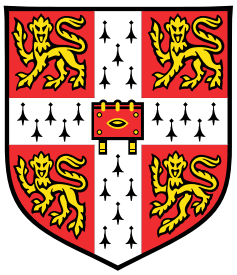
My PhD Thesis

My PhD subtitle



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 ...

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 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Omar El Garwany
November 2023

Acknowledgements

And I would like to acknowledge ...

Abstract

This is where you write your abstract ...

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 *Naure* 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 is not clear between an exposure and an outcome.

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 variants do not suffer from the same problems of observational data. Genetic variants are rarely confounded by social, behavioural or environmental factors. Moreover, genetic variants are determined at conception and do not therefore suffer from reverse causation in the same way that observed 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 unique 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 confounding and 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 genetic variants that are enriched in disease cases compared to healthy controls. These efforts have revealed that disease-associated genetic variants are significantly enriched in non-coding sequences such as enhancers, open chromatin regions, and chromatin markers [4–6].

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 is understanding the molecular effects of disease-associated variants. At the phenotype end is to understand how genetic variation predisposes to various disease subphenotypes. In this context, the aim of this thesis it 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 contexts. At the disease subphenotype level, a better understanding of the genetic determinants of disease subphenotypes will help us explain the heterogeneity of disease manifestations in complex diseases.

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 [7]. This has made the interpretation of their downstream functional effects difficult. To help bridging 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 start by gene expression and end with translation to effector proteins and different post-translation modifications. Moreover, gene expression is regulated at the DNA level by various epigenetic modifiers such as differential chromatin accessibility and histone marker modifications. 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 has also been shown to improve the 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, overl 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]. But 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 organimal 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 processess. *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* ensure 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 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 [40]. 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 to over 12 million transcripts in the latest RefSeq release (September 2023; [41]).

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 [42]. Others have shown that the majority of genes have a single dominant protein isoform in most tissues [43, 44]. 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 [45]. Similar collaborative efforts that catalogue

splice variants in an unbiased manner are warranted given the central role of AS in human health and disease.

1.2.3 Genetic regulation of alternative splicing

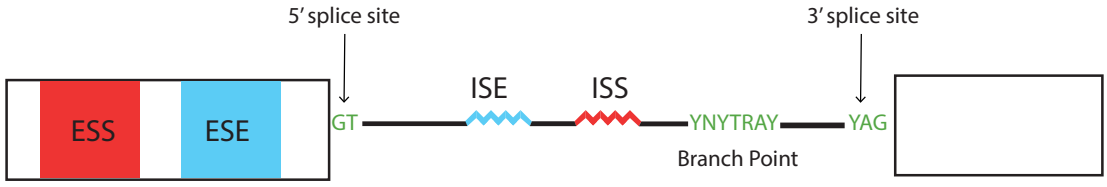


Fig. 1.1 Cis-acting splicing motifs in eukaryotes shown in an exon-intron-exon junction. The accptor and donor dinucleoties are indicated at the 5' and 3' ends of the intron. ESS and ESE = exonic splicing enhancers and silencers. ISE and ISS = intron splicing enhancers and silencers. Y = pyrimidines.

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. Despite the apparent complexity of the splicing code [46], 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 [47]). 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 [48]. The disruption of this tight regulation underpins several diseases. Spinal muscular atrophy, a debilitating motor neuron disease, is caused by the skipping of exon7 in *SMN1*. Exon 7 skipping is caused by a single nucleotide substition that alters the ESE sequence and results in a non-functional *SMN1* protein isoform [49].

1.2.4 Technological limitations

Several reasons 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 thir 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 successfuly 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.

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

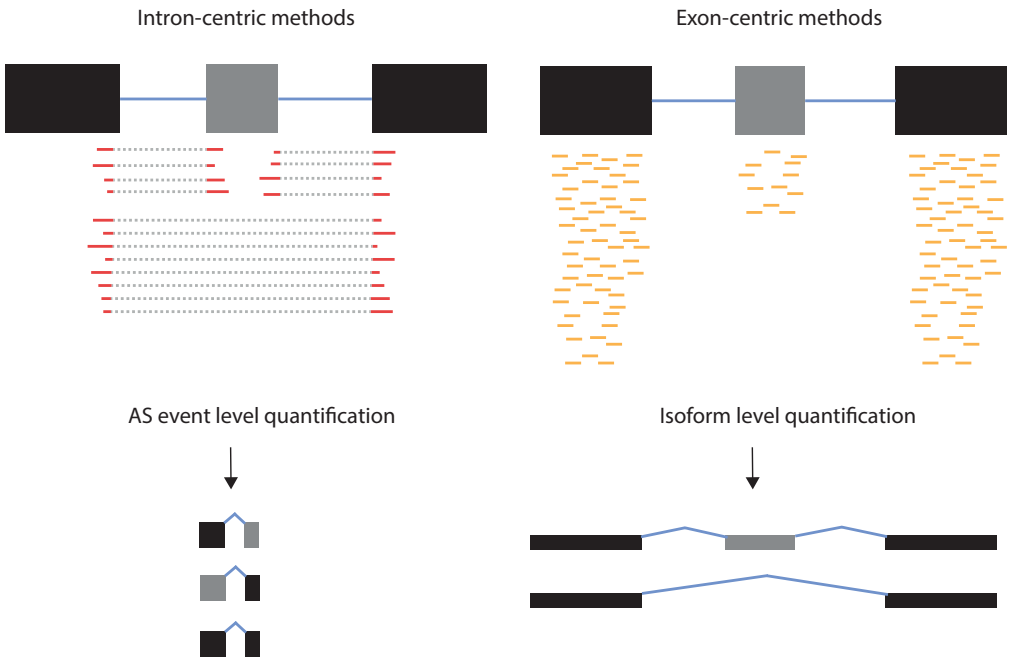


Fig. 1.2

reads, intron-centric methods use observed split reads at each splice junction to directly quantify local AS events (Figure 1.2). The obivious advantage of intron-centric method 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 much 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 undrirected 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).

General outline of QTL mapping

QTL mapping pipelines are relatively well-established. Typicially, 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 qunatitative 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. This step ensures 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 interepeted 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 moecular feature has equal variance across all samples in a study, satisfying the homoskedasticity condition. Second, an inverse normal transformation ensures 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 features, thousands of genetic variants are usually tested. Compared to GWASes where all variants are tested genome-wide, the number of tests in QTL mapping performed 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 feature and genotype values are permuted hundreds or even thousands of time and the association test is 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 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.

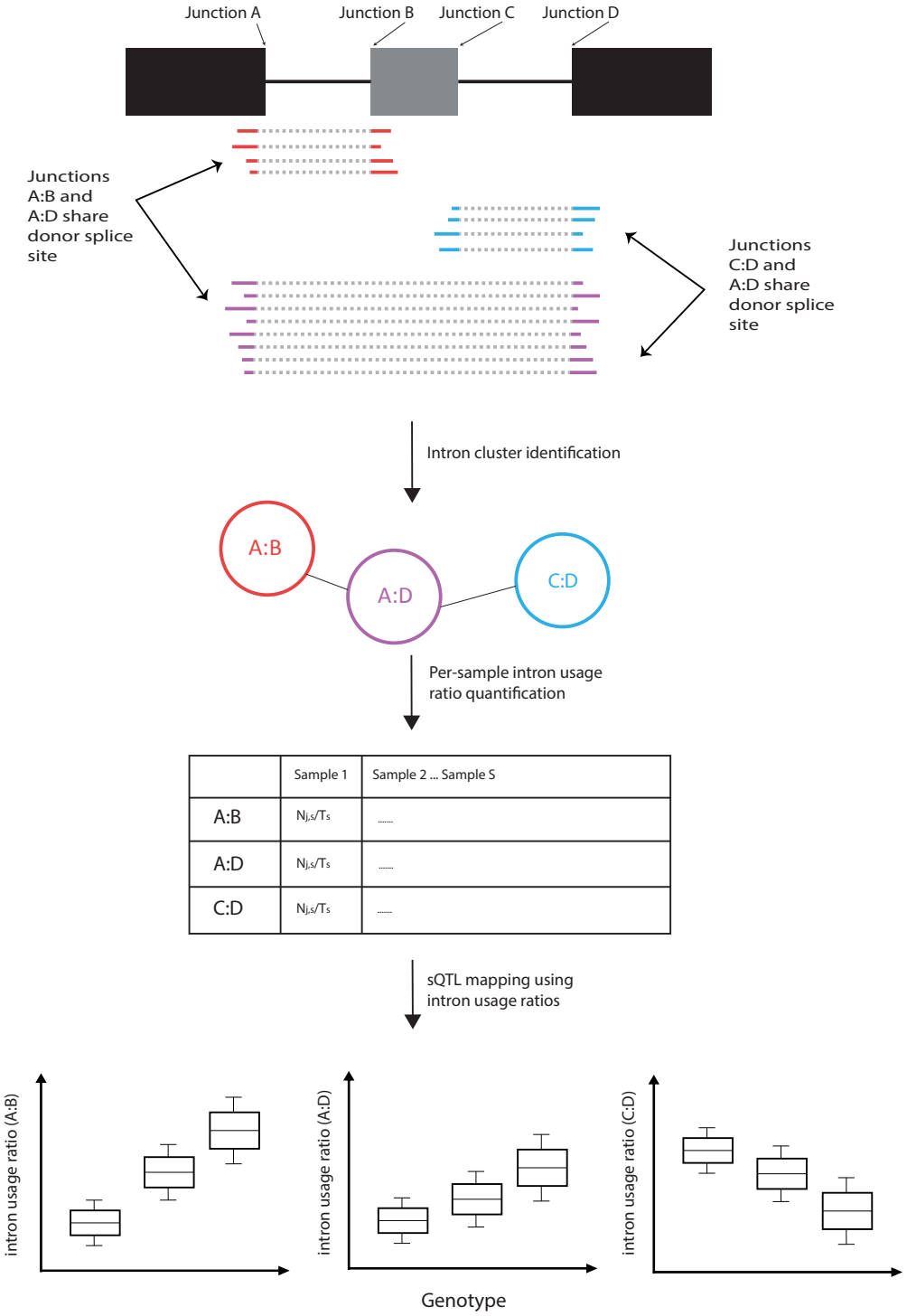


Fig. 1.3 Conceptual outline of sQTL mapping using intron usage ratios as quantitative traits. Intron clusters are identified from all pooled samples in a study. Quantification is then performed for each intron per sample. For a sample S, and an intron cluster with total number of reads T_S, intron usage 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 strucuture 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 was not enough statistical power to suggest heterogeneity of effect sizes. Additionally, when the 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 showing 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 disequillibrium 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 understand gene regulation, a major objective of QTL studies is to integrate QTL effects with GWAS data. 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]. Threfore, 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 disequillibrium with the true causal variant. Colocalisation methods leverages the linkage disequillbrium 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 studies being compared are derived from population with the same ancestry, which is an important consideration when comparing two association signals. Additionally, standard colocalisation approaches only test the hypothesis that a *single* shared variant underpins the two association signals. Many QTL studies have shown that for many genes secondary and even tertiary association signals are discovered for several genes, 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 idenitfy 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 [68].

1.3 Part II Introduction

1.3.1 Genome-wide association studies

Complex disease risk is determined by a multitude of genetic and environmental factors. Over the last 16 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) has 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 determinant 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 genetic analyses. The continuous growth of sample sizes has increased our ability to detect genome-wide significant loci associated with disease risk. These efforts have also revelaed the extensively polygenic nature of most complex diseases, whereby several genetic loci increase or decrease disease risk with small effect sizes. The complexity of the genetic architecture of most common disease has made it more challenging to draw biological insights from GWAS results. Although most GWAS results were initially puzzling, over the last few years massive GWASes have revealed biological insights about common diseases [69, 70]. This increased understanding was facilitated by the availability of functional genomic datasets as well as methodolgical 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 affects. 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 cotribution to IBD [71].

IBD is broadly classified into two broad categories based on radiological, clinical and endoscopic features: ulcerative colitis (UC) and Crohn's disease (CD). The two classes show differences in terms of disease behaviour and location, clinical manifestatios and prognosis. CD can affect any part of the GI tract from mouth to anus 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. UC usually starts near the rectum and diffuses proximally to different parts of the colon. Unlike CD, UC inflammation occurs in a continuous manner, and is often characterised by chronic mucosal inflammation and leukocyte infiltration [72]. However, not all IBD cases fall into these distinct categories, and approximately 6-13% are classified as IBD unclassified (IBDU) [73].

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 [74]. 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 santiation, 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 UC and decreasing risk of CD. However, the mechanism of this paradoxical association between smoking and IBD is not completely understood [75]. Other non-dietary factors include oral contraceptive pill intake, which was shown to increase both

CD and UC risk [76], and appendectomy which was associated with reduced UC risk [77].

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 [78]. Similarly, alcohol and coffee consumption were not conclusively linked to IBD development or progression [79]. However, obesity has been shown to independently worsen IBD behaviour and increase likelihood of relapse [80]. 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 [81].

1.3.3 Crohn's disease genetics

The genetic component of CD has been recognised for over 70 years via family studies on monozygotic and dizygotic twins. Family studies have shown that monozygotic twins are more likely to co-inherit CD comapred to dizygotic twins, often with similar disease behaviour, location and progression [82]. Over the last decade, several GWASes have identified over 250 loci associated with CD susceptibility [83–86]. The largest CD GWAS studies have focussed on discovering both common and rare genetic variants associated with CD susceptibility. These studies have revealed several key mechanisms in the pathogenesis of CD including autophagy, host-microbe interactions, intestinal innate immune response, and impaired epithelial barrier function [87, 83]. These pathways seem to converge on a CD pathogenesis model whereby impaired intestinal permeability, leads to microbial infilatration 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 CD susceptibility.

Fewer GWAS studies have dissected other clinical aspects of CD. 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 [88]. Understanding the genetic determinants of the different clinical aspects of CD is therefore crucial for a more nuanced biological insight into what drives disease course.

GWASes of disease subphenotypes and progression have generally lagged behind susceptibility GWASes, due to the difficulty of obtaining deep phenotypic or longitudinal data. It has been previously suggested that the same genetic variants driving both disease susceptibility and disease subphenotypes. However, evidence in relatively smaller subphenotype GWASes suggests that the genetic variants that underpin disease susceptibility and disease subphenotype may be distinct [89, 90]. Both paradigms raise interesting questions about the genetic architecture of disease susceptibility and subphenotypes. Under the former paradigm, it will be particularly important to understand the relationship between the effect of each variant on suscpetibility and subphenotype risks. For example, for a given CD-associated variant, is the susceptibility risk truly driven by subphenotype risk? As subphenotype GWASes become more commonplace, it will be particularly interesting to compare the effects sizes of each variant on both susceptibility and subphenotypes. This may lead to better stratification of disease risk based on distinct subphenotype risk profiles. Under the latter paradigm, it is important to understand what are the distinct biological pathways involved in disease subphenotypes? Do they interact with disease susceptibility pathways? For example, the etiology of fistulising CD has been hypothesised to start as an epithelial-to-mesenchymal transformation, whereby stationary epithelial cell gain migratory features. Is this transformation dependent on the impaired intestinal barrier and subsequent immune activation that likely underpins CD susceptibility?

Subphenotype GWASes

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] 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.

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

[42] 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.

[43] 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.

[44] 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.

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

[46] 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.

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

[48] 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.

[49] 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.

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

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

[76] 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.

[77] 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.

[78] 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.

[79] 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.

[80] 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.

[81] 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.

[82] 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.

[83] 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.

[84] 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.

[85] 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.

[86] 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.

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

[88] 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.

[89] 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.

[90] 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.