped and imputed data were obtained under UK Biobank Application 45669. UKBB participants were genotyped using either UK Biobank Axiom Array [**?**] or the Affymetrix UK BiLEVE Axiom array [**?**]. Sample processing, genotyping and quality control were performed at the UK Biobank, Affymetrix and the Wellcome Trust Centre for Human Genetics [**?**]. The imputation process has been previously described [**?**] and consists of imputing directly genotype data to the Haplotype Reference Consortium (HRC) and UK10K, resulting in 96 million variants. Imputed data was obtained as BGEN v1.2 files [**?**].

4.3.2 Defining pAD case control cohorts

Case inclusion criteria

To define the case cohort, I identified all individuals with ICD-10 code K60 or ICD-9 code 565. In total, 5,257 UKBB participants had at least a single visit where they received either a primary or secondary pAD diagnosis or its corresponding ICD-9 code ("anal fissure and fistula"; 565). Six level-2 codes are nested within K60, representing two broad categories of pAD: fissures and fistulas. Three codes are used for acute and chronic fissures and three codes for acute and chronic fistulas. 92% of patients (4,858) presented with either K60.1, K60.2 or K60.3 ("chronic anal fissure", "anal fissure, unspecified" and "anal fistula", respectively; Table 4.1).

Control exclusion criteria

To avoid contamination of controls with lower digestive tract disorders that may be true pAD cases that were incorrectly diagnosed, I applied a set of control exclusion criteria. Specifically, I excluded from the control set any individuals who had an ICD-10 hospital diagnosis of K55-K64 or their corresponding ICD-9 codes as outlined in Table 4.2 (collectively grouped

as "Other diseases of intestines" in ICD). These ICD codes cover disorders with symptoms that may resemble pAD symptoms upon presentation, and include ano-rectal bleeding (K55 vascular disorders of the intestine, K57 diverticular disease of intestine and K64 Haemorrhoids and perianal venous thrombosis), or a change in bowel habits (K56 Paralytic ileus and K58 Irritable bowel syndrome), perianal fistula or abscess (K60 fissure and fistula of the anal region and K61 abscess of the anal and rectal region), any ano-rectal abnormalities (K62), or proximal fistulas or abscesses (K63). In total, I excluded 128,319 individuals from the control cohort (26.7%), resulting in 353,437 controls (per-code number of individuals in Table 4.1).

Table 4.1 Number of UKBB participants with a primary or secondary K60 diagnosis. Number of patients with each level-2 K60 diagnosis are shown. K60.0=Acute Anal Fissure; K60.1=Chronic Anal Fissure; K60.2=Anal Fissure; unspecified; K60.3=Anal Fistula; K60.4=Rectal Fistula; K60.5= Anorectal Fistula

ICD-10 code	K60.0	K60.1	K60.2	K60.3	K60.4	K60.5
Number of individuals	144	788	2,624	1,954	76	122

Table 4.2 pAD control set exclusion criteria. All ICD-10 codes had corresponding ICD-9 codes except K56, K62 and K63. For those, ICD-9 codes were obtained manually by inspecting level-2 ICD-10 codes and searching for their corresponding level-2 ICD-9 codes.

ICD-10 code	**ICD-10 meaning**	**ICD-9 code**	**ICD-9 meaning**	**N**
K55	Vascular disorders of intestine	557	Vascular insufficiency of intestine	2,923
K56	Paralytic ileus and intestinal obstruction without hernia	5600, 5601, 5602, 5603, 5608A, 5608, 5609	Intussusception, Paralytic ileus, Volvulus, Impaction of intestine,Other specified intestinal obstruction, Unspecified intestinal obstruction	9257
K57	Diverticular disease of intestine	562	Diverticula of intestine	61,519
K58	Irritable bowel syndrome	5641	Irritable bowel syndrome	12418
K59	Other functional intestinal disorders	564	Functional digestive disorders not elsewhere classified	30,087
K60	Fissure and fistula of anal and rectal regions	565	Anal fissure and fistula	5,079
K61	Abscess of anal and rectal regions	566	Abscess of anal and rectal regions	2,178
K62	Other diseases of anus and rectum	5690, 5691, 5692, 5693, 5694	Anal and rectal polyp, Rectal prolapse, Stenosis of rectum and anus, Hemorrhage of rectum and anus, Other specified disorders of rectum and anus	39,191
K63	Other diseases of intestine	5695, 5696, 5697, 5698, 5699	Abscess of intestine, Colostomy and enterostomy complications, Complications of intestinal pouch, Other specified disorders of intestine, Unspecified disorder of intestine	33,307
K64	Hemorrhoids and perianal venous thrombosis	455	Hemorrhoids	19,060

4.3.3 ICD code enrichment in pAD cases versus controls

The availability of a large number of clinical diagnoses and phenotypes for UKB participants enables a thorough characterisation of the pAD case cohort. I aimed to understand the cohort composition by identifying which ICD-10 codes are enriched in cases versus controls. For each ICD-10 code, I compared the prevalence in pAD cases versus controls, and I formally tested the enrichment of 1,693 codes using Fisher's exact test. For this test, I did not apply the control exclusion criteria outlined in Table 4.2.

4.3.4 UKBB genotype quality control

Genotyping array data from the UK Biobank dataset underwent quality control as part of the International IBD Genetics Consortium GWAS that is being undertaken in the laboratory, which resulted in 419,871 variants being retained. QC was performed using a combination of PLINK (v1.9 and v2) [**?**], bcftools (v1.16) [**?**], and KING (v2.2.4) [**?**]. Variants that met the following criteria were excluded:

- Low call rate (<0.95 for variants with minor allele frequency (MAF) > 0.01 or < 0.98 for variants with MAF ≤ 0.01).
- Significant difference in genotype call rate (P-value $< 10^{-4}$) between IBD cases and controls.
- Large allele frequency (AF) differences between UKBB and Gnomad (Non-Finnish Europeans), or TOPMed (global) using the following formula: $\frac{(P_1-P_0)^2}{(P_1+P_0)(2-P_1-P_0)} > \varepsilon$, where $\varepsilon = 0.025$ or 0.125, for Gnomad and TOPMed respectively, P_0 is the minor allele frequency (MAF) in Gnomad or TOPMed and P_1 is the UKBB MAF.

Genotypic principal components (PC) were estimated for all participants, using a set of genotyped variants that were also available in the 1000 Genomes Project (1000GP; excluding variants associated with IBD susceptibility (P-value < 10^{-4}), and variants in long LD regions (as defined in [**?**]). This final list of variants was pruned with the following parameters: window size = 50 kbp; step size = 5; R^2 = 0.2. PCs were then projected to 1000GP PCs. Samples within the European ancestry group were retained for the subsequent analyses.

4.3.5 UKBB GWAS using REGENIE

All genome-wide association analyses were performed using REGENIE v3.2.5 [**?**] following a 2-step approach. This approach achieves higher computational efficiency compared to

linear mixed model, which are normally used in GWAS methods to account for cryptic relatedness. Briefly, in step 1, a whole-genome regression model is fitted using a subset of high-quality genome-wide variants in order to estimate a set of genome-wide predictors that capture a large fraction of phenotypic variance. These predictors are then used in step 2 in a single-variant association testing model, where a larger set of variants of interest are tested for association. I used post-QC genotyped variants in step 1 as recommended by REGENIE documentation (N=419,871), and both genotyped and imputed variants in step 2, testing all autosomal chromosomes (N=9,705,089). Additionally, I enabled a Firth correction of effect sizes for all variants with P-value < 0.01 in order to account for the case-control imbalance in the pAD cohort (`--firth --approx --pThresh 0.01`). The Firth test corrects biased effect sizes and P-values obtained from highly unbalanced case-control designs, where such an imbalance causes unreliable P-values, inflating Type I error. Approximate Firth logistic regression is a variant of the Firth test that is more computationally tractable and is implemented in REGENIE.

With the pAD case control cohort defined above, I used REGENIE to perform a pAD GWAS with White British UKBB participants [**?**]. In addition to genotype QC described earlier, I excluded individuals with missing genotypes or with discordant reported and genetically-inferred sex. After this filtering a total of 4,606 pAD cases and 332,234 pAD controls remained (see Methods for genotype data quality control and imputation). In order to account for cryptic population stratification, I used 10 European-ancestry genotypic principal components, as well as sex and genotyping array as covariates in the REGENIE model. After filtering out variants with low imputation quality (INFO < 0.4) and minor allele frequency (MAF) < 0.01, a total of 9,705,089 variants were tested. After running REGENIE, I found that the summary statistics exhibited moderate genomic inflation (median $\chi^2 = 0.48$; $\lambda_{GC} = 1.06$).

4.3.6 LD calculation from 1000GP

Reference LD panels obtained from the 1000 Genomes Project High Coverage project [**?**] were used for different analysis, including genome-wide loci identification using LD clumping, and post-GWAS checks to study the relationship between LD and association strength at genome-wide significant loci. R^2 values were calculated between the index variant and all variants in each locus (loci were defined using LD clumping as described in the next section). I downloaded VCFs from the 1000GP high coverage and used PLINK v1.9 to compute LD between all variants and the index variant at each locus. For each GWAS check, I used unrelated individuals assigned to the relevant reference population: NFE and GBR for

UKBB and FE for FinnGen (N=426, 99 and 90 respectively). Only one sample was retained from the trios found in 1000GP. Relevant samples were included in the LD calculation using the following PLINK command:

```
plink --r2 --keep EUR.samples --ld-window-r2 0
```

4.3.7 Defining genome-wide significant loci in UKBB

I defined genome-wide significant loci from the UKBB pAD GWAS summary statistics using PLINK v1.9 via a clumping procedure. Briefly, LD clumping identifies the most significant variant in a user-defined window to represent each locus (termed index variant). It then proceeds to define the locus boundaries by clumping neighbouring correlated variants. Specifically, any variants within the predefined window that are correlated with the index variant are considered to belong to the same locus represented by the index variant (i.e. variants in high LD). I used VCFs downloaded from the 1000GP, which are used to compute LD, and set a maximum P-value of 5×10^{-8} for defining a genome-wide significant locus, with default values for the rest of the parameters: variants with $R^2 < 0.5$, variants outside a window of 250 kbp, or variants that have a P-value > 0.01 are not clumped with the index variant.

```
plink --clump-p1 0.00000005 --clump-r2 0.50 --clump-kb 250
--clump-p2 0.01
```

PLINK outputs each locus' index variant along with any variants that meet the clumping criteria outlined above. I then defined each locus' boundaries by sorting the clumped variants within each locus according to their genomic location: the most downstream variant defined the 5' boundary and the most upstream variant defined the 3' boundary.

4.3.8 FinnGen summary statistics preprocessing

Publicly available FinnGen GWAS summary statistics (data freeze 7) were downloaded from the FinnGen results website `finngen.fi/en/access_results`. Similar to UKBB, variants with MAF < 0.01 were removed, but imputation quality information were not available, so I was not able to filter out variants with low imputation quality. The association summary statistics showed evidence of moderate genomic inflation (λ_{GC}=1.089). I used LD clumping with a 1000GP-derived FE LD reference panel to identify genome-wide significant loci.

4.3.9 Meta-analysis of UKBB and FinnGen

I used METAL to perform the meta-analysis between UKBB and FinnGen GWAS summary statistics. METAL can perform fixed-effects meta-analysis using one of two different well-established schemes: P-values and effective sample size, or effect sizes and standard errors. The P-value scheme is implemented to enable meta-analysis of GWAS summary statistics that do not report the effect allele, while the effect sizes scheme can be used when each variant's effect size and effect allele are reported. Both my UKBB analysis and FinnGen's summary statistics report the effect allele, so I used the effect size scheme of METAL (`SCHEME STDERR`).

After filtering out variants with MAF < 0.01 and with low imputation score (INFO < 0.4), the two GWAS summary statistics had an intersection of 7,663,827 variants and a total of 11,096,129 variants across the two cohorts. Of these, 2,041,145 variants were specific to UKBB and 1,390,527 were specific to FinnGen. Given that 31% of variants were unique to one of the two GWAS, I did not remove them from their respective summary statistics file. It is important to note, however, that this choice may favour variants that are available in both studies. Additionally, I enabled METAL's `GENOMICCONTROL ON` option to correct genomic inflation in each of the two summary statistics before performing the meta-analysis. The resulting meta-analysed summary statistics showed no evidence of genomic inflation (λ_{GC}=1.02).

For each variant, METAL outputs the effect allele, meta-analysed effect size, standard error, and P-value. After performing meta-analyses, it is important to compare the effect sizes between the meta-analysed cohorts. Comparison of both the direction and magnitude of effect sizes gives an indication on how similar the estimated effects of meta-analysed genetic variants are, which is an import post-GWAS quality control check. To formally test this, METAL uses Cochran's Q test to test for effect size heterogeneity. Cochran's Q test assesses two or more effect size estimates and their corresponding standard errors and reports a χ^2 statistic that quantifies the deviation from the null hypothesis that the meta-analysed effect sizes are similar. Depending on the number of meta-analysed studies (in this case 2), a P-value is derived from a theoretical χ^2 distribution with $N-1$ degrees of freedom, where N is the number of meta-analysed studies (heterogeneity of effect P-value P_{het}). I used P_{het} to test if index variants at genome-wide significant loci demonstrate heterogeneity of effect size between the two cohorts. To account for multiple index variants being tested, I set a Bonferroni-corrected P-value threshold for rejecting the null hypothesis that the effect sizes

are similar between the two studies (P-value < $\frac{0.05}{k}$, where k is the number index variants tested).

4.3.10 Defining genome-wide significant loci in the UKBB/FinnGen meta-analysis

Meta-analysis associations are derived from two different populations with different LD structures (NFE and FE). For these associations, a representative LD reference panel that captures the true underlying LD pattern in the meta-analysis is not easy to obtain because LD will be dependent on the contribution of each population to the association signal. However, an LD reference panel is needed both to define genome-wide significant loci based on LD clumping, but also to perform post-GWAS checks. To this end, I identified genome-wide significant loci similar to UKBB and FinnGen, except that I performed LD clumping of the meta-analysis genome-wide significant loci once with an LD reference panel derived from NFE in 1000GP, and once with an LD reference panel derived from FE in 1000GP. Both LD clumping procedures identified the same 18 genome-wide significant loci, but their boundaries differed slightly. I defined consensus loci boundaries as the union of loci boundaries defined by NFE- and FE-based LD clumping.

4.3.11 Quality control of meta-analysis genome-wide significant loci

For all 18 genome-wide significant loci defined via the initial LD clumping procedure, I investigated the LD friends of each index variant. Specifically, for each index variant, I checked if the P-values of its LD friends decay linearly with their LD values with the index variant. In a meta-analysis between different subpopulations, it is important to ensure that P-values match expectation under LD in all the constituent cohorts. I therefore computed the correlation between P-values and LD in both NFE and FE for all 18 loci. Six loci showed that the index variant had no LD friends in NFE or FE, or that there was weak or non-existent correlation beween P-values and LD of the index variant's LD friends in NFE or FE (Pearson correlation coefficient between $-log_{10}(P)$ and (R^2 $\rho < 0.2$). I removed these six loci from all downstream analysis.

4.3.12 Genetic correlation analysis

Genetic correlation analysis is commonly employed to understand the genetic similarities between two phenotypes or diseases of interest. I used genetic correlation to understand the

overall genetic similarity between pAD and haemorrhoids. Linkage disequilibrium score regression is a common method to compute pairwise genome-wide genetic correlations (LDSC [**?**]) as it leverages association summary statistics between a pair of traits to compute a genetic correlation estimate (r_g). The availability of GWAS summary statistics for large numbers of traits and diseases makes genetic correlation a feasible exploratory analysis to discover genetic similarities between traits.

The fundamental concept of LDSC is that there is a linear relationship between the Z-score product ($z_1 z_2$) and the LD score of SNPs, where LD score is defined as the sum of R^2 values for all SNP in a pre-defined window (the default is 1 cM) [**?**]. The rationale behind this relationship is that SNPs with a higher LD score are more likely to tag the causal variant at each locus, and will therefore have a larger $z_1 z_2$ value. Regressing the LD score for all SNPs can give an estimate of the overall genetic covariance between two traits. Specifically, the regression slope quantifies the genetic covariance, which can then be normalised by the sample sizes of the two traits and the number of SNPs to obtain a genetic correlation estimate between the two traits (r_g).

The accuracy of r_g rests on the assumption that the GWAS population matches the population from which LD scores are derived. By default, LD scores are provided by LDSC, which are computed from the HapMap3 European-ancestry reference panel [**?**]. It is also important to note that although r_g is a genome-wide measure, its computation is based on a predefined set of SNPs. The estimation of r_g is based on a set of high-quality common SNPs in HapMap3, which is also provided by LDSC (MAF $\geq$ 0.05; N=1,217,312).

Genetic correlation between pAD and haemorrhoids

I used LDSC to compute r_g between my pAD meta-analysis and two haemorrhoids GWASes (Zheng et al. 2021 [**?**] and the Pan-UKBB GWAS: `pan.ukbb.broadinstitute.org/`). I downloaded the Zheng et al. 2021 summary statistics via the GWAS catalogue website (study accession: GCST90014033; N_{cases}=218,920; $N_{controls}$=725,213). Since the Pan-UKBB performed the haemorrhoids GWAS using different ancestries, I used the summary statistics from the European-ancestry GWAS only (ICD-10 code: I84; N_{cases}=26,348; $N_{controls}$=394,183).

After downloading the two haemorrhoids summary statistics, I preprocessed both of them using the LDSC script `munge_sumstats.py`. The script filters the SNPs and aligns their alleles to the HapMap3 SNP list using the flag `--merge_alleles hm3.snplist`. This script also takes as input a signed summary statistic column which I provided using the flag

`--signed-sumstats effect_size,0`, where the first argument specifies the column name (effect size column) and the second argument specifies the expected value of the signed summary statistic. Each of the two "munged" summary statistics files were then provided as input to the `ldsc.py --rg`, along with the pAD summary statistics file, and r_g is then computed between pAD and each of the two haemorrhoids GWASes.

4.3.13 Colocalisation analysis

In order to link the meta-analysis genome-wide significant loci to effector genes, I performed statistical colocalisation with a set of expression and splicing QTLs from the Genotype Expression Project (GTEx v8). Colocalisation analysis is a statistical approach that uses summary statistics from two association studies in order to make an inference about whether the two association signals are likely to be driven by a shared causal variant. In this regard, five different hypothesis regarding the relationship between the two association signals are tested:

- H_0: none of the two signals are associated with their corresponding traits
- H_1: only the first signal is associated with its corresponding trait
- H_2: only the second signal is associated with it corresponding trait
- H_3: the two signals are associated with their corresponding traits, with different underlying genetic variants
- H_4: the two signals are associated with their corresponding traits, and share a single underlying genetic variant.

Certainty about each of these hypotheses is quantified as a posterior probability. Therefore, colocalisation analysis outputs five different posterior probabilities: PP_0, PP_1, PP_2, PP_3, and PP_4. Statistical colocalisation is implemented in the R package `coloc v5.1.2`.

To maximise the ability of `coloc` to identify effector genes, I downloaded summary statistics from GTEx v8, a large compendium of expression and splicing quantitative trait loci (eQTLs and sQTLs) mapped from RNA-seq samples obtained from 49 human tissues, ranging in sample sizes from 73-706 individuals. eQTLs and sQTLs were mapped in a 1mbp window centred around the transcript start site (TSS) of each gene (cis-eQTLs and cis-sQTLs).

Within each genome-wide significant locus, I identified a list of genes and splice junctions for eQTL and sQTL colocalisation, respectively. To achieve this, for each locus I defined a 1 mbp window around the index variant at each locus, and created a set of genes and splice junctions whose respective TSS are located within this window. Next, I performed colocalisation analysis between the meta-analysis summary statistics and each eQTLs and sQTLs summary statistics for each gene in the window. I used the `coloc.abf()`, which takes as input effect sizes and standard errors of each variant from the meta-analysis and gene or splice junction being tested. Importantly, `coloc.abf()` does not require the effect sizes to be aligned to the same effect allele, as the Bayes Factor calculation implemented in `coloc` relies on the Z^2 statistic to compute the posterior probabilities. Finally, I used the default priors implemented in `coloc.abf()`: prior probability a SNP is associated with pAD=10^{-4}, prior probability a SNP is associated with the eQTL/sQTL=10^{-4}.

Table 4.3 Number of eQTL and sQTL genes tested for colocalisation across all GTEx v8 tissues. All genes and splice junctions in a 1mbp around each index variant were tested.

Index variant	Number of tested eQTL genes	Number of tested sQTL genes
3:53034026_C_T	41	18
5:64868326_TTTC_T	13	4
6:1775202_G_A	12	5
6:31121854_C_T	89	45
6:31253340_T_C	98	50
6:133260944_G_A	18	7
7:2524404_G_A	21	13
8:70735125_A_G	15	6
8:70993166_AAGTT_A	10	6
9:22124505_A_T	19	5
11:10356352_C_A	20	11
12:114235969_T_C	18	10

4.4 Results

4.4.1 pAD cases are enriched in multiple disorders compared to pAD controls

In order to understand the composition of the pAD case cohort, I tested the enrichment of 1,693 ICD-10 codes in pAD cases versus controls in UKBB. Overall, the tested enrichment odds ratios were inflated (median odds ratio=1.36), likely as a consequence of sampling a disease cohort within a healthy population cohort. In total, 198 codes were significantly enriched among pAD cases versus controls (Fisher's exact P-value $< 3 \times 10^{-5}$ for 1,693 tested codes; top 20 enriched ICD-10 codes are shown in Figure 4.1). Within digestive systems disorders, ICD-10 code K61 (abscess of anal and rectal regions), followed by K50 (Crohn's disease) were the most significantly enriched (log odds ratio=4.04 and 1.96 respectively). This is expected as many perianal fistulas start as abscess of the perianal region and later extend to form a fistula connecting the anorectal canal to the perianal skin [**?**], and perianal fissures and fistulas are known subphenotypes of CD [**?**]. Other examples include K57 (Diverticular disease of the intestine; log odds ratio=0.74; P-value=1.2×10^{-87}), which is also a known cause of pAD [**?**]. Among non-digestive codes, haemorrhoids (I84) was the most significantly enriched diagnosis (P-value=6×10^{-98}; 38% in pAD cases versus 6% in controls; log odds ratio=2.3). This enrichment could indicate a shared pathogenesis between haemorrhoids and pAD. However, it could also be due to a higher likelihood of diagnosing haemorrhoids in patients with more serious ano-rectal disorders such as pAD, compared to the general population where haemorrhoids patients with no other ano-rectal manifestations are less likely to seek medical advice, and may therefore remain undiagnosed.

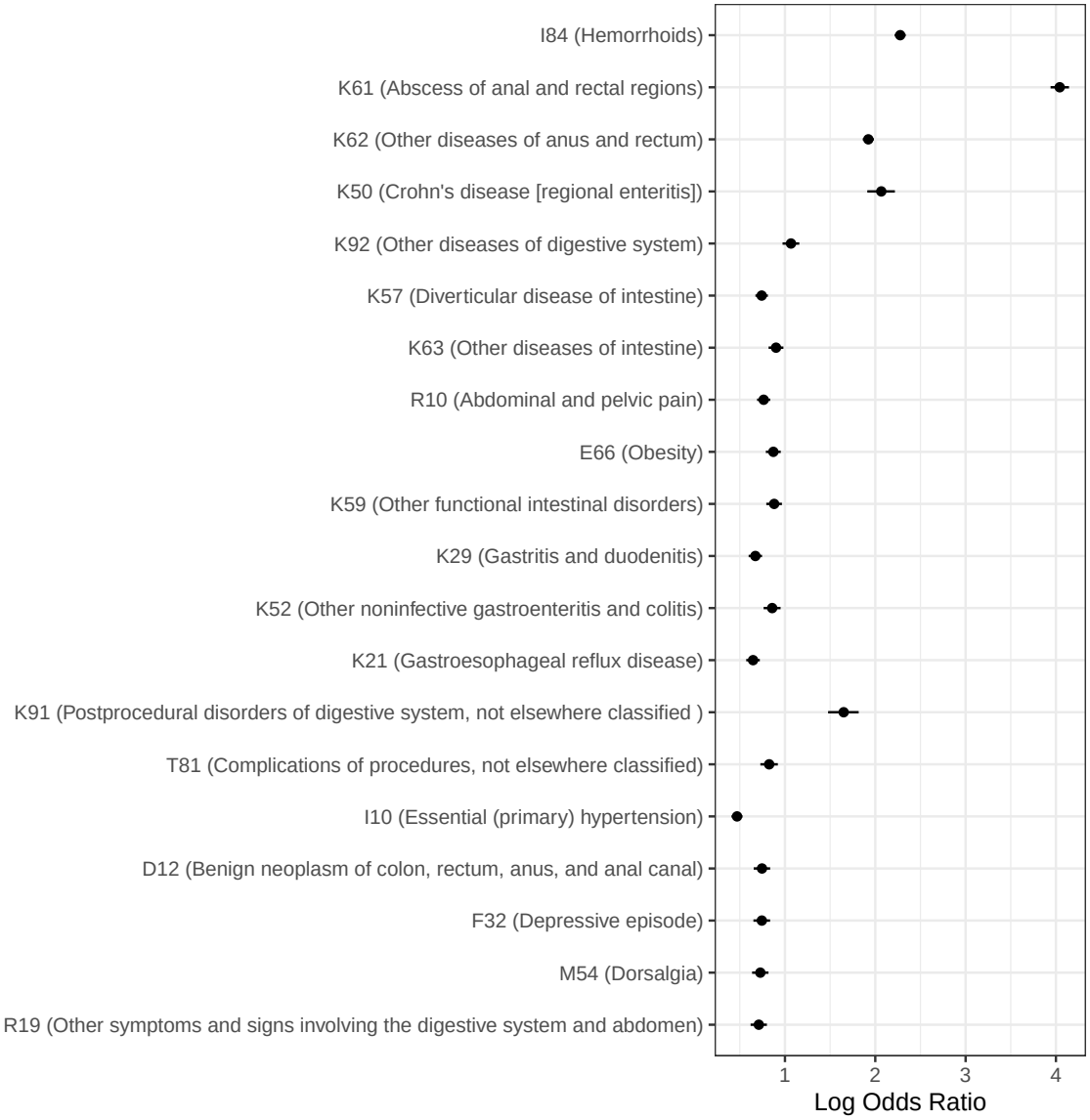


Fig. 4.1 Top 20 ICD-10 codes enriched in pAD cases versus pAD controls ordered by Fisher's exact test P-value. Odds ratios and their 95% confidence intervals were calculated using Fisher's exact test based on the number of pAD cases and controls with and without each particular ICD-10. Note that the control exclusion criteria in Table 4.2 were not applied in this analysis.

4.4.2 Identifying genome-wide significant loci

Defining genome-wide significant loci in GWAS studies is often complicated by the widespread correlation between proximal genetic variants (i.e. linkage disequilibrium or LD). To identify pAD-associated loci, I used an LD clumping procedure, which outputs a set of *index variants*,

each representing a set of highly correlated variants in a locus. Additionally, LD clumping identifies nominally-associated variants that are highly correlated with the index variant at each locus (which I will refer to as LD friends; $R^2 > 0.5$ and P-value < 0.01; Methods). In total, seven independent loci achieved genome-wide significant association (P-value < 5×10^{-8}). All index variants were well-imputed (INFO $\geq$ 0.99). I also compared the index variants MAFs to population MAFs to ensure that they did not significantly deviate from expected MAFs in non-Finnish Europeans (NFE). All index variants' MAFs matched MAFs obtained from 1000 Genomes Project (1000GP; Table 4.4; see Methods for how MAF deviation from the general population was formally assessed).

Table 4.4 Genome-wide significant index variants in the UKBB analysis. Odds ratio and their 95% confidence intervals are shown. Minor allele frequencies (MAF) in UKBB and 1000GP (NFE) are shown in the last two columns.

Chromosome	Position (b38)	Effect Allele	Odds Ratio	P-value	MAF UKBB	MAF 1000GP
3	52,992,368	T	1.13 (1.08 - 1.17)	1.5×10^{-8}	0.42	0.44
6	31,044,486	G	1.13 (1.08 - 1.18)	2.2×10^{-8}	0.37	0.37
6	31,113,288	C	1.13 (1.08 - 1.18)	1.1×10^{-8}	0.41	0.44
6	31,113,923	A	1.12 (1.08 - 1.17)	3.2×10^{-8}	0.49	0.49
6	31,148,469	A	1.12 (1.08 - 1.17)	2.6×10^{-8}	0.44	0.45
9	22,119,196	T	0.89 (0.85 - 0.93)	2.7×10^{-8}	0.48	0.47
11	10,356,352	C	0.88 (0.84 - 0.92)	7.3×10^{-9}	0.29	0.30

Four of the seven loci were located in the major histocompatibility complex region (MHC; 6p21.33), and one locus in each of 3p21.1, 9p21.3 and 11p15.4. The MHC region is known to be highly polymorphic and to exhibit complex and long-range LD patterns, which complicate the definition of independent loci. For example, two of the four MHC loci overlapped, with their two independent index variant located less than 700 bp apart. One of the two index variants (6:31113288_T_C) tagged a large number of variants in the locus, while the other tagged no variants (6:31113923_A_G; $R^2 > 0.8$; Figure 4.3). Compared to the four MHC loci, the three non-MHC loci had less complex LD patterns. All three non-MHC index variants tagged a large number of variants and there were no overlapping independent loci in any of them (Figure 4.2). Given the complexity of the MHC loci, I performed a number of post-GWAS checks to better understand the LD structure of all seven loci, which I will describe in the next section.

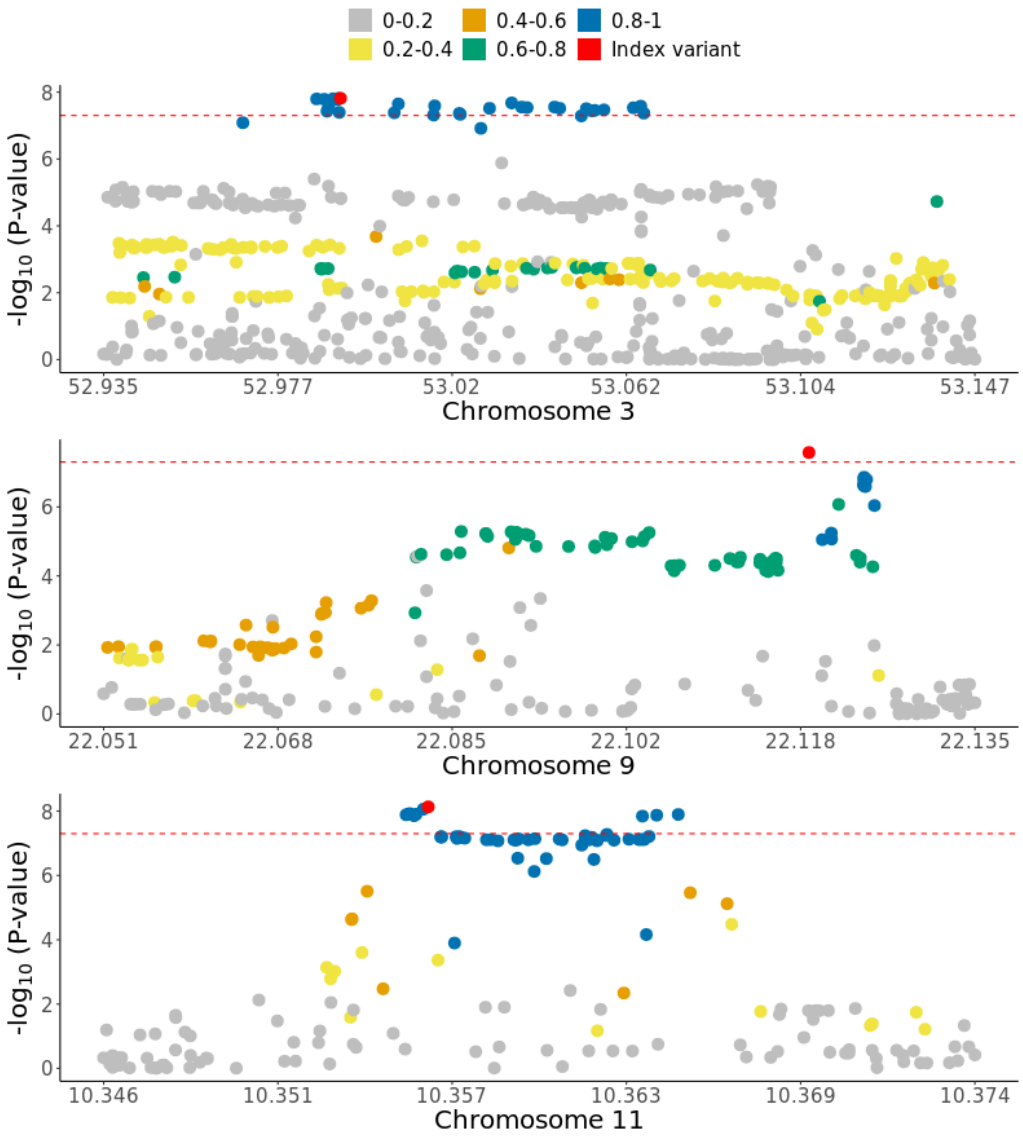


Fig. 4.2 Regional association plots for the three non-MHC loci, with position (build 38) plotted on the x-axis and $-log_{10}$ P-values shown on the y-axis for each variant. Colors indicate the R^2 between each variant and the index variant, and the red horizontal line indicates genome-wide significance (P-value = 5×10^{-8}).

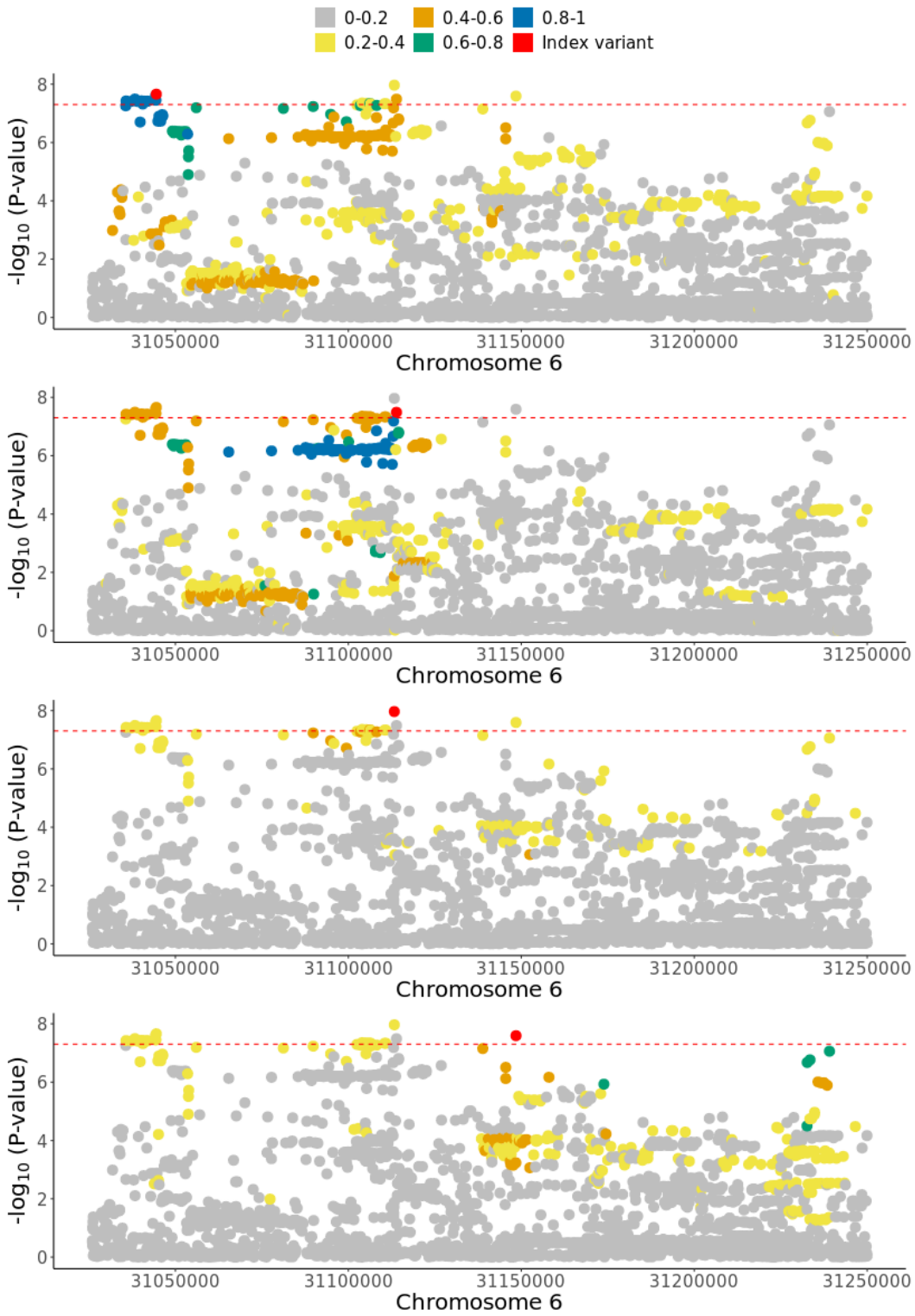


Fig. 4.3 Regional association plots for the four MHC loci, with position (build 38) plotted on the x-axis and $-log_{10}$ P-values shown on the y-axis for each variant. Colors indicate the R^2 between each variant and the index variant, and the red horizontal line indicates genome-wide significance (P-value = 5×10^{-8}). The second and third plots show the two overlapping MHC loci (index variants: 6:31113288_T_C and 6:31113923_A_G, respectively).

4.4.3 Post-GWAS quality checks

Spurious associations can seriously affect the validity of any significant results in GWAS studies. At the level of a single locus, spurious associations can be diagnosed by assessing the relationship between the index variant and its LD friends. For a given variant in a genome-wide significant locus, the lower its LD with the index variant, the weaker its association is expected to be. Loci where the association strength of LD friends does not "decay" as expected given their LD with the index variant are therefore particularly problematic. Specifically, such a mismatch would suggest that the LD structure that drives the observed association strength of LD friends does not match the general population LD. A possible source of this mismatch may be cryptic subpopulation stratification, which often contributes to false positive associations in GWAS [**?**].

I investigated the seven genome-wide significant loci to ensure the association signal follows the expected LD pattern in the general population. For this check to be valid, LD needs to be computed from a suitable matching reference panel such as 1000GP. Additionally, each index variant needs to have a number variants LD friends. To this end, I computed R^2 between each variant and the index variant at each locus using NFE individuals in 1000GP as a reference panel. For each pAD-associated locus, I quantified the correlation between R^2 and P-values (on the $-log_{10}$ scale) of each index variant's LD friends. Additionally, I performed two follow-up assessments for the loci where this correlation is weak ($\rho < 0.2$) or cannot be computed due to a lack of LD friends.

4.4.4 Relationship between P-value and LD

Index variants in 3p21.1, 9p21.3 and 11p15.4 had a large number of LD friends (N=63, 66, and 49, respectively), and the P-values for each index variants' LD friends were highly correlated with R^2 (ρ = 0.98, 0.74, 0.83, respectively), indicating that the P-values closely match the expected LD pattern in NFE. Two of the MHC loci also showed a similar LD decay pattern (index variants 6:31044486_G_C and 6:31148469_G_A in Figure 4.4), with a strong correlation between P-values and R^2 (Figure 4.4). However, this correlation did not hold for the two other overlapping MHC loci mentioned earlier, which motivated me to further investigate these two loci (index variants: 6:31113288_T_C and 6:31113923_A_G).

A complex LD pattern at two MHC loci

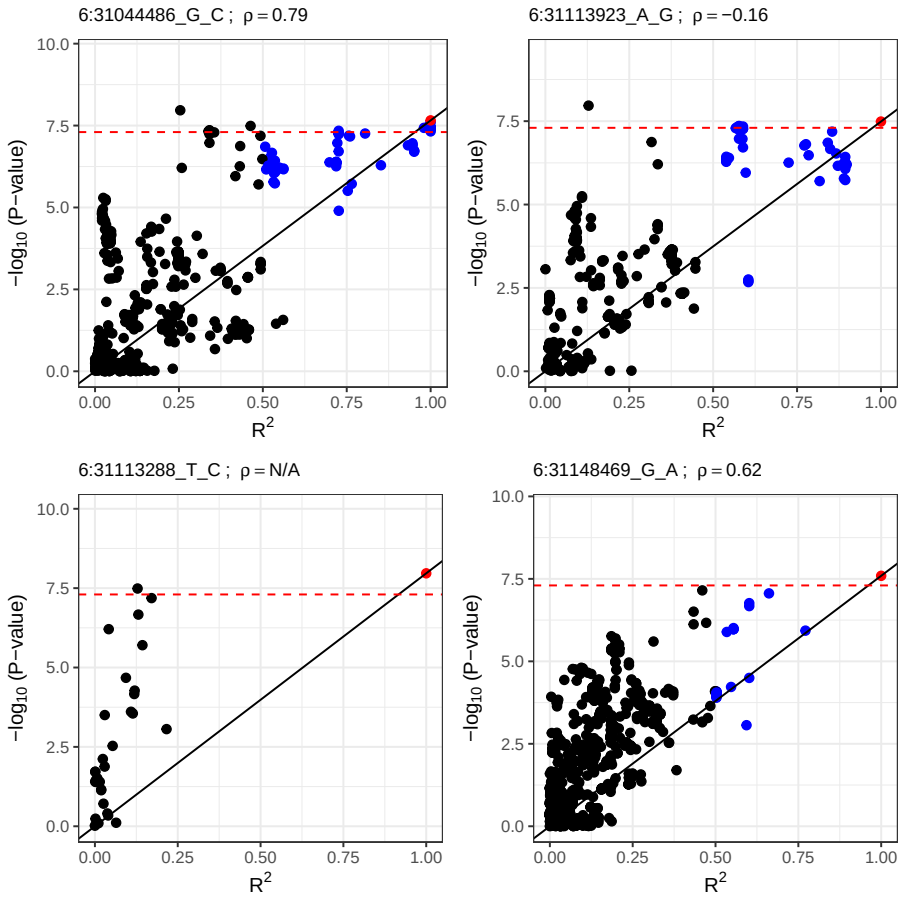


Fig. 4.4 LD decay plots showing association P-values for the four genome-wide significant loci in the MHC region (x-axis) and each variant's R^2 with the index variant, derived from NFE in 1000GP (y-axis). Red dots and titles indicate the index variant in each locus. Blue dots indicate each index variant's LD friends, and the red horizontal line indicates genome-wide significance level (P-value $< 5 \times 10^{-8}$). The black line is fitted to the origin (0,0), and to the point $(1, -log_{10}(P_{index\_variant}))$, and shows the expected association strength given the LD with the index variant.

First, one of MHC index variants at the two overlapping MHC loci did not tag any LD friends and therefore the correlation between P-value and R^2 could not be assessed (6:31113288_T_C in Figure 4.4). It is unclear whether the absence of LD friends for 6:31113288_T_C suggests that it is a truly independent variant, or whether it is driven by a mismatch between the LD patterns in UKBB and 1000GP. Such a mismatch may lead to an underestimation of LD between the index variant and its LD friends. To answer this question, I recalculated the LD values in 1000GP using only British individuals (GBR; N=90), and found that the

index variant did not tag any LD friends in 1000GP GBR as well (Figure 4.5). Given that 6:31113288_T_C is well-imputed (INFO=0.99) and common and that it is not well-tagged in both the NFE and GBR subpopulations in 1000GP, it is unlikely that the its association is driven by British-ancestry-specific LD. However, it is important to note that this does not rule out possible subpopulation stratification at this locus, which could potentially drive this association.

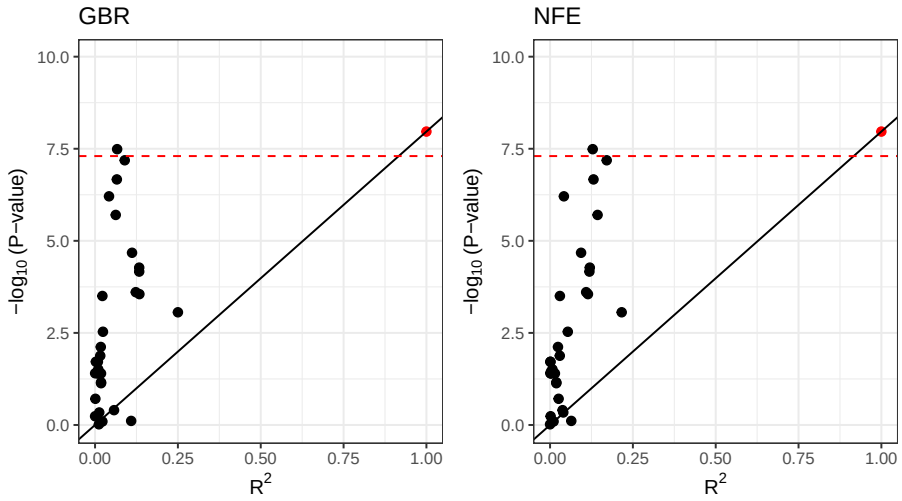


Fig. 4.5 LD decay plots showing association P-values for the locus around index variant 6:31113288_T_C. Each variant's R^2 with the index variant, derived from NFE and GBR in 1000GP is shown on the x-axis and $-log_{10}$ P-values on the y-axis, showing that the index variant does not tag any LD friends in both NFE and GBR. Red dots indicate the index variant and the red horizontal line indicates genome-wide significance level (P-value $< 5 \times 10^{-8}$). The black line is fitted to the origin (0,0), and to the point $(1, -log_{10}(P_{index\_variant}))$, and shows the expected association strength given the LD with the index variant.

Second, for the other overlapping MHC locus, R^2 and P-values showed an inverse correlation (index variant 6:31113923_A_G in Figure 4.4; ρ=-0.16). This inverse correlation between R^2 and P-values suggest that not all the LD friends' P-values conform to their expected P-values given their LD with the index variant. I hypothesised that the reversal of correlation may be caused by the subset of LD friends with R^2 close to the value used for defining LD friends. This subset could lead to an inverse correlation due to a stronger-than-expected P-value given their LD with the index variant. To this effect, I found that 10 LD friends had a genome-wide significant P-value ($< 5 \times 10^{-8}$) despite all having an R^2 of 0.58 with the index variant (Figure 4.4). When I repeated the LD clumping procedure at this locus with a higher clumping R^2 cutoff (=0.6), I found that this subset of variants constituted a new genome-wide significant locus. This suggests that the identification of independent loci at

this region is sensitive to the choice of LD clumping R^2 cutoff, which further complicates the identification of independent loci at this region.

4.4.5 FinnGen GWAS

Similar to UKBB, other national biobanks with genetic, clinical and phenotypic data are available. Although most national biobanks limit access to their individual-level genotype and phenotype data to approved researchers only, results from secondary analyses, including GWAS summary statistics, are made publicly available.

FinnGen is a national biobank whose aim is to collect genetic and phenotypic data for 500,000 Finnish individuals. The latest data freeze (Data Freeze 9) has genotype data for over 377,000 individuals and GWASes for over 2,200 clinical endpoints were carried out. FinnGen uses a different clinical coding system from ICD to organise phenotypes into endpoints (FinnGen endpoints). There are two main differences between UKBB and FinnGen in terms of their clinical code structure. First, most FinnGen endpoints have parallel ICD codes, but additional FinnGen endpoints are created at request. Bespoke endpoints define certain inclusion or exclusion criteria based on ICD codes, or sometimes combine codes from different ICD chapters to create a new endpoint. Second, FinnGen endpoints are curated by experts in each field and are constantly reviewed in different FinnGen data freezes. They are broadly classified as *core endpoints*, or *non-core endpoints*. Basic statistics such as prevalence and gender ratio are calculated for all FinnGen endpoints, while GWAS is conducted only for core endpoints only.

ICD-10 code K60 corresponds to FinnGen endpoint K11_FISSANAL (Fissure and fistula of anal and rectal regions). K11_FISSANAL defines cases and controls similar to my UKBB cohort definition outlined in Table 4.2. However, K11_FISSANAL was considered a core endpoint only until Data freeze 7, and GWAS summary statistics for K11_FISSANAL are therefore unavailable in later data freezes.

4.4.6 Identification of genome-wide significant loci in FinnGen

In order to investigate if the seven UKBB genome-wide significant loci replicated in an independent cohort and to identify additional loci, I downloaded GWAS summary statistics for FinnGen's clinical endpoint K11_FISSANAL. As of data freeze 7, FinnGen reports 6,610 pAD cases and 253,186 controls. There was no further information regarding the subtypes of pAD (e.g. numbers of fissure and fistula cases), and it is therefore unclear if the composition

of FinnGen's pAD case cohort is similar to the UKBB pAD case cohort. Understanding the differences in subphenotype composition of each cohort is important to understand if differences in association at genome-wide significant loci is driven by genetic factors (e.g. differences in MAFs or LD structure) or by phenotypic differences between the cohorts.

After I filtered out variants with MAF < 0.01, a total of 9,054,355 variants remained. There was an acceptable level of genomic inflation (median χ^2=0.495; λ_{GC}=1.089). To identify genome-wide significant loci, I used an LD clumping approach similar to the UKBB analysis, with the only difference being that I calculated LD from Finnish Europeans in 1000GP (FE; N=99). I found three genome-wide significant non-MHC loci: 1p34.2, 6p25.3 and 12q24.21 (P-value < 5×10^{-8}). Imputation quality information was not available in the downloaded summary statistics, so I was not able to confirm if the index variants were well-imputed. However, the index variants' MAFs matched MAFs derived from FE in 1000GP, suggesting that they are imputed or genotyped with high accuracy (Table 4.5). Furthermore, I performed similar post-GWAS checks to UKBB to ensure the P-value of the index variants' LD friends match their expected values given their LD with the index variant. All three showed a good decay of P-values with LD (ρ=0.92, 0.74 and 0.44, respectively; Figure 4.6)

Table 4.5 Genome-wide significant index variants in the FinnGen GWAS. Odds ratio and their 95% confidence intervals are shown. Minor allele frequencies (MAF) in FinnGen and 1000GP (FE) are shown in the last two columns.

Chromosome	Position (b38)	Effect Allele	Odds Ratio	P-value	MAF FinnGen	MAF 1000GP
1	39,817,036	T	1.14 (1.09 - 1.19)	7.2×10^{-10}	0.21	0.22
6	1,771,278	T	0.9 (0.87 - 0.93)	6.7×10^{-9}	0.42	0.39
12	114,235,969	T	1.11 (1.07 - 1.15)	7.0×10^{-9}	0.47	0.47

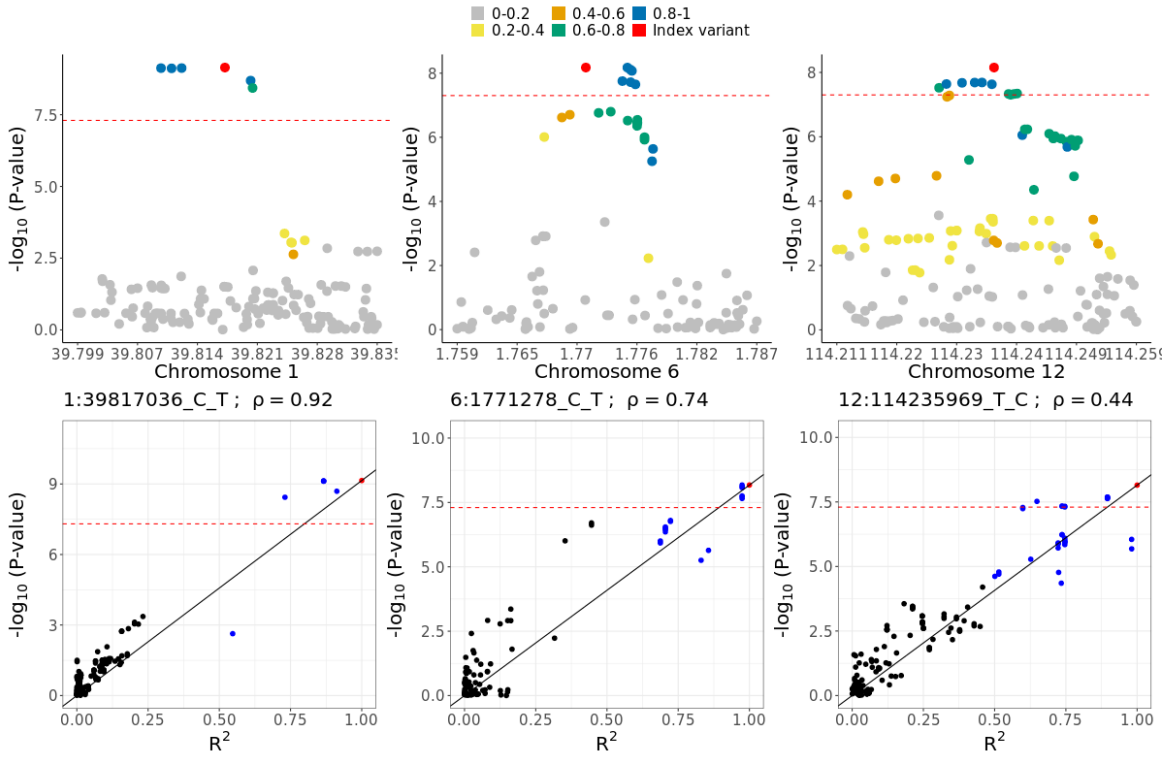


Fig. 4.6 (Top) Regional association plots for the three FinnGen loci, with position plotted on the x-axis and $-log_{10}$ P-values shown on the y-axis for each variant. Colors indicate the LD value of each variant with the index variant, and the red horizontal line indicates genome-wide significance (P-value = 5×10^{-8}). (Bottom) LD decay plots showing association P-values for the three genome-wide significant loci in FinnGen (x-axis) and each variant's R^2 with the index variant, derived from FE in 1000GP (y-axis). Red dots and titles indicate the index variant in each locus. Blue dots indicate each index variant's LD friends. The red horizontal line indicates genome-wide significance level, and the black line is fitted to the origin (0,0), and to the point $(1,-log_{10}(P_{index\_variant}))$, and shows the expected association strength given the LD with the index variant.

4.4.7 Replication of UKBB loci in FinnGen

LD pattern in Finnish Europeans

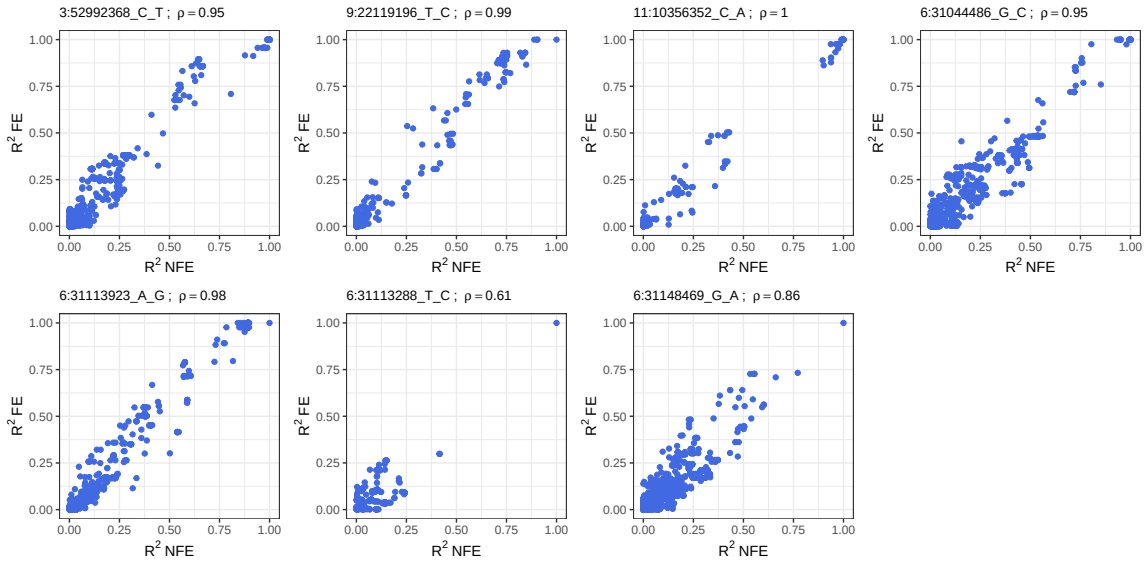
In GWAS studies, the true causal variant at a given associated locus is often unknown due to LD between variants. Moreover, it is often the case that the true causal variant may not even be genotyped in array-based GWAS studies, or may not be imputed due to different imputation protocols and QC metrics being used in different GWAS. When assessing replication of a GWAS locus between two cohorts, it is therefore important to ensure that variants that are genotyped or imputed in the two cohorts have similar LD structures. Indeed, a lack of GWAS

hit replication is sometimes driven by a difference in LD patterns between the two studies under comparison, one of which may not have genotyped or imputed any variants that tag the true causal variant in its respective population [**?**]. Finnish and non-Finnish Europeans are known to exhibit systematic difference in their LD structure, which may affect the ability to replicate the pAD-associated loci discovered in the UKBB. To compare the LD pattern between FE and NFE at the pAD-associated loci, I computed R^2 values between the index variant and all variants in each locus using the FE and NFE subpopulations of 1000GP as references panels. Additionally, I compared MAF for each variant in UKBB and FinnGen.

I found that MAF was nearly perfectly correlated in NFE and FE in all seven pAD-associated loci ($\rho > 0.94$; Figure 4.7a). R^2 were also strongly correlated in all loci ($\rho > 0.86$; Figure 4.7b). Notably, despite a strong R^2 correlation, 6:31113288_T_C did not have any LD friends in FE, similar to GBR and NFE. Overall, with the exception of 6:31113288_T_C which did not have any LD friends in FE, both MAF and the LD structure were consistent across all pAD-associated loci between NFE and FE. Replication of UKBB hits can therefore be reasonably assessed in the FinnGen GWAS.

I found that two UKBB non-MHC index variants replicated in FinnGen (FinnGen P-value $< 7 \times 10^{-3}$ for seven variants; Table 4.6), and that neither of them showed evidence of heterogeneity fo effect size ($P_{het} < 7 \times 10^{-3}$; Table 4.6). Additionally, all the five index variants that failed to replicate also showed evidence of heterogeneity of effect sizes.

(a)



(b)

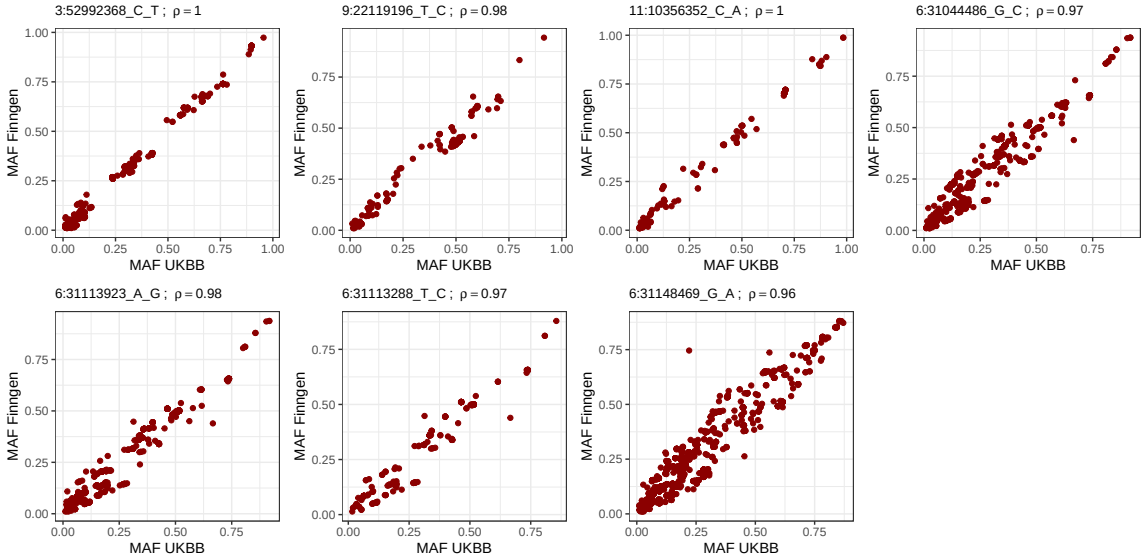


Fig. 4.7 (a) R^2 between all variants within each locus' boundaries and the index variant in the seven genome-wide significant loci identified in UKBB. R^2 are derived from non-Finnish Europeans (x-axis) and Finnish Europeans (y-axis) in 1000GP. Pearson correlation coefficients and index variants are indicated on top of each figure. (b) MAF of all variants in the UKBB (x-axis) and FinnGen (y-axis).

Table 4.6 Replication of the UKBB genome-wide significant index variants in FinnGen. Odds ratio and their 95% confidence intervals are shown for both cohorts. The heterogeneity of effect P-value is shown in the last column. Only two index variants passed the replication threshold (3:52992368_C_T and 11:10356352_C_A; P-value < 7×10^{-3} for seven variants).

Index variant	P-value UKBB	P-value FinnGen	OR UKBB	OR FinnGen	P_{het}
3:52992368_C_T	1.5×10^{-08}	1.2×10^{-03}	1.13 (1.08 - 1.17)	1.06 (1.02 - 1.1)	0.03
6:31044486_G_C	2.2×10^{-08}	5.7×10^{-02}	1.13 (1.08 - 1.18)	1.04 (1 - 1.07)	1.9×10^{-03}
6:31113923_A_G	3.2×10^{-08}	3.9×10^{-02}	1.12 (1.08 - 1.17)	1.04 (1 - 1.07)	3.6×10^{-03}
6:31113288_T_C	1.1×10^{-08}	1.5×10^{-01}	1.13 (1.08 - 1.18)	1.03 (0.99 - 1.07)	7.2×10^{-04}
6:31148469_G_A	2.6×10^{-08}	9.9×10^{-01}	1.12 (1.08 - 1.17)	1 (0.97 - 1.04)	2.0×10^{-05}
9:22119196_T_C	2.7×10^{-08}	2.0×10^{-02}	0.89 (0.85 - 0.93)	0.96 (0.93 - 0.99)	6.0×10^{-03}
11:10356352_C_A	7.3×10^{-09}	3.4×10^{-07}	0.88 (0.84 - 0.92)	0.91 (0.87 - 0.94)	0.27

4.4.8 Replication of FinnGen loci in UKBB

Following the same replication approach, I tested the replication of FinnGen's three genome-wide significant loci in the UKBB GWAS. I found evidence of replication for all three index variants in the UKBB (UKBB P-value < 0.017 for three variants), and none of the variants showed evidence of heterogeneity of effect sizes (P_{het} < 0.017; Table 4.7).

Table 4.7 Replication of the FinnGen genome-wide significant index variants in UKBB. Odds ratio and their 95% confidence intervals are shown for both cohorts. The heterogeneity of effect P-value is shown in the last column. All three variants passed the replication threshold (P-value < 0.017 for three variants).

Index variant	P-value UKBB	P-value FinnGen	OR UKBB	OR FinnGen	P_{het}
1:39817036_C_T	1.6×10^{-06}	7.2×10^{-10}	1.13 (1.08 - 1.19)	1.14 (1.09 - 1.19)	0.77
6:1771278_C_T	1.5×10^{-04}	6.7×10^{-09}	0.91 (0.87 - 0.96)	0.9 (0.87 - 0.93)	0.63
12:114235969_T_C	1.3×10^{-03}	7.0×10^{-09}	1.07 (1.03 - 1.12)	1.11 (1.07 - 1.15)	0.2

4.4.9 Meta-analysis of UKBB and FinnGen

Meta-analysis between GWAS cohorts is commonly used to increase statistical power to identify genome-wide significant loci. Practically, meta-analysis is carried out when the are constraints on sharing individual-level data, or when genotype data from several studies cannot be combined [**?**]. In these cases, meta-analysis of association summary statistics is the preferred analytical approach, and there is ample evidence that it achieves similar statistical power as combining genotype data from several studies [**?**].

Meta-analysis and identification of genome-wide significant loci

I performed a fixed-effects meta-analysis between UKBB and FinnGen effect sizes and standard errors using METAL (see Methods for more details). Because I performed a meta-analysis between two GWAS summary statistics from Finnish and Non-Finnish Europeans, I performed LD clumping separately with an LD panel of NFE and FE in 1000GP, and found 18 genome-wide significant loci (P-value $< 5 \times 10^{-8}$). I tested whether the index variants' effect size estimates were consistent between UKBB and FinnGen using Cochran's Q test, which is implemented in METAL. A strong deviation from the null hypothesis that effect sizes are similar between UKBB and FinnGen reflects uncertainty around the meta-analysed effect size estimate. To this end, I found no evidence of heterogeneity for any of the 18 index variants ($P_{het} < 3 \times 10^{-3}$ for 18 variants; Table 4.8). Furthermore, I measured the correlation between the P-values and R^2 values at each locus, similar to the previous LD decay check. Six of these loci showed either weak or inverse correlation between P-values and R^2 derived from either NFE and FE ($\rho < 0.2$), and were therefore removed from the rest of the downstream analyses (more details in Methods).

Table 4.8 Meta-analysis genome-wide significant loci (P-value < 5×10^{-8}), showing the index variant at each locus, the meta-analysis P-value, and the odds ratio in UKBB, FinnGen, and in the meta-analysis. 95% confidence intervals are shown for each odds ratio value. MAF is shown for UKBB and FinnGen. The last column shows the P-value of the effect size heterogeneity test, where $P_{het} < 3 \times 10^{-3}$ suggests evidence of heterogeneity of effects. The six loci that failed the LD decay test are highlighted in bold.

Index variant	Meta-analysis P-value	OR UKBB	OR FinnGen	OR Meta-analysis	MAF UKBB	MAF FinnGen	P_{het}
1:39809417_A_T	7.4×10^{-15}	1.13 (1.08 - 1.19)	1.14 (1.09 - 1.19)	1.14 (1.1 - 1.17)	0.25	0.22	0.84
1:39836225_G_C	4.1×10^{-08}	1.09 (1.04 - 1.15)	1.11 (1.06 - 1.16)	1.1 (1.07 - 1.14)	0.19	0.17	0.63
3:53034026_C_T	7.5×10^{-10}	1.13 (1.08 - 1.17)	1.07 (1.03 - 1.11)	1.09 (1.06 - 1.12)	0.42	0.38	0.05
5:64868326_TTTC_T	2.0×10^{-08}	0.89 (0.85 - 0.93)	0.94 (0.91 - 0.98)	0.92 (0.89 - 0.95)	0.34	0.35	0.05
6:1775202_G_A	1.0×10^{-11}	0.91 (0.87 - 0.95)	0.9 (0.87 - 0.93)	0.9 (0.88 - 0.93)	0.28	0.43	0.80
6:31121854_C_T	4.2×10^{-08}	1.11 (1.07 - 1.16)	1.06 (1.02 - 1.1)	1.08 (1.05 - 1.11)	0.43	0.34	0.08
6:31253340_T_C	3.8×10^{-08}	1.1 (1.06 - 1.15)	1.07 (1.03 - 1.1)	1.08 (1.05 - 1.11)	0.50	0.42	0.24
6:133008360_T_A	2.7×10^{-08}	1.12 (1.07 - 1.19)	1.1 (1.05 - 1.15)	1.11 (1.07 - 1.15)	0.17	0.16	0.47
6:133260944_G_A	4.7×10^{-08}	1.11 (1.06 - 1.17)	1.08 (1.04 - 1.13)	1.1 (1.06 - 1.13)	0.22	0.20	0.42
6:133267939_T_C	4.5×10^{-08}	1.09 (1.05 - 1.14)	1.07 (1.03 - 1.11)	1.08 (1.05 - 1.11)	0.49	0.55	0.42
7:2524404_G_A	4.1×10^{-08}	1.14 (1.07 - 1.22)	1.13 (1.06 - 1.2)	1.14 (1.09 - 1.19)	0.11	0.09	0.76
8:70735125_A_G	3.9×10^{-11}	0.83 (0.77 - 0.9)	0.82 (0.76 - 0.89)	0.83 (0.78 - 0.87)	0.09	0.06	0.94
8:70993166_AAGTT_A	1.2×10^{-10}	0.83 (0.77 - 0.9)	0.82 (0.75 - 0.88)	0.82 (0.78 - 0.87)	0.08	0.05	0.75
9:21995045_T_G	4.3×10^{-08}	1.41 (1.25 - 1.59)	NA	1.41 (1.25 - 1.6)	0.02	NA	1.00
9:22124505_A_T	2.1×10^{-08}	0.9 (0.86 - 0.93)	0.95 (0.91 - 0.98)	0.92 (0.9 - 0.95)	0.49	0.43	0.05
10:61661180_A_G	2.0×10^{-08}	1.09 (1.04 - 1.14)	1.08 (1.05 - 1.12)	1.09 (1.06 - 1.12)	0.68	0.56	0.90
11:10356352_C_A	1.3×10^{-13}	0.88 (0.84 - 0.92)	0.91 (0.87 - 0.94)	0.89 (0.87 - 0.92)	0.71	0.72	0.27
12:114235969_T_C	4.2×10^{-10}	1.07 (1.03 - 1.12)	1.11 (1.07 - 1.15)	1.09 (1.06 - 1.12)	0.48	0.53	0.20

4.4.10 Disentangling the genetic effect of pAD-associated variants on haemorrhoids

In section 4.4.1, I analysed the composition of the pAD case cohort and showed that it is significantly enriched with 198 ICD-10 clinical codes compared to pAD controls. Haemorrhoids was the most strongly enriched phenotype in pAD cases versus controls. I hypothesised that

this enrichment was also reflected at the level of genetic risk predisposition. To confirm this, I carried out a genetic correlation analysis between the pAD meta-analysis summary statistics and a UKBB-based haemorrhoids GWAS performed as part of the Pan UKBB case-control analysis (pan.ukbb.broadinstitute.org). I found strong evidence of high genetic correlation (ICD-10 code I84; P-value=5.37×10^{-26}; r_g=0.66). To validate this correlation, I repeated the genetic correlation analysis using a larger haemorrhoids GWAS of over 900,000 individuals by Zheng et al. 2021 [**?**]. I found a similar genetic correlation estimate that was even more significant than the estimate from the Pan-UKBB analysis (r_g=0.63; P-value=10^{-62}).

The existence of a strong genetic correlation and enrichment of haemorrhoids could be explained by several factors. First, pAD could be a co-morbidity of haemorrhoids, in a similar way that Type 2 diabetes and obesity are co-morbidities. This could be a result of the same risk factors (genetic or otherwise) underlying both diseases, potentially with varying effect sizes. Alternatively, clinical diagnostic factors could also account for this overlap. Both diseases are among the differential diagnoses for patients presenting with rectal pain, swelling, bleeding and discharge. Therefore, a patient suffering from inflamed haemorrhoids is more likely to be diagnosed if they also suffer from pAD (e.g. after performing rectal examination).

Bias introduced by clinical diagnostic factors cannot be completely addressed with observational data, as this will require constructing pAD case-control cohorts where haemorrhoids cases are balanced in both cases and controls. However, the impact of such bias could also be assessed by performing a pAD GWAS where haemorrhoids cases are excluded from cases and controls (pADexclHaem), and a haemorrhoids GWAS where pAD cases are excluded from cases and controls (HaemexclpAD). Comparing the effect sizes of the previously reported 12 index variants between pADexclHaem and HaemexclpAD may give an indication as to which genetic variants are likely to underlie both diseases and which are likely to be specific to pAD.

Constructing the two cohorts requires access to a individual-level phenotypic data in both UKBB and FinnGen. Since I do not have access to FinnGen's phenotype data, I tested the hypothesis that effect sizes are different between haemorrhoids and pAD in the UKBB only. To construct the HaemexclpAD case and control cohorts, I selected individuals who have been diagnosed with ICD-10 code I84 or ICD-9 code 455 in at least one inpatient episode as cases and excluded individuals with ICD-10 code K60 or ICD-9 code 565 from both cases and controls. Similarly, for pADexclHaem, I selected individuals who have been diagnosed with ICD-10 code K60 or ICD-9 code 565 in at least one inpatient episode and

excluded individuals with ICD-10 code I84 or ICD-9 code 455 from both cases and controls. Additionally, I applied the same control exclusion criteria in Table 4.2.

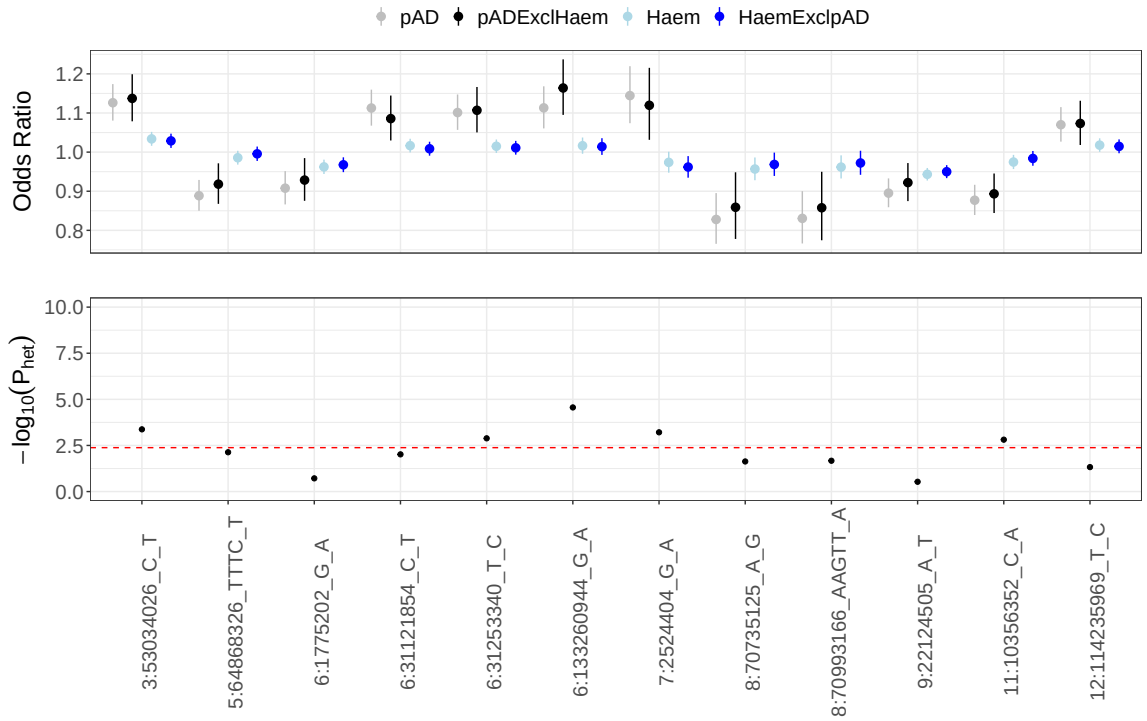


Fig. 4.8 Top plot shows the ffect sizes of the 12 pAD-associated index variants from four UKBB case/control cohorts: pAD in grey (N_{case}=4,606), pADExclHaem in black (N_{case}=2,799), Haem in light blue (N_{case}=29,285), and HaemExclpAD in blue (N_{case}=27,477). Bottom plot shows the heterogeneity of effect-size P-value (P_{het}) between the two disjoint GWAS analyses: pADExclHaem and HaemexclpAD. The red dotted line shows P_{het} significance threshold ($P_{het} < 4 \times 10^{-3}$).

I tested the association of each of the 12 pAD-associated index variants with each of the four phenotypes described above. First, I examined if any of the index variants were associated with the two haemorrhoids phenotypes Haem and HaemExclpAD. Since I performed a targeted association test, I set a more permissive association threshold for declaring significance than normally used to declare genome-wide associations (P-value $< 4 \times 10^{-3}$ for 12 variants). Despite the large difference in statistical power between the two haemorrhoids cohorts and the two pAD cohorts, I found that only three of the tested index variants achieved significant association in the better-powered haemorrhoids GWASes (index variants: 3:53034026_C_T, 6:1775202_G_A, and 9:22124505_A_T). Additionally, all three variants were significant in both Haem and HaemExclpAD, suggesting that the exclusion of pAD cases from the haemorrhoids cohort has little impact on their association. Moreover,

3:53034026_C_T showed significant evidence of effect size heterogeneity between pADExclHaem and HaemExclpAD ($P_{het} < 3 \times 10^{-3}$; Figure 4.8), with a significantly larger effect on pADExclHaem ($OR_{pADExclHaem} = 1.14 - 1.2$ and $OR_{HaemExclpAD} = 1.03 - 1.05$).

Although the other 11 index variants did not show a significant association with the two haemorrhoids definitions, four of them had significantly smaller effect sizes in HaemexclpAD than in pADexclHaem ($P_{het} < 4 \times 10^{-3}$; Figure 4.8). Moreover, five additional variants had suggestive evidence of heterogeneity of effect ($P_{het} < 0.05$), and for all five variants the effect size was larger in pADExclHaem than in HaemExclpAD.

Two conclusions can be made from this analysis. First, despite a much larger sample size in favour of the haemorrhoids cohorts, only three pAD-associated index variants were also associated with haemorrhoids, even with a relatively lenient threshold for association. Second, despite their nominal association with haemorrhoids, these three variants (and indeed all other variants) had a consistently smaller effect size on HaemExclpAD than pADExclHaem, and for 3:53034026_C_T that difference in effect size was significant. More importantly, all 12 index variants had a smaller effect size on haemorrhoids compared to pAD, although the P_{het} values did not always confirm a significant effect heterogeneity of effects. Overall, this confirms that the 12 discovered variants likely have more significant effects on pAD than haemorrhoids. Performing a similar 'disentanglement' analysis in both FinnGen and UKBB, and subsequently identifying which variants have a significantly larger effect size on pAD than haemorrhoids is a plausible way to validate this pattern. Such validation would more strongly establish these variants as bona fidae pAD-associated variants, with a significantly smaller effect size on haemorrhoids.

4.4.11 Replication of pAD-associated variants in the pCD meta-analysis

In order to understand the relationship between the genetic underpinnings of sporadic perianal manifestations and perianal manifestations within the context of Crohn's disease, I attempted to replicate the 12 pAD-associated variants in the pCD meta-analysis. I found no evidence that any of the 12 variants replicated in the pCD meta-analysis (P-value $< 4 \times 10^{-3}$ for 12 variants). However, it is worth noting that the pCD meta-analysis is much less powered to detect the effects that were detected in the pAD meta-analysis due to its smaller sample size. For a variant with a MAF of 0.49, the pCD meta-analysis has 75% power to detect a variant with odds ratio > 1.1 at a significance level of 4×10^{-3}. Although two variants fell within this MAF value (6:31253340_T_C and 9:22124505_A_T), neither of them passed the replication

P-value threshold. This suggests that they are either not associated with pCD or that their effect sizes are too small to detect with the statistical power of the pCD meta-analysis.

Table 4.9 Replication of the 12 pAD-associated variants in the pCD meta-analysis. P-values and odds ratios and their 95% confidence intervals from the pCD meta-analysis summary statistics are shown. None of the variants passed the replication threshold (P-value $< 4 \times 10^{-3}$ for 12 variants).

Chromosome	Position (b38)	Effect Allele	Odds Ratio	P-value	MAF
3	53,034,026	T	1.02 (0.96 - 1.08)	0.58	0.42
5	64,868,326	T	0.98 (0.92 - 1.04)	0.52	0.33
6	1,775,202	A	1 (0.94 - 1.07)	0.92	0.28
6	31,121,854	T	0.98 (0.93 - 1.04)	0.58	0.42
6	31,253,340	C	1.03 (0.97 - 1.09)	0.34	0.49
6	133,260,944	A	1.05 (0.98 - 1.13)	0.16	0.22
7	2,524,404	A	1.06 (0.96 - 1.17)	0.25	0.11
8	70,735,125	G	0.94 (0.84 - 1.04)	0.23	0.09
8	70,993,166	A	0.95 (0.85 - 1.06)	0.38	0.09
9	22,124,505	T	0.93 (0.88 - 0.99)	0.03	0.49
11	10,356,352	A	0.99 (0.93 - 1.06)	0.81	0.30
12	114,235,969	C	0.96 (0.91 - 1.02)	0.24	0.47

4.4.12 Identification of effector genes via colocalisation analysis

Many GWAS loci that have been uncovered over the last 15 years are located in non-coding regions. This complicates the task of understanding their downstream effects and linking them to effector genes. Over the last ten years, large-scale studies that map genetic variants associated with transcriptomic variation have improved our understanding of the downstream effects of disease-associated genetic variants. For example, the Genotype Expression Project (GTEx) has mapped genetic variants associated with individual variation in overall levels of gene expression (eQTL) and splicing (sQTL). Additionally, statistical methods that are able to integrate association signals from different studies have been applied to GWAS and QTL data in order to investigate which effector genes likely underpin disease-associated GWAS signals. Colocalisation analysis, for example, quantifies the probability that two association

signals are driven by a single causal variant (PP_4) and can therefore be used to compare GWAS and QTL association signals (more details in the Methods section).

I carried out colocalisation analysis between the 12 pAD-associated loci and eQTL and sQTL signals from GTEx v8 in a 1 mbp window centered around each locus' index variant. Across all 49 GTEx tissues, I performed the colocalisation with a total of 293 genes and all their splice junctions (see Methods for the number of genes and splice junctions tested at each locus).

Overall, I found high-confidence colocalisation evidence for seven loci, where at least one eQTL or sQTL colocalised with the association signal ($PP_4 > 0.8$; Table 4.10). All seven loci had at least a single colocalisation with an sQTL (12 sQTL genes), while five loci colocalised with at least one eQTL signal (eight eQTL genes), implicating a total of 15 genes. At many loci where both an eQTL and sQTL colocalisation were detected, distinct eQTL and sQTL genes were implicated. Moreover, QTLs in different tissues often implicated different genes. For example, the locus at index variant 7:2524404_G_A colocalised with two different genes (*BRAT1* in the liver and thyroid gland, and *LFNG* in the skin and whole blood). In fact, only three of the seven colocalised loci implicated a single gene, and only one locus implicated the same gene with high confidence in multiple tissues (index variant 5:64868326_TTTC_T and *CWC27*). The pleiotropic nature of genetic effects on gene expression is well documented in GTEx [**?**], and even in other organisms [**?** **?**]. This pleiotropy is often attributed to the widespread gene co-expression patterns, whereby the expression of multiple genes is controlled by a single locus, sometimes termed "QTL hotspots" [**?**]. Co-expressed genes are often found to be functionally related via shared biological pathways [**?** **?**]. To explore this, I performed a gene set enrichment analysis in four databases: Reactome [**?**], the Gene Ontlogy (GO) Molecular Function database, GO Cellular Component and GO Biological Processes [**?**]. I did not find any significantly enriched pathways in any of the the three GO databases or the Reactome database. Notably, 6 of the 15 genes were not found in the Reacome database, reflecting the lack of knowledge of their biological functions.

Table 4.10 Colocalisation analysis for the 12 pAD-associated index variants. The first column shows the index variants and the second and third columns shows the tissues and genes with high colocalisation PP_4 (> 0.8). Genes and their PP_4 values are shown in parentheses.

Index SNP	Tissues (eQTL)	Tissues (sQTL)
3:53034026_C_T	Kidney Cortex (ITIH4: 0.97), Colon Transverse (SFMBT1: 0.98), Esophagus Gastroesophageal Junction (SFMBT1: 0.95), Esophagus Muscularis (TMEM110: 0.82), Pancreas (TMEM110: 0.98)	Artery Aorta (ITIH4: 0.9), Artery Tibial (ITIH4: 0.97), Liver (ITIH4: 0.96), Nerve Tibial (ITIH4: 0.93), Liver (MUSTN1: 0.87), Esophagus Muscularis (RFT1: 0.84)
5:64868326_TTTC_T	Testis (CWC27: 0.86)	Esophagus Muscularis (CWC27: 0.81), Testis (CWC27: 0.84)
6:31121854_C_T	Lung (HLA-B: 0.86), Thyroid (HLA-B: 0.87), Adrenal Gland (POU5F1: 0.85), Brain Cerebellar Hemisphere (POU5F1: 0.89), Brain Cerebellum (POU5F1: 0.91), Brain Hypothalamus (POU5F1: 0.84)	Skin Not Sun Exposed Suprapubic (FLOT1: 0.85), Skin Not Sun Exposed Suprapubic (MICA: 0.81), Colon Transverse (PSORS1C1: 0.98), Lung (PSORS1C1: 0.99), Small Intestine Terminal Ileum (PSORS1C1: 0.89)

Table 4.10 *(continued)*

Index SNP	Tissues (eQTL)	Tissues (sQTL)
6:31253340_T_C	Lung (HLA-B: 0.86), Thyroid (HLA-B: 0.87), Adrenal Gland (POU5F1: 0.85), Brain Cerebellar Hemisphere (POU5F1: 0.89), Brain Cerebellum (POU5F1: 0.91), Brain Hypothalamus (POU5F1: 0.84)	Skin Not Sun Exposed Suprapubic (MICA: 0.81), Colon Transverse (PSORS1C1: 0.98), Lung (PSORS1C1: 0.99), Small Intestine Terminal Ileum (PSORS1C1: 0.89)
7:2524404_G_A	Liver (BRAT1: 0.94), Skin Not Sun Exposed Suprapubic (LFNG: 0.92), Skin Sun Exposed Lower leg (LFNG: 0.99), Whole Blood (LFNG: 0.85)	Thyroid (BRAT1: 0.95), Skin Not Sun Exposed Suprapubic (LFNG: 0.86), Skin Sun Exposed Lower leg (LFNG: 0.95), Whole Blood (LFNG: 0.99)
11:10356352_C_A	NA	Adrenal Gland (AMPD3: 0.85)
12:114235969_T_C	NA	Vagina (RBM19: 0.81)

Only one locus consistently implicated a single gene across various tissues and with both eQTL and sQTL colocalisation evidence (*CWC27*; index variant 5:64868326_TTTC_T; $PP_4 > 0.8$). The protective allele of the index variant (odds ratio=0.91) increases the expression of *CWC27* (eQTL effect size in testis=0.29; Figure 4.9), and also changes usage of five *CWC27* splice junctions in testis. *CWC27* codes for a spliceosomoal complex component. Although little is known about its role in common complex diseases, rare *CWC27* variants are known to be associated with a degenerative eye disorder called retinitis pigmentosa, which may or may not be accompanied with skeletal deformities. Indeed, Xu et al. [**?**] performed whole-exome sequencing of ten individuals from seven unrelated families, nine of which suffered from

retinitis pigmentosa, either alone or with a range of structural disorders such as bradydactyly, short stature, and craniofacial defects. In all seven families, rare protein-truncating variants in CWC27 were found, establishing *CWC27* as the effector gene for this group of rare disorders.

It is not obvious how common pAD-associated *CWC27* variants and rare *CWC27* mutations that cause skeletal abnormalities converge on the same molecular pathways. However, I found that another colocalised gene, *LFNG*, is associated with skeletal abnormalities. Two reports have shown that missense variants in *LFNG* are associated with spondylocostal dysostosis, a congenital disorder characterised by short neck, short stature and scoliosis [**?** **?**]. *LFNG* codes for a member of the glycosyltransferase family and plays an important role in vertebral formation during embryogenesis [**?**]. Notably, the two reported variants affected the active site of the protein product, rendering it functionally inactive. Additionally, the pAD-associated signal at 3:53034026_C_T colocalised with very high confidence with a *SFMBT1* eQTL in transverse colon. *SFMBT1* has been linked to idiopathic normal pressure hydrocephalus (iNPH), a late-onset structural disorder characterised by enlarged cerebral ventricles and increased accumulation of the cerebrospinal fluid (CSF) in the brain. Cerebral ventricles are a network of channels responsible for circulating the cerebrospinal fluid. Copy number variants in intron 2 of *SFMBT1* have been linked to iNPH in Finnish, Norwegian and Japanese populations [**?** **?**]. Although it is not understood how *SFMBT1* variants cause iNPH, the SFMBT1 protein has been detected in anatomical structures involved in CSF circulation, including smooth muscle cells of arteries [**?**]. The most well-characterised function of *SFMBT1* is its transcriptional repressor activity. The expression of *SFMBT1* has been found to halt the myogenic differentiation of progenitor myoblasts into muscle cells [**?**].

A compelling hypothesis emerges from the evidence that links *CWC27*, *LFNG* and *SFMBT1* loss-of-function variants to structural or skeletal disorders. Most relevant literature cites the cryptoglandular theory to account for the origin of anal fistulas, whereby perianal abscess that starts in the proctodeal glands extends to form fistulas [**?**]. But little is known so far about how genetic predisposition affects this progression and which biological pathways are responsible for this progression. In this regard, the implication of *CWC27*, *LFNG* and *SFMBT1* as effector genes hints at a possible role for pathways responsible for normal skeletal development. These dysregulated pathways might be responsible for the development, extension, and branching of fistulas. However, more robust genetic evidence is needed to identify effector genes, especially at the loci where colocalisation evidence implicates multiple genes in different tissues.

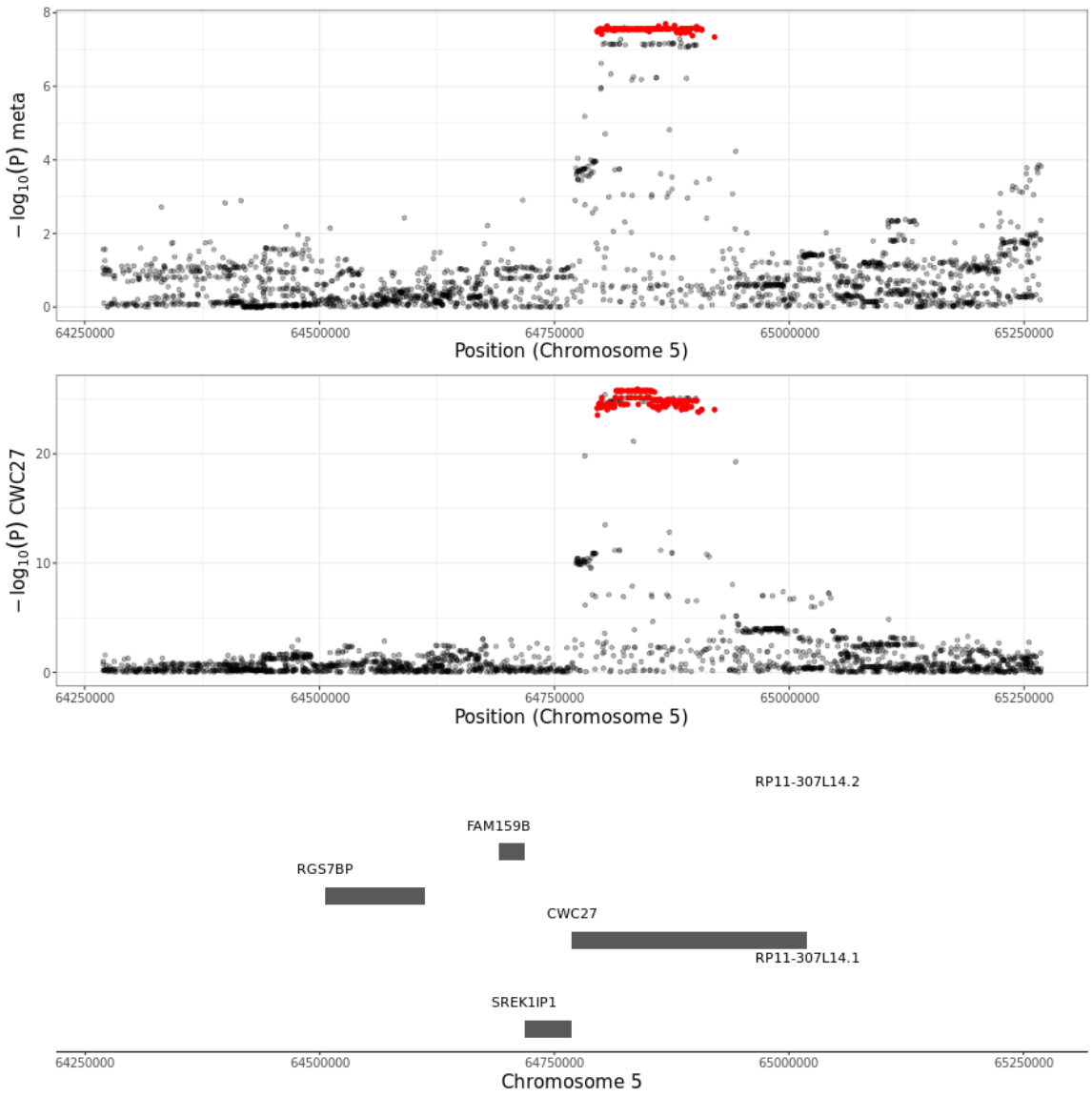


Fig. 4.9 Regional association plot for the pAD-associated signal at index variant 5:64868326_TTTC_T (top) and *CWC27* eQTL in testis (middle), showing position on the x-axis and $-log_{10}$ (P-value) on the y-axis. Red dots show pAD-associated genome-wide significant variants. Protein-coding and lincRNA gene positions are also shown (bottom).

4.5 Discussion

In this chapter, I have performed several analysis to map genetic variants associated with perianal disease (pAD), defined as anal and rectal fissures and fistulas. I have leveraged two large-scale national biobanks, UKBB and FinnGen, with a total of 11,216 cases and 585,420 controls. First, I performed a separate UKBB GWAS and found seven genome-wide significant loci, that I subjected to a number of post-GWAS quality checks to ensure their veracity. Next, I downloaded FinnGen GWAS summary statistics for the same phenotype and performed similar post-GWAS quality checks. I also attempted to cross-replicate each GWAS' findings, and found that three of the seven UKBB loci replicated in FinnGen, while all FinnGen loci replicated in the UKBB. Interestingly, all the UKBB loci that failed to replicate in FinnGen were located in the MHC region, which is known to be highly polymorphic and exhibits complex LD patterns that tend to be population-specific.

There are several possible explanations for the absence of replication at the MHC loci. First, the four MHC loci discovered in the UKBB GWAS may simply be spurious associations. As I showed in Section 4.4.4, association strength and R^2 were not correlated in at least two of the MHC loci, which suggests that the underlying LD structure does not match general population LD and may lead to spurious associations at these loci. This could possibly result from cryptic population stratification, but it could also result from poor genotyping or imputation. Second, it is possible that neither the index variants nor their LD friends tag the true causal variant in FE. This is likely to be the case for the MHC loci where the index variant tags few or no LD friends.

Despite the increased power afforded by the meta-analysis, several limitations should be noted. First, although case and control inclusion criteria are similar across UKBB and FinnGen, it is not obvious if the composition of the pAD case cohort is also similar. Similar to the ICD codes used in the UKBB analysis to identify anal fissures and fistula cases, FinnGen's clinical endpoint covers two broad clinical diagnoses: anal fissures and fistulas. I have shown in Table 4.1 that the proportion of anal fissures and fistula cases in the UKBB case cohort is roughly 2:1. Since FinnGen's individual-level data are not publicly available, I could not confirm if the proportion of anal fissure and fistula cases are similar in FinnGen. Possible compositional differences may introduce heterogeneity and decrease the power of the meta-analysis to discover pAD-associated variants. Second, it is unclear if FinnGen's case cohort is enriched in any other clinical endpoints compared to FinnGen's control cohort. Showing that the pAD cases are enriched in the same disorders (e.g. haemorrhoids and anal abscess) can serve as an important phenotypic quality control check to ensure that both

cohorts are as similar as possible, and maximises the ability of a meta-analysis.

The availability of individual-level phenotype data in the UKBB allowed me to disentangle the effects of the pAD-associated variants on pAD and haemorrhoids. This analysis showed that the effects of these variants were stronger and more significant on pAD than haemorrhoids despite a large difference in statistical power in favour of haemorrhoids. However, this analysis was limited to UKBB participant, and it is plausible that better powered GWAS of haemorrhoids may reveal that a larger proportion of the 12 pAD-associated variants are also associated with haemorrhoids. Indeed, when I replicated the index variants in the largest haemorrhoids GWAS (Zheng et al. [**?**]), I found that six of the 12 variants showed genome-wide significant association (P-value $< 5 \times 10^{-8}$). However, similar to the UKBB analysis, they all had a concordant but smaller effect sizes on haemorrhoids than pAD. A compelling interpretation of this shared genetic risk is that pAD may be a more severe form or manifestation of haemorrhoids, with the same genetic variants underlying both and with stronger effect sizes on pAD. But this observation may not be true for all haemorrhoids-associated loci. Over 100 haemorrhoids-associated loci were identified by Zheng et al., and it is plausible that most of these loci will not be associated with pAD if a genome-wide significant comparison of effect sizes was performed between the two diseases. Therefore, any conclusions made regarding the difference in effect sizes should be limited to these 12 loci.

Finally, I aimed to identify effector genes at each locus using colocalisation analysis. Although I identified several genes that colocalised with high confidence, evidence at most loci was conflicting, implicating several genes in several tissues. The role of these genes in many of the tissues where the colocalisations were detected, such as testis, thyroid and liver, was difficult to interpret given our knowledge of the pathogenesis of pAD. Although I showed evidence that three of these genes, *LFNG*, *CWC27*, and *SFMBT1* were necessary for normal musculoskeletal development, this does not confidently constitute sufficient evidence for a novel insight into pAD pathogenesis. To this end, more follow-up work should be conducted to better interpret these loci. First, more robust methods need to be employed to establish a causal link between pAD-associated loci and effector genes (e.g. Mendelian Randomisation methods [**?**]). Additionally, QTL studies from more relevant tissues need to be used. As discussed in 4.4.12, it is well-known that genetic variants affect the expression of different genes in different tissues. Therefore, QTLs derived from anorectal tissues will provide the best colocalisation and mendelian randomisation evidence for effector genes. However, the anal region is composed of several tissues and cell types, and it is plausible that

colocalisation with single-cell QTLs derived from anorectal biopsies will also potentially implicate different genes in different cell types. Therefore, the first step to identify the most likely effector genes is to identify the most relevant cell type via a heritability enrichment analysis (e.g. LDSC-SEG [**?**]). In conclusion, establishing a bona fide set of effector genes for these loci in relevant tissues will provide much stronger evidence that points to biological pathways implicated by pAD loci effector genes.

Chapter 5

Future Directions

5.1 Context and future directions of alternative splicing regulation in innate immunity

In chapter 2, I analysed genetic and gene expression data from macrophages, an important and highly responsive cell type in innate immunity, exposed to a wide range of stimuli. I have particularly focussed on the genetic regulation of an understudied layer of gene expression: alternative splicing. I linked sQTLs to immune-mediated disease loci via colocalisation analysis and demonstrated several examples of how IMD risk loci may dysregulate alternative splicing of a number of genes (*PTPN2*, *DENND1B* and *LRRK2*) in macrophages. In over half of the colocalisation events, it appears that low-usage splice junction may play a role in IMD risk, as evident in the example of an IBD-associated locus that increases the usage of a rare *PTPN2* splice junction.

Although this work represents an important step towards uncovering the role of alternative splicing in complex diseases, several avenues need to be explored in order to better understand the complex landscape of alternative splicing in different biological contexts and in complex disease risk. In my view, efforts should focus on three main objectives:

- Before hypotheses can be generated about the role of particular gene isoforms (e.g. the role of *PTPN2-205*) in macrophage response and disease risk, accurate characterisation of these isoforms needs to be established in both physiological conidtions and in response to environmental stimuli. This is particularly true for examples where subtle changes in isoform proportions is hypothesised to play a role in disease risk. Current short-read-based RNA-seq methods as well as alternative splicing quantification methods provide measurements that are either uncertain or challenging to interpret. In my

analysis, I used Leafcutter, which quantifies intron usage ratios, but does not provide complete isoform-level quantification. Arguably, precise isoform-level measurement is only attainable via long-read RNA-sequencing.

- Second, the function of different gene isoforms needs to be understood. Establishing the presence of particular isoforms and their differential splicing upon macrophage stimulation is not enough to understand their exact role in innate immune response.
- Third, the regulatory landscape of isoform expression needs to be established. The choice of alternative splice sites as well as the splicing factors that govern activation and/or suppression of particular splice choices is determined by complex cis- and trans-acting factors. How do these regulatory factors respond to external stimuli? How do they determine the choice of alternative splice sites and how do they regulate the expression of different isoforms in the context of innate immunity?

5.1.1 Long-read sequencing

Large-scale efforts to identify splicing variants of genes such as GENCODE, RefSeq and APRIS have improved our understanding of the diversity of the transcriptome. The same cannot be said about the tissue, cell-type and biological context distribution of splice variants. The tissue- and cell-type-specific functions of isoforms remain largely unknown due to the difficulty of quantifying alternative splicing at the single-cell level. Single-cell long-read RNA-seq (scLR RNA-seq) is one of the most promising technologies for studying the distribtion of splice variants. Although scLR RNA-seq methods have been slowly developing over the last decade, recent experimental developments and reduced cost make the study of alternative splicing in depth at the single cell level feasible and cost-effective. Until recently, low-throughput, high cost, and higher error rate of LRS technologies have been a major obstacles facing scLR RNA-seq, and rendering many of them unsuitable for population-level single-cell RNA-seq studies.

A recent scLR RNA-seq library preparation method developed by Al'Khafaji et al. [**?**] leverages the circular nature of SMRT-sequencing, Pacbio's proprietary sequencing method which greatly increases long-read sequencing. In previous Pacbio RNA-seq methods, the same cDNA molecule is sequenced up to 60 times in repetitive circular passes, greatly increasing sequencing accuracy. However, due to the relative short length of most RNA molecules (and their reverse-transcribed cDNAs), most transcripts were "over-sequenced". Although this achieved acceptable sequencing accuracy, it came at the cost of high throughput. Higher throughput and low cost was achieved by concatenating multiple cDNA molecules

from different transcripts, which achieved a balance between multiple circular sequencing passes, cost and accuracy. As a result, this method was successfully applied to generate up to 60 million full-length reads per sample [**?** **?**].

Other scLR RNA-seq competitors have also developed similar methods. For example, Oxford Nanopore Technologies (ONT) developed a single-cell RNA-seq library preparation method that enables full-length sequencing of both reverse-transcribed cDNA molecules as well as direct sequencing of RNA molecules. Direct RNA sequencing (DRS) is a particularly exciting development as it enables the detection of native epigenetic markers such as methylation. However, DRS methods still suffer from inaccurate basecalling, and high input RNA requirements [**?**]. Overall, recent developments in scLR RNA-seq methods should be leveraged in future studies to improve our understanding of alternative splicing regulation.

5.1.2 Lack of understanding of functional impact

Splicing QTL studies aim to identify genetic variants associated with variation in isoform usage or relative usage of splice junctions. Although useful to implicate alternative splicing in complex disease risk, sQTL studies do not provide a deep understanding of the functional consequences of implicated alternative splicing events. The biggest hurdle is understanding the function of each gene's isoforms.

Gene knockout and knockdown studies have improved our understanding of gene functions. This understanding has been aided by the development of technologies such as RNA interference (RNAi) and more recently CRISPR-Cas9 KO and KD methods. Following gene KO, any number of desired cellular assays can be applied to understand the functional effect of knocking out the gene under investigation (e.g. cell survival, proliferation, migration, defence against pathogens, or any particular function of interest). Optimisation of the targeting capabilities of these methods has led to a targeting efficiency of up to 80% [**?** **?**]. However, most of the available methods target overall mRNA levels, and do not differentiate between different gene isoforms.

More recently, RNAi- and CRISPR-based methods have been modified to enable them to target either individual exons or individual exon-exon junctions. For example, Schertzer et al. [**?**] developed a CRISPR-Cas13d system coupled with a guide RNA that targets different types of alternative splicing events including exon skipping, and alternative acceptor and donor splice sites. gRNAs that target all splicing events within a specific isoform under investigation can be constructed with the aim of targeting a specific isoform. This strategy

can be coupled with cellular functional assays to understand the impact of targeting different types of splicing events as well as specific isoforms. As a newly developed method, it remains to be seen how feasible it is to use this method in large scale isoform essentiality scans. Similar to other RNA targeting system, CRISPR-Cas13d still suffers from sequence-dependent targeting efficiency. Moreover, Schertzer et al. only tested the efficiency of their method on HEK293 cell lines rather than primary cells, another limitation of RNA targeting methods. Still, this method represents a significant advance compared to previous attempts to target specific gene isoforms, which targeted specific types of splicing events (most commonly exon cassettes), and/or were not scalable [**?** **?**].

5.1.3 Regulators of alternative splicing

Defining the role of alternative splicing regulators in defining cell identity and their role in different biological contexts is crucial to understanding the context-specificity of alternative splicing. RNA-binding proteins (RBP) are major alternative splicing regulators that bind to cis-acting splicing regulatory elements on pre-mRNA to either suppress or activate splicing. In addition to their role in regulating alternative splicing, RBPs have been shown to regulate gene expression, and transcript polyadenylation, transport, localisation and translation. Additionally, mutations in genes coding for RBPs have been linked to several Mendelian and complex diseases [**?** **?**]. Therefore, methods to characterise RBPs and their binding targets in different physiological contexts have been developed [**?** **?** **?**]. Of particular interest are high-throughput methods that are able to identify transcriptome-wide RNA binding sites for particular RBPs of interest such as eCLIP [**?**]. In the context of macrophages, for example, RBPs such as TTP, HUR, TIAR and hnRNP K have been shown to regulate macrophage response to LPS stimulation [**?**]. Moreover, RBPs that play tissue-specific roles in tissue-resident macrophages, such as RBP-J, have been identified [**?**]. Knocking out RBP-J in colonic macrophages resulted in reduced macrophage-Th17 cross-talk, which in turn disrupted the ability of macrophages to clear bacterial pathogens.

More systematically, Wagner et al. [**?**] identified macrophage splicing regulatory networks for 10 distinct splicing factors. Knocking out each of these factors revealed a vast network of genes whose splicing patterns become profoundly dysregulated when macrophages were challenged with *Salmonella*. What remains missing in this puzzle is a mechanism whereby these splicing factors control this vast number of splicing choices. Wagner et al. have ruled out the possibility of gene downregulation via the differential inclusion of poison exons as a possible regulatory mode exerted by splicing factors. They have speculated that higher order interactions between the 10 splicing factors and other regu-

lators of innate immune response may be behind the extensive regulatory roles of splicing factors. In any case, the systematic identification of the targets of as many splicing factors as possible is key to better understand the splicing regulatory networks that dictate innate immune response [**?**]. A second puzzle is how genetic variation perturbs innate immunity splicing networks. This will likely be a more substantial challenge as most RBPs are known to recognise short degenerate RNA motifs that cannot be easily predicted from sequence features alone [**?**]. In this regard, trans-sQTL mapping can be useful. Trans-sQTL mapping can identify trans-acting genetic variants associated with individual variation in splicing of distant genes. The power of trans-sQTL mapping can be greatly increased by testing only genetic variants that likely impact the expression of a bona fide set of splicing factors. Such an "informed" trans-sQTL mapping analysis assumes that variants that affect the expression of splicing factors in *cis* will affect, in *trans*, the splicing patterns of distant genes. Identifying a set of splicing factor whose expression is affected in cis is a two-fold problem. First, the splicing factors themselves are not easy to identify, but focussing on the ones identified by Wagner et al (ref [**?**]) is a good start. Second, the "correct" variants need to be identified in the correct context. In the context of macrophages, MacroMap significant eQTL variants in splicing factor genes can be reasonable candidates as drivers of trans-sQTL effects.

Although drivers of splicing regulation in macrophages remain understudied, even less is known about the dynamics of splicing regulation by RBPs in macrophages. The established role of splicing in neuro-developmental diseases has prompted some investigations into how some established RBPs subtly control splicing in their target exons and introns. Interestingly, some RBPs control alternative splicing in a dose-sensitive manner. For example, MBNL1 affects the splicing of several MBNL1 target exons in a manner that is proportional to the concentration of MBNL1 [**?**]. In future studies, it will be interesting to see if splicing targets of innate immune response genes respond to different dosages of splicing factor concentrations, and if that plays a role in determining the severity of immune response.

5.1.4 Antisense oligonucleotides

Correcting aberrant alternative splicing has more recently emerged as a promising therapeutic modality. The recent improvements in mRNA therapeutics chemistry and improved delivery [**? ? ?**] meant that dysregulated pathways can be targeted at the mRNA level, rather than at the protein level. Targeting RNA molecules via antisense oligonucletides (ASO) offers more flexibility than targeting small proteins. ASOs are 15-20 base oligonucleotides that bind to their target mRNA via Watson-Crick base pairing, and can therefore be easily adapted to

target any RNA sequence. Upon binding to mRNA molecules, ASOs exert therapeutic effects by causing mRNA degradation, blocking mRNA-ribosomal interactions or by modulating splicing [**?**].

The flexibility of ASOs has led to the development of Nusinersen, a splice-switching ASO that increase the rate of exon 7 inclusion in the pre-mRNA of *SMN2*. Nusinersen has been approved in 2016 to treat spinal muscular atropy, a motor neuron disease that is caused by the exclusion of exon 7 from the two spinal motor neuron proteins SMN1 and SMN2 [**?**]. Other splice-switching ASOs are being actively developed to correct splicing abberations in cancer, cardiomyopathies, Alzheimer's disease and a number of rare neurological disorders (reviewed in [**?**]). These splice-switching ASOs correct splicing by targeting a wide range of splicing motifs including 5' and 3' splice sites, as well as exonic and intronic splicing enhancers/silencers [**?**]. Thus, understanding the genetic regulation of alternative splicing in disease-relevant contexts opens up several opportunities to build on the promise of splice-switching ASOs.

5.2 Sub-phenotype GWAS

In chapters 3 and 4, I turned my focus to another problem in population genetics, namely how genetic variation affects disease outcomes using perianal Crohn's disease as an example (pCD). I started by exploring the genetic determinants of perianal CD in two well-phenotyped IBD cohorts. My main research aim was to understand its genetic architecture, and particularly if it revealed specific biological pathways that may shed light on the distinct pathogenesis of pCD. I was motivated by broader interest in answering the question of whether disease susceptibility and sub-phenotypes are driven by the same or by distinct dysregulated pathways. A natural follow-up question is how the same sub-phenotype may arise in different contexts, and whether the occurrence of the same sub-phenotype within the general population is driven by the same genetic underpinnings that drive it within a disease population. In my view, these two important aspects of disease sub-phenotypes are important to understand.

5.2.1 The burden of sporadic perianal manifestations is likely underappreciated

My choice to focus on sporadic perianal disease (pAD) was driven by its phenotypic similarity to pCD. The most burdensome feature of both is the development of perianal fistulas. Compared to CD-associated perianal manifestations, sporadic perianal manifestations are

less well-studied. This is reflected, for example, in conflicting estimates for sporadic perianal fistula prevalence. Hokannen et al [**?**] reports a 1-2 per 10,000 prevalence in the UK based on the THIN database, which contains primary care data on approximately 6% of the UK population. Moreover, Garcia-Olmo et al. [**?**] reported a prevalence of 1.69 per 10,000 in Europe based on a meta-analysis of six epidemiological studies. Based on these estimates, the UKBB is expected to report approximately 50-100 pAD cases. The real number of UKB perianal fistula cases as indicated by ICD-10 codes K60.3-K60.5 was surprisingly much higher (> 2,200 cases). A similarly high number of anal fissure and fistula patients were found in FinnGen (N=6,600 for the K60-equivalent FinnGen code K11_FISSANAL). This large number of UKB anal fistulas is unexpected as UKB participants are generally known to be more "health-conscious" than the general population. Previous work even cautions against generalising prevalence estimates from UKB as they are generally considered *lower* than general population estimates [**?**].

Data source differences may partially explain this discrepancy. The UK population estimate by Hokannen et al is based on primary healthcare records, which does not include patients who receive a diagnosis in hospital settings. In the European estimate, only a single Finnish study from the late 1980s reported anal fistula prevalence based on hospital records. The UKB and FinnGen clinical data are entirely based on hospital inpatient episodes, where pAD may be more frequently diagnosed. Overall, the burden of sporadic pAD is likely under-estimated.

5.2.2 Perianal manifestations: different mechanisms in different contexts?

The finding that none of the pAD-associated index variants replicate in the pCD GWAS was surprising. Given the phenotypic similarity between the two phenotypes, albeit in different contexts, it is reasonable that a shared genetic background exists. The interpretation of lack of replication should, however, be interpreted with caution. It is worth noting that there is a substantial difference in statistical power between the two meta-analysis (11,216 pAD cases versus 3,967 pCD cases). It is true that the pCD meta-analysis is not well-powered to detect the pAD associations at a genome-wide significant level. But their association with pCD did not even pass a more lenient replication threshold, which makes the "distinct-genetics" hypothesis at least plausible. Additionally, it is unclear how the composition of the constituent cohorts in each GWAS affects the discovery of genome-wide association signals. Over half

the IBD-BR patients with pCD manifestations report fistulas, but the proportion of UKBB pAD cases with fistulas is only 37%. The lack of more granular data for both UKIBDGC and FinnGen makes an overall comparison more difficult. Moreover, the pCD case cohort includes a large number of perianal abscess, a phenotype that was not included in the pAD meta-analysis. Overall, these compositional differences cannot be completely ruled out as factors that may contribute to the apparently distinct genetic underpinnings of pCD and pAD.

With these limitations in mind, disorders that exhibit phenotypic similarity do not always share molecular and/or genetic similarity and vice versa. In line with this, Zhou et al. [**?**] integrated protein-protein interaction data with GWAS data for thousands of ICD codes to re-define clusters of ICD codes that likely share molecular and genetic profiles. They found that clusters of seemingly unrelated illnesses are often more related than expected (e.g. Alzheimer's disease and lipoprotein deficiencies converge on *APOE*), and diseases traditionally classified within the same category exhibit diversity at the molecular level. Thus, it is reasonable that pAD and pCD diverge at the molecular and genetic levels despite substantial overlap at the clinical level.

5.2.3 The shared genetics of pAD and haemorrhoids needs to be explored

The enrichment of pAD cases in a number of lower intestinal conditions is particularly revealing regarding the nature of the disease. Of particular interest is the pAD cases enrichment in harmorrhoids cases, as well as a strong and significant genetic correlation between pAD and haemorrhoids, confirmed in a largely independent dataset. The co-occurence of a number of ano-rectal disorders including haemorrhoids, anal fissures and fistulas, and rectal prolapse is recognised in clinical practice, and many of them typically require only conservative treatment [**? ?**]. In this regard, my genetic correlation analysis suggests that pAD and haemorrhoids likely share underlying genetic predisposition. In future studies, characterising the relationship between haemorrhoids and pAD is warranted. I attempted to understand this relationship by showing that the effect sizes of several pAD-associated variants are significantly smaller than their effects on haemorrhoids. But this is not sufficient to rule out that their effect on pAD risk is not mediated via their effect on haemorrhoids risk. Gaining a more complete understanding of the relationship between haemorrhoids and pAD requires applying methods that systematically discover shared and distinct genetic variants in genetically correlated groups of disorders such as genomic structural equation modelling (gSEM [**?**]). These methods have been previously applied to psychiatric disorders, which are

known to exhibit high degrees of genetic correlation [**?**]. Particularly, it will be interesting to identify which genetic variants confer risk to pAD only and which genetic variants confer risk to both haemorrhoids and pAD.

5.2.4 The genetics of pCD and CD

One of the aims of this thesis is to understand the relationship between the genetic underpinnings of pCD and CD. The most obvious way to understand this relationship is to investigate pCD-associated variants and identify the biological pathways or functions that they likely dysregulate. This can only be achieved if a substantial number of pCD-associated loci and effector genes are identified. My pCD meta-analysis only identified a single pCD-associated locus, which was insufficient to derive any meaningful biological understanding of the relationship between pCD and CD.

The relationship between CD and pCD can still be studied with few or no genome-wide significant variants. It will be interesting to compare CD polygenic risk scores (PRS) between individuals with and without perianal manifestations. Although PRS have little predictive power in general, they can suggest if the heritability of pCD is captured by PRS constructed from CD-associated genetic variants. Indeed, Cleynen et al. [95] showed that small but significant differences in PRS can differentiate between groups of IBD patients with different disease locations. On the other hand, Lee et al [96] did not find differences in PRS between patients with the best and the worst CD outcomes. But a difference in PRS should be ruled out with caution as the power of PRS to differentiate between groups of patients is not only limited by the heritability captured by the genetic variants used to construct PRS, but also by the number of patients being compared. Nonetheless, future analyses should explore if PRS can meaningfully differentiate between the pCD+ and pCD- cohorts as a possible answer to the question of whether CD and pCD share genetic risk profiles.

5.3 Future Outlook

The boundaries of scientific understanding have often been pushed by two factors: better tools and better collaboration. Better technologies have allowed us to appreciate the unseen complexity of our molecular machinery, and have, time and again, put us in a better position to fight nature's misfortunes. In the near future, I believe technological advancements will only provide us with a better appreciation of how pervasive and impactful alternative splicing is. It is fascinating to think about alternative splicing in light of our long evolutionary history:

as a simple, albeit error-laden, eukaryotic invention that has bestowed so much diversity upon us and continues to enable new functions to continuously emerge [**? ? ? ?**]. It is even more exciting to think that we can now exploit recent therapeutic developments to correct the undesired side effects of this rich evolutionary toolkit.

At the turn of this millenium, international collaboration has enabled us to crack the human genome. Collaborative efforts have also shown us the benefit of building national biobanks and deeply phenotyped disease cohorts. In the near future, these collaborative efforts are poised to uncover the links between the human genome and phenome, with all the intricacies of disease sub-phenotypes. In conclusion, these two pillars of scientific progress will enable a better understanding of both the invisible and visible complexities of human health.

References

[1] Ruth E Malone and Kenneth E Warner. Tobacco control at twenty: reflecting on the past, considering the present and developing the new conversations for the future. *Tob. Control*, 21(2):74–76, March 2012.

[2] Thomas A Glass, Steven N Goodman, Miguel A Hernán, and Jonathan M Samet. Causal inference in public health. *Annu. Rev. Public Health*, 34(1):61–75, January 2013.

[3] George Davey Smith, Debbie A Lawlor, Roger Harbord, Nic Timpson, Ian Day, and Shah Ebrahim. Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology. *PLoS Med.*, 4(12):e352, December 2007.

[4] Tiina Ahonen, Juha Saltevo, Markku Laakso, Hannu Kautiainen, Esko Kumpusalo, and Mauno Vanhala. Gender differences relating to metabolic syndrome and proinflammation in finnish subjects with elevated blood pressure. *Mediators Inflamm.*, 2009:959281, August 2009.

[5] Jacob F Degner, Athma A Pai, Roger Pique-Regi, Jean-Baptiste Veyrieras, Daniel J Gaffney, Joseph K Pickrell, Sherryl De Leon, Katelyn Michelini, Noah Lewellen, Gregory E Crawford, Matthew Stephens, Yoav Gilad, and Jonathan K Pritchard. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*, 482(7385):390–394, February 2012.

[6] Gosia Trynka, Cynthia Sandor, Buhm Han, Han Xu, Barbara E Stranger, X Shirley Liu, and Soumya Raychaudhuri. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.*, 45(2):124–130, February 2013.

[7] Lucia A Hindorff, Praveen Sethupathy, Heather A Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri A Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.*, 106(23):9362–9367, June 2009.

[8] Meritxell Oliva, Kathryn Demanelis, Yihao Lu, Meytal Chernoff, Farzana Jasmine, Habibul Ahsan, Muhammad G Kibriya, Lin S Chen, and Brandon L Pierce. DNA methylation QTL mapping across diverse human tissues provides molecular links between genetic variation and complex traits. *Nat. Genet.*, 55(1):112–122, January 2023.

[9] Eilis Hannon, Helen Spiers, Joana Viana, Ruth Pidsley, Joe Burrage, Therese M Murphy, Claire Troakes, Gustavo Turecki, Michael C O'Donovan, Leonard C Schalkwyk, Nicholas J Bray, and Jonathan Mill. Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. *Nat. Neurosci.*, 19(1):48–54, January 2016.

[10] Jarrett D Morrow, Kimberly Glass, Michael H Cho, Craig P Hersh, Victor Pinto-Plata, Bartolome Celli, Nathaniel Marchetti, Gerard Criner, Raphael Bueno, George Washko, Augustine M K Choi, John Quackenbush, Edwin K Silverman, and Dawn L DeMeo. Human lung DNA methylation quantitative trait loci colocalize with chronic obstructive pulmonary disease genome-wide association loci. *Am. J. Respir. Crit. Care Med.*, 197(10):1275–1284, May 2018.

[11] D Leland Taylor, Anne U Jackson, Narisu Narisu, Gibran Hemani, Michael R Erdos, Peter S Chines, Amy Swift, Jackie Idol, John P Didion, Ryan P Welch, Leena Kinnunen, Jouko Saramies, Timo A Lakka, Markku Laakso, Jaakko Tuomilehto, Stephen C J Parker, Heikki A Koistinen, George Davey Smith, Michael Boehnke, Laura J Scott, Ewan Birney, and Francis S Collins. Integrative analysis of gene expression, DNA methylation, physiological traits, and genetic variation in human skeletal muscle. *Proc. Natl. Acad. Sci. U. S. A.*, 116(22):10883–10888, May 2019.

[12] Tianxiao Huan, Roby Joehanes, Ci Song, Fen Peng, Yichen Guo, Michael Mendelson, Chen Yao, Chunyu Liu, Jiantao Ma, Melissa Richard, Golareh Agha, Weihua Guan, Lynn M Almli, Karen N Conneely, Joshua Keefe, Shih-Jen Hwang, Andrew D Johnson, Myriam Fornage, Liming Liang, and Daniel Levy. Genome-wide identification of DNA methylation QTLs in whole blood highlights pathways for cardiovascular disease. *Nat. Commun.*, 10(1):4267, September 2019.

[13] Shan V Andrews, Shannon E Ellis, Kelly M Bakulski, Brooke Sheppard, Lisa A Croen, Irva Hertz-Picciotto, Craig J Newschaffer, Andrew P Feinberg, Dan E Arking, Christine Ladd-Acosta, and M Daniele Fallin. Cross-tissue integration of genetic and epigenetic data offers insight into autism spectrum disorder. *Nat. Commun.*, 8(1), October 2017.

[14] Kaur Alasoo, HIPSCI Consortium, Julia Rodrigues, Subhankar Mukhopadhyay, Andrew J Knights, Alice L Mann, Kousik Kundu, Christine Hale, Gordon Dougan, and Daniel J Gaffney. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nat. Genet.*, 50(3):424–431, March 2018.

[15] Kevin W Currin, Michael R Erdos, Narisu Narisu, Vivek Rai, Swarooparani Vadlamudi, Hannah J Perrin, Jacqueline R Idol, Tingfen Yan, Ricardo D'oliveira Albanus, K Alaine Broadaway, Amy S Etheridge, Lori L Bonnycastle, Peter Orchard, John P Didion, Amarjit S Chaudhry, NISC Comparative Sequencing Program, Federico Innocenti, Erin G Schuetz, Laura J Scott, Stephen C J Parker, Francis S Collins, and Karen L Mohlke. Genetic effects on liver chromatin accessibility identify disease regulatory variants. *Am. J. Hum. Genet.*, 108(7):1169–1189, July 2021.

[16] The GTEx Consortium, François Aguet, Shankara Anand, Kristin G Ardlie, Stacey Gabriel, Gad A Getz, Aaron Graubert, Kane Hadley, Robert E Handsaker, Katherine H

Huang, Seva Kashin, Xiao Li, Daniel G MacArthur, Samuel R Meier, Jared L Nedzel, Duyen T Nguyen, Ayellet V Segrè, Ellen Todres, Brunilda Balliu, Alvaro N Barbeira, Alexis Battle, Rodrigo Bonazzola, Andrew Brown, Christopher D Brown, Stephane E Castel, Donald F Conrad, Daniel J Cotter, Nancy Cox, Sayantan Das, Olivia M de Goede, Emmanouil T Dermitzakis, Jonah Einson, Barbara E Engelhardt, Eleazar Eskin, Tiffany Y Eulalio, Nicole M Ferraro, Elise D Flynn, Laure Fresard, Eric R Gamazon, Diego Garrido-Martín, Nicole R Gay, Michael J Gloudemans, Roderic Guigó, Andrew R Hame, Yuan He, Paul J Hoffman, Farhad Hormozdiari, Lei Hou, Hae Kyung Im, Brian Jo, Silva Kasela, Manolis Kellis, Sarah Kim-Hellmuth, Alan Kwong, Tuuli Lappalainen, Xin Li, Yanyu Liang, Serghei Mangul, Pejman Mohammadi, Stephen B Montgomery, Manuel Muñoz-Aguirre, Daniel C Nachun, Andrew B Nobel, Meritxell Oliva, Yoson Park, Yongjin Park, Princy Parsana, Abhiram S Rao, Ferran Reverter, John M Rouhana, Chiara Sabatti, Ashis Saha, Matthew Stephens, Barbara E Stranger, Benjamin J Strober, Nicole A Teran, Ana Viñuela, Gao Wang, Xiaoquan Wen, Fred Wright, Valentin Wucher, Yuxin Zou, Pedro G Ferreira, Gen Li, Marta Melé, Esti Yeger-Lotem, Mary E Barcus, Debra Bradbury, Tanya Krubit, Jeffrey A McLean, Liqun Qi, Karna Robinson, Nancy V Roche, Anna M Smith, Leslie Sobin, David E Tabor, Anita Undale, Jason Bridge, Lori E Brigham, Barbara A Foster, Bryan M Gillard, Richard Hasz, Marcus Hunter, Christopher Johns, Mark Johnson, Ellen Karasik, Gene Kopen, William F Leinweber, Alisa McDonald, Michael T Moser, Kevin Myer, Kimberley D Ramsey, Brian Roe, Saboor Shad, Jeffrey A Thomas, Gary Walters, Michael Washington, Joseph Wheeler, Scott D Jewell, Daniel C Rohrer, Dana R Valley, David A Davis, Deborah C Mash, Philip A Branton, Laura K Barker, Heather M Gardiner, Maghboeba Mosavel, Laura A Siminoff, Paul Flicek, Maximilian Haeussler, Thomas Juettemann, W James Kent, Christopher M Lee, Conner C Powell, Kate R Rosenbloom, Magali Ruffier, Dan Sheppard, Kieron Taylor, Stephen J Trevanion, Daniel R Zerbino, Nathan S Abell, Joshua Akey, Lin Chen, Kathryn Demanelis, Jennifer A Doherty, Andrew P Feinberg, Kasper D Hansen, Peter F Hickey, Farzana Jasmine, Lihua Jiang, Rajinder Kaul, Muhammad G Kibriya, Jin Billy Li, Qin Li, Shin Lin, Sandra E Linder, Brandon L Pierce, Lindsay F Rizzardi, Andrew D Skol, Kevin S Smith, Michael Snyder, John Stamatoyannopoulos, Hua Tang, Meng Wang, Latarsha J Carithers, Ping Guan, Susan E Koester, A Roger Little, Helen M Moore, Concepcion R Nierras, Abhi K Rao, Jimmie B Vaught, and Simona Volpi. The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, September 2020.

[17] Urmo Võsa, Annique Claringbould, Harm-Jan Westra, Marc Jan Bonder, Patrick Deelen, Biao Zeng, Holger Kirsten, Ashis Saha, Roman Kreuzhuber, Seyhan Yazar, Harm Brugge, Roy Oelen, Dylan H de Vries, Monique G P van der Wijst, Silva Kasela, Natalia Pervjakova, Isabel Alves, Marie-Julie Favé, Mawussé Agbessi, Mark W Christiansen, Rick Jansen, Ilkka Seppälä, Lin Tong, Alexander Teumer, Katharina Schramm, Gibran Hemani, Joost Verlouw, Hanieh Yaghootkar, Reyhan Sönmez Flitman, Andrew Brown, Viktorija Kukushkina, Anette Kalnapenkis, Sina Rüeger, Eleonora Porcu, Jaanika Kronberg, Johannes Kettunen, Bernett Lee, Futao Zhang, Ting Qi, Jose Alquicira Hernandez, Wibowo Arindrarto, Frank Beutner, BIOS Consortium, i2QTL Consortium, Julia Dmitrieva, Mahmoud Elansary, Benjamin P Fairfax, Michel Georges, Bastiaan T Heijmans, Alex W Hewitt, Mika Kähönen, Yungil Kim, Julian C Knight, Peter Kovacs, Knut Krohn, Shuang Li, Markus Loeffler, Urko M Marigorta, Hailang Mei, Yukihide Momozawa, Martina Müller-Nurasyid, Matthias Nauck,

Michel G Nivard, Brenda W J H Penninx, Jonathan K Pritchard, Olli T Raitakari, Olaf Rotzschke, Eline P Slagboom, Coen D A Stehouwer, Michael Stumvoll, Patrick Sullivan, Peter A C 't Hoen, Joachim Thiery, Anke Tönjes, Jenny van Dongen, Maarten van Iterson, Jan H Veldink, Uwe Völker, Robert Warmerdam, Cisca Wijmenga, Morris Swertz, Anand Andiappan, Grant W Montgomery, Samuli Ripatti, Markus Perola, Zoltan Kutalik, Emmanouil Dermitzakis, Sven Bergmann, Timothy Frayling, Joyce van Meurs, Holger Prokisch, Habibul Ahsan, Brandon L Pierce, Terho Lehtimäki, Dorret I Boomsma, Bruce M Psaty, Sina A Gharib, Philip Awadalla, Lili Milani, Willem H Ouwehand, Kate Downes, Oliver Stegle, Alexis Battle, Peter M Visscher, Jian Yang, Markus Scholz, Joseph Powell, Greg Gibson, Tõnu Esko, and Lude Franke. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.*, 53(9):1300–1310, September 2021.

[18] Nurlan Kerimov, James D Hayhurst, Kateryna Peikova, Jonathan R Manning, Peter Walter, Liis Kolberg, Marija Samoviča, Manoj Pandian Sakthivel, Ivan Kuzmin, Stephen J Trevanion, Tony Burdett, Simon Jupp, Helen Parkinson, Irene Papatheodorou, Andrew D Yates, Daniel R Zerbino, and Kaur Alasoo. A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat. Genet.*, 53(9):1290–1299, September 2021.

[19] Ting Qi, Yang Wu, Hailing Fang, Futao Zhang, Shouye Liu, Jian Zeng, and Jian Yang. Genetic control of RNA splicing and its distinct role in complex trait variation. *Nat. Genet.*, 54(9):1355–1363, September 2022.

[20] Chen Yao, George Chen, Ci Song, Joshua Keefe, Michael Mendelson, Tianxiao Huan, Benjamin B Sun, Annika Laser, Joseph C Maranville, Hongsheng Wu, Jennifer E Ho, Paul Courchesne, Asya Lyass, Martin G Larson, Christian Gieger, Johannes Graumann, Andrew D Johnson, John Danesh, Heiko Runz, Shih-Jen Hwang, Chunyu Liu, Adam S Butterworth, Karsten Suhre, and Daniel Levy. Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nat. Commun.*, 9(1):3268, August 2018.

[21] Benjamin B Sun, Joseph C Maranville, James E Peters, David Stacey, James R Staley, James Blackshaw, Stephen Burgess, Tao Jiang, Ellie Paige, Praveen Surendran, Clare Oliver-Williams, Mihir A Kamat, Bram P Prins, Sheri K Wilcox, Erik S Zimmerman, An Chi, Narinder Bansal, Sarah L Spain, Angela M Wood, Nicholas W Morrell, John R Bradley, Nebojsa Janjic, David J Roberts, Willem H Ouwehand, John A Todd, Nicole Soranzo, Karsten Suhre, Dirk S Paul, Caroline S Fox, Robert M Plenge, John Danesh, Heiko Runz, and Adam S Butterworth. Genomic atlas of the human plasma proteome. *Nature*, 558(7708):73–79, June 2018.

[22] Daria V Zhernakova, Patrick Deelen, Martijn Vermaat, Maarten van Iterson, Michiel van Galen, Wibowo Arindrarto, Peter van 't Hof, Hailiang Mei, Freerk van Dijk, Harm-Jan Westra, Marc Jan Bonder, Jeroen van Rooij, Marijn Verkerk, P Mila Jhamai, Matthijs Moed, Szymon M Kielbasa, Jan Bot, Irene Nooren, René Pool, Jenny van Dongen, Jouke J Hottenga, Coen D A Stehouwer, Carla J H van der Kallen, Casper G Schalkwijk, Alexandra Zhernakova, Yang Li, Ettje F Tigchelaar, Niek de Klein, Marian Beekman, Joris Deelen, Diana van Heemst, Leonard H van den Berg, Albert Hofman, André G Uitterlinden, Marleen M J van Greevenbroek, Jan H Veldink, Dorret I

Boomsma, Cornelia M van Duijn, Cisca Wijmenga, P Eline Slagboom, Morris A Swertz, Aaron Isaacs, Joyce B J van Meurs, Rick Jansen, Bastiaan T Heijmans, Peter A C 't Hoen, and Lude Franke. Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.*, 49(1):139–145, January 2017.

[23] Zepeng Mu, Wei Wei, Benjamin Fair, Jinlin Miao, Ping Zhu, and Yang I Li. The impact of cell type and context-dependent regulatory variants on human immune traits. *Genome Biol.*, 22(1):122, April 2021.

[24] Halit Ongen, GTEx Consortium, Andrew A Brown, Olivier Delaneau, Nikolaos I Panousis, Alexandra C Nica, and Emmanouil T Dermitzakis. Estimating the causal tissues for complex traits and diseases. *Nat. Genet.*, 49(12):1676–1683, December 2017.

[25] Ana Viñuela, Arushi Varshney, Martijn van de Bunt, Rashmi B Prasad, Olof Asplund, Amanda Bennett, Michael Boehnke, Andrew A Brown, Michael R Erdos, João Fadista, Ola Hansson, Gad Hatem, Cédric Howald, Apoorva K Iyengar, Paul Johnson, Ulrika Krus, Patrick E MacDonald, Anubha Mahajan, Jocelyn E Manning Fox, Narisu Narisu, Vibe Nylander, Peter Orchard, Nikolay Oskolkov, Nikolaos I Panousis, Anthony Payne, Michael L Stitzel, Swarooparani Vadlamudi, Ryan Welch, Francis S Collins, Karen L Mohlke, Anna L Gloyn, Laura J Scott, Emmanouil T Dermitzakis, Leif Groop, Stephen C J Parker, and Mark I McCarthy. Genetic variant effects on gene expression in human pancreatic islets and their implications for T2D. *Nat. Commun.*, 11(1):4912, September 2020.

[26] Edward Mountjoy, Ellen M Schmidt, Miguel Carmona, Jeremy Schwartzentruber, Gareth Peat, Alfredo Miranda, Luca Fumis, James Hayhurst, Annalisa Buniello, Mohd Anisul Karim, Daniel Wright, Andrew Hercules, Eliseo Papa, Eric B Fauman, Jeffrey C Barrett, John A Todd, David Ochoa, Ian Dunham, and Maya Ghoussaini. An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nat. Genet.*, 53(11):1527–1533, November 2021.

[27] P A Sharp. Split genes and RNA splicing. *Cell*, 77(6):805–815, June 1994.

[28] A A Mironov, J W Fickett, and M S Gelfand. Frequent alternative splicing of human genes. *Genome Res.*, 9(12):1288–1293, December 1999.

[29] Qun Pan, Ofer Shai, Leo J Lee, Brendan J Frey, and Benjamin J Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, 40(12):1413–1415, December 2008.

[30] A Krämer. The structure and function of proteins involved in mammalian pre-mRNA splicing. *Annu. Rev. Biochem.*, 65(1):367–409, 1996.

[31] Eugene V Koonin. The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? *Biol. Direct*, 1(1):22, August 2006.

[32] David G Knowles and Aoife McLysaght. High rate of recent intron gain and loss in simultaneously duplicated arabidopsis genes. *Mol. Biol. Evol.*, 23(8):1548–1557, August 2006.

[33] Landen Gozashti, Scott W Roy, Bryan Thornlow, Alexander Kramer, Manuel Ares, Jr, and Russell Corbett-Detig. Transposable elements drive intron gain in diverse eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.*, 119(48):e2209766119, November 2022.

[34] J Messing. Do plants have more genes than humans? *Trends Plant Sci.*, 6(5):195–196, May 2001.

[35] Stephen J Bush, Lu Chen, Jaime M Tovar-Corona, and Araxi O Urrutia. Alternative splicing and the evolution of phenotypic novelty. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 372(1713):20150474, February 2017.

[36] Luciano E Marasco and Alberto R Kornblihtt. The physiology of alternative splicing. *Nat. Rev. Mol. Cell Biol.*, 24(4):242–254, April 2023.

[37] Daisuke Hattori, S Sean Millard, Woj M Wojtowicz, and S Lawrence Zipursky. Dscam-mediated cell recognition regulates neural circuit formation. *Annu. Rev. Cell Dev. Biol.*, 24(1):597–620, 2008.

[38] Woj M Wojtowicz, John J Flanagan, S Sean Millard, S Lawrence Zipursky, and James C Clemens. Alternative splicing of drosophila dscam generates axon guidance receptors that exhibit isoform-specific homophilic binding. *Cell*, 118(5):619–633, September 2004.

[39] Luiz O F Penalva and Lucas Sánchez. RNA binding protein sex-lethal (sxl) and control of drosophila sex determination and dosage compensation. *Microbiol. Mol. Biol. Rev.*, 67(3):343–59, table of contents, September 2003.

[40] Kishore Jaganathan, Sofia Kyriazopoulou Panagiotopoulou, Jeremy F McRae, Siavash Fazel Darbandi, David Knowles, Yang I Li, Jack A Kosmicki, Juan Arbelaez, Wenwu Cui, Grace B Schwartz, Eric D Chow, Efstathios Kanterakis, Hong Gao, Amirali Kia, Serafim Batzoglou, Stephan J Sanders, and Kyle Kai-How Farh. Predicting splicing from primary sequence with deep learning. *Cell*, 176(3):535–548.e24, January 2019.

[41] Peter J Shepard and Klemens J Hertel. The SR protein family. *Genome Biol.*, 10(10):242, October 2009.

[42] Thomas Geuens, Delphine Bouhy, and Vincent Timmerman. The hnRNP family: insights into their role in health and disease. *Hum. Genet.*, 135(8):851–867, August 2016.

[43] U R Monani, C L Lorson, D W Parsons, T W Prior, E J Androphy, A H Burghes, and J D McPherson. A single nucleotide difference that alters splicing patterns distinguishes the SMA gene SMN1 from the copy gene SMN2. *Hum. Mol. Genet.*, 8(7):1177–1183, July 1999.

[44] Nuala A O'Leary, Mathew W Wright, J Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, Alexander Astashyn, Azat Badretdin, Yiming Bao, Olga Blinkova, Vyacheslav Brover, Vyacheslav Chetvernin, Jinna Choi, Eric Cox, Olga Ermolaeva, Catherine M Farrell, Tamara Goldfarb, Tripti Gupta, Daniel Haft, Eneida Hatcher, Wratko Hlavina,

Vinita S Joardar, Vamsi K Kodali, Wenjun Li, Donna Maglott, Patrick Masterson, Kelly M McGarvey, Michael R Murphy, Kathleen O'Neill, Shashikant Pujar, Sanjida H Rangwala, Daniel Rausch, Lillian D Riddick, Conrad Schoch, Andrei Shkeda, Susan S Storz, Hanzhen Sun, Francoise Thibaud-Nissen, Igor Tolstoy, Raymond E Tully, Anjana R Vatsan, Craig Wallin, David Webb, Wendy Wu, Melissa J Landrum, Avi Kimchi, Tatiana Tatusova, Michael DiCuccio, Paul Kitts, Terence D Murphy, and Kim D Pruitt. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, 44(D1):D733–45, January 2016.

[45] RefSeq. Refseq ftp site, 2023. Accessed on 02/11/2023.

[46] Eric T Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F Kingsmore, Gary P Schroth, and Christopher B Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, November 2008.

[47] Iakes Ezkurdia, Jose Manuel Rodriguez, Enrique Carrillo-de Santa Pau, Jesús Vázquez, Alfonso Valencia, and Michael L Tress. Most highly expressed protein-coding genes have a single dominant isoform. *J. Proteome Res.*, 14(4):1880–1887, April 2015.

[48] Mar Gonzàlez-Porta, Adam Frankish, Johan Rung, Jennifer Harrow, and Alvis Brazma. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biol.*, 14(7):R70, July 2013.

[49] HCA. Human cell atlas resources, 2023. Accessed on 02/11/2023.

[50] Vincent Lacroix, Michael Sammeth, Roderic Guigo, and Anne Bergeron. Exact transcriptome reconstruction from short sequence reads. In *Lecture Notes in Computer Science*, Lecture notes in computer science, pages 50–63. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[51] Michael Hagemann-Jensen, Christoph Ziegenhain, Ping Chen, Daniel Ramsköld, Gert-Jan Hendriks, Anton J M Larsson, Omid R Faridani, and Rickard Sandberg. Single-cell RNA counting at allele and isoform resolution using smart-seq3. *Nat. Biotechnol.*, 38(6):708–714, June 2020.

[52] Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and cufflinks. *Nat. Protoc.*, 7(3):562–578, March 2012.

[53] Shihao Shen, Juw Won Park, Zhi-Xiang Lu, Lan Lin, Michael D Henry, Ying Nian Wu, Qing Zhou, and Yi Xing. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci. U. S. A.*, 111(51):E5593–601, December 2014.

[54] Yarden Katz, Eric T Wang, Edoardo M Airoldi, and Christopher B Burge. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods*, 7(12):1009–1015, December 2010.

[55] Jorge Vaquero-Garcia, Alejandro Barrera, Matthew R Gazzara, Juan Gonzalez-Vallinas, Nicholas F Lahens, John B Hogenesch, Kristen W Lynch, and Yoseph Barash. A new view of transcriptome complexity and regulation through the lens of local splicing variations. *Elife*, 5, February 2016.

[56] Towfique Raj, Katie Rothamel, Sara Mostafavi, Chun Ye, Mark N Lee, Joseph M Replogle, Ting Feng, Michelle Lee, Natasha Asinovski, Irene Frohlich, Selina Imboywa, Alina Von Korff, Yukinori Okada, Nikolaos A Patsopoulos, Scott Davis, Cristin McCabe, Hyun-Il Paik, Gyan P Srivastava, Soumya Raychaudhuri, David A Hafler, Daphne Koller, Aviv Regev, Nir Hacohen, Diane Mathis, Christophe Benoist, Barbara E Stranger, and Philip L De Jager. Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science*, 344(6183):519–523, May 2014.

[57] James E Peters, Paul A Lyons, James C Lee, Arianne C Richard, Mary D Fortune, Paul J Newcombe, Sylvia Richardson, and Kenneth G C Smith. Insight into genotype-phenotype associations through eQTL mapping in multiple cell types in health and immune-mediated disease. *PLoS Genet.*, 12(3):e1005908, March 2016.

[58] Benjamin P Fairfax, Seiko Makino, Jayachandran Radhakrishnan, Katharine Plant, Stephen Leslie, Alexander Dilthey, Peter Ellis, Cordelia Langford, Fredrik O Vannberg, and Julian C Knight. Genetics of gene expression in primary immune cells identifies cell type–specific master regulators and roles of HLA alleles. *Nat. Genet.*, 44(5):502–510, May 2012.

[59] Mark N Lee, Chun Ye, Alexandra-Chloé Villani, Towfique Raj, Weibo Li, Thomas M Eisenhaure, Selina H Imboywa, Portia I Chipendo, F Ann Ran, Kamil Slowikowski, Lucas D Ward, Khadir Raddassi, Cristin McCabe, Michelle H Lee, Irene Y Frohlich, David A Hafler, Manolis Kellis, Soumya Raychaudhuri, Feng Zhang, Barbara E Stranger, Christophe O Benoist, Philip L De Jager, Aviv Regev, and Nir Hacohen. Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science*, 343(6175):1246980, March 2014.

[60] Benjamin D Umans, Alexis Battle, and Yoav Gilad. Where are the disease-associated eQTLs? *Trends Genet.*, 37(2):109–124, February 2021.

[61] Bryce van de Geijn, Hilary Finucane, Steven Gazal, Farhad Hormozdiari, Tiffany Amariuta, Xuanyao Liu, Alexander Gusev, Po-Ru Loh, Yakir Reshef, Gleb Kichaev, Soumya Raychauduri, and Alkes L Price. Annotations capturing cell type-specific TF binding explain a large fraction of disease heritability. *Hum. Mol. Genet.*, 29(7):1057–1067, May 2020.

[62] Eric R Gamazon, GTEx Consortium, Ayellet V Segrè, Martijn van de Bunt, Xiaoquan Wen, Hualin S Xi, Farhad Hormozdiari, Halit Ongen, Anuar Konkashbaev, Eske M Derks, François Aguet, Jie Quan, Dan L Nicolae, Eleazar Eskin, Manolis Kellis, Gad Getz, Mark I McCarthy, Emmanouil T Dermitzakis, Nancy J Cox, and Kristin G Ardlie. Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nat. Genet.*, 50(7):956–967, July 2018.

[63] Timothée Flutre, Xiaoquan Wen, Jonathan Pritchard, and Matthew Stephens. A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genet.*, 9(5):e1003486, May 2013.

[64] Gen Li, Andrey A Shabalin, Ivan Rusyn, Fred A Wright, and Andrew B Nobel. An empirical bayes approach for multiple tissue eQTL analysis. *Biostatistics*, 19(3):391–406, July 2018.

[65] Jae Hoon Sul, Buhm Han, Chun Ye, Ted Choi, and Eleazar Eskin. Effectively identifying eQTLs from multiple tissues by combining mixed model and meta-analytic approaches. *PLoS Genet.*, 9(6):e1003491, June 2013.

[66] Sarah M Urbut, Gao Wang, Peter Carbonetto, and Matthew Stephens. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat. Genet.*, 51(1):187–195, January 2019.

[67] Boxiang Liu, Michael J Gloudemans, Abhiram S Rao, Erik Ingelsson, and Stephen B Montgomery. Abundant associations with gene expression complicate GWAS follow-up. *Nat. Genet.*, 51(5):768–769, May 2019.

[68] Nikolaos I Panousis, Omar El Garwany, Andrew Knights, Jesse Cheruiyot Rop, Natsuhiko Kumasaka, Maria Imaz, Lorena Boquete Vilarino, Anthi Tsingene, Alice Barnett, Celine Gomez, Carl A Anderson, and Daniel J Gaffney. Gene expression QTL mapping in stimulated iPSC-derived macrophages provides insights into common complex diseases. May 2023.

[69] Niek de Klein, Ellen A Tsai, Martijn Vochteloo, Denis Baird, Yunfeng Huang, Chia-Yen Chen, Sipko van Dam, Roy Oelen, Patrick Deelen, Olivier B Bakker, Omar El Garwany, Zhengyu Ouyang, Eric E Marshall, Maria I Zavodszky, Wouter van Rheenen, Mark K Bakker, Jan Veldink, Tom R Gaunt, Heiko Runz, Lude Franke, and Harm-Jan Westra. Brain expression quantitative trait locus and network analyses reveal downstream effects and putative drivers for brain-related diseases. *Nat. Genet.*, 55(3):377–388, March 2023.

[70] Chris Wallace. A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLoS Genet.*, 17(9):e1009440, September 2021.

[71] Xueyi Dong, Mei R M Du, Quentin Gouil, Luyi Tian, Jafar S Jabbari, Rory Bowden, Pedro L Baldoni, Yunshun Chen, Gordon K Smyth, Shanika L Amarasinghe, Charity W Law, and Matthew E Ritchie. Benchmarking long-read RNA-sequencing analysis tools using in silico mixtures. *Nat. Methods*, 20(11):1810–1821, November 2023.

[72] Diego Garrido-Martín, Beatrice Borsari, Miquel Calvo, Ferran Reverter, and Roderic Guigó. Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. *Nat. Commun.*, 12(1):727, February 2021.

[73] Angli Xue, Yang Wu, Zhihong Zhu, Futao Zhang, Kathryn E Kemper, Zhili Zheng, Loic Yengo, Luke R Lloyd-Jones, Julia Sidorenko, Yeda Wu, Mawussé Agbessi, Habibul Ahsan, Isabel Alves, Anand Andiappan, Philip Awadalla, Alexis Battle, Frank Beutner, Marc Jan Bonder, Dorret Boomsma, Mark Christiansen, Annique

Claringbould, Patrick Deelen, Tõnu Esko, Marie-Julie Favé, Lude Franke, Timothy Frayling, Sina Gharib, Gregory Gibson, Gibran Hemani, Rick Jansen, Mika Kähönen, Anette Kalnapenkis, Silva Kasela, Johannes Kettunen, Yungil Kim, Holger Kirsten, Peter Kovacs, Knut Krohn, Jaanika Kronberg-Guzman, Viktorija Kukushkina, Zoltan Kutalik, Bernett Lee, Terho Lehtimäki, Markus Loeffler, Urko M Marigorta, Andres Metspalu, Lili Milani, Martina Müller-Nurasyid, Matthias Nauck, Michel Nivard, Brenda Penninx, Markus Perola, Natalia Pervjakova, Brandon Pierce, Joseph Powell, Holger Prokisch, Bruce Psaty, Olli Raitakari, Susan Ring, Samuli Ripatti, Olaf Rotzschke, Sina Ruëger, Ashis Saha, Markus Scholz, Katharina Schramm, Ilkka Seppälä, Michael Stumvoll, Patrick Sullivan, Alexander Teumer, Joachim Thiery, Lin Tong, Anke Tönjes, Jenny van Dongen, Joyce van Meurs, Joost Verlouw, Uwe Völker, Urmo Võsa, Hanieh Yaghootkar, Biao Zeng, Allan F McRae, Peter M Visscher, Jian Zeng, Jian Yang, and eQTLGen Consortium. Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nat. Commun.*, 9(1), July 2018.

[74] Krishna G Aragam, Tao Jiang, Anuj Goel, Stavroula Kanoni, Brooke N Wolford, Deepak S Atri, Elle M Weeks, Minxian Wang, George Hindy, Wei Zhou, Christopher Grace, Carolina Roselli, Nicholas A Marston, Frederick K Kamanu, Ida Surakka, Loreto Muñoz Venegas, Paul Sherliker, Satoshi Koyama, Kazuyoshi Ishigaki, Bjørn O Åsvold, Michael R Brown, Ben Brumpton, Paul S de Vries, Olga Giannakopoulou, Panagiota Giardoglou, Daniel F Gudbjartsson, Ulrich Güldener, Syed M Ijlal Haider, Anna Helgadottir, Maysson Ibrahim, Adnan Kastrati, Thorsten Kessler, Theodosios Kyriakou, Tomasz Konopka, Ling Li, Lijiang Ma, Thomas Meitinger, Sören Mucha, Matthias Munz, Federico Murgia, Jonas B Nielsen, Markus M Nöthen, Shichao Pang, Tobias Reinberger, Gavin Schnitzler, Damian Smedley, Gudmar Thorleifsson, Moritz von Scheidt, Jacob C Ulirsch, Biobank Japan, EPIC-CVD, David O Arnar, Noël P Burtt, Maria C Costanzo, Jason Flannick, Kaoru Ito, Dong-Keun Jang, Yoichiro Kamatani, Amit V Khera, Issei Komuro, Iftikhar J Kullo, Luca A Lotta, Christopher P Nelson, Robert Roberts, Gudmundur Thorgeirsson, Unnur Thorsteinsdottir, Thomas R Webb, Aris Baras, Johan L M Björkegren, Eric Boerwinkle, George Dedoussis, Hilma Holm, Kristian Hveem, Olle Melander, Alanna C Morrison, Marju Orho-Melander, Loukianos S Rallidis, Arno Ruusalepp, Marc S Sabatine, Kari Stefansson, Pierre Zalloua, Patrick T Ellinor, Martin Farrall, John Danesh, Christian T Ruff, Hilary K Finucane, Jemma C Hopewell, Robert Clarke, Rajat M Gupta, Jeanette Erdmann, Nilesh J Samani, Heribert Schunkert, Hugh Watkins, Cristen J Willer, Panos Deloukas, Sekar Kathiresan, Adam S Butterworth, and CARDIoGRAMplusC4D Consortium. Discovery and systematic characterization of risk variants and genes for coronary artery disease in over a million participants. *Nat. Genet.*, 54(12):1803–1815, December 2022.

[75] Siew C Ng, Charles N Bernstein, Morten H Vatn, Peter Laszlo Lakatos, Edward V Loftus, Jr, Curt Tysk, Colm O'Morain, Bjorn Moum, Jean-Frédéric Colombel, and Epidemiology and Natural History Task Force of the International Organization of Inflammatory Bowel Disease (IOIBD). Geographical variability and environmental risk factors in inflammatory bowel disease. *Gut*, 62(4):630–649, April 2013.

[76] Lauren E Thurgate, Daniel A Lemberg, Andrew S Day, and Steven T Leach. An overview of inflammatory bowel disease unclassified in children. *Inflamm. Intest. Dis.*,

4(3):97–103, August 2019.

[77] Barbara A Hendrickson, Ranjana Gokhale, and Judy H Cho. Clinical aspects and pathophysiology of inflammatory bowel disease. *Clin. Microbiol. Rev.*, 15(1):79–94, January 2002.

[78] Robert T Lewis and David J Maron. Efficacy and complications of surgery for crohn's disease. *Gastroenterol. Hepatol. (N. Y.)*, 6(9):587–596, September 2010.

[79] Charles N Bernstein and Fergus Shanahan. Disorders of a modern lifestyle: reconciling the epidemiology of inflammatory bowel diseases. *Gut*, 57(9):1185–1191, September 2008.

[80] C E Richardson, J M Morgan, B Jasani, J T Green, J Rhodes, G T Williams, J Lindstrom, S Wonnacott, S Peel, and G A O Thomas. Effect of smoking and transdermal nicotine on colonic nicotinic acetylcholine receptors in ulcerative colitis. *QJM*, 96(1):57–65, January 2003.

[81] Julie A Cornish, Emile Tan, Constantinos Simillis, Susan K Clark, Julian Teare, and Paris P Tekkis. The risk of oral contraceptives in the etiology of inflammatory bowel disease: a meta-analysis. *Am. J. Gastroenterol.*, 103(9):2394–2400, September 2008.

[82] I E Koutroubakis and I G Vlachonikolis. Appendectomy and the development of ulcerative colitis: results of a metaanalysis of published case-control studies. *Am. J. Gastroenterol.*, 95(1):171–176, January 2000.

[83] Jacob J Rozich, Ariela Holmer, and Siddharth Singh. Effect of lifestyle factors on outcomes in patients with inflammatory bowel diseases. *Am. J. Gastroenterol.*, 115(6):832–840, June 2020.

[84] Yanhua Yang, Lili Xiang, and Jianhua He. Beverage intake and risk of crohn disease: A meta-analysis of 16 epidemiological studies. *Medicine (Baltimore)*, 98(21):e15795, May 2019.

[85] Animesh Jain, Nghia H Nguyen, James A Proudfoot, Christopher F Martin, William J Sandborn, Michael D Kappelman, Millie D Long, and Siddharth Singh. Impact of obesity on disease activity and Patient-Reported outcomes measurement information system (PROMIS) in inflammatory bowel diseases. *Am. J. Gastroenterol.*, 114(4):630–639, April 2019.

[86] S Reif, I Klein, F Lubin, M Farbstein, A Hallak, and T Gilat. Pre-illness dietary factors in inflammatory bowel disease. *Gut*, 40(6):754–760, June 1997.

[87] Siew C Ng, Susannah Woodrow, Nisha Patel, Javaid Subhani, and Marcus Harbord. Role of genetic and environmental factors in british twins with inflammatory bowel disease. *Inflamm. Bowel Dis.*, 18(4):725–736, April 2012.

[88] Luke Jostins, The International IBD Genetics Consortium (IIBDGC), Stephan Ripke, Rinse K Weersma, Richard H Duerr, Dermot P McGovern, Ken Y Hui, James C Lee, L Philip Schumm, Yashoda Sharma, Carl A Anderson, Jonah Essers, Mitja Mitrovic, Kaida Ning, Isabelle Cleynen, Emilie Theatre, Sarah L Spain, Soumya

Raychaudhuri, Philippe Goyette, Zhi Wei, Clara Abraham, Jean-Paul Achkar, Tariq Ahmad, Leila Amininejad, Ashwin N Ananthakrishnan, Vibeke Andersen, Jane M Andrews, Leonard Baidoo, Tobias Balschun, Peter A Bampton, Alain Bitton, Gabrielle Boucher, Stephan Brand, Carsten Büning, Ariella Cohain, Sven Cichon, Mauro D'Amato, Dirk De Jong, Kathy L Devaney, Marla Dubinsky, Cathryn Edwards, David Ellinghaus, Lynnette R Ferguson, Denis Franchimont, Karin Fransen, Richard Gearry, Michel Georges, Christian Gieger, Jürgen Glas, Talin Haritunians, Ailsa Hart, Chris Hawkey, Matija Hedl, Xinli Hu, Tom H Karlsen, Limas Kupcinskas, Subra Kugathasan, Anna Latiano, Debby Laukens, Ian C Lawrance, Charlie W Lees, Edouard Louis, Gillian Mahy, John Mansfield, Angharad R Morgan, Craig Mowat, William Newman, Orazio Palmieri, Cyriel Y Ponsioen, Uros Potocnik, Natalie J Prescott, Miguel Regueiro, Jerome I Rotter, Richard K Russell, Jeremy D Sanderson, Miquel Sans, Jack Satsangi, Stefan Schreiber, Lisa A Simms, Jurgita Sventoraityte, Stephan R Targan, Kent D Taylor, Mark Tremelling, Hein W Verspaget, Martine De Vos, Cisca Wijmenga, David C Wilson, Juliane Winkelmann, Ramnik J Xavier, Sebastian Zeissig, Bin Zhang, Clarence K Zhang, Hongyu Zhao, Mark S Silverberg, Vito Annese, Hakon Hakonarson, Steven R Brant, Graham Radford-Smith, Christopher G Mathew, John D Rioux, Eric E Schadt, Mark J Daly, Andre Franke, Miles Parkes, Severine Vermeire, Jeffrey C Barrett, and Judy H Cho. Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, 491(7422):119–124, November 2012.

[89] Katrina M de Lange, Loukas Moutsianas, James C Lee, Christopher A Lamb, Yang Luo, Nicholas A Kennedy, Luke Jostins, Daniel L Rice, Javier Gutierrez-Achury, Sun-Gou Ji, Graham Heap, Elaine R Nimmo, Cathryn Edwards, Paul Henderson, Craig Mowat, Jeremy Sanderson, Jack Satsangi, Alison Simmons, David C Wilson, Mark Tremelling, Ailsa Hart, Christopher G Mathew, William G Newman, Miles Parkes, Charlie W Lees, Holm Uhlig, Chris Hawkey, Natalie J Prescott, Tariq Ahmad, John C Mansfield, Carl A Anderson, and Jeffrey C Barrett. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.*, 49(2):256–261, February 2017.

[90] Jimmy Z Liu, Suzanne van Sommeren, Hailiang Huang, Siew C Ng, Rudi Alberts, Atsushi Takahashi, Stephan Ripke, James C Lee, Luke Jostins, Tejas Shah, Shifteh Abedian, Jae Hee Cheon, Judy Cho, Naser E Daryani, Lude Franke, Yuta Fuyuno, Ailsa Hart, Ramesh C Juyal, Garima Juyal, Won Ho Kim, Andrew P Morris, Hossein Poustchi, William G Newman, Vandana Midha, Timothy R Orchard, Homayon Vahedi, Ajit Sood, Joseph J Y Sung, Reza Malekzadeh, Harm-Jan Westra, Keiko Yamazaki, Suk-Kyun Yang, Jeffrey C Barrett, Andre Franke, Behrooz Z Alizadeh, Miles Parkes, Thelma, Mark J Daly, Michiaki Kubo, Carl A Anderson, Rinse K Weersma, International Multiple Sclerosis Genetics Consortium, and International IBD Genetics Consortium. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.*, 47(9):979–986, September 2015.

[91] Yang Luo, Katrina M de Lange, Luke Jostins, Loukas Moutsianas, Joshua Randall, Nicholas A Kennedy, Christopher A Lamb, Shane McCarthy, Tariq Ahmad, Cathryn Edwards, Eva Goncalves Serra, Ailsa Hart, Chris Hawkey, John C Mansfield, Craig Mowat, William G Newman, Sam Nichols, Martin Pollard, Jack Satsangi, Alison

Simmons, Mark Tremelling, Holm Uhlig, David C Wilson, James C Lee, Natalie J Prescott, Charlie W Lees, Christopher G Mathew, Miles Parkes, Jeffrey C Barrett, and Carl A Anderson. Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7. *Nat. Genet.*, 49(2):186–192, February 2017.

[92] Bernard Khor, Agnès Gardet, and Ramnik J Xavier. Genetics and pathogenesis of inflammatory bowel disease. *Nature*, 474(7351):307–317, June 2011.

[93] Hirotaka Iwaki, Cornelis Blauwendraat, Hampton L Leonard, Jonggeol J Kim, Ganqiang Liu, Jodi Maple-Grødem, Jean-Christophe Corvol, Lasse Pihlstrøm, Marlies van Nimwegen, Samantha J Hutten, Khanh-Dung H Nguyen, Jacqueline Rick, Shirley Eberly, Faraz Faghri, Peggy Auinger, Kirsten M Scott, Ruwani Wijeyekoon, Vivianna M Van Deerlin, Dena G Hernandez, J Raphael Gibbs, International Parkinson's Disease Genomics Consortium, Kumaraswamy Naidu Chitrala, Aaron G Day-Williams, Alexis Brice, Guido Alves, Alastair J Noyce, Ole-Bjørn Tysnes, Jonathan R Evans, David P Breen, Karol Estrada, Claire E Wegel, Fabrice Danjou, David K Simon, Ole Andreassen, Bernard Ravina, Mathias Toft, Peter Heutink, Bastiaan R Bloem, Daniel Weintraub, Roger A Barker, Caroline H Williams-Gray, Bart P van de Warrenburg, Jacobus J Van Hilten, Clemens R Scherzer, Andrew B Singleton, and Mike A Nalls. Genomewide association study of parkinson's disease clinical biomarkers in 12 longitudinal patients' cohorts. *Mov. Disord.*, 34(12):1839–1850, December 2019.

[94] Severe Covid-19 GWAS Group, David Ellinghaus, Frauke Degenhardt, Luis Bujanda, Maria Buti, Agustín Albillos, Pietro Invernizzi, Javier Fernández, Daniele Prati, Guido Baselli, Rosanna Asselta, Marit M Grimsrud, Chiara Milani, Fátima Aziz, Jan Kässens, Sandra May, Mareike Wendorff, Lars Wienbrandt, Florian Uellendahl-Werth, Tenghao Zheng, Xiaoli Yi, Raúl de Pablo, Adolfo G Chercoles, Adriana Palom, Alba-Estela Garcia-Fernandez, Francisco Rodriguez-Frias, Alberto Zanella, Alessandra Bandera, Alessandro Protti, Alessio Aghemo, Ana Lleo, Andrea Biondi, Andrea Caballero-Garralda, Andrea Gori, Anja Tanck, Anna Carreras Nolla, Anna Latiano, Anna Ludovica Fracanzani, Anna Peschuck, Antonio Julià, Antonio Pesenti, Antonio Voza, David Jiménez, Beatriz Mateos, Beatriz Nafria Jimenez, Carmen Quereda, Cinzia Paccapelo, Christoph Gassner, Claudio Angelini, Cristina Cea, Aurora Solier, David Pestaña, Eduardo Muñiz-Diaz, Elena Sandoval, Elvezia M Paraboschi, Enrique Navas, Félix García Sánchez, Ferruccio Ceriotti, Filippo Martinelli-Boneschi, Flora Peyvandi, Francesco Blasi, Luis Téllez, Albert Blanco-Grau, Georg Hemmrich-Stanisak, Giacomo Grasselli, Giorgio Costantino, Giulia Cardamone, Giuseppe Foti, Serena Aneli, Hayato Kurihara, Hesham ElAbd, Ilaria My, Iván Galván-Femenia, Javier Martín, Jeanette Erdmann, Jose Ferrusquía-Acosta, Koldo Garcia-Etxebarria, Laura Izquierdo-Sanchez, Laura R Bettini, Lauro Sumoy, Leonardo Terranova, Leticia Moreira, Luigi Santoro, Luigia Scudeller, Francisco Mesonero, Luisa Roade, Malte C Rühlemann, Marco Schaefer, Maria Carrabba, Mar Riveiro-Barciela, Maria E Figuera Basso, Maria G Valsecchi, María Hernandez-Tejero, Marialbert Acosta-Herrera, Mariella D'Angiò, Marina Baldini, Marina Cazzaniga, Martin Schulzky, Maurizio Cecconi, Michael Wittig, Michele Ciccarelli, Miguel Rodríguez-Gandía, Monica Bocciolone, Monica Miozzo, Nicola Montano, Nicole Braun, Nicoletta Sacchi, Nilda Martínez, Onur Özer, Orazio Palmieri, Paola Faverio, Paoletta Preatoni,

Paolo Bonfanti, Paolo Omodei, Paolo Tentorio, Pedro Castro, Pedro M Rodrigues, Aaron Blandino Ortiz, Rafael de Cid, Ricard Ferrer, Roberta Gualtierotti, Rosa Nieto, Siegfried Goerg, Salvatore Badalamenti, Sara Marsal, Giuseppe Matullo, Serena Pelusi, Simonas Juzenas, Stefano Aliberti, Valter Monzani, Victor Moreno, Tanja Wesse, Tobias L Lenz, Tomas Pumarola, Valeria Rimoldi, Silvano Bosari, Wolfgang Albrecht, Wolfgang Peter, Manuel Romero-Gómez, Mauro D'Amato, Stefano Duga, Jesus M Banales, Johannes R Hov, Trine Folseraas, Luca Valenti, Andre Franke, and Tom H Karlsen. Genomewide association study of severe covid-19 with respiratory failure. *N. Engl. J. Med.*, 383(16):1522–1534, October 2020.

[95] Isabelle Cleynen, Gabrielle Boucher, Luke Jostins, L Philip Schumm, Sebastian Zeissig, Tariq Ahmad, Vibeke Andersen, Jane M Andrews, Vito Annese, Stephan Brand, Steven R Brant, Judy H Cho, Mark J Daly, Marla Dubinsky, Richard H Duerr, Lynnette R Ferguson, Andre Franke, Richard B Gearry, Philippe Goyette, Hakon Hakonarson, Jonas Halfvarson, Johannes R Hov, Hailang Huang, Nicholas A Kennedy, Limas Kupcinskas, Ian C Lawrance, James C Lee, Jack Satsangi, Stephan Schreiber, Emilie Théâtre, Andrea E van der Meulen-de Jong, Rinse K Weersma, David C Wilson, Miles Parkes, Severine Vermeire, John D Rioux, John Mansfield, Mark S Silverberg, Graham Radford-Smith, Dermot P B McGovern, Jeffrey C Barrett, and Charlie W Lees. Inherited determinants of crohn's disease and ulcerative colitis phenotypes: a genetic association study. *Lancet*, 387(10014):156–167, January 2016.

[96] James C Lee, Daniele Biasci, Rebecca Roberts, Richard B Gearry, John C Mansfield, Tariq Ahmad, Natalie J Prescott, Jack Satsangi, David C Wilson, Luke Jostins, Carl A Anderson, UK IBD Genetics Consortium, James A Traherne, Paul A Lyons, Miles Parkes, and Kenneth G C Smith. Genome-wide association study identifies distinct genetic contributions to prognosis and susceptibility in crohn's disease. *Nat. Genet.*, 49(2):262–268, February 2017.

[97] Jeffrey C Barrett, Sarah Hansoul, Dan L Nicolae, Judy H Cho, Richard H Duerr, John D Rioux, Steven R Brant, Mark S Silverberg, Kent D Taylor, M Michael Barmada, Alain Bitton, Themistocles Dassopoulos, Lisa Wu Datta, Todd Green, Anne M Griffiths, Emily O Kistner, Michael T Murtha, Miguel D Regueiro, Jerome I Rotter, L Philip Schumm, A Hillary Steinhart, Stephan R Targan, Ramnik J Xavier, NIDDK IBD Genetics Consortium, Cécile Libioulle, Cynthia Sandor, Mark Lathrop, Jacques Belaiche, Olivier Dewit, Ivo Gut, Simon Heath, Debby Laukens, Myriam Mni, Paul Rutgeerts, André Van Gossum, Diana Zelenika, Denis Franchimont, Jean-Pierre Hugot, Martine de Vos, Severine Vermeire, Edouard Louis, Belgian-French IBD Consortium, Wellcome Trust Case Control Consortium, Lon R Cardon, Carl A Anderson, Hazel Drummond, Elaine Nimmo, Tariq Ahmad, Natalie J Prescott, Clive M Onnie, Sheila A Fisher, Jonathan Marchini, Jilur Ghori, Suzannah Bumpstead, Rhian Gwilliam, Mark Tremelling, Panos Deloukas, John Mansfield, Derek Jewell, Jack Satsangi, Christopher G Mathew, Miles Parkes, Michel Georges, and Mark J Daly. Genome-wide association defines more than 30 distinct susceptibility loci for crohn's disease. *Nat. Genet.*, 40(8):955–962, August 2008.

[98] Adil Harroud, Pernilla Stridh, Jacob L. McCauley, Janna Saarela, Aletta M. van den Bosch, Hendrik J. Engelenburg, Ashley H. Beecham, Lars Alfredsson, Katayoun Alikhani, Lilyana Amezcua, and et al. Locus for severity implicates cns resilience in progression of multiple sclerosis. *Nature*, 619(7969):323–331, 2023.

[99] Doug Speed, Clive Hoggart, Slave Petrovski, Ioanna Tachmazidou, Alison Coffey, Andrea Jorgensen, Hariklia Eleftherohorinou, Maria De Iorio, Marian Todaro, Tisham De, David Smith, Philip E Smith, Margaret Jackson, Paul Cooper, Mark Kellett, Stephen Howell, Mark Newton, Raju Yerra, Meng Tan, Chris French, Markus Reuber, Graeme E Sills, David Chadwick, Munir Pirmohamed, David Bentley, Ingrid Scheffer, Samuel Berkovic, David Balding, Aarno Palotie, Anthony Marson, Terence J O'Brien, and Michael R Johnson. A genome-wide association study and biological pathway analysis of epilepsy prognosis in a prospective cohort of newly treated epilepsy. *Hum. Mol. Genet.*, 23(1):247–258, January 2014.

[100] Matthew T Maurano, Richard Humbert, Eric Rynes, Robert E Thurman, Eric Haugen, Hao Wang, Alex P Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, Anthony Shafer, Fidencio Neri, Kristen Lee, Tanya Kutyavin, Sandra Stehling-Sun, Audra K Johnson, Theresa K Canfield, Erika Giste, Morgan Diegel, Daniel Bates, R Scott Hansen, Shane Neph, Peter J Sabo, Shelly Heimfeld, Antony Raubitschek, Steven Ziegler, Chris Cotsapas, Nona Sotoodehnia, Ian Glass, Shamil R Sunyaev, Rajinder Kaul, and John A Stamatoyannopoulos. Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337(6099):1190–1195, September 2012.

[101] Marc A Schaub, Alan P Boyle, Anshul Kundaje, Serafim Batzoglou, and Michael Snyder. Linking disease associations with regulatory information in the human genome. *Genome Res.*, 22(9):1748–1759, September 2012.

[102] Sung Chun, Alexandra Casparino, Nikolaos A Patsopoulos, Damien C Croteau-Chonka, Benjamin A Raby, Philip L De Jager, Shamil R Sunyaev, and Chris Cotsapas. Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nat. Genet.*, 49(4):600–605, April 2017.

[103] Hakhamanesh Mostafavi, Jeffrey P Spence, Sahin Naqvi, and Jonathan K Pritchard. Limited overlap of eQTLs and GWAS hits due to systematic differences in discovery. May 2022.

[104] Yang I Li, Bryce van de Geijn, Anil Raj, David A Knowles, Allegra A Petti, David Golan, Yoav Gilad, and Jonathan K Pritchard. RNA splicing is a primary link between genetic variation and disease. *Science*, 352(6285):600–604, April 2016.

[105] Sarah Kim-Hellmuth, François Aguet, Meritxell Oliva, Manuel Muñoz-Aguirre, Silva Kasela, Valentin Wucher, Stephane E Castel, Andrew R Hamel, Ana Viñuela, Amy L Roberts, Serghei Mangul, Xiaoquan Wen, Gao Wang, Alvaro N Barbeira, Diego Garrido-Martín, Brian B Nadel, Yuxin Zou, Rodrigo Bonazzola, Jie Quan, Andrew Brown, Angel Martinez-Perez, José Manuel Soria, Gad Getz, Emmanouil T Dermitzakis, Kerrin S Small, Matthew Stephens, Hualin S Xi, Hae Kyung Im, Roderic Guigó, Ayellet V Segrè, Barbara E Stranger, Kristin G Ardlie, Tuuli Lappalainen, François Aguet, Shankara Anand, Kristin G Ardlie, Stacey Gabriel, Gad A Getz, Aaron Graubert, Kane Hadley, Robert E Handsaker, Katherine H Huang, Seva Kashin, Xiao Li, Daniel G MacArthur, Samuel R Meier, Jared L Nedzel, Duyen T Nguyen, Ayellet V Segrè, Ellen Todres, Brunilda Balliu, Alvaro N Barbeira, Alexis Battle, Rodrigo Bonazzola, Andrew Brown, Christopher D Brown, Stephane E Castel, Donald F

Conrad, Daniel J Cotter, Nancy Cox, Sayantan Das, Olivia M de Goede, Emmanouil T Dermitzakis, Jonah Einson, Barbara E Engelhardt, Eleazar Eskin, Tiffany Y Eulalio, Nicole M Ferraro, Elise D Flynn, Laure Fresard, Eric R Gamazon, Diego Garrido-Martín, Nicole R Gay, Michael J Gloudemans, Roderic Guigó, Andrew R Hame, Yuan He, Paul J Hoffman, Farhad Hormozdiari, Lei Hou, Hae Kyung Im, Brian Jo, Silva Kasela, Manolis Kellis, Sarah Kim-Hellmuth, Alan Kwong, Tuuli Lappalainen, Xin Li, Yanyu Liang, Serghei Mangul, Pejman Mohammadi, Stephen B Montgomery, Manuel Muñoz-Aguirre, Daniel C Nachun, Andrew B Nobel, Meritxell Oliva, Yoson Park, Yongjin Park, Princy Parsana, Abhiram S Rao, Ferran Reverter, John M Rouhana, Chiara Sabatti, Ashis Saha, Matthew Stephens, Barbara E Stranger, Benjamin J Strober, Nicole A Teran, Ana Viñuela, Gao Wang, Xiaoquan Wen, Fred Wright, Valentin Wucher, Yuxin Zou, Pedro G Ferreira, Gen Li, Marta Melé, Esti Yeger-Lotem, Mary E Barcus, Debra Bradbury, Tanya Krubit, Jeffrey A McLean, Liqun Qi, Karna Robinson, Nancy V Roche, Anna M Smith, Leslie Sobin, David E Tabor, Anita Undale, Jason Bridge, Lori E Brigham, Barbara A Foster, Bryan M Gillard, Richard Hasz, Marcus Hunter, Christopher Johns, Mark Johnson, Ellen Karasik, Gene Kopen, William F Leinweber, Alisa McDonald, Michael T Moser, Kevin Myer, Kimberley D Ramsey, Brian Roe, Saboor Shad, Jeffrey A Thomas, Gary Walters, Michael Washington, Joseph Wheeler, Scott D Jewell, Daniel C Rohrer, Dana R Valley, David A Davis, Deborah C Mash, Philip A Branton, Laura K Barker, Heather M Gardiner, Maghboeba Mosavel, Laura A Siminoff, Paul Flicek, Maximilian Haeussler, Thomas Juettemann, W James Kent, Christopher M Lee, Conner C Powell, Kate R Rosenbloom, Magali Ruffier, Dan Sheppard, Kieron Taylor, Stephen J Trevanion, Daniel R Zerbino, Nathan S Abell, Joshua Akey, Lin Chen, Kathryn Demanelis, Jennifer A Doherty, Andrew P Feinberg, Kasper D Hansen, Peter F Hickey, Farzana Jasmine, Lihua Jiang, Rajinder Kaul, Muhammad G Kibriya, Jin Billy Li, Qin Li, Shin Lin, Sandra E Linder, Brandon L Pierce, Lindsay F Rizzardi, Andrew D Skol, Kevin S Smith, Michael Snyder, John Stamatoyannopoulos, Hua Tang, Meng Wang, Latarsha J Carithers, Ping Guan, Susan E Koester, A Roger Little, Helen M Moore, Concepcion R Nierras, Abhi K Rao, Jimmie B Vaught, Simona Volpi, and GTEx Consortium. Cell type–specific genetic regulation of gene expression across human tissues. *Science*, 369(6509):eaaz8528, September 2020.

[106] Helena Kilpinen, Angela Goncalves, Andreas Leha, Vackar Afzal, Kaur Alasoo, Sofie Ashford, Sendu Bala, Dalila Bensaddek, Francesco Paolo Casale, Oliver J Culley, Petr Danecek, Adam Faulconbridge, Peter W Harrison, Annie Kathuria, Davis McCarthy, Shane A McCarthy, Ruta Meleckyte, Yasin Memari, Nathalie Moens, Filipa Soares, Alice Mann, Ian Streeter, Chukwuma A Agu, Alex Alderton, Rachel Nelson, Sarah Harper, Minal Patel, Alistair White, Sharad R Patel, Laura Clarke, Reena Halai, Christopher M Kirton, Anja Kolb-Kokocinski, Philip Beales, Ewan Birney, Davide Danovi, Angus I Lamond, Willem H Ouwehand, Ludovic Vallier, Fiona M Watt, Richard Durbin, Oliver Stegle, and Daniel J Gaffney. Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature*, 546(7658):370–375, June 2017.

[107] Hao Zhao, Zhifu Sun, Jing Wang, Haojie Huang, Jean-Pierre Kocher, and Liguo Wang. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*, 30(7):1006–1007, April 2014.

[108] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, 38(8):904–909, August 2006.

[109] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, January 2013.

[110] Bryce van de Geijn, Graham McVicker, Yoav Gilad, and Jonathan K Pritchard. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods*, 12(11):1061–1063, November 2015.

[111] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009.

[112] Kelsy C Cotto, Yang-Yang Feng, Avinash Ramu, Megan Richters, Sharon L Freshour, Zachary L Skidmore, Huiming Xia, Joshua F McMichael, Jason Kunisaki, Katie M Campbell, Timothy Hung-Po Chen, Emily B Rozycki, Douglas Adkins, Siddhartha Devarakonda, Sumithra Sankararaman, Yiing Lin, William C Chapman, Christopher A Maher, Vivek Arora, Gavin P Dunn, Ravindra Uppaluri, Ramaswamy Govindan, Obi L Griffith, and Malachi Griffith. Integrated analysis of genomic and transcriptomic data for the discovery of splice-associated variants in cancer. *Nat. Commun.*, 14(1):1589, March 2023.

[113] Yang I Li, David A Knowles, Jack Humphrey, Alvaro N Barbeira, Scott P Dickinson, Hae Kyung Im, and Jonathan K Pritchard. Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.*, 50(1):151–158, January 2018.

[114] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform manifold approximation and projection. *J. Open Source Softw.*, 3(29):861, September 2018.

[115] Olivier Delaneau, Halit Ongen, Andrew A Brown, Alexandre Fort, Nikolaos I Panousis, and Emmanouil T Dermitzakis. A complete tool set for molecular QTL discovery and analysis. *Nat. Commun.*, 8(1):15452, May 2017.

[116] John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A.*, 100(16):9440–9445, August 2003.

[117] Annalisa Buniello, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, Daniel Suveges, Olga Vrousgou, Patricia L Whetzel, Ridwan Amode, Jose A Guillen, Harpreet S Riat, Stephen J Trevanion, Peggy Hall, Heather Junkins, Paul Flicek, Tony Burdett, Lucia A Hindorff, Fiona Cunningham, and Helen Parkinson. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, 47(D1):D1005–D1012, January 2019.

[118] Wei Zhou, Jonas B Nielsen, Lars G Fritsche, Rounak Dey, Maiken E Gabrielsen, Brooke N Wolford, Jonathon LeFaive, Peter VandeHaar, Sarah A Gagliano, Aliya Gifford, Lisa A Bastarache, Wei-Qi Wei, Joshua C Denny, Maoxuan Lin, Kristian Hveem, Hyun Min Kang, Goncalo R Abecasis, Cristen J Willer, and Seunggeun Lee. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.*, 50(9):1335–1341, September 2018.

[119] Hongfei Liu, Paolo A Lorenzini, Fan Zhang, Shaohai Xu, Mei Su M Wong, Jie Zheng, and Xavier Roca. Alternative splicing analysis in human monocytes and macrophages reveals MBNL1 as major regulator. *Nucleic Acids Res.*, 46(12):6069–6086, July 2018.

[120] Harald Stenmark. Rab GTPases as coordinators of vesicle traffic. *Nat. Rev. Mol. Cell Biol.*, 10(8):513–525, August 2009.

[121] Ayuko Sakane, Ahmed Alamir Mahmoud Abdallah, Kiyoshi Nakano, Kazufumi Honda, Wataru Ikeda, Yumiko Nishikawa, Mitsuru Matsumoto, Natsuki Matsushita, Toshio Kitamura, and Takuya Sasaki. Rab13 small G protein and junctional rab13-binding protein (JRAB) orchestrate actin cytoskeletal organization during epithelial junctional development. *J. Biol. Chem.*, 287(51):42455–42468, December 2012.

[122] Mitsutoshi Yoneyama, Mika Kikuchi, Kanae Matsumoto, Tadaatsu Imaizumi, Makoto Miyagishi, Kazunari Taira, Eileen Foy, Yueh-Ming Loo, Michael Gale, Jr, Shizuo Akira, Shin Yonehara, Atsushi Kato, and Takashi Fujita. Shared and unique functions of the DExD/H-box helicases RIG-I, MDA5, and LGP2 in antiviral innate immunity. *J. Immunol.*, 175(5):2851–2858, September 2005.

[123] Hiroki Kato, Shintaro Sato, Mitsutoshi Yoneyama, Masahiro Yamamoto, Satoshi Uematsu, Kosuke Matsui, Tohru Tsujimura, Kiyoshi Takeda, Takashi Fujita, Osamu Takeuchi, and Shizuo Akira. Cell type-specific involvement of RIG-I in antiviral response. *Immunity*, 23(1):19–28, July 2005.

[124] Céline Castanier, Naima Zemirli, Alain Portier, Dominique Garcin, Nicolas Bidère, Aimé Vazquez, and Damien Arnoult. MAVS ubiquitination by the E3 ligase TRIM25 and degradation by the proteasome is involved in type I interferon production after activation of the antiviral RIG-I-like receptors. *BMC Biol.*, 10(1):44, May 2012.

[125] Mariya Al Hamrashdi and Gareth Brady. Regulation of IRF3 activation in human antiviral signaling pathways. *Biochem. Pharmacol.*, 200(115026):115026, June 2022.

[126] Danish Saleh, Malek Najjar, Matija Zelic, Saumil Shah, Shoko Nogusa, Apostolos Polykratis, Michelle K Paczosa, Peter J Gough, John Bertin, Michael Whalen, Katherine A Fitzgerald, Nikolai Slavov, Manolis Pasparakis, Siddharth Balachandran, Michelle Kelliher, Joan Mecsas, and Alexei Degterev. Kinase activities of RIPK1 and RIPK3 can direct IFN-β synthesis induced by lipopolysaccharide. *J. Immunol.*, 198(11):4435–4447, June 2017.

[127] Michaela Ulrike Gack, Randy Allen Albrecht, Tomohiko Urano, Kyung-Soo Inn, I-Chueh Huang, Elena Carnero, Michael Farzan, Satoshi Inoue, Jae Ung Jung, and Adolfo García-Sastre. Influenza a virus NS1 targets the ubiquitin ligase TRIM25 to evade recognition by the host viral RNA sensor RIG-I. *Cell Host Microbe*, 5(5):439–449, May 2009.

[128] Kuo-Chieh Liao and Mariano A Garcia-Blanco. Role of alternative splicing in regulating host response to viral infection. *Cells*, 10(7):1720, July 2021.

[129] Sonya P Lad, Guang Yang, David A Scott, Ta-Hsiang Chao, Jean da Silva Correia, Juan Carlos de la Torre, and Erguang Li. Identification of MAVS splicing variants that interfere with RIGI/MAVS pathway signaling. *Mol. Immunol.*, 45(8):2277–2287, April 2008.

[130] Allison R Wagner, Haley M Scott, Kelsi O West, Krystal J Vail, Timothy C Fitzsimons, Aja K Coleman, Kaitlyn E Carter, Robert O Watson, and Kristin L Patrick. Global transcriptomics uncovers distinct contributions from splicing regulatory proteins to the macrophage innate immune response. *Front. Immunol.*, 12:656885, July 2021.

[131] Haroon Kalam, Mary F Fontana, and Dhiraj Kumar. Alternate splicing of transcripts shape macrophage response to mycobacterium tuberculosis infection. *PLoS Pathog.*, 13(3):e1006236, March 2017.

[132] Duygu Unuvar Purcu, Asli Korkmaz, Sinem Gunalp, Derya Goksu Helvaci, Yonca Erdal, Yavuz Dogan, Asli Suner, Gerhard Wingender, and Duygu Sag. Effect of stimulation time on the expression of human macrophage polarization markers. *PLoS One*, 17(3):e0265196, March 2022.

[133] Omar Sharif, Viacheslav N Bolshakov, Stephanie Raines, Peter Newham, and Neil D Perkins. Transcriptional profiling of the LPS induced NF-kappaB response in macrophages. *BMC Immunol.*, 8(1):1, January 2007.

[134] Eric T Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F Kingsmore, Gary P Schroth, and Christopher B Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, November 2008.

[135] Tuuli Lappalainen, The Geuvadis Consortium, Michael Sammeth, Marc R Friedländer, Peter A C 't Hoen, Jean Monlong, Manuel A Rivas, Mar Gonzàlez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G Ferreira, Matthias Barann, Thomas Wieland, Liliana Greger, Maarten van Iterson, Jonas Almlöf, Paolo Ribeca, Irina Pulyakhina, Daniela Esser, Thomas Giger, Andrew Tikhonov, Marc Sultan, Gabrielle Bertier, Daniel G MacArthur, Monkol Lek, Esther Lizano, Henk P J Buermans, Ismael Padioleau, Thomas Schwarzmayr, Olof Karlberg, Halit Ongen, Helena Kilpinen, Sergi Beltran, Marta Gut, Katja Kahlem, Vyacheslav Amstislavskiy, Oliver Stegle, Matti Pirinen, Stephen B Montgomery, Peter Donnelly, Mark I McCarthy, Paul Flicek, Tim M Strom, Hans Lehrach, Stefan Schreiber, Ralf Sudbrak, Ángel Carracedo, Stylianos E Antonarakis, Robert Häsler, Ann-Christine Syvänen, Gert-Jan van Ommen, Alvis Brazma, Thomas Meitinger, Philip Rosenstiel, Roderic Guigó, Ivo G Gut, Xavier Estivill, and Emmanouil T Dermitzakis. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, September 2013.

[136] Claudia Giambartolomei, Damjan Vukcevic, Eric E Schadt, Lude Franke, Aroon D Hingorani, Chris Wallace, and Vincent Plagnol. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.*, 10(5):e1004383, May 2014.

[137] Sofie Symoens, Fransiska Malfait, Philip Vlummens, Trinh Hermanns-Lê, Delfien Syx, and Anne De Paepe. A novel splice variant in the n-propeptide of COL5A1 causes an EDS phenotype with severe kyphoscoliosis and eye involvement. *PLoS One*, 6(5):e20121, May 2011.

[138] Anna Abramowicz and Monika Gos. Splicing mutations in human genetic disorders: examples, detection, and confirmation. *J. Appl. Genet.*, 59(3):253–268, August 2018.

[139] David J Sanz, Jennifer A Hollywood, Martina F Scallan, and Patrick T Harrison. Cas9/gRNA targeted excision of cystic fibrosis-causing deep-intronic splicing mutations restores normal splicing of CFTR mRNA. *PLoS One*, 12(9):e0184009, September 2017.

[140] Abhinav Nellore, Andrew E Jaffe, Jean-Philippe Fortin, José Alquicira-Hernández, Leonardo Collado-Torres, Siruo Wang, Robert A Phillips, III, Nishika Karbhari, Kasper D Hansen, Ben Langmead, and Jeffrey T Leek. Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the sequence read archive. *Genome Biol.*, 17(1), December 2016.

[141] Joseph K Pickrell, Athma A Pai, Yoav Gilad, and Jonathan K Pritchard. Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet.*, 6(12):e1001236, December 2010.

[142] Marianna Parlato, Qing Nian, Fabienne Charbit-Henrion, Frank M Ruemmele, Fernando Rodrigues-Lima, Nadine Cerf-Bensussan, and Immunobiota Study Group. Loss-of-function mutation in PTPN2 causes aberrant activation of JAK signaling via STAT and very early onset intestinal inflammation. *Gastroenterology*, 159(5):1968–1971.e4, November 2020.

[143] Kelly A Pike and Michel L Tremblay. Protein tyrosine phosphatases: Regulators of CD4 T cells in inflammatory bowel disease. *Front. Immunol.*, 9:2504, October 2018.

[144] Marianne R Spalinger, Roberto Manzini, Larissa Hering, Julianne B Riggs, Claudia Gottier, Silvia Lang, Kirstin Atrott, Antonia Fettelschoss, Florian Olomski, Thomas M Kündig, Michael Fried, Declan F McCole, Gerhard Rogler, and Michael Scharl. PTPN2 regulates inflammasome activation and controls onset of intestinal inflammation and colon cancer. *Cell Rep.*, 22(7):1835–1848, February 2018.

[145] Marianne R Spalinger, Meli'sa Crawford, Sarah D Bobardt, Jiang Li, Anica Sayoc-Becerra, Alina N Santos, Ali Shawki, Pritha Chatterjee, Meera G Nair, and Declan F McCole. Loss of protein tyrosine phosphatase non-receptor type 2 reduces IL-4-driven alternative macrophage activation. *Mucosal Immunol.*, 15(1):74–83, January 2022.

[146] Sara Sigismund, Letizia Lanzetti, Giorgio Scita, and Pier Paolo Di Fiore. Endocytosis in the context-dependent regulation of individual and collective cell properties. *Nat. Rev. Mol. Cell Biol.*, 22(9):625–643, September 2021.

[147] Martin Steger, Francesca Tonelli, Genta Ito, Paul Davies, Matthias Trost, Melanie Vetter, Stefanie Wachter, Esben Lorentzen, Graham Duddy, Stephen Wilson, Marco A S Baptista, Brian K Fiske, Matthew J Fell, John A Morrow, Alastair D Reith, Dario R Alessi, and Matthias Mann. Phosphoproteomics reveals that parkinson's disease kinase LRRK2 regulates a subset of rab GTPases. *Elife*, 5, January 2016.

[148] Eun-Jin Bae, Dong-Kyu Kim, Changyoun Kim, Michael Mante, Anthony Adame, Edward Rockenstein, Ayse Ulusoy, Michael Klinkenberg, Ga Ram Jeong, Jae Ryul Bae, Cheolsoon Lee, He-Jin Lee, Byung-Dae Lee, Donato A Di Monte, Eliezer Masliah, and Seung-Jae Lee. LRRK2 kinase regulates α-synuclein propagation via RAB35 phosphorylation. *Nat. Commun.*, 9(1):3465, August 2018.

[149] Luis Bonet-Ponce, Alexandra Beilina, Chad D Williamson, Eric Lindberg, Jillian H Kluss, Sara Saez-Atienzar, Natalie Landeck, Ravindran Kumaran, Adamantios Mamais, Christopher K E Bleck, Yan Li, and Mark R Cookson. LRRK2 mediates tubulation and vesicle sorting from lysosomes. *Sci. Adv.*, 6(46):eabb2454, November 2020.

[150] Ho-Su Lee, Evy Lobbestael, Séverine Vermeire, João Sabino, and Isabelle Cleynen. Inflammatory bowel disease and parkinson's disease: common pathophysiological links. *Gut*, 70(2):408–417, February 2021.

[151] Paulina S Mrozowska and Mitsunori Fukuda. Regulation of podocalyxin trafficking by rab small GTPases in epithelial cells. *Small GTPases*, 7(4):231–238, October 2016.

[152] Riko Kinoshita, Yuta Homma, and Mitsunori Fukuda. Rab35-GEFs, DENND1A and folliculin differentially regulate podocalyxin trafficking in two- and three-dimensional epithelial cell cultures. *J. Biol. Chem.*, 295(11):3652–3663, March 2020.

[153] Eugene Melamud and John Moult. Structural implication of splicing stochastics. *Nucleic Acids Res.*, 37(14):4862–4872, August 2009.

[154] David J Wright, Nicola A L Hall, Naomi Irish, Angela L Man, Will Glynn, Arne Mould, Alejandro De Los Angeles, Emily Angiolini, David Swarbreck, Karim Gharbi, Elizabeth M Tunbridge, and Wilfried Haerty. Long read sequencing reveals novel isoforms and insights into splicing regulation during cell state changes. *BMC Genomics*, 23(1):42, January 2022.

[155] Maxime Rotival, Hélène Quach, and Lluis Quintana-Murci. Defining the genetic and evolutionary architecture of alternative splicing in response to infection. *Nat. Commun.*, 10(1):1671, April 2019.

[156] Pavel V Mazin, Philipp Khaitovich, Margarida Cardoso-Moreira, and Henrik Kaessmann. Alternative splicing during mammalian organ development. *Nat. Genet.*, 53(6):925–934, June 2021.

[157] Prescott Deininger. Alu elements: know the SINEs. *Genome Biol.*, 12(12):236, December 2011.

[158] Jan Attig, Igor Ruiz de los Mozos, Nejc Haberman, Zhen Wang, Warren Emmett, Kathi Zarnack, Julian König, and Jernej Ule. Splicing repression allows the gradual emergence of new alu-exons in primate evolution. *Elife*, 5, November 2016.

[159] Silke S Singer, Daniela N Männel, Thomas Hehlgans, Jürgen Brosius, and Jürgen Schmitz. From "junk" to gene: curriculum vitae of a primate receptor isoform gene. *J. Mol. Biol.*, 341(4):883–886, August 2004.

[160] Eugenio Mercuri, Basil T Darras, Claudia A Chiriboga, John W Day, Craig Campbell, Anne M Connolly, Susan T Iannaccone, Janbernd Kirschner, Nancy L Kuntz, Kayoko Saito, Perry B Shieh, Már Tulinius, Elena S Mazzone, Jacqueline Montes, Kathie M Bishop, Qingqing Yang, Richard Foster, Sarah Gheuens, C Frank Bennett, Wildon Farwell, Eugene Schneider, Darryl C De Vivo, and Richard S Finkel. Nusinersen versus sham control in later-onset spinal muscular atrophy. *N. Engl. J. Med.*, 378(7):625–635, February 2018.

[161] Thomas C Roberts, Robert Langer, and Matthew J A Wood. Advances in oligonucleotide drug delivery. *Nat. Rev. Drug Discov.*, 19(10):673–694, October 2020.

[162] Yonit Lavin, Deborah Winter, Ronnie Blecher-Gonen, Eyal David, Hadas Keren-Shaul, Miriam Merad, Steffen Jung, and Ido Amit. Tissue-resident macrophage enhancer landscapes are shaped by the local microenvironment. *Cell*, 159(6):1312–1326, December 2014.

[163] Aziz M Al'Khafaji, Jonathan T Smith, Kiran V Garimella, Mehrtash Babadi, Moshe Sade-Feldman, Michael Gatzen, Siranush Sarkizova, Marc A Schwartz, Victoria Popic, Emily M Blaum, Allyson Day, Maura Costello, Tera Bowers, Stacey Gabriel, Eric Banks, Anthony A Philippakis, Genevieve M Boland, Paul C Blainey, and Nir Hacohen. High-throughput RNA isoform sequencing using programmable cDNA concatenation. October 2021.

[164] Shanika L Amarasinghe, Shian Su, Xueyi Dong, Luke Zappia, Matthew E Ritchie, and Quentin Gouil. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.*, 21(1):30, February 2020.

[165] Yu Hu, Li Fang, Xuelian Chen, Jiang F Zhong, Mingyao Li, and Kai Wang. LIQA: long-read isoform quantification and analysis. *Genome Biol.*, 22(1):182, June 2021.

[166] Wee Khoon Ng, Sunny H Wong, and Siew C Ng. Changing epidemiological trends of inflammatory bowel disease in asia. *Intest. Res.*, 14(2):111–119, April 2016.

[167] Charlène Brochard, Marie-Laure Rabilloud, Stéphanie Hamonic, Emma Bajeux, Maël Pagenault, Alain Dabadie, Agathe Gerfaud, Jean-François Viel, Isabelle Tron, Michel Robaszkiewicz, Jean-François Bretagne, Laurent Siproudhis, Guillaume Bouguen, and Groupe ABERMAD. Natural history of perianal crohn's disease: Long-term follow-up of a population-based cohort. *Clin. Gastroenterol. Hepatol.*, 20(2):e102–e110, February 2022.

[168] Tsunekazu Mizushima, Mihoko Ota, Yasushi Fujitani, Yuya Kanauchi, and Ryuichi Iwakiri. Diagnostic features of perianal fistula in patients with crohn's disease: Analysis of a japanese claims database. *Crohns Colitis 360*, 3(3):otab055, July 2021.

[169] Javier Salgado Pogacnik and Gervasio Salgado. Perianal crohn's disease. *Clin. Colon Rectal Surg.*, 32(5):377–385, September 2019.

[170] Pauline Wils, Ariane Leroyer, Mathurin Fumery, Alonso Fernandez-Nistal, Corinne Gower-Rousseau, and Benjamin Pariente. Fistulizing perianal lesions in a french population with crohn's disease. *Dig. Liver Dis.*, 53(5):661–665, May 2021.

[171] Tim W Eglinton, Murray L Barclay, Richard B Gearry, and Frank A Frizelle. The spectrum of perianal crohn's disease in a population-based cohort. *Dis. Colon Rectum*, 55(7):773–777, July 2012.

[172] Samuel O Adegbola, Lesley Dibley, Kapil Sahnan, Tiffany Wade, Azmina Verjee, Rachel Sawyer, Sameer Mannick, Damian McCluskey, Nuha Yassin, Robin K S Phillips, Philip J Tozer, Christine Norton, and Ailsa L Hart. Burden of disease and adaptation to life in patients with crohn's perianal fistula: a qualitative exploration. *Health Qual. Life Outcomes*, 18(1):370, November 2020.

[173] Julian Panes, Walter Reinisch, Ewa Rupniewska, Shahnaz Khan, Joan Forns, Javaria Mona Khalid, Daniela Bojic, and Haridarshan Patel. Burden and outcomes for complex perianal fistulas in crohn's disease: Systematic review. *World J. Gastroenterol.*, 24(42):4821–4834, November 2018.

[174] G C Braithwaite, M J Lee, D Hind, and S R Brown. Prognostic factors affecting outcomes in fistulating perianal crohn's disease: a systematic review. *Tech. Coloproctol.*, 21(7):501–519, July 2017.

[175] Laurent Peyrin-Biroulet, Edward V Loftus, Jr, Jean-Frederic Colombel, and William J Sandborn. The natural history of adult crohn's disease in population-based cohorts. *Am. J. Gastroenterol.*, 105(2):289–297, February 2010.

[176] Michael Scharl, Gerhard Rogler, and Luc Biedermann. Fistulizing crohn's disease. *Clin. Transl. Gastroenterol.*, 8(7):e106, July 2017.

[177] Christoph Gasche, Jurgen Scholmerich, Jorn Brynskov, Geert D'Haens, Stephen B Hanauer, Jan E Irvine, Derek P Jewell, Daniel Rachmilewitz, David B Sachar, William J Sandborn, and Lloyd R Sutherland. A simple classification of crohn's disease: Report of the working party for the world congresses of gastroenterology, vienna 1998. *Inflamm. Bowel Dis.*, 6(1):8–15, February 2000.

[178] Tim W Eglinton, Murray L Barclay, Richard B Gearry, and Frank A Frizelle. The spectrum of perianal crohn's disease in a population-based cohort. *Dis. Colon Rectum*, 55(7):773–777, July 2012.

[179] Raghu Kalluri and Robert A Weinberg. The basics of epithelial-mesenchymal transition. *J. Clin. Invest.*, 119(6):1420–1428, June 2009.

[180] Michael Scharl, Sandra Frei, Theresa Pesch, Silvia Kellermeier, Joba Arikkat, Pascal Frei, Michael Fried, Achim Weber, Ekkehard Jehle, Anne Rühl, and Gerhard Rogler. Interleukin-13 and transforming growth factor β synergise in the pathogenesis of human intestinal fistulae. *Gut*, 62(1):63–72, January 2013.

[181] Frauke Bataille, Christian Rohrmeier, Richard Bates, Achim Weber, Florian Rieder, Julia Brenmoehl, Ulrike Strauch, Stefan Farkas, Alois Fürst, Ferdinand Hofstädter, Jürgen Schölmerich, Hans Herfarth, and Gerhard Rogler. Evidence for a role of epithelial mesenchymal transition during pathogenesis of fistulae in crohn's disease. *Inflamm. Bowel Dis.*, 14(11):1514–1527, November 2008.

[182] F Bataille, F Klebl, P Rümmele, J Schroeder, S Farkas, P-J Wild, A Fürst, F Hofstädter, J Schölmerich, H Herfarth, and G Rogler. Morphological characterisation of crohn's disease fistulae. *Gut*, 53(9):1314–1321, September 2004.

[183] Eddie Cano-Gamez and Gosia Trynka. From GWAS to function: Using functional genomics to identify the mechanisms underlying complex diseases. *Front. Genet.*, 11:424, May 2020.

[184] Manreet Kaur, Deepa Panikkath, Xiaofei Yan, Zhenqiu Liu, Dror Berel, Dalin Li, Eric A Vasiliauskas, Andrew Ippoliti, Marla Dubinsky, David Q Shih, Gil Y Melmed, Talin Haritunians, Phillip Fleshner, Stephan R Targan, and Dermot P B McGovern. Perianal crohn's disease is associated with distal colonic disease, stricturing disease behavior, IBD-associated serologies and genetic variation in the JAK-STAT pathway. *Inflamm. Bowel Dis.*, 22(4):862–869, April 2016.

[185] Xiaoyi Hu, Jing Li, Maorong Fu, Xia Zhao, and Wei Wang. The JAK/STAT signaling pathway: from bench to clinic. *Signal Transduct. Target. Ther.*, 6(1):402, November 2021.

[186] Farhad Seif, Majid Khoshmirsafa, Hossein Aazami, Monireh Mohsenzadegan, Gholamreza Sedighi, and Mohammadali Bahar. The role of JAK-STAT signaling pathway and its regulators in the fate of T helper cells. *Cell Commun. Signal.*, 15(1), December 2017.

[187] Marzieh Akhlaghpour, Talin Haritunians, Shyam K More, Lisa S Thomas, Dalton T Stamps, Shishir Dube, Dalin Li, Shaohong Yang, Carol J Landers, Emebet Mengesha, Hussein Hamade, Ramachandran Murali, Alka A Potdar, Andrea J Wolf, Gregory J Botwin, Michelle Khrom, International IBD Genetics Consortium, Ashwin N Ananthakrishnan, William A Faubion, Bana Jabri, Sergio A Lira, Rodney D Newberry, Robert S Sandler, R Balfour Sartor, Ramnik J Xavier, Steven R Brant, Judy H Cho, Richard H Duerr, Mark G Lazarev, John D Rioux, L Philip Schumm, Mark S Silverberg, Karen Zaghiyan, Phillip Fleshner, Gil Y Melmed, Eric A Vasiliauskas, Christina Ha, Shervin Rabizadeh, Gaurav Syal, Nirupama N Bonthala, David A Ziring, Stephan R Targan, Millie D Long, Dermot P B McGovern, and Kathrin S Michelsen. Genetic coding variant in complement factor B (CFB) is associated with increased risk for perianal crohn's disease and leads to impaired CFB cleavage and phagocytosis. *Gut*, April 2023.

[188] The IBD BioResource. The ibd bioresource protocol version 8, 2021.

[189] The IBD BioResource. The ibd bioresource questionnaire version 7, 2021.

[190] The IBD BioResource. What is the ibd bioresource?, 2022.

[191] The UK IBD Genetics Consortium. Uk ibd genetics consortium aims, 2023.

[192] T W Eglinton, R Roberts, J Pearson, M Barclay, T R Merriman, F A Frizelle, and R B Gearry. Clinical and genetic risk factors for perianal crohn's disease in a population-based cohort. *Am. J. Gastroenterol.*, 107(4):589–596, April 2012.

[193] A Latiano, O Palmieri, S Cucchiara, M Castro, R D'Incà, G Guariso, B Dallapiccola, M R Valvano, T Latiano, A Andriulli, and V Annese. Polymorphism of the IRGM gene might predispose to fistulizing behavior in crohn's disease. *Am. J. Gastroenterol.*, 104(1):110–116, January 2009.

[194] Philip J Tozer, Kevin Whelan, Robin K S Phillips, and Ailsa L Hart. Etiology of perianal crohn's disease: Role of genetic, microbiological, and immunological factors. *Inflamm. Bowel Dis.*, 15(10):1591–1598, October 2009.

[195] PLINK. Plink qc high ld regions, 2023. Accessed on 20/10/2023.

[196] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, October 2015.

[197] Ani Manichaikul, Josyf C Mychaleckyj, Stephen S Rich, Kathy Daly, Michèle Sale, and Wei-Min Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873, November 2010.

[198] Joelle Mbatchou, Leland Barnard, Joshua Backman, Anthony Marcketta, Jack A Kosmicki, Andrey Ziyatdinov, Christian Benner, Colm O'Dushlaine, Mathew Barber, Boris Boutkov, Lukas Habegger, Manuel Ferreira, Aris Baras, Jeffrey Reid, Goncalo Abecasis, Evan Maxwell, and Jonathan Marchini. Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.*, 53(7):1097–1103, July 2021.

[199] Marta Byrska-Bishop, Uday S Evani, Xuefang Zhao, Anna O Basile, Haley J Abel, Allison A Regier, André Corvelo, Wayne E Clarke, Rajeeva Musunuri, Kshithija Nagulapalli, Susan Fairley, Alexi Runnels, Lara Winterkorn, Ernesto Lowy, Human Genome Structural Variation Consortium, Paul Flicek, Soren Germer, Harrison Brand, Ira M Hall, Michael E Talkowski, Giuseppe Narzisi, and Michael C Zody. High-coverage whole-genome sequencing of the expanded 1000 genomes project cohort including 602 trios. *Cell*, 185(18):3426–3440.e19, September 2022.

[200] Takashi Shiina, Kazuyoshi Hosomichi, Hidetoshi Inoko, and Jerzy K Kulski. The HLA genomic loci map: expression, interaction, diversity and disease. *J. Hum. Genet.*, 54(1):15–39, January 2009.

[201] Meral Beksac, editor. *Bone marrow and stem cell transplantation*. Methods in molecular biology (Clifton, N.J.). Humana Press, New York, NY, 2 edition, January 2014.

[202] Paul I W de Bakker, Gil McVean, Pardis C Sabeti, Marcos M Miretti, Todd Green, Jonathan Marchini, Xiayi Ke, Alienke J Monsuur, Pamela Whittaker, Marcos Delgado, Jonathan Morrison, Angela Richardson, Emily C Walsh, Xiaojiang Gao, Luana Galver, John Hart, David A Hafler, Margaret Pericak-Vance, John A Todd, Mark J Daly, John Trowsdale, Cisca Wijmenga, Tim J Vyse, Stephan Beck, Sarah Shaw Murray, Mary Carrington, Simon Gregory, Panos Deloukas, and John D Rioux. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat. Genet.*, 38(10):1166–1172, October 2006.

[203] Alienke J Monsuur, Paul I W de Bakker, Alexandra Zhernakova, Dalila Pinto, Willem Verduijn, Jihane Romanos, Renata Auricchio, Ana Lopez, David A van Heel, J Bart A Crusius, and Cisca Wijmenga. Effective detection of human leukocyte antigen risk alleles in celiac disease using tag single nucleotide polymorphisms. *PLoS One*, 3(5):e2270, May 2008.

[204] Xiaoming Jia, Buhm Han, Suna Onengut-Gumuscu, Wei-Min Chen, Patrick J Concannon, Stephen S Rich, Soumya Raychaudhuri, and Paul I W de Bakker. Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS One*, 8(6):e64683, June 2013.

[205] X Zheng, J Shen, C Cox, J C Wakefield, M G Ehm, M R Nelson, and B S Weir. HIBAG–HLA genotype imputation with attribute bagging. *Pharmacogenomics J.*, 14(2):192–200, April 2014.

[206] Seungho Cook, Wanson Choi, Hyunjoon Lim, Yang Luo, Kunhee Kim, Xiaoming Jia, Soumya Raychaudhuri, and Buhm Han. Accurate imputation of human leukocyte antigens with CookHLA. *Nat. Commun.*, 12(1):1264, February 2021.

[207] Tatsuhiko Naito, Ken Suzuki, Jun Hirata, Yoichiro Kamatani, Koichi Matsuda, Tatsushi Toda, and Yukinori Okada. A deep learning method for HLA imputation and trans-ethnic MHC fine-mapping of type 1 diabetes. *Nat. Commun.*, 12(1):1639, March 2021.

[208] Tatsuhiko Naito and Yukinori Okada. HLA imputation and its application to genetic and molecular fine-mapping of the MHC region in autoimmune diseases. *Semin. Immunopathol.*, 44(1):15–28, January 2022.

[209] Xiuwen Zheng and Bruce S. Weir. Hibag - hla genotype imputation with attribute bagging, 01/03/2017. Accessed on 25/10/2023.

[210] Dana Duricova, Johan Burisch, Tine Jess, Corinne Gower-Rousseau, Peter L Lakatos, and ECCO-EpiCom. Age-related differences in presentation and course of inflammatory bowel disease: an update on the population-based literature. *J. Crohns. Colitis*, 8(11):1351–1361, November 2014.

[211] Miguel Regueiro and Houssam Mardini. Treatment of perianal fistulizing crohn's disease with infliximab alone or as an adjunct to exam under anesthesia with seton placement. *Inflamm. Bowel Dis.*, 9(2):98–103, March 2003.

[212] Paulo Gustavo Kotze, Idblan Carvalho de Albuquerque, André da Luz Moreira, Wanessa Bertrami Tonini, Marcia Olandoski, and Claudio Saddy Rodrigues Coy. Perianal complete remission with combined therapy (seton placement and anti-TNF agents) in crohn's disease: a brazilian multicenter observational study. *Arq. Gastroenterol.*, 51(4):284–289, October 2014.

[213] A Haennig, G Staumont, B Lepage, P Faure, L Alric, L Buscail, B Bournet, and J Moreau. The results of seton drainage combined with anti-TNFα therapy for anal fistula in crohn's disease. *Colorectal Dis.*, 17(4):311–319, April 2015.

[214] Wolfgang B Gaertner, Alejandra Decanini, Anders Mellgren, Ann C Lowry, Stanley M Goldberg, Robert D Madoff, and Michael P Spencer. Does infliximab infusion impact results of operative treatment for crohn's perianal fistulas? *Dis. Colon Rectum*, 50(11):1754–1760, November 2007.

[215] David A Schwartz, Laurent Peyrin-Biroulet, Karen Lasch, Shashi Adsul, and Silvio Danese. Efficacy and safety of 2 vedolizumab intravenous regimens for perianal fistulizing crohn's disease: ENTERPRISE study. *Clin. Gastroenterol. Hepatol.*, 20(5):1059–1067.e9, May 2022.

[216] J H Jones and J E Lennard-Jones. Corticosteroids and corticotrophin in the treatment of crohn's disease. *Gut*, 7(2):181–187, April 1966.

[217] Samuel Adegbola. Medical and surgical management of perianal crohn's disease. *Ann. Gastroenterol.*, 2018.

[218] Sang Hyoung Park, Satimai Aniwan, W Scott Harmsen, William J Tremaine, Amy L Lightner, William A Faubion, and Edward V Loftus. Update on the natural course of fistulizing perianal crohn's disease in a population-based cohort. *Inflamm. Bowel Dis.*, 25(6):1054–1060, May 2019.

[219] Charlène Brochard, Marie-Laure Rabilloud, Stéphanie Hamonic, Emma Bajeux, Maël Pagenault, Alain Dabadie, Agathe Gerfaud, Jean-François Viel, Isabelle Tron, Michel Robaszkiewicz, Jean-François Bretagne, Laurent Siproudhis, Guillaume Bouguen, and Groupe ABERMAD. Natural history of perianal crohn's disease: Long-term follow-up of a population-based cohort. *Clin. Gastroenterol. Hepatol.*, 20(2):e102–e110, February 2022.

[220] Annecarin Brückner, Katharina J Werkstetter, Jan de Laffolie, Claudia Wendt, Christine Prell, Tanja Weidenhausen, Klaus P Zimmer, and Sibylle Koletzko. Incidence and risk factors for perianal disease in pediatric crohn disease patients followed in CEDATA-GPGE registry. *J. Pediatr. Gastroenterol. Nutr.*, 66(1):73–78, January 2018.

[221] Kevin W A Göttgens, Steven F G Jeuring, Rosel Sturkenboom, Mariëlle J L Romberg-Camps, Liekele E Oostenbrug, Daisy M A E Jonkers, Laurents P S Stassen, Ad A M Masclee, Marieke J Pierik, and Stéphanie O Breukink. Time trends in the epidemiology and outcome of perianal fistulizing crohn's disease in a population-based cohort. *Eur. J. Gastroenterol. Hepatol.*, 29(5):595–601, May 2017.

[222] Lester Tsai, Jeffrey D McCurdy, Christopher Ma, Vipul Jairath, and Siddharth Singh. Epidemiology and natural history of perianal crohn's disease: A systematic review and meta-analysis of population-based cohorts. *Inflamm. Bowel Dis.*, 28(10):1477–1484, October 2022.

[223] Vasiliki Matzaraki, Vinod Kumar, Cisca Wijmenga, and Alexandra Zhernakova. The MHC locus and genetic susceptibility to autoimmune and infectious diseases. *Genome Biol.*, 18(1), December 2017.

[224] S G E Marsh, E D Albert, W F Bodmer, R E Bontrop, B Dupont, H A Erlich, M Fernández-Viña, D E Geraghty, R Holdsworth, C K Hurley, M Lau, K W Lee, B Mach, M Maiers, W R Mayr, C R Müller, P Parham, E W Petersdorf, T Sasazuki,

J L Strominger, A Svejgaard, P I Terasaki, J M Tiercy, and J Trowsdale. Nomenclature for factors of the HLA system, 2010. *Tissue Antigens*, 75(4):291–455, April 2010.

[225] Anthony Nolan Research Institute. Nomenclature for factors of the hla system, 20/09/2019. Accessed on 24/10/2023.

[226] Anthony Nolan Research Institute. Full list of class i proteins, 06/10/2023. Accessed on 24/10/2023.

[227] Manreet Kaur, Deepa Panikkath, Xiaofei Yan, Zhenqiu Liu, Dror Berel, Dalin Li, Eric A Vasiliauskas, Andrew Ippoliti, Marla Dubinsky, David Q Shih, Gil Y Melmed, Talin Haritunians, Phillip Fleshner, Stephan R Targan, and Dermot P B McGovern. Perianal crohn's disease is associated with distal colonic disease, stricturing disease behavior, IBD-associated serologies and genetic variation in the JAK-STAT pathway. *Inflamm. Bowel Dis.*, 22(4):862–869, April 2016.

[228] Yong Huang, Peter M Krein, Daniel A Muruve, and Brent W Winston. Complement factor B gene regulation: synergistic effects of TNF-alpha and IFN-gamma in macrophages. *J. Immunol.*, 169(5):2627–2635, September 2002.

[229] Kim Goring, Yong Huang, Connie Mowat, Caroline Léger, Teik-How Lim, Raza Zaheer, Dereck Mok, Lee Anne Tibbles, David Zygun, and Brent W Winston. Mechanisms of human complement factor B induction in sepsis and inhibition by activated protein C. *Am. J. Physiol. Cell Physiol.*, 296(5):C1140–50, May 2009.

[230] Philippe Goyette, International Inflammatory Bowel Disease Genetics Consortium, Gabrielle Boucher, Dermot Mallon, Eva Ellinghaus, Luke Jostins, Hailiang Huang, Stephan Ripke, Elena S Gusareva, Vito Annese, Stephen L Hauser, Jorge R Oksenberg, Ingo Thomsen, Stephen Leslie, Mark J Daly, Kristel Van Steen, Richard H Duerr, Jeffrey C Barrett, Dermot P B McGovern, L Philip Schumm, James A Traherne, Mary N Carrington, Vasilis Kosmoliaptsis, Tom H Karlsen, Andre Franke, and John D Rioux. 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.*, 47(2):172–179, February 2015.

[231] Biljana Klimenta, Hilada Nefic, Nenad Prodanovic, Radivoj Jadric, and Fatima Hukic. Association of biomarkers of inflammation and HLA-DRB1 gene locus with risk of developing rheumatoid arthritis in females. *Rheumatol. Int.*, 39(12):2147–2157, December 2019.

[232] Vincent van Drongelen and Joseph Holoshitz. Human leukocyte antigen–disease associations in rheumatoid arthritis. *Rheum. Dis. Clin. North Am.*, 43(3):363–376, August 2017.

[233] Tianju Wang, Chunmei Shen, Hengxin Li, Liping Chen, Sheng Liu, and Jun Qi. High resolution HLA-DRB1 analysis and shared molecular amino acid signature of DRβ1 molecules in occult hepatitis B infection. *BMC Immunol.*, 23(1):22, April 2022.

[234] Julio E Molineros, Loren L Looger, Kwangwoo Kim, Yukinori Okada, Chikashi Terao, Celi Sun, Xu-Jie Zhou, Prithvi Raj, Yuta Kochi, Akari Suzuki, Shuji Akizuki, Shuichiro Nakabo, So-Young Bang, Hye-Soon Lee, Young Mo Kang, Chang-Hee Suh,

Won Tae Chung, Yong-Beom Park, Jung-Yoon Choe, Seung-Cheol Shim, Shin-Seok Lee, Xiaoxia Zuo, Kazuhiko Yamamoto, Quan-Zhen Li, Nan Shen, Lauren L Porter, John B Harley, Kek Heng Chua, Hong Zhang, Edward K Wakeland, Betty P Tsao, Sang-Cheol Bae, and Swapan K Nath. Amino acid signatures of HLA Class-I and II molecules are strongly associated with SLE susceptibility and autoantibody production in eastern asians. *PLoS Genet.*, 15(4):e1008092, April 2019.

[235] Soumya Raychaudhuri, Cynthia Sandor, Eli A Stahl, Jan Freudenberg, Hye-Soon Lee, Xiaoming Jia, Lars Alfredsson, Leonid Padyukov, Lars Klareskog, Jane Worthington, Katherine A Siminovitch, Sang-Cheol Bae, Robert M Plenge, Peter K Gregersen, and Paul I W de Bakker. Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat. Genet.*, 44(3):291–296, January 2012.

[236] Andrew D Grotzinger, Mijke Rhemtulla, Ronald de Vlaming, Stuart J Ritchie, Travis T Mallard, W David Hill, Hill F Ip, Riccardo E Marioni, Andrew M McIntosh, Ian J Deary, Philipp D Koellinger, K Paige Harden, Michel G Nivard, and Elliot M Tucker-Drob. Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nat. Hum. Behav.*, 3(5):513–525, May 2019.

[237] Jennifer Sam Beaty and M Shashidharan. Anal fissure. *Clin. Colon Rectal Surg.*, 29(1):30–37, March 2016.

[238] Douglas W Mapel, Michael Schum, and Ann Von Worley. The epidemiology and treatment of anal fissures in a population-based cohort. *BMC Gastroenterol.*, 14(1):129, July 2014.

[239] Michael R B Keighley and Norman S Williams, editors. *Keighley & Williams' surgery of the anus, rectum and colon, fourth edition*. CRC Press, London, England, 4 edition, November 2018.

[240] P McDonald, A M Driscoll, and R J Nicholls. The anal dilator in the conservative management of acute anal fissures. *Br. J. Surg.*, 70(1):25–26, January 1983.

[241] M R Lock and J P Thomson. Fissure-in-ano: the initial management and prognosis. *Br. J. Surg.*, 64(5):355–358, May 1977.

[242] NICE NICE. Scenario: Management of an anal fissure, Apr 2021.

[243] Erica B Sneider and Justin A Maykel. Anal abscess and fistula. *Gastroenterol. Clin. North Am.*, 42(4):773–784, December 2013.

[244] Zubing Mei, Qingming Wang, Yi Zhang, Peng Liu, Maojun Ge, Peixin Du, Wei Yang, and Yazhou He. Risk factors for recurrence after anal fistula surgery: A meta-analysis. *Int. J. Surg.*, 69:153–164, September 2019.

[245] Carlo Zanotti, Carmen Martinez-Puente, Isabel Pascual, María Pascual, Dolores Herreros, and Damián García-Olmo. An assessment of the incidence of fistula-in-ano in four countries of the european union. *Int. J. Colorectal Dis.*, 22(12):1459–1462, December 2007.

[246] Chiara Eberspacher, Domenico Mascagni, Iulia Catalina Ferent, Enrico Coletta, Rossella Palma, Cristina Panetta, Anna Esposito, Stefano Arcieri, and Stefano Pontone. Mesenchymal stem cells for cryptoglandular anal fistula: Current state of art. *Front. Surg.*, 9:815504, February 2022.

[247] A G Parks. Pathogenesis and treatment of fistula-in-ano. *BMJ*, 1(5224):463–460, February 1961.

[248] Marcin Włodarczyk, Jakub Włodarczyk, Aleksandra Sobolewska-Włodarczyk, Radzisław Trzciński, Łukasz Dziki, and Jakub Fichna. Current concepts in the pathogenesis of cryptoglandular perianal fistula. *J. Int. Med. Res.*, 49(2):300060520986669, February 2021.

[249] Haig Dudukgian and Herand Abcarian. Why do we have so much trouble treating anal fistula? *World J. Gastroenterol.*, 17(28):3292–3296, July 2011.

[250] M J Johnston, G M Robertson, and F A Frizelle. Management of late complications of pelvic radiation in the rectum and anus. *Dis. Colon Rectum*, 46(2):247–259, February 2003.

[251] Roland Assi, Peter W Hashim, Vikram B Reddy, Hulda Einarsdottir, and Walter E Longo. Sexually transmitted infections of the anus and rectum. *World J. Gastroenterol.*, 20(41):15262–15268, November 2014.

[252] Jonathan Alastair Simpson, Ayan Banerjea, and John Howard Scholefield. Management of anal fistula. *BMJ*, 345(oct15 4):e6705, October 2012.

[253] Antonio Arroyo, Juan Pérez-Legaz, Pedro Moya, Laura Armañanzas, Javier Lacueva, Francisco Pérez-Vicente, Fernando Candela, and Rafael Calpena. Fistulotomy and sphincter reconstruction in the treatment of complex fistula-in-ano: long-term clinical and manometric results. *Ann. Surg.*, 255(5):935–939, May 2012.

[254] Bhupendra Kumar Jain, Kumar Vaibhaw, Pankaj Kumar Garg, Sanjay Gupta, and Debajyoti Mohanty. Comparison of a fistulectomy and a fistulotomy with marsupialization in the management of a simple anal fistula: a randomized, controlled pilot trial. *J. Korean Soc. Coloproctol.*, 28(2):78–82, April 2012.

[255] UK Biobank. Ukbb data showcase, 2023. Accessed on 20/10/2023.

[256] UK Biobank. Integrating electronic health records into the uk biobank resource version 1.0, January 2014. Accessed on 20/10/2023.

[257] FinnGen. Finngen project, 2023. Accessed on 20/10/2023.

[258] FinnGen. Finngen data freeze 9, 2023. Accessed on 20/10/2023.

[259] Centers for disease control and prevention. International classification of diseases, (icd-10-cm/pcs) transition - background, November 2015. Accessed on 20/10/2023.

[260] FinnGen. Data available in finngen, 2023. Accessed on 20/10/2023.

[261] Thermo Fisher Scientific. Axiom™ genotyping solution data analysis user guide, 18 April 2023. Accessed on 20/10/2023.

[262] Affymetrix. Uk biobank 500k samples genotyping data generation by the affymetrix research services laboratory, April 2015. Accessed on 20/10/2023.

[263] UK Biobank. Description of sample processing workflow and preparation of dna for genotyping, April 2015. Accessed on 20/10/2023.

[264] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. The UK biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, October 2018.

[265] Gavin Band and Jonathan Marchini. BGEN: a binary file format for imputed genotype and haplotype data. April 2018.

[266] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A R Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I W de Bakker, Mark J Daly, and Pak C Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, 81(3):559–575, September 2007.

[267] Heng Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993, November 2011.

[268] Brendan K Bulik-Sullivan, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.*, 47(3):291–295, March 2015.

[269] Brendan Bulik-Sullivan, Hilary K Finucane, Verneri Anttila, Alexander Gusev, Felix R Day, Po-Ru Loh, ReproGen Consortium, Psychiatric Genomics Consortium, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3, Laramie Duncan, John R B Perry, Nick Patterson, Elise B Robinson, Mark J Daly, Alkes L Price, and Benjamin M Neale. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.*, 47(11):1236–1241, November 2015.

[270] The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861, October 2007.

[271] Tenghao Zheng, David Ellinghaus, Simonas Juzenas, François Cossais, Greta Burmeister, Gabriele Mayr, Isabella Friis Jørgensen, Maris Teder-Laving, Anne Heidi Skogholt, Sisi Chen, Peter R Strege, Go Ito, Karina Banasik, Thomas Becker, Frank Bokelmann, Søren Brunak, Stephan Buch, Hartmut Clausnitzer, Christian Datz, DBDS Consortium, Frauke Degenhardt, Marek Doniec, Christian Erikstrup, Tõnu Esko, Michael Forster, Norbert Frey, Lars G Fritsche, Maiken Elvestad Gabrielsen, Tobias Gräßle, Andrea Gsur, Justus Gross, Jochen Hampe, Alexander Hendricks, Sebastian Hinz, Kristian Hveem, Johannes Jongen, Ralf Junker, Tom Hemming Karlsen, Georg Hemmrich-Stanisak, Wolfgang Kruis, Juozas Kupcinskas, Tilman Laubert, Philip C Rosenstiel, Christoph Röcken, Matthias Laudes, Fabian H Leendertz, Wolfgang Lieb, Verena Limperger, Nikolaos Margetis, Kerstin Mätz-Rensing, Christopher Georg

Németh, Eivind Ness-Jensen, Ulrike Nowak-Göttl, Anita Pandit, Ole Birger Pedersen, Hans Günter Peleikis, Kenneth Peuker, Cristina Leal Rodriguez, Malte Christoph Rühlemann, Bodo Schniewind, Martin Schulzky, Jurgita Skieceviciene, Jürgen Tepel, Laurent Thomas, Florian Uellendahl-Werth, Henrik Ullum, Ilka Vogel, Henry Volzke, Lorenzo von Fersen, Witigo von Schönfels, Brett Vanderwerff, Julia Wilking, Michael Wittig, Sebastian Zeissig, Myrko Zobel, Matthew Zawistowski, Vladimir Vacic, Olga Sazonova, Elizabeth S Noblin, 23andMe Research Team, Gianrico Farrugia, Arthur Beyder, Thilo Wedel, Volker Kahlke, Clemens Schafmayer, Mauro D'Amato, and Andre Franke. Genome-wide analysis of 944 133 individuals provides insights into the etiology of haemorrhoidal disease. *Gut*, 70(8):1538–1549, April 2021.

[272] American Society of Colon and Rectal Surgery. Abscess and fistula, 2023. Accessed on 21/10/2023.

[273] Manuela Marzo, Carla Felice, Daniela Pugliese, Gianluca Andrisani, Giammarco Mocci, Alessandro Armuzzi, and Luisa Guidi. Management of perianal fistulas in crohn's disease: an up-to-date review. *World J. Gastroenterol.*, 21(5):1394–1403, February 2015.

[274] American Society of Colon and Rectal Surgery. Diverticular disease expanded information, 2023. Accessed on 21/10/2023.

[275] Jacklyn N Hellwege, Jacob M Keaton, Ayush Giri, Xiaoyi Gao, Digna R Velez Edwards, and Todd L Edwards. Population stratification in genetic association studies. *Curr. Protoc. Hum. Genet.*, 95(1):1.22.1–1.22.23, October 2017.

[276] Peter Kraft, Eleftheria Zeggini, and John P A Ioannidis. Replication in genome-wide association studies. *Stat. Sci.*, 24(4):561–573, November 2009.

[277] Evangelos Evangelou and John P A Ioannidis. Meta-analysis methods for genome-wide association studies and beyond. *Nat. Rev. Genet.*, 14(6):379–389, June 2013.

[278] Center for Statistical Genetics. Metal documentation, December 2017. Accessed on 21/10/2023.

[279] Diogo M Ribeiro, Simone Rubinacci, Anna Ramisch, Robin J Hofmeister, Emmanouil T Dermitzakis, and Olivier Delaneau. The molecular basis, genetic control and pleiotropic effects of local gene co-expression. *Nat. Commun.*, 12(1):4842, August 2021.

[280] Rachel B Brem, Gaël Yvert, Rebecca Clinton, and Leonid Kruglyak. Genetic dissection of transcriptional regulation in budding yeast. *Science*, 296(5568):752–755, April 2002.

[281] Eric E Schadt, Stephanie A Monks, Thomas A Drake, Aldons J Lusis, Nam Che, Veronica Colinayo, Thomas G Ruff, Stephen B Milligan, John R Lamb, Guy Cavet, Peter S Linsley, Mao Mao, Roland B Stoughton, and Stephen H Friend. Genetics of gene expression surveyed in maize, mouse and man. *Nature*, 422(6929):297–302, March 2003.

[282] Jianan Tian, Mark P Keller, Aimee Teo Broman, Christina Kendziorski, Brian S Yandell, Alan D Attie, and Karl W Broman. The dissection of expression quantitative trait locus hotspots. *Genetics*, 202(4):1563–1574, April 2016.

[283] Sipko van Dam, Urmo Võsa, Adriaan van der Graaf, Lude Franke, and João Pedro de Magalhães. Gene co-expression analysis for functional classification and gene–disease predictions. *Brief. Bioinform.*, page bbw139, January 2017.

[284] Harm-Jan Westra, Marjolein J Peters, Tõnu Esko, Hanieh Yaghootkar, Claudia Schurmann, Johannes Kettunen, Mark W Christiansen, Benjamin P Fairfax, Katharina Schramm, Joseph E Powell, Alexandra Zhernakova, Daria V Zhernakova, Jan H Veldink, Leonard H Van den Berg, Juha Karjalainen, Sebo Withoff, André G Uitterlinden, Albert Hofman, Fernando Rivadeneira, Peter A C 't Hoen, Eva Reinmaa, Krista Fischer, Mari Nelis, Lili Milani, David Melzer, Luigi Ferrucci, Andrew B Singleton, Dena G Hernandez, Michael A Nalls, Georg Homuth, Matthias Nauck, Dörte Radke, Uwe Völker, Markus Perola, Veikko Salomaa, Jennifer Brody, Astrid Suchy-Dicey, Sina A Gharib, Daniel A Enquobahrie, Thomas Lumley, Grant W Montgomery, Seiko Makino, Holger Prokisch, Christian Herder, Michael Roden, Harald Grallert, Thomas Meitinger, Konstantin Strauch, Yang Li, Ritsert C Jansen, Peter M Visscher, Julian C Knight, Bruce M Psaty, Samuli Ripatti, Alexander Teumer, Timothy M Frayling, Andres Metspalu, Joyce B J van Meurs, and Lude Franke. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.*, 45(10):1238–1243, October 2013.

[285] Marc Gillespie, Bijay Jassal, Ralf Stephan, Marija Milacic, Karen Rothfels, Andrea Senff-Ribeiro, Johannes Griss, Cristoffer Sevilla, Lisa Matthews, Chuqiao Gong, Chuan Deng, Thawfeek Varusai, Eliot Ragueneau, Yusra Haider, Bruce May, Veronica Shamovsky, Joel Weiser, Timothy Brunson, Nasim Sanati, Liam Beckman, Xiang Shao, Antonio Fabregat, Konstantinos Sidiropoulos, Julieth Murillo, Guilherme Viteri, Justin Cook, Solomon Shorser, Gary Bader, Emek Demir, Chris Sander, Robin Haw, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The reactome pathway knowledgebase 2022. *Nucleic Acids Res.*, 50(D1):D687–D692, January 2022.

[286] Paul D Thomas, Dustin Ebert, Anushya Muruganujan, Tremayne Mushayahama, Laurent-Philippe Albou, and Huaiyu Mi. PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Sci.*, 31(1):8–22, January 2022.

[287] Mingchu Xu, Yajing Angela Xie, Hana Abouzeid, Christopher T Gordon, Alessia Fiorentino, Zixi Sun, Anna Lehman, Ihab S Osman, Rachayata Dharmat, Rosa Riveiro-Alvarez, Linda Bapst-Wicht, Darwin Babino, Gavin Arno, Virginia Busetto, Li Zhao, Hui Li, Miguel A Lopez-Martinez, Liliana F Azevedo, Laurence Hubert, Nikolas Pontikos, Aiden Eblimit, Isabel Lorda-Sanchez, Valeria Kheir, Vincent Plagnol, Myriam Oufadem, Zachry T Soens, Lizhu Yang, Christine Bole-Feysot, Rolph Pfundt, Nathalie Allaman-Pillet, Patrick Nitschké, Michael E Cheetham, Stanislas Lyonnet, Smriti A Agrawal, Huajin Li, Gaëtan Pinton, Michel Michaelides, Claude Besmond, Yumei Li, Zhisheng Yuan, Johannes von Lintig, Andrew R Webster, Hervé Le Hir, Peter Stoilov, UK Inherited Retinal Dystrophy Consortium, Jeanne Amiel, Alison J Hardcastle, Carmen Ayuso, Ruifang Sui, Rui Chen, Rando Allikmets, and Daniel F Schorderet. Mutations in the spliceosome component CWC27 cause retinal degeneration with or

without additional developmental anomalies. *Am. J. Hum. Genet.*, 100(4):592–604, April 2017.

[288] Nao Otomo, Shuji Mizumoto, Hsing-Fang Lu, Kazuki Takeda, Belinda Campos-Xavier, Lauréane Mittaz-Crettol, Long Guo, Kazuharu Takikawa, Masaya Nakamura, Shuhei Yamada, Morio Matsumoto, Kota Watanabe, and Shiro Ikegawa. Identification of novel LFNG mutations in spondylocostal dysostosis. *J. Hum. Genet.*, 64(3):261–264, March 2019.

[289] D B Sparrow, G Chapman, M A Wouters, N V Whittock, S Ellard, D Fatkin, P D Turnpenny, K Kusumi, D Sillence, and S L Dunwoodie. Mutation of the LUNATIC FRINGE gene in humans causes spondylocostal dysostosis with a severe vertebral phenotype. *Am. J. Hum. Genet.*, 78(1):28–37, January 2006.

[290] Katrin Serth, Karin Schuster-Gossler, Ralf Cordes, and Achim Gossler. Transcriptional oscillation of lunatic fringe is essential for somitogenesis. *Genes Dev.*, 17(7):912–925, April 2003.

[291] Ville E Korhonen, Seppo Helisalmi, Aleksi Jokinen, Ilari Jokinen, Juha-Matti Lehtola, Minna Oinas, Kimmo Lönnrot, Cecilia Avellan, Anna Kotkansalo, Janek Frantzen, Jaakko Rinne, Antti Ronkainen, Mikko Kauppinen, Antti Junkkari, Mikko Hiltunen, Hilkka Soininen, Mitja Kurki, Juha E Jääskeläinen, Anne M Koivisto, Hidenori Sato, Takeo Kato, Anne M Remes, Per Kristian Eide, and Ville Leinonen. Copy number loss in SFMBT1 is common among finnish and norwegian patients with iNPH. *Neurol. Genet.*, 4(6):e291, December 2018.

[292] Hidenori Sato, Yoshimi Takahashi, Luna Kimihira, Chifumi Iseki, Hajime Kato, Yuya Suzuki, Ryosuke Igari, Hiroyasu Sato, Shingo Koyama, Shigeki Arawaka, Toru Kawanami, Masakazu Miyajima, Naoyuki Samejima, Shinya Sato, Masahiro Kameda, Shinya Yamada, Daisuke Kita, Mitsunobu Kaijima, Isao Date, Yukihiko Sonoda, Takamasa Kayama, Nobumasa Kuwana, Hajime Arai, and Takeo Kato. A segmental copy number loss of the SFMBT1 gene is a genetic risk for shunt-responsive, idiopathic normal pressure hydrocephalus (iNPH): A case-control study. *PLoS One*, 11(11):e0166615, November 2016.

[293] Takeo Kato, Hidenori Sato, Mitsuru Emi, Tomomi Seino, Shigeki Arawaka, Chifumi Iseki, Yoshimi Takahashi, Manabu Wada, and Toru Kawanami. Segmental copy number loss of SFMBT1 gene in elderly individuals with ventriculomegaly: a community-based study. *Intern. Med.*, 50(4):297–303, February 2011.

[294] Shuibin Lin, Huangxuan Shen, Jian-Liang Li, Shaojun Tang, Yumei Gu, Zirong Chen, Chengbin Hu, Judd C Rice, Jianrong Lu, and Lizi Wu. Proteomic and functional analyses reveal the role of chromatin reader SFMBT1 in regulating epigenetic silencing and the myogenic gene program. *J. Biol. Chem.*, 288(9):6238–6247, March 2013.

[295] Marcin Włodarczyk, Jakub Włodarczyk, Aleksandra Sobolewska-Włodarczyk, Radzisław Trzciński, Łukasz Dziki, and Jakub Fichna. Current concepts in the pathogenesis of cryptoglandular perianal fistula. *J. Int. Med. Res.*, 49(2):300060520986669, February 2021.

[296] Eleanor Sanderson, M Maria Glymour, Michael V Holmes, Hyunseung Kang, Jean Morrison, Marcus R Munafò, Tom Palmer, C Mary Schooling, Chris Wallace, Qingyuan Zhao, and George Davey Smith. Mendelian randomization. *Nat. Rev. Methods Primers*, 2(1), February 2022.

[297] Hilary K Finucane, Yakir A Reshef, Verneri Anttila, Kamil Slowikowski, Alexander Gusev, Andrea Byrnes, Steven Gazal, Po-Ru Loh, Caleb Lareau, Noam Shoresh, Giulio Genovese, Arpiar Saunders, Evan Macosko, Samuela Pollack, John R B Perry, Jason D Buenrostro, Bradley E Bernstein, Soumya Raychaudhuri, Steven McCarroll, Benjamin M Neale, Alkes L Price, and The Brainstorm Consortium. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.*, 50(4):621–629, April 2018.

[298] Camille Moore. Application note: Mas-seq for single-cell isoform sequencing, Jan 2023. Accessed on 14/11/2023.

[299] Miten Jain, Robin Abu-Shumays, Hugh E Olsen, and Mark Akeson. Advances in nanopore direct RNA sequencing. *Nat. Methods*, 19(10):1160–1164, October 2022.

[300] Haiyong Han. RNA interference to knock down gene expression. In *Methods in Molecular Biology*, Methods in molecular biology (Clifton, N.J.), pages 293–302. Springer New York, New York, NY, 2018.

[301] Akiko Seki and Sascha Rutz. Optimized RNP transfection for highly efficient CRISPR/Cas9-mediated gene knockout in primary T cells. *J. Exp. Med.*, 215(3):985–997, March 2018.

[302] Megan D Schertzer, Andrew Stirn, Keren Isaev, Laura Pereira, Anjali Das, Claire Harbison, Stella H Park, Hans-Hermann Wessels, Neville E Sanjana, and David A Knowles. Cas13d-mediated isoform-specific RNA knockdown with a unified computational and experimental toolbox. *bioRxivorg*, September 2023.

[303] James D Thomas, Jacob T Polaski, Qing Feng, Emma J De Neef, Emma R Hoppe, Maria V McSharry, Joseph Pangallo, Austin M Gabel, Andrea E Belleville, Jacqueline Watson, Naomi T Nkinsi, Alice H Berger, and Robert K Bradley. RNA isoform screens uncover the essentiality and tumor-suppressor activity of ultraconserved poison exons. *Nat. Genet.*, 52(1):84–94, January 2020.

[304] Thomas Gonatopoulos-Pournatzis, Mingkun Wu, Ulrich Braunschweig, Jonathan Roth, Hong Han, Andrew J Best, Bushra Raj, Michael Aregger, Dave O'Hanlon, Jonathan D Ellis, John A Calarco, Jason Moffat, Anne-Claude Gingras, and Benjamin J Blencowe. Genome-wide CRISPR-Cas9 interrogation of splicing networks reveals a mechanism for recognition of autism-misregulated neuronal microexons. *Mol. Cell*, 72(3):510–524.e12, November 2018.

[305] Julia K Nussbacher, Ranjan Batra, Clotilde Lagier-Tourenne, and Gene W Yeo. RNA-binding proteins in neurodegeneration: Seq and you shall receive. *Trends Neurosci.*, 38(4):226–236, April 2015.

[306] Alfredo Castello, Bernd Fischer, Matthias W Hentze, and Thomas Preiss. RNA-binding proteins in mendelian disease. *Trends Genet.*, 29(5):318–327, May 2013.

[307] Yang Wang, Meng Ma, Xinshu Xiao, and Zefeng Wang. Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules. *Nat. Struct. Mol. Biol.*, 19(10):1044–1052, October 2012.

[308] Yang Wang, Xinshu Xiao, Jianming Zhang, Rajarshi Choudhury, Alex Robertson, Kai Li, Meng Ma, Christopher B Burge, and Zefeng Wang. A complex network of factors with overlapping affinities represses splicing through intronic elements. *Nat. Struct. Mol. Biol.*, 20(1):36–45, January 2013.

[309] Eric L Van Nostrand, Gabriel A Pratt, Alexander A Shishkin, Chelsea Gelboin-Burkhart, Mark Y Fang, Balaji Sundararaman, Steven M Blue, Thai B Nguyen, Christine Surka, Keri Elkins, Rebecca Stanton, Frank Rigo, Mitchell Guttman, and Gene W Yeo. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods*, 13(6):508–514, June 2016.

[310] Dirk H Ostareck and Antje Ostareck-Lederer. RNA-binding proteins in the control of LPS-induced macrophage response. *Front. Genet.*, 10:31, February 2019.

[311] Lan Kang, Xiang Zhang, Liangliang Ji, Tiantian Kou, Sinead M Smith, Baohong Zhao, Xiaohuan Guo, Inés Pineda-Torra, Li Wu, and Xiaoyu Hu. The colonic macrophage transcription factor RBP-J orchestrates intestinal immunity against bacterial pathogens. *J. Exp. Med.*, 217(4), April 2020.

[312] Stacey D Wagner, Adam J Struck, Riti Gupta, Dylan R Farnsworth, Amy E Mahady, Katy Eichinger, Charles A Thornton, Eric T Wang, and J Andrew Berglund. Dose-dependent regulation of alternative splicing by MBNL proteins reveals biomarkers for myotonic dystrophy. *PLoS Genet.*, 12(9):e1006316, September 2016.

[313] Matteo Cereda, Uberto Pozzoli, Gregor Rot, Peter Juvan, Anthony Schweitzer, Tyson Clark, and Jernej Ule. RNAmotifs: prediction of multivalent RNA motifs that control alternative splicing. *Genome Biol*, 15(1):R20, 2014.

[314] Andrew J Clark and Mark E Davis. Increased brain uptake of targeted nanoparticles by adding an acid-cleavable linkage between transferrin and the nanoparticle core. *Proc. Natl. Acad. Sci. U. S. A.*, 112(40):12486–12491, October 2015.

[315] Cornelia Lorenzer, Mehrdad Dirin, Anna-Maria Winkler, Volker Baumann, and Johannes Winkler. Going beyond the liver: progress and challenges of targeted delivery of siRNA therapeutics. *J. Control. Release*, 203:1–15, April 2015.

[316] Jung Soo Suk, Qingguo Xu, Namho Kim, Justin Hanes, and Laura M Ensign. PEGylation as a strategy for improving nanoparticle-based drug and gene delivery. *Adv. Drug Deliv. Rev.*, 99(Pt A):28–51, April 2016.

[317] Karishma Dhuri, Clara Bechtold, Elias Quijano, Ha Pham, Anisha Gupta, Ajit Vikram, and Raman Bahal. Antisense oligonucleotides: An emerging area in drug discovery and development. *J. Clin. Med.*, 9(6):2004, June 2020.

[318] Annemieke Aartsma-Rus. FDA approval of nusinersen for spinal muscular atrophy makes 2016 the year of splice modulating oligonucleotides. *Nucleic Acid Ther.*, 27(2):67–69, April 2017.

[319] Mallory A Havens and Michelle L Hastings. Splice-switching antisense oligonucleotides as therapeutic drugs. *Nucleic Acids Res.*, 44(14):6549–6563, August 2016.

[320] Suvi Rk Hokkanen, Naomi Boxall, Javaria Mona Khalid, Dimitri Bennett, and Haridarshan Patel. Prevalence of anal fistula in the united kingdom. *World J. Clin. Cases*, 7(14):1795–1804, July 2019.

[321] Damián García-Olmo, Gert Van Assche, Ignacio Tagarro, Mary Carmen Diez, Marie Paule Richard, Javaria Mona Khalid, Marc van Dijk, Dimitri Bennett, Suvi R K Hokkanen, and Julián Panés. Prevalence of anal fistulas in europe: Systematic literature reviews and population-based database analysis. *Adv. Ther.*, 36(12):3503–3518, December 2019.

[322] Anna Fry, Thomas J Littlejohns, Cathie Sudlow, Nicola Doherty, Ligia Adamska, Tim Sprosen, Rory Collins, and Naomi E Allen. Comparison of sociodemographic and health-related characteristics of UK biobank participants with those of the general population. *Am. J. Epidemiol.*, 186(9):1026–1034, November 2017.

[323] Xuezhong Zhou, Lei Lei, Jun Liu, Arda Halu, Yingying Zhang, Bing Li, Zhili Guo, Guangming Liu, Changkai Sun, Joseph Loscalzo, Amitabh Sharma, and Zhong Wang. A systems approach to refine disease taxonomy by integrating phenotypic and molecular networks. *EBioMedicine*, 31:79–91, May 2018.

[324] Richelle J F Felt-Bersma and Joep F Bartelsman. Haemorrhoids, rectal prolapse, anal fissure, peri-anal fistulae and sexually transmitted diseases. *Best Pract. Res. Clin. Gastroenterol.*, 23(4):575–592, 2009.

[325] Amy E Foxx-Orenstein, Sarah B Umar, and Michael D Crowell. Common anorectal disorders. *Gastroenterol. Hepatol. (N. Y.)*, 10(5):294–301, May 2014.

[326] Andrew D Grotzinger, Travis T Mallard, Wonuola A Akingbuwa, Hill F Ip, Mark J Adams, Cathryn M Lewis, Andrew M McIntosh, Jakob Grove, Søren Dalsgaard, Klaus-Peter Lesch, Nora Strom, Sandra M Meier, Manuel Mattheisen, Anders D Børglum, Ole Mors, Gerome Breen, iPSYCH, Tourette Syndrome and Obsessive Compulsive Disorder Working Group of the Psychiatric Genetics Consortium, Bipolar Disorder Working Group of the Psychiatric Genetics Consortium, Major Depressive Disorder Working Group of the Psychiatric Genetics Consortium, Schizophrenia Working Group of the Psychiatric Genetics Consortium, Phil H Lee, Kenneth S Kendler, Jordan W Smoller, Elliot M Tucker-Drob, and Michel G Nivard. Genetic architecture of 11 major psychiatric disorders at biobehavioral, functional genomic and molecular genetic levels of analysis. *Nat. Genet.*, 54(5):548–559, May 2022.

[327] Prescott Deininger. Alu elements: know the SINEs. *Genome Biol.*, 12(12):236, December 2011.

[328] Kathi Zarnack, Julian König, Mojca Tajnik, Iñigo Martincorena, Sebastian Eustermann, Isabelle Stévant, Alejandro Reyes, Simon Anders, Nicholas M Luscombe, and Jernej Ule. Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of alu elements. *Cell*, 152(3):453–466, January 2013.

[329] Yi Xing and Christopher Lee. Alternative splicing and RNA selection pressure–evolutionary consequences for eukaryotic genomes. *Nat. Rev. Genet.*, 7(7):499–509, July 2006.

[330] Larisa Fedorova. *Genetica*, 118(2/3):123–131, 2003.