A multivariate approach to study the genetic determinants of phenotypic traits

Diego Garrido Martín

TESI DOCTORAL UPF / 2019

THESIS SUPERVISORS

Roderic Guigó Serra & Miquel Calvo Llorca

DEPARTMENT OF BIOINFORMATICS AND GENOMICS AT
CENTRE FOR GENOMIC REGULATION (CRG)

Universitat Pompeu Fabra Barcelona

CRG Centre for Genomic Regulation

Abstract

We have developed an efficient and reproducible pipeline for the identification of genetic variants affecting splicing (splicing quantitative trait loci or sQTLs), based on an approach that captures the intrinsically multivariate nature of this phenomenon. We employed it to study the multi-tissue transcriptome GTEx dataset, generating a comprehensive catalogue of sQTLs in the human genome. Downstream analyses of this catalogue provide novel insights into the mechanisms underlying alternative splicing regulation and its contribution to human complex traits and diseases. To facilitate the visualization of splicing events in GTEx and other large-scale RNA-seq studies, we developed a software to generate sashimi plots, which supports the aggregated representation of hundreds of samples. Given the growing interest in efficient methods to identify genetic effects on multiple traits, we extended the statistical framework employed for sQTL mapping (Anderson test) to accommodate any quantitative multivariate phenotype and experimental design. We derived the limiting distribution of the test statistic, allowing to compute asymptotic p values. We further demonstrated the advantages and applicability of our approach to GWAS and QTL mapping analyses using simulated and real datasets.

Resumen

Hemos desarrollado un método computacional eficiente y reproducible, que permite la identificación de variantes genéticas que afectan al *splicing* (*splicing quantitative trait loci* o sQTLs), y que es capaz de capturar la naturaleza multivariante de este fenómeno. Lo hemos empleado para estudiar el conjunto de datos GTEx, que contiene información sobre el transcriptoma en múltiples tejidos, generando un catálogo completo de sQTLs en el genoma humano. El análisis de dicho catálogo proporciona nuevos conocimientos sobre los mecanismos que subyacen a la regulación del *splicing* alternativo, así como sobre su contribución a los rasgos complejos y enfermedades humanas. Con el objetivo de facilitar la visualización de eventos de *splicing* en GTEx y otros estudios de secuenciación de ARN a gran escala, hemos desarrollado un software para generar gráficos de tipo *sashimi*, que permite la representación agregada de cientos de muestras. En vista del creciente interés por métodos capaces de analizar efectos genéticos en múltiples rasgos de manera eficiente, hemos extendido el marco estadístico empleado para la identificación de sQTLs (test de Anderson) para acomodar cualquier fenotipo multivariante cuantitativo y diseño experimental. Hemos derivado la distribución límite del estadístico, lo que nos permite calcular *p* valores asintóticos. Además, demostramos las ventajas y la aplicabilidad de nuestro método en GWAS y análisis de QTLs, empleando conjuntos de datos tanto simulados como reales.

Preface

In the era of precision medicine, it is crucial to identify the genetic determinants of human complex traits and diseases. Genome-wide association studies (GWAS) have greatly contributed to this task, by enabling the discovery of tens of thousands of statistical associations between genetic variants and human phenotypes. However, in most of the cases the causal variants, the target genes and the biological mechanisms through which they act remain largely unknown. Hence, characterizing the impact of regulatory variation on molecular phenotypes, along the path that goes from DNA to proteins, is essential to shed light upon the underpinnings of disease susceptibility. Albeit the genetic effects on transcriptional and epigenetic regulation are considered major drivers of phenotypic variability, the relevance of genetic variants affecting RNA splicing (i.e. splicing quantitative trait loci or sQTLs) has only recently been acknowledged.

In this context, the Genotype-Tissue Expression (GTEx) Project has emerged as an unprecedented resource, with RNA sequencing (RNA-seq) data available across multiple tissues in a large cohort of genotyped individuals. By leveraging this dataset, we have generated the most comprehensive catalogue to date of sQTLs in the human genome (Chapter 1). Notably, to capture the strongly correlated nature of the alternative transcript isoforms that can arise from a given gene, we model alternative splicing as a multivariate outcome (Chapters 1 and 3). Our analyses revealed that sQTLs tend to be shared across multiple tissues and target global splicing patterns, rather than individual splicing events. In addition, a substantial fraction of sQTLs also affects gene expression, although not always of the same gene. This reflects the tight association between splicing and transcription, while uncovering unexpected complexity underlying the regulation of both processes. Furthermore, we found stronger regulation of post-transcriptional compared to co-transcriptional splicing (Chapter 1). We observed that genetic effects on splicing are not restricted to splice sites, since many sQTLs act as modifiers of RNA-binding protein (RBP) binding. Moreover, we show that sQTLs can have a phenotypic impact comparable or even stronger than variants affecting expression, in particular those altering RBP binding sites (Chapter 1).

The study of alternative splicing often requires the efficient visualization of splicing events from RNA-seq. In the context of sQTL analyses, for instance, it is useful to illustrate and compare the splicing patterns of a gene between different genotype groups. The most common representation, known as *sashimi* plot, displays the read coverage plus the support of each splicing junction in a given genomic region (Chapter 2). However, currently available implementations present several limitations which narrow their applicability (e.g. annotation-dependence, inefficiency, poor visualization when long introns are present, etc.). Most importantly, they represent each RNA-seq experiment on a separate line, and this hinders the comparison of more than a few samples precluding their usage in large datasets. Aiming to improve visualization for our splicing analyses using GTEx data, we have developed a fast command-line implementation of the *sashimi* plot (Chapter 2), which solves many of the current flaws and presents novel features that enhance visualization, supporting the aggregated representation of hundreds of samples.

Our work with multivariate vectors of proportions –relative isoform abundances– in the context of sQTL mapping (Chapter 1), sparked our interest in extending the non-parametric statistical framework employed for association testing (Anderson test), in order to accommodate any quantitative multivariate phenotype and experimental design. Certainly, the increasing availability of human phenotypic data, both at organismic and molecular level, requires methods capable of leveraging multiple traits, while accounting for potential confounders in complex designs. Moreover, computational efficiency is key to perform millions of statistical tests in reasonable computing times. The limiting distribution of Anderson test statistic under the null hypothesis of no effects was long known for the one-way case. However, in complex designs, permutation tests were still required to assess significance, becoming unfeasible in large datasets such as GTEx. Here, we derive the limiting distribution of the Anderson test statistic for any complex design, and provide a methodology to compute asymptotic p values (Chapter 3). Using a comprehensive set of simulations, we show that the asymptotic test has controlled type I error rates and high power, outperforming parametric alternatives in several settings. We also demonstrate its utility by applying it to two distinct real-case scenarios: condition-specific splicing QTL mapping across tissues, using data from the GTEx (Genotype-Tissue Expression) project, and GWAS of MRI-derived volumes of hippocampal subfields in the ADNI (Alzheimer's Disease Neuroimaging Initiative) cohort (Chapter 3).

Altogether, the work presented in this thesis represents a useful contribution to understanding genetic effects on alternative splicing, and constitutes a valuable resource for the field from the methodological standpoint, providing enhanced statistical approaches and analysis tools.

List of publications during the thesis:

1. Chen, L., Ge, B., Casale, F. P., Vasquez, L., Kwan, T., **Garrido-Martín, D.**, et al. (2016). Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell*, *167*(5), 1398-1414
2. Astle, W. J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A. L., et al. (2016). The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell*, *167*(5), 1415-1429.
3. GTEx Consortium (2017) Genetic effects on gene expression across human tissues. *Nature* 550, 204–213
4. **Garrido-Martín, D.**, Palumbo, E., Guigó R. and Breschi, A. (2018) ggsashimi: Sashimi plot revised for browser- and annotation-independent splicing visualization. *PLoS computational biology*, *14*(8), e1006360
5. **Garrido-Martín, D.**, Borsari, B., Calvo, M., Reverter, F., Guigó R. (2019) Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. *Submitted*
6. **Garrido-Martín, D.**, Calvo, M., Reverter, F., Guigó R. (2019) A fast non-parametric test of association for multivariate phenotypes. *In preparation*

Contents

INTRODUCTION

Genome-wide association studies: a success story?

Over the past decade, genome-wide association studies (GWAS) have led to the discovery of thousands of associations between genetic variants on one side, and human diseases and complex traits on the other (MacArthur et al., 2017). This experimental design has been very successful in the identification of novel disease susceptibility loci, genes and pathways (de Lange et al., 2017, Demenais et al., 2018, Li et al., 2017, Michailidou et al., 2017). In addition, GWAS associations have been highly replicated across different studies and populations (Marigorta et al., 2018), and the number of risk loci keeps increasing as larger sample sizes are used, with no evidence of saturation for any trait (Visscher et al., 2017). Although GWAS do not directly point to causal genes or mechanisms, the identified loci often comprise genes of unknown function or unexpected relevance, whose experimental follow-up may uncover new biological mechanisms underlying diseases (Tam et al., 2019). Some well-known examples include the association between the CFH gene and age-related macular degeneration (Klein et al., 2005), or between the major histocompatibility complex locus and schizophrenia (Sekar et al., 2016). Furthermore, GWAS for some diseases, such as type II diabetes, dyslipidemia or rheumatoid arthritis, have guided the development of new candidate drugs currently tested in clinical trials or already employed in clinical care (Visscher et al., 2017).

GWAS have demonstrated that most complex traits are highly polygenic, that is, a large set of mutations targeting multiple genes contributes to the trait variation in the population. However, on average, the proportion of variance explained by individual variants and their effect sizes are small (Visscher et al., 2017). Nonetheless, the gap between the amount of heritability explained and the amount observed in family studies, i.e. the *missing* heritability (Manolio et al., 2009), has been reduced as larger sample sizes have led to new discoveries. As an example, while in 2008 there were only 40 genome-wide significant variants for height, explaining 5% of the heritability, in 2014 this number was around 700, explaining 20% of the heritability (Visscher et al., 2017). Furthermore, small effect size variants below the genome-wide statistical significance threshold often account for a substantial fraction of heritability (Yang et al., 2010). Another lesson learnt from GWAS is the pleiotropic nature of genetic variants: most GWAS loci are significantly associated with several traits (Pickrell et al., 2016, Sivakumaran et al., 2011). A paradigmatic case are autoimmune diseases, where shared causal variants seem to drive the associations across single disorders (Ellinghaus et al., 2016, Parkes et al., 2013).

In GWAS, genotyping can be done using multiple technologies, including SNP arrays (combined with imputation of unobserved genotypes from population reference panels) and whole-genome sequencing (WGS). SNP arrays are currently the most widely used, mainly because of their reduced cost and high reliability (Tam et al., 2019). GWAS based on SNP arrays rely on existing reference panels of genetic variants derived from sequencing studies. As a consequence, most genetic variants surveyed to date are relatively common (Minor Allele Frequency, MAF, above 1%) (Visscher et al., 2017). However, as larger reference panels become available (e.g. the one recently created by the Haplotype Reference Consortium (Consortium et al., 2016)), SNP arrays are able to interrogate relatively low-frequency variants, achieving reasonable accuracy for imputation of MAFs down to 0.1% (Consortium et al., 2016). Still, the distinct linkage disequilibrium patterns across different ethnic groups hinder their usage beyond populations that have been sequenced at high coverage (Rosenberg et al., 2010). Moreover, the study of ultra-rare variants is feasible only through WGS, which is already the gold standard in the field, and will become widely used as sequencing costs drop (Tam et al., 2019, Visscher et al., 2017).

Despite clear successes, GWAS have faced strong criticism. A major concern is the modest fraction of the *missing* heritability explained for many complex traits (Manolio et al., 2009). Although this may be alleviated with new discoveries in larger experimental samples, as pointed out before, it is unlikely that GWAS identify all the genetic determinants of a given trait. Some of the reasons include the difficulty in assessing the contribution of rare and ultra-rare variants, as well as the presence of complex interactions (e.g. epistasis, environment, etc.), largely invisible to GWAS (Tam et al., 2019). Another key aspect is the correlated structure of the genome. While linkage disequilibrium definitely helps to find genotype-phenotype associations, it precludes the identification of the actual causal variant(s) and target genes (Schaid et al., 2018). Indeed, the vast majority of GWAS loci falls in non-coding regions (Maurano et al., 2012), which hinders their interpretation, and functional characterization is essential to understand the biological mechanisms beyond the statistical associations.

Moreover, for most complex traits GWAS loci perform poorly at classifying patients with and without the disease (Loos and Janssens, 2017, Marigorta et al., 2018), limiting their utility for clinical applications. Nonetheless, summarizing the risk of multiple loci at the individual level through polygenic risk scores has been a useful strategy to stratify the population in different groups according to the disease risk (e.g. in cancer) (Torkamani et al., 2018). Given the huge number of associations identified to date, some authors have suggested that GWAS may eventually involve most of the genome (Boyle et al., 2017, Goldstein, 2009), becoming therefore uninformative. Furthermore, population stratification (i.e. differences in allele frequencies due to differences in ancestry rather than to the association with a given trait) may lead to spurious associations: actually, it has even been suggested that most of the weak GWAS signals might be due to cryptic population stratification (McClellan and King, 2010), although this is probably unrealistic. Additional limitations include the reduced range of phenotypes explored (typically single-trait, easy-to-measure phenotypes) and the multiple testing burden (GWAS often use Bonferroni correction to achieve a false-positive rate of 5% –assuming 1 million independent tests for common genetic variation, $p < 5{\cdot}10^{-8}$–, which results in a lack of power to detect true associations (Tam et al., 2019)).

The post-GWAS Era and the study of molecular Quantitative Trait Loci (QTL)

To bridge the gap between the genotypes and the associated organismal phenotypes identified by GWAS, it is essential to fully understand the flow of biological information that underlies complex traits. Only by characterizing the molecular functions of the causal variants and target genes, as well as the biological mechanisms through which they act, we will be able to shed light on their contribution to disease susceptibility. The generation of a large number of publicly available catalogs of regulatory elements, across a broad range of tissues and cell types (such as ENCODE (Consortium, 2012), FANTOM (Andersson et al., 2014), or Roadmap Epigenome (Kundaje et al., 2015)), is critical to achieve this goal. GWAS loci have consistently shown enrichments in regulatory regions, suggesting that they may indeed play a role in gene regulation (Maurano et al., 2012, Schaub et al., 2012). In addition, understanding the three-dimensional organization of chromatin may help to identify the target genes. For instance, conformation capture experiments were key to determine that the intronic variants associated with obesity at the FTO locus were in fact interacting with the promoter of IRX3, a gene whose expression is related to the regulation of body mass (Smemo et al., 2014).

However, besides exploiting the overlap with regulatory elements to prioritize GWAS variants, we can directly study the impact of genetic variation on molecular phenotypes. The latter are tightly connected to changes in the DNA sequence, and are therefore more informative about the regulatory architecture of the human genome. Recent technological developments enable genome-wide profiling of virtually all the steps in the molecular path from DNA to proteins, including DNA methylation (Bisulfite-seq), chromatin accessibility (DNAse-seq, FAIRE-seq, ATAC-seq), histone modifications, transcription factor binding (ChIP-seq), RNA abundance and processing (RNA-seq), etc. Most of these molecular phenotypes can be treated as quantitative traits. Hence, measuring them in large cohorts of genotyped individuals (either by SNP arrays or WGS), allows to map molecular quantitative trait loci (QTLs), i.e. genomic regions containing one or more variants associated with the molecular trait (Albert and Kruglyak, 2015). Molecular QTLs often explain a substantial fraction of the heritability of complex traits (Gamazon et al., 2018), and weak GWAS associations might actually be strong molecular QTLs, boosting their clinical relevance (Tam et al., 2019).

Overall, QTLs can be classified as *local* or *distal* according to their relative location with respect to the element (gene, splicing event, histone mark, etc.) that they affect. However, the distance required for a QTL to be *distal* is often arbitrary, and it may change depending on the molecular trait considered. As *local* QTLs tend to act in *cis* and *distal* QTLs in *trans*, the classification based on the mechanism through which the QTL affects the trait (*cis* or *trans*) is often used instead of the former based on location (*local* or *distal*), although they are not completely interchangeable (e.g. *local* QTLs may have *trans* effects) (Albert and Kruglyak, 2015, Rockman and Kruglyak, 2006). Note also that *cis* QTLs act in an allele-specific manner, and therefore heterozygous individuals display different levels of the molecular trait at each allele. For this reason, allele-specific expression (ASE) analyses often accompany *cis* expression QTL (eQTL) mapping (Chen et al., 2016, Consortium, 2017). In contrast, *trans* QTLs modify the abundance, structure or activity of a diffusible factor, thus being able to affect both alleles of the target gene. To date, most QTL mapping studies have focused on *cis* QTLs, as the stringent multiple testing correction required hinders the identification of *trans* QTLs, which generally display smaller effect sizes and more tissue-specific patterns (Albert and Kruglyak, 2015).

The most widely studied molecular phenotype is gene expression. In a variety of experimental settings, eQTL mapping has confirmed that genetic variants influence the expression of most human genes (Battle et al., 2014, Chen et al., 2016, Consortium, 2017, Lappalainen et al., 2013). eQTLs have helped to link GWAS variants with the genes and biological pathways that they actually affect, suggesting plausible causal mechanisms. For instance, obesity-associated SNPs in the FTO locus, mentioned above, affect the expression of IRX3 (but not FTO) in brain (Smemo et al., 2014). Another example is the 1p13 locus, associated with myocardial infarction (MI). A study showed that a SNP in this locus creates a TF binding site and alters the hepatic expression of the SORT1 gene in humans. Moreover, it revealed that SORT1 affects in mice the low-density lipoprotein (LDL) levels, a well-known risk factor for MI. Altogether, these observations point to SORT1, rather than to the gene in which the SNP is located, as the actual causal gene, and to its pathway as a promising new target for the reduction of LDL levels and MI prevention (Musunuru et al., 2010).

Integrating eQTL analyses with regulatory information across different tissues and cell types has revealed an enrichment of eQTLs in open chromatin regions and TF binding sites, as well as loci decorated by promoter- and enhancer-related histone marks (Consortium, 2017, Lappalainen et al., 2013). Indeed, increasing evidence supports the fact that a substantial fraction of eQTLs might affect gene expression through *cis* effects on TF binding (Albert and Kruglyak, 2015). However, it is unclear whether changes in the chromatin landscape drive TF binding or it is the opposite case (Henikoff and Shilatifard, 2011). To help in this task, and to achieve a better understanding of the control of transcriptional activation and repression, DNaseI sensitivity QTLs (dsQTLs) and histone QTLs (hQTL) have also been identified (Degner et al., 2012, McVicker et al., 2013). This has been further complemented with the study of genetic variants affecting DNA methylation at CpG sites (methylation QTLs, meQTLs) (Banovich et al., 2014). Downstream analyses of these QTL catalogs suggest that genetic variants lead to coordinated changes across different molecular phenotypes, and point to variation in TF binding as a primary driver of such changes (Banovich et al., 2014, Kilpinen et al., 2013).

Moreover, besides the genetic effects on chromatin or RNA abundance, the identification of QTLs affecting post-transcriptional mechanisms, such as alternative splicing (AS), is key to bridge the gap between genotypes and organismal phenotypes (Li et al., 2016). In this sense, protein QTL (pQTL) mapping has also proven valuable (Wu et al., 2013), especially given the generally weak correlation between transcript and protein levels (Battle et al., 2015), and despite the technical limitations to assess protein abundances proteome-wide (Chandramouli and Qian, 2009). Nevertheless, the comprehensive analysis and integration of all these layers of molecular information is still challenging, and approaches to model higher-order relationships between different molecular phenotypes (also between these and environmental exposures) are required to fully understand their interaction and relative contribution to the organismal phenotype (Civelek and Lusis, 2014).

As with GWAS, in QTL mapping the correlated structure of the genome hinders the identification of the causal variant(s) controlling the molecular trait of interest. To overcome this limitation, in addition to functional characterization, statistical fine-mapping (i.e. assigning probabilities of causality to each variant in a given locus) is commonly applied (Spain and Barrett, 2015). This requires high quality genotyping (or high confidence imputation) of the variants in the region, and large sample sizes to distinguish between variants in high LD. WGS definitely helps in this task, and it is more feasible in QTL mapping studies, where the number of samples required to achieve reasonable power is much smaller than in GWAS (Brown et al., 2017). Most available methods for statistical fine-mapping rely on bayesian approaches that use summary statistics and LD information to assign causal posterior probabilities. These often allow multiple causal variants per locus and provide causal *credible sets* (groups of variants that account for a large fraction, e.g. 95%, of the posterior probability in a given locus) (Hormozdiari et al., 2014, Wen et al., 2016). A recent alternative, `CaVeMan` (Brown et al., 2017), proposes to use non-parametric resampling to estimate causal probabilities.

A related problem is to determine whether two association signals at the same locus (e.g. GWAS and eQTL) are independent or correspond to a shared causal variant (i.e. colocalization) (Giambartolomei et al., 2014). Identifying a variant associated with different traits is not sufficient to confirm colocalization, as this situation cannot be distinguished from a scenario in which there are two distinct causal variants in LD. Most colocalization methods rely on Bayesian approaches that use summary association statistics from the two traits of interest, plus information at the locus level, such as LD structure or minimum allele frequencies (Giambartolomei et al., 2014, Hormozdiari et al., 2016, Wen et al., 2017). Moreover, colocalization has been recently extended to multiple traits (Giambartolomei et al., 2018). This allows to integrate GWAS with multiple molecular QTL data, providing valuable insights into the mechanisms underlying GWAS associations.

The Genotype-Tissue Expression (GTEx) Project

Analyzing multiple tissues is an important aspect in the study of the genetic basis of complex traits and diseases for several reasons. First, complex diseases often involve several tissues, being difficult to identify the causal one(s) (Dermitzakis, 2012). Second, the interpretation of the functional consequences of disease-associated variants needs to be conducted in a disease-relevant cellular context (Lonsdale et al., 2013). Third, the molecular phenotypes of interest, such as gene expression, RNA splicing or those derived from epigenomic assays, often display distinct patterns and are differentially regulated across tissues (Kundaje et al., 2015, Merkin et al., 2012).

The Genotype-Tissue Expression (GTEx) Project has emerged as an unprecedented resource to study the genetic effects on gene expression and other molecular phenotypes across a large panel of (non-diseased) reference tissues (Consortium, 2015, 2017, Lonsdale et al., 2013). In its current release (v7), it provides RNA-seq data for more than 10,000 samples collected from 53 tissue sites (spanning solid-organ tissues –including several brain subregions–, whole blood, and two derived cell lines) across more than 600 deceased donors. In addition, peripheral blood samples are employed as DNA source for WGS-based genotyping. Samples (from donors of both sexes, any ancestry, aged between 21 and 70) are characterized by a short post-mortem interval (biospecimen collection starts within 24h from death), which ensures high-quality nucleic acids and robust gene expression measurements (Lonsdale et al., 2013). GTEx raw data is publicly available at the database of Genotypes and Phenotypes (dbGaP).

RNA splicing: a key step in the path from genotype to phenotype

It is well established that the genetic effects on transcriptional and epigenetic regulation, often leading to changes in gene expression, are major drivers of the phenotypic variability among individuals (Chen et al., 2016, Consortium, 2017). In addition, recent studies indicate an emerging role of genetic variants affecting RNA splicing, which may contribute to complex traits at a comparable level to variants affecting gene expression, and often through independent mechanisms (Li et al., 2016). This is not unexpected, given the central position of splicing in the gene regulation cascade.

The majority of human genes contain introns, i.e. segments that should be removed from the transcribed pre-mRNA during splicing. Moreover, beyond this *constitutive* processing, more than 90% of human genes undergo alternative splicing (AS), producing multiple transcript isoforms from a single pre-mRNA (Wang et al., 2008). This dramatically increases the coding capacity of the human genome, and provides countless opportunities for regulation. Splicing (and especially AS) is subject to a tight regulation, often tissue-, cell type- or condition-specific (e.g. in response to signaling triggered by external stimuli), involving a wide range of *cis*-acting regulatory sequences (e.g. binding sites, RNA secondary structure) and *trans*-acting factors (e.g. RNA-binding proteins) (Chen and Manley, 2009, Fu and Ares, 2014). All these elements become potential targets of genetic variation affecting complex traits and diseases.

Splicing occurs by specific cleavage at the 5’ and 3’ splice sites, highly conserved sequences that contain the first and last two nucleotides of each intron (consensus GU/AG dinucleotides). These, together with the *branchpoint* sequence, which lies 18 to 40 nucleotides upstream of the 3’ splice site, are strictly required by the spliceosome (a large ribonucleoprotein complex) for exon recognition and splicing catalysis (Herzel et al., 2017, Shi, 2017). Therefore, mutations disrupting them often lead to severe phenotypic consequences (Manning and Cooper, 2017). Classical examples include splice site mutations in the DMD gene (dystrophin), which result in frameshifts and subsequent nonsense-mediated decay (NMD) degradation of the transcripts, causing Duchenne muscular dystrophy (Scotti and Swanson, 2016), or polymorphisms in the polypyrimidine tract (a pyrimidine rich sequence located between the branchpoint and the 3’ splice site) before exon 9 of the CFTR gene, which modify the severity of cystic fibrosis (Chu et al., 1993). Activation of cryptic splice sites may also cause disease, as in the case of Hutchinson-Gilford progeria syndrome, where a mutation activating a cryptic 5’ splice site causes a deletion of exon 11 of the LMNA gene (Eriksson et al., 2003). Nevertheless, the majority of genetic variants affecting splicing (i.e. splicing QTLs, sQTLs) do not disrupt splice sites, leading to more frequent and subtler phenotypic effects: for instance, they may modulate disease susceptibility or response to therapy (Manning and Cooper, 2017).

Splice sites can be considered strong or weak depending on how much they differ from the consensus sequences, as this determines their affinity for spliceosomal components and other splicing factors. While strong splice sites are generally used, leading to constitutive splicing, the alternative usage of weak splice sites mainly depends on the cellular context and is controlled by other *cis*-acting regulatory elements (Kornblihtt et al., 2013). Hence, another potential mechanism through which sQTLs may affect splicing patterns is the impact on these auxiliary elements. They are typically short sequences (6-8 nucleotides) that correspond to binding sites of *trans*-acting RNA-binding proteins (RBPs). Depending on their location and influence on the usage of the associated splice site(s), they can be classified as exonic splicing enhancers (ESEs), intronic splicing enhancers (ISEs), exonic splicing silencers (ESSs) or intronic splicing silencers (ISSs) (Kornblihtt et al., 2013). The relative position of these elements with respect to each other and the splice sites is key for their function (Fu and Ares, 2014), and it has been shown that distal regulatory elements are as relevant as those in the vicinity of splice sites (Lovci et al., 2013). RNA local structure at these positions may also be an important determinant of RBP binding (Manning and Cooper, 2017).

To understand how alterations in the sequence of *cis* regulatory elements impact splicing, it is crucial to determine the *trans* factors that actually bind to them. However, this is not always straightforward, as individual *cis* elements have little information content, and RBPs often recognize variable motifs (Manning and Cooper, 2017). Recent technical developments, such as the enhanced crosslinking and immunoprecipitation (eCLIP) methodology, which enables transcriptome-wide identification of RBP binding sites, have helped to build more comprehensive catalogues of splicing regulatory elements (Van Nostrand et al., 2016). They have also facilitated allele-specific binding studies to link the function of a given genetic variant to the role of the RBP that displays allele-specific binding patterns (Yang et al., 2019). Of note, many RBPs can act both as splicing activators or inhibitors depending on the sequence and position of their binding sites (Ule et al., 2006).

Some classical splicing factors include Ser/Arg rich (SR) proteins and heterogeneous nuclear ribonucleoproteins (hnRNPs), often viewed as positive and negative regulators of splicing, respectively (although this is not always the case) (Fu and Ares, 2014). For example, SR proteins bind to ESEs and promote exon inclusion by recruiting spliceosomal components and other splicing factors to splice sites in the first stages of spliceosome assembly (Zhou and Fu, 2013), while some hnRNPs, such as hnRNP A/B or polypyrimidine tract-binding protein (PTB), antagonize the function of SR proteins (Okunola and Krainer, 2009) or interfere with the spliceosome assembly (Sharma et al., 2008), respectively. Other well-characterized RBPs regulating splicing include tissue-specific factors such as NOVA (Ule et al., 2006) and RBFOX (Yeo et al., 2009). Interestingly, many splicing factors are involved in positive and/or negative autoregulation and cross-regulation mechanisms, such as alternative splicing coupled to nonsense-mediated decay (AS-NMD) (Jangi and Sharp, 2014).

In addition to *cis* effects, sQTLs can also have *trans* effects, that is, they can affect splicing of the target genes by altering expression, splicing, stability, etc. of *trans*-acting factors. Indeed, mutations in the core spliceosomal proteins such as the pre-mRNA processing factor (PRPF) proteins have been related to several diseases, including retinitis pigmentosa (Tanackovic et al., 2011) and cancer

(Kurtovic-Kozaric et al., 2015). Another example are genetic variants that impact the expression of RBFOX1, which has been associated with misregulation of AS in brains of individuals with autism spectrum disorders (Voineagu et al., 2011). However, as stated in the previous section, *trans* QTL mapping still presents some limitations.

To add another layer of complexity, RNA splicing (as well as other RNA processing steps such as 5’ end capping, 3’ end cleavage, polyadenylation or editing) is generally coupled with transcription (Bentley, 2014, Kornblihtt et al., 2013). This has two main implications: i) splicing often occurs co-transcriptionally (introns are removed prior to transcription termination) and ii) the two processes influence each other through coordinated mechanisms. For instance, splicing factors can be recruited by the transcription machinery (de la Mata and Kornblihtt, 2006, Huang et al., 2012), or kinetics of transcriptional elongation may determine exon inclusion (e.g. slow elongation is often associated with higher inclusion rates) (Ip et al., 2011). This considerably expands the set of potential mechanisms through which sQTLs might act, and provides an interesting basis to explore the interplay between sQTLs and eQTLs. However, despite the co-transcriptional nature of splicing, sQTL mapping studies evidence that the majority of genetic variants that affect splicing differ from those affecting gene expression (Lappalainen et al., 2013, Li et al., 2016).

The fact that splicing is generally coupled with transcription inevitably opens to an additional level of information, provided by the epigenetic landscape (Naftelberg et al., 2015). In this sense, a number of histone modifications, such as H3K36me3 and H3K79me2, has been related to splicing. For example, SETD2, the histone methyltransferase that specifically tri-methylates lysine 36 of H3, has been shown to regulate exon inclusion by helping to recruit RBPs (Luco et al., 2010, Pradeepa et al., 2012), while the distribution of the H3K36me3 mark seems itself sensitive to alterations in splicing (de Almeida et al., 2011, Kim et al., 2011). As for nucleosome positioning, it has been suggested that this feature may help the splicing machinery to find exons (where nucleosomes are preferentially located), acting as ’bumps’ that pause transcriptional elongation (Hodges et al., 2009). Remarkably, although splicing generally occurs co-transcriptionally, there is a group of transcripts, mainly lncRNAs and often alternatively spliced, that tend to be processed post-transcriptionally (Tilgner et al., 2012).

Alternative splicing visualization

Visualization of splicing events from RNA-seq data is often required for the study of alternative splicing. For instance, in the context of sQTL analyses, it is useful to represent and compare the splicing patterns of a gene between individuals with different alleles at a particular SNP. However, the fact that splice sites are not contiguous on the genome sequence (they can be even hundreds of kilobases away from each other), complicates such task. The sashimi plot, a very effective and established splicing visualization strategy, solves this issue by drawing curves that connect splice sites to illustrate the presence of a splice junction supported by RNA-seq. These connective elements are displayed in combination with information of read coverage in the form of a signal track (Katz et al., 2015).

A tool for drawing sashimi plots was initially developed as part of the MISO suite (Katz et al., 2010), a software that quantifies and compares alternative splicing from different RNA-seq experiments. In addition to this stand-alone utility available specifically for MISO-indexed splicing events (Katz et al., 2015), the Integrative Genomics Viewer (IGV, (Thorvaldsdottir et al., 2013)) offers its own built-in. However, both implementations present several limitations that significantly hinder their applicability. For instance, the former relies on a proper compatible annotation of the event, and the latter requires IGV installation and time-consuming uploading of voluminous alignment files. Moreover, in both cases the comparison of splicing events is restricted to a few samples, since each RNA-seq experiment is represented on a separate line. This supposes a major limitation for the analysis of large-scale RNA-seq datasets such as GTEx or ENCODE.

Additional sources of transcriptional diversity

Alternative usage of splice sites gives rise to a variety of alternative splicing *events*, such as alternative cassette exon inclusion, mutually exclusive cassette exons, alternative 5’ or 3’ splice site usage or intron retention (Kornblihtt et al., 2013). As previously stated, combinations of these events result in an increased repertoire of transcripts, proteins and functions encoded by the human genome. However, in addition to alternative splicing, there are other relevant sources of transcriptional diversity. For instance, about 30% of human genes have alternative transcription start sites (TSS), and over 70% display multiple polyadenylation sites (Manning and Cooper, 2017). Indeed, recent studies suggest that alternative transcription start and termination drive most transcript isoform differences across human tissues (Reyes and Huber, 2018). Hence, genetic variation with effects on these processes should be also considered an important determinant of phenotypic variability. Nevertheless, the alternative usage of promoters, splice sites and polyadenylation sites is highly interleaved, and the structure of alternative isoforms often results from combinations of all three (de Klerk and 't Hoen, 2015).

Isoforms originated from alternative transcription start and termination contain different 5' and 3' untranslated regions (UTRs), which often carry *cis*-acting elements involved in the regulation of RNA stability, secondary structure, localization or translation (Gupta et al., 2014, Wang et al., 2016). In particular, there is evidence of 3' UTRs being especially relevant from the functional standpoint (Manning and Cooper, 2017), and some studies have reported that a substantial fraction of genetic variants affecting transcript isoform levels is located at 3' UTRs (Lappalainen et al., 2013). A potential mechanism through which genetic variation in 3' UTRs might affect human traits is the disruption of polyadenylation signals, which include the AAUAAA hexanucleotide, 20-40 nucleotides upstream of the 3' cleavage site, and a GU-rich sequence, within 50 nucleotides downstream of the cleavage site. Changes in these motifs can reduce dramatically the efficiency of the 3' end formation, or determine the use of distal versus local alternative polyadenylation sites, leading to disease (Manning and Cooper, 2017). For instance, mutations in the polyadenylation hexanucleotide of the gene HBA2 are long known to cause α-thalassemia (Higgs et al., 1983). In addition, genetic effects on polyadenylation are an important risk factor for several diseases, including systemic lupus erythematosus (Hellquist et al., 2007) and cancer (Stacey et al., 2011).

As concerns the genetic variants regulating alternative TSS usage, recent studies highlight their markedly context-specific nature and report colocalization with GWAS hits associated to a wide variety of complex traits (Alasoo et al., 2019). Alternative TSS usage often results from changes in the chromatin state or transcription factor binding, not only at the level of promoters but also at enhancers (de Klerk and 't Hoen, 2015). Hence, regulatory variation affecting alternative

TSS usage might act through these processes. Of note, leveraging the CAGE (Cap Analysis of Gene Expression) (Kodzius et al., 2006) data generated by the FANTOM consortium (Consortium et al., 2014) has suggested that variants with opposite effects on different transcript isoforms reflect compensatory mechanisms between alternative promoters (Garieri et al., 2017).

Approaches to splicing QTL mapping

In recent years, several approaches have been proposed to identify genetic variants that affect alternative splicing (AS), i.e. splicing QTLs or sQTLs (Lappalainen et al., 2013, Li et al., 2018, Monlong et al., 2014). In a typical sQTL mapping study, RNA-seq is performed in a large cohort of individuals ($n > 100$), genotyped using either arrays or WGS. Then, AS phenotypes derived from RNA-seq are tested for association with nearby (*cis* sQTL mapping) or distant (*trans* sQTL mapping) genetic variants, often using linear regression or generalized linear models. A crucial step is the definition of the AS phenotype as a quantitative trait, which relates to how AS can be quantified from RNA-seq data. This is a complex task, and the nature and properties of the resulting AS phenotype (distribution, single- or multi-trait, etc.) largely impact the choice of the statistical method employed for the mapping.

Although it represents a different application, sQTL mapping can be related to differential splicing (DS) analyses, in which AS is compared between two or more experimental conditions (Hooper, 2014). In this context, it can be interpreted as a particular case of DS between groups defined by genotypes. In fact, the two types of analyses share the definition of the AS phenotype, and in some cases even the statistical framework to assess differences (Li et al., 2018, Nowicka and Robinson, 2016). Nevertheless, most methods employed for DS cannot be scaled up to deal with many replicates and perform millions of tests in reasonable computation times, or simply have not been adapted for sQTL mapping.

Another relevant aspect, general to all QTL mapping approaches independently of the molecular trait of interest, is the multiple testing burden: multiple genetic variants are tested per phenotype and multiple phenotypes are tested genome-wide (Ongen et al., 2016). Hence, to reduce the number of false discoveries, several approaches have been proposed, ranging from classical family-wise error rate (FWER) controlling procedures to complex permutation-based schemes.

Defining the splicing phenotype

RNA-seq technology allows to study AS at an unprecedented resolution, by producing millions of reads derived from the transcriptome (Pan et al., 2008, Wang et al., 2008). However, limited coverage depth, experimental biases (e.g. reads are not evenly distributed along the transcripts due to differences in the GC content, positional biases, etc.), and, in the case of short-read RNA-seq, reads spanning only a small fraction of the alternatively spliced portions of the transcripts, pose several challenges for AS quantification (Alamancos et al., 2014, Vaquero-Garcia et al., 2016). As the exact structure of the different isoforms cannot be directly derived from short-read RNA-seq data, two alternative strategies have been proposed to quantify AS.

A first approach aims to probabilistically estimate the abundance of full-length transcript isoforms. This generally involves mapping the reads to a reference genome or transcriptome, followed by the probabilistic assignment of reads to isoforms (e.g. maximum likelihood (ML) estimation by expectation maximization in methods such as Cufflinks (Trapnell et al., 2010) or RSEM (Li and Dewey, 2011)). In addition, pseudo-alignment algorithms allow fast transcript quantification without prior read mapping (Sailfish (Patro et al., 2014), Kallisto (Bray et al., 2016) or Salmon (Patro et al., 2017)). Furthermore, although transcript quantification is generally annotation-dependent, some methods are able to reconstruct isoforms *de novo* (Cufflinks, StringTie (Pertea et al., 2015)). The expression of each isoform is generally given in terms of R/FPKM (Reads/Fragments Per Kilobase of transcript per Million mapped reads) or TPM (Transcripts Per Million), and its contribution to the overall gene expression constitutes the AS phenotype. Indeed, several studies have employed transcript ratios (transcript expression divided by total gene expression) for sQTL mapping (Battle et al., 2014, Lappalainen et al., 2013, Monlong et al., 2014, Montgomery et al., 2010, Ye et al., 2018).

The main drawback of the transcript-based approach is the high complexity of inferring full-length isoform abundances from short reads, given that most reads cannot be unambiguously assigned to individual transcripts (Park et al., 2018). Uncertainty in this assignment should be appropriately modeled (Alamancos et al., 2014), taking into account experimental biases and read overdispersion (inflation of variance, read counts are often more variable than what is expected according to a Poisson distribution) (Nowicka and Robinson, 2016). Moreover, the estimation is highly sensitive to coverage –especially for lowly expressed transcripts– and the choice of transcript annotations, and the set of most probable transcripts derived may not be unique. In addition, it is often difficult to attribute changes in isoform abundances to changes at specific exons or splice sites, especially for genes with multiple alternatively spliced regions (Park et al., 2018).

A second strategy exploits local information on the read distribution at the exon or junction level to directly measure specific AS events. This approach avoids the complex estimation of transcript abundances, assuming that the differences in transcript usage can be tracked locally, and has been widely used in DS analyses (Alamancos et al., 2014). The splicing phenotype can be defined as exon usage (as in DEXSeq (Anders et al., 2012), for DS analyses), splice junction usage (as in Altrans (Ongen and Dermitzakis, 2015), for sQTL mapping) or more commonly, as *percent spliced in*, Ψ or PSI (as in MISO (Katz et al., 2010) or rMATS (Shen et al., 2014), for DS analyses, and GLIMMPs (Zhao et al., 2013), for sQTL mapping).

For a given AS event, the PSI can be computed as the fraction of reads supporting the inclusion of a specific exon or splice site. For example, in the case of a cassette exon, the reads that align to the body of this exon, or to splice junctions involving it, support its inclusion, while the reads joining the two adjacent exons support its exclusion (Katz et al., 2010). This framework can be extended to capture other simple AS events, such as alternative 3' and 5' splice sites, mutually exclusive exons or intron retention, in a straightforward manner.

Despite differences in AS event definition or read-counting procedures, event-based approaches tend to produce highly concordant PSI estimates on the same set of events (Park et al., 2018). Furthermore, these approaches display a good agreement with experimental results (real-time PCR) (Alamancos et al., 2014). As with transcript-based approaches, here coverage is a key determinant of reliability, and an appropriate modeling of the confidence of PSI estimates improves downstream analyses (Katz et al., 2010). Read overdispersion should be taken into account (Zhao et al., 2013). It is worth mentioning that PSI values at the event level can also be derived from transcript estimates (as in SUPPA (Trincado et al., 2018) or MISO). Although this potentially uses more information to compute the PSI estimates, it presents the limitations of both event- and transcript-level approaches.

A drawback of some event-based methods is that they rely on preexisting annotations of transcripts or AS events (e.g. MISO). This can be particularly restrictive when comparing AS between healthy and diseased individuals (e.g. some isoforms may be disease-specific) (Li et al., 2018), or when genetic variants lead to splicing events only in a subset of individuals (Stein et al., 2015). In addition, even when annotations are complete, it is not easy to quantify complex AS events (Vaquero-Garcia et al., 2016). To overcome these limitations, some methods integrate annotated transcriptomes with novel splice junctions, using split-reads to identify local splicing variations (LSVs). Generally, these approaches build splicing graphs where nodes are exons and edges represent shared splice junctions between two exons. In this case, simple AS events correspond to particular cases of binary graph splits, whereas LSVs are able to capture complex splits involving two or more junctions. This strategy is employed by MAJIQ in DS analyses (Vaquero-Garcia et al., 2016).

An analogous setting, but with an intron-centric perspective, is taken by LeafCutter both for DS and sQTL mapping analyses (Li et al., 2018). In this case, the AS phenotype is defined by each cluster of alternatively excised introns derived from the splicing graph. However, this approach has received some criticism, including the lack of interpretability of intron ratios, which do not correspond directly to any known AS-related biological entity, or the fact that it does not model intron retention, known to be a highly relevant event (Vaquero-Garcia et al., 2018).

Currently, transcript quantifications tend to be noisier and less robust than local measurements of AS (Alamancos et al., 2014). On the other hand, local methods, even when accounting for complex events, may miss information at global level. For instance, alternative first and last exons are very rarely accounted for, although they contain the 5’ and 3’ UTRs of the genes, well-known targets of regulation (Gupta et al., 2014, Wang et al., 2016). Of note, these can result not only from AS but also from alternative promoter usage or alternative cleavage and polyadenylation signals, which despite being relevant sources of transcript diversity (Reyes and Huber, 2018), may not be strictly considered AS (Kornblihtt et al., 2013). Overall, global and local views of AS should be considered complementary.

This may change in the next few years, as long-read RNA sequencing becomes cost-effective for sQTL mapping studies. To date, third generation long-read RNA sequencing, mostly using Pacific Biosciences (PacBio) (Rhoads and Au, 2015) and Oxford Nanopore (Feng et al., 2015) technologies, has been successfully employed to study AS in a variety of experimental settings (Bolisetty et al., 2015, Sharon et al., 2013). Thanks to long reads, these approaches allow the direct resolution of isoform structure. Yet this comes at the cost of higher error rates and lower throughput, and whereas aligners can overcome the former by leveraging the information of long reads, the latter is still a major barrier for accurate isoform quantification (Park et al., 2018).

Another relevant aspect of AS phenotype definition is whether AS is considered a *single-trait* (i.e. univariate) or a *multi-trait* (i.e. multivariate) phenotype. This is tightly linked to how the AS phenotype will be modeled to detect differences in AS. To date, most sQTL mapping approaches rely on single-trait AS phenotypes: abundances of individual transcript isoforms (Battle et al., 2014, Lappalainen et al., 2013, Ye et al., 2018), PSI of individual splicing events (Rotival et al., 2019, Takata et al., 2017), etc. However, studying AS-related phenotypes independently ignores the strongly correlated structure of the different AS events occuring in a given gene, and may result in a loss of power to detect splicing differences (Monlong et al., 2014). Hence, some recent approaches have proposed to study AS as a multivariate phenotype, built from either event-level (Li et al., 2018) or isoform-level traits (Monlong et al., 2014, Nowicka and Robinson, 2016).

Furthermore, some approaches use feature (e.g. transcript, exon) abundances as AS phenotypes (Nowicka and Robinson, 2016), while others rely on feature ratios (Li et al., 2018, Monlong et al., 2014) to account for the overall expression of the gene. The latter may result in greater power to detect differences, but ignores the uncertainty of isoform expression (higher for lowly expressed transcripts). Therefore, when the AS phenotype is a ratio, additional steps may be required to filter out very lowly expressed genes (Nowicka and Robinson, 2016).

Testing for association with genetic variants

Initial approaches to assess statistical associations between genotypes and AS were inherited from expression QTL mapping. As a result, AS traits were often modeled as univariate phenotypes using linear regression. In this framework, a *t*-test is employed to evaluate whether the β coefficient corresponding to the genotype differs significantly from zero. Linear regression has become widely used in sQTL analyses, especially since the development of highly efficient implementations of the *ordinary least squares* (OLS) method (such as MatrixeQTL (Shabalin, 2012), based on efficient large matrix operations), which allow to perform millions of statistical tests in reasonable computation times. Indeed, the regression approach has been employed in a variety of experimental settings, using either transcript ratios (Lappalainen et al., 2013, Pickrell et al., 2010, Ye et al., 2018) or event-based quantifications (Li et al., 2016, Rotival et al., 2019, Takata et al., 2017) as response variables. Spearman's rank correlation between transcript ratios and genetic variants has been occasionally used as an alternative to linear regression (Battle et al., 2014, Montgomery et al., 2010, Ongen and Dermitzakis, 2015). Generally, the genotype is modelled as a continuous variable (0,1,2), rather than as a categorical variable in an analysis of variance (ANOVA) framework. This assumes a dosage model, where each copy of the minor allele has an additive (linear) effect on the phenotype, while ANOVA would allow both additive and dominant effects (Shabalin, 2012).

The popularity of the linear model relies on its simplicity, computational efficiency, and ability to account for potential confounders including them as covariates. In some sQTL analyses, random effect terms accounting for polygenic signal or sample relatedness have been included, extending linear regression to linear mixed models (LMMs) (Chen et al., 2016, Kahles et al., 2018, Zhang et al., 2015). Here, the statistical significance of the association is generally assessed using likelihood-ratio test statistics, which compare the likelihood of the full model to a null model without the genotype component, treated as a fixed effect.

However, both linear regression and LMMs have strong assumptions regarding the normality of the distribution of the residuals. In contrast, usual AS phenotypes, such as transcript ratios or PSI values, are likely to depart substantially from normality (e.g. the distribution of PSI values resembles a convex beta distribution with preference for extreme values (Kakaradov et al., 2012)). As a result, normalization procedures such as rank-based inverse quantile transformation –a methodology that replaces the sample quantiles by quantiles from the standard normal distribution– are commonly applied (Kahles et al., 2018, Rotival et al., 2019). Nevertheless, it is unclear whether these transformations result in higher power and lower type I error rates compared to modeling the untransformed data (Beasley et al., 2009).

Hence, linear models (LMs) have been sometimes replaced by the more flexible framework provided by generalized linear models (GLMs), especially in the case of event-based AS traits. The simplest scenario would be using a GLM with a logit link function, assuming that the exon inclusion reads (y) for a given exon with PSI Ψ follow a binomial distribution $y \sim Binomial(n, \Psi)$ and $logit(\Psi)$ is linearly modeled by the genotype effect (Zhao et al., 2013). However, due to overdispersion, the variance of read counts is higher than expected. To take that into account, GLIMMPs (Generalized Linear Mixed Model Prediction of sQTL) proposed to model the extra variance of Ψ as a random effect for each individual. This generalized linear mixed model approach was shown to outperform both LM and GLM approaches (Zhao et al., 2013). As in LMMs, likelihood-ratio tests can be used to assess the significance of the association between the AS phenotypes and genetic variants.

Despite the fact that they are widely used, univariate approaches (both LM and GLM based) ignore the strongly correlated structure of the different AS events occuring in a given gene. Moreover, testing independently each AS-derived trait leads to a substantially larger number of tests, which results in a more stringent multiple testing correction. Altogether, this may translate into a loss of power to detect changes in the splicing patterns (Monlong et al., 2014, Nowicka and Robinson, 2016). To overcome these limitations, several recent methods have proposed to model AS as a multivariate outcome. For example, DRIMSeq (Nowicka and Robinson, 2016) models transcript abundances through the multinomial distribution, assuming that transcript proportions follow a (conjugate) Dirichlet distribution to account for overdispersion. This Dirichlet-multinomial framework has been implemented as a multivariate GLM in Leafcutter (Li et al., 2018), where the AS phenotype is represented by intron excision ratios.

Both DRIMSeq and LeafCutter are suitable for either type of analysis (DS or sQTL mapping). However, published works to date using LeafCutter for sQTL mapping rely on univariate linear regression or LMMs with individual intron excision values as response variable (Knowles et al., 2018, Li et al., 2016, Raj et al., 2018).

Another recent approach, sQTLseekeR (Monlong et al., 2014), tests for association between genotypes and relative isoform abundances using an approach analogous to multivariate analysis of variance (MANOVA), without assuming any probabilistic distribution (Anderson, 2001). This non-parametric strategy appears superior to its parametric alternatives, potentially leading to higher statistical power. For example, sQTLseekeR detected, on the same dataset, a larger number of sQTLs than DRIMSeq (Nowicka and Robinson, 2016). However, as any ANOVA-like strategy, it assumes homoscedasticity, i.e. it requires the variance-covariance structure of the response variables to be approximately equal between the groups compared (Anderson, 2001). This is especially problematic in unbalanced situations, as it is the case of sQTL mapping. Analogously, heteroscedasticity is also a problem in linear regression settings, although this assumption may be relaxed in GLMs or mixed models.

Remarkably, AS phenotypes derived from RNA-seq data using any method (including methods suited for DS analysis) can be potentially employed to map sQTLs with the different statistical approaches available for this purpose.

Covariate correction

In large-scale RNA-seq studies, confounding factors are common and can have a substantial impact on the transcriptome. This may lead to an increased false positive rate (FPR) and reduced power in sQTL mapping (Dahl et al., 2019, Leek and Storey, 2007). Known technical or biological effects (e.g. batch, sex, age, etc.) are often accounted for by including them as covariates in the model used to test for association between the AS phenotype and the genotype. This is straightforward for most sQTL mapping methods (as they are LM- or GLM-based) except for DRIMSeq and sQTLseekeR, whose current implementations cannot deal with covariates.

In addition, there may be other relevant sources of unwanted variability that cannot be directly modeled, as they are not known (or not measured). To take them into account, a common practice is to compute the principal components (PCs) of the phenotype matrix as proxies for these *hidden factors*, and include them as covariates in the model (Chen et al., 2016, Lappalainen et al., 2013, Raj et al., 2018). To perform this task, approaches like surrogate variable analysis (SVA) (Leek and Storey, 2007) or probabilistic estimation of expression residuals (PEER) (Stegle et al., 2010) are widely used in QTL studies. These differ in their assumptions regarding the confounder structure, and generally outperform PCA (Dahl et al., 2019).

The aim of covariate correction is to increase the signal-to-noise ratio between the response (here, the AS phenotype) and the candidate predictor (here, a genetic variant). However, when the predictor is associated with the confounding factors, modelling them as covariates may lead to both false positives and false negatives, depending on the underlying causal structure of the data (Aschard et al., 2017). This is often the case of PCs computed on the phenotype matrix: if a genetic variant affects an individual AS phenotype, the inferred PCs may partially capture the genotype effect (a phenomenon known as *collider effect*) (Dahl et al., 2019). Analogous problems result from conditioning on heritable covariates in other contexts (Aschard et al., 2015, Day et al., 2016). Even when correlations between covariates and predictors are modest, the FPR inflation may be dramatic (e.g. $\rho \approx 0.01 \Rightarrow 10 \times \text{FPR}$). Furthermore, these false positives can be largely replicated (Dahl et al., 2019).

Nevertheless, in the case of *cis* sQTL studies, focused on local genomic windows and where most genetic effects are expected to be small and restricted to nearby exons or splice sites, the bias induced by conditioning on PCs derived from the AS phenotypes is probably small. Yet there may be situations in which this is still problematic, for example when a genetic variant affects AS of several co-expressed genes, or in the case of *trans* sQTL mapping (Dahl et al., 2019). A potential workaround would be to remove the effect of the genotype from the AS phenotype matrix prior to the inference of hidden confounders. However, this can be unfeasible when testing millions of variants genome-wide. Alternatively, benchmarks of the available methods can help to select the most appropriate one, although recent works suggest that accounting only for the effect of known factors is often a better strategy (Somekh et al., 2019).

Multiple testing correction

sQTL mapping requires to perform millions of statistical tests to assess association of all possible AS phenotype-variant pairs (in *cis* analyses this is restricted to the variants located within a given window around the AS phenotype), resulting in an equivalent number of nominal *p* values. Selecting a global significance threshold in this case is especially difficult, due to the variable nature of each tested genomic region in terms of allele frequency and linkage disequilibrium (LD). This problem is common to all QTL mapping methods, independently of the molecular phenotype of interest. Overall, we distinguish two multiple testing levels that should be considered: i) multiple genetic variants are tested per phenotype and ii) multiple phenotypes are tested genome-wide (Ongen et al., 2016).

To account for the fact that multiple genetic variants are tested per AS phenotype, a common strategy is to correct the nominal *p* values for the number of tested variants using Bonferroni’s method (Battle et al., 2014, Chen et al., 2016). This approach assumes that all the tested genetic variants are independent. Nonetheless, they tend to be correlated due to LD and, therefore, the number of effective tests can be much smaller than the number of variants. As a result, Bonferroni correction is overly conservative, leading to many false negatives and reduced power (Ongen et al., 2016). Alternatively, some approaches aim to estimate the number of effective tests to correct the nominal *p* values, using the eigenvalues of the genotype correlation matrix (e.g. eigenMT (Davis et al., 2016)). Still, a single parameter is unlikely to completely recapitulate the correlation structure among genetic variants (Conneely and Boehnke, 2007).

The gold-standard approach to account for the fact that multiple genetic variants are tested per AS phenotype empirically characterizes the null distribution of associations using permutations (Consortium, 2017, Montgomery et al., 2010). Typically, AS phenotypes are randomly permuted, leaving the genotype data unchanged to preserve the LD structure (hence, this approach takes LD into account). In each permutation, the smallest nominal *p* value, p'_{min} (or equivalently, the largest test statistic), is stored. Altogether, p'_{min} values are used to build the expected distribution of the strongest associations under the null hypothesis. Finally, an adjusted empirical *p* value is computed as the fraction of p'_{min} values that are smaller than the smallest observed nominal *p* value, p_{min} (Ongen et al., 2016, Sul et al., 2015). A major drawback of this approach is its large computational burden. Moreover, adjusted *p* values can be approximated only to a threshold limited by the number of permutations (e.g. to achieve *p* values down to 10^{-8}, the number of permutations required is 10^{8}) (Sul et al., 2015).

To overcome these limitations, a possible alternative is offered by adaptive permutation schemas, in which the number of permutations carried out depends on the significance of the association (e.g. the algorithm permutes until a given number of null associations are stronger than the observed one, so that many permutations are required only for highly significant associations, saving computation time) (Hubner et al., 2005). Other approaches have proposed to approximate the distribution obtained by permutations using multivariate normal (MVN) sampling (e.g. eGeneMVN (Sul et al., 2015)). This assumes that the test statistics under the null hypothesis asymptotically follow a MVN with mean 0 and variance defined by the variant correlation matrix. However, these assumptions do not hold for some test statistics (Anderson, 2001).

A third possibility, probably the most widely used, combines an adaptive permutation schema with an approximation of the distribution obtained by permutations (implemented in FastQTL (Ongen et al., 2016) and QTLtools (Delaneau et al., 2017)). It takes advantage of the fact that order statistics of independently and uniformly distributed random variables are beta distributed (Jones, 2009). Thus, the smallest nominal *p* values coming from m tests follow a beta distribution with parameters 1 and m. As the number of independent tests is generally lower than m due to LD, instead of fixing the parameters a priori, they are estimated from a reduced set of permutations by maximum likelihood (ML). Adjusted *p* values are then approximated from the ML-fitted beta distribution as $P(p'_{min} < p_{min})$ (Ongen et al., 2016).

To account for the fact that multiple AS phenotypes are tested genome-wide, a common approach is to apply FDR correction (Chen et al., 2016, Li et al., 2016). FDR estimates the proportion of false positives, by comparing the number of significant associations found to the number expected by chance. Of note, in this case any of the FWER corrections (e.g. Bonferroni) could also be applied, although they tend to be too stringent. To compute FDR, several approaches are available, including Benjamini-Hochberg (Benjamini and Hochberg, 1995) and Storey-Tibshirani (Storey and Tibshirani, 2003) procedures. The latter seems to be more suitable in the sQTL mapping scenario, as it assumes that the set of association tests comes from a mixture of both null and alternative hypotheses (in a variable proportion, learnt from the data), rather than being all null (Ongen et al., 2016).

Finally, to recover all significant sQTLs, the applied procedure is often the following: first, the empirical *p* value closest to the FDR threshold (usually 0.05) defines a genome-wide AS phenotype-level threshold p_t. Then, for each AS phenotype, a nominal *p* value threshold is computed, based on the expected distribution of minimum nominal *p* values, $f(p_{min})$, as $F^{-1}(p_t)$, where F^{-1} is the inverse cumulative distribution. Finally, for each AS phenotype, variants with a nominal *p* value below the AS phenotype-level threshold are considered significant (Consortium, 2017).

Beyond alternative splicing: analysis of multivariate phenotypes in Biology

Most GWAS and QTL analyses test for association with genetic variants a single phenotype at a time, even when multiple phenotypes are available (Consortium, 2017, Li et al., 2016, Natarajan et al., 2018, van der Meer et al., 2018). However, univariate approaches ignore the correlation, shared risk factors or clinical overlap between different traits (O'Reilly et al., 2012). In addition, summarizing the variety of biological processes that lead to complex diseases in a single phenotype is a difficult task, although less problematic for quantitative traits (e.g. body-mass index: BMI $= \frac{\text{weight}}{\text{height}^2}$) (O'Reilly et al., 2012). Certainly, using multiple traits helps to better capture the underlying biology. For example, genetic variants at the FTO locus, known to confer risk of obesity, have been reported several orders of magnitude more significantly associated to BMI than to weight (O'Reilly et al., 2012, Thorleifsson et al., 2009). Still, the most effective approach to study genetic effects on multiple traits is to model them jointly in a multivariate framework (Stephens, 2013).

In the last few years, the availability of phenotype data in large human cohorts has dramatically increased, as a result of considerable efforts to build comprehensive phenotype resources, such as biobanks. To cite an example, the UK biobank (Sudlow et al., 2015) has collected extensive phenotypic information (physical measurements, sample assays, multimodal imaging, longitudinal electronic health records etc.) from up to 500,000 participants. Moreover, recent technological developments have enabled genome-wide profiling of a wide variety of molecular phenotypes: DNA methylation, chromatin accessibility, histone modifications, transcription factor binding, RNA levels, alternative RNA processing, protein abundances, etc. (Consortium, 2012, Kundaje et al., 2015, Lonsdale et al., 2013) in addition to genotypes. As a result, there is a growing interest in multivariate analyses to study the genetic basis of multi-trait phenotypes, both at organismic and molecular level (Elliott et al., 2018, Nowicka and Robinson, 2016). Indeed, intrinsically multivariate phenotypes are widespread in Biology. Some examples include the size and connectivity of brain regions, the levels of blood lipids (LDL, HDL, triglycerides), the cellular composition of a tissue, the expression of genes in the same pathway, the abundances of the alternative spliced isoforms of a gene, etc.

Multivariate approaches present several advantages over their univariate counterparts. Overall, joint analysis of multiple traits offers increased statistical power to detect genetic associations (Galesloot et al., 2014, Porter and O'Reilly, 2017). Many complex traits and diseases share genetic and environmental influences, which may be reflected in the correlation structure of the traits, and therefore captured by multivariate analyses (Casale, 2016, Korte et al., 2012). Indeed, pleiotropy is a common phenomenon in the human genome: genetic variants tend to have multiple distinct phenotypic effects (Pickrell et al., 2016). Well-known examples include autoimmune or psychiatric diseases, where shared causal variants seem to drive the associations between individual disorders (Consortium et al., 2018, Parkes et al., 2013). Remarkably, multivariate analyses can also increase power when not all the phenotypes are affected by the genetic variants tested (Stephens, 2013).

As genetic effects on phenotypes may be context-specific, another scenario in which multivariate approaches are valuable is when the same trait is measured in different conditions (Casale, 2016). Indeed, multivariate analyses have been used to characterize context-dependent genetic effects across tissues (Sul et al., 2013), environmental exposures (Moore et al., 2019) or developmental stages (Francesconi and Lehner, 2014). Multivariate analyses have also proven valuable in longitudinal studies (Ning et al., 2019). In addition, multivariate approaches provide a unique framework to study the molecular mechanisms through which genetic variants contribute to organismal phenotypes, by enabling joint analyses across multiple molecular layers (Giambartolomei et al., 2018). Last but not least, when compared to single-trait approaches, multi-trait analyses reduce the number of individual tests performed, and with this also the multiple testing burden.

Available approaches to assess association between genetic variants and multiple phenotypes can be grouped into four broad categories: i) meta-analysis approaches, which exploit summary statistics from univariate tests (van der Sluis et al., 2013, Zhu et al., 2015) ii) methods that build linear combinations of phenotypes, such as principal component analysis (PCA) (Aschard et al., 2014) or canonical correlation analysis (CCA) (Ferreira and Purcell, 2009), iii) multivariate (generalized) linear models (Zhang et al., 2017), including MANOVA (Liu et al., 2012) and mixed models (Casale et al., 2015, Joo et al., 2016, Korte et al., 2012, Zhou and Stephens, 2014) (we could also add here generalized estimating equations (Zhang et al., 2014), as well as non-parametric alternatives such as multivariate distance matrix regression (Anderson, 2001, Zapala and Schork, 2012)) and iv) Bayesian approaches (Stephens, 2013). Among them, multivariate linear mixed models (mvLMMs) have become very popular, due to their ability to handle relatedness between individuals (i.e. population stratification) (Price et al., 2010).

A potential criticism of multivariate approaches is the lack of interpretability, as they do not directly yield the individual phenotypes that are associated with the genetic variants, which usually represents the main interest (Stephens, 2013). This is particularly problematic in methods that deal with linear combinations of phenotypes, such as PCA. In addition, approaches based on multivariate linear models tend to make strong assumptions regarding the distribution of the phenotypes (e.g. multivariate normality is required in MANOVA or mvLMMs), which often do not hold (Monlong et al., 2014). Furthermore, fitting mvLMMs is computationally intensive, becoming unfeasible in very large datasets, despite continuous implementation enhancements (Furlotte and Eskin, 2015, Zhou and Stephens, 2014). Computational inefficiency is also a relevant concern in multivariate distance matrix regression (Anderson, 2001), which relies on permutations to assess significance, and often in Bayesian approaches

(Stephens, 2013). Finally, note that independent variables in the depicted multivariate approaches do not need to be restricted to genotypes, as there may be other predictors of interest (e.g. gene expression, age, gender, etc.).

CHAPTER 1

Genetic effects on splicing across human tissues

sQTL mapping and characterization is paramount to achieve a complete understanding of the mechanisms underlying alternative splicing and its contribution to human phenotypes. However, despite the tissue-specific nature of splicing events, most sQTL analyses to date are restricted to a single tissue or cell type. Moreover, alternative splicing is generally treated as a univariate phenotype, ignoring the strongly correlated structure of the alternative transcript isoforms produced from a gene. In light of this, we have employed a multivariate strategy to leverage the Genotype-Tissue Expression (GTEx) dataset (RNA-seq data over 50 tissue sites across hundreds of deceased donors with available whole genome sequences), generating the most comprehensive sQTL catalogue to date in the human genome. Extensive analyses of GTEx sQTLs provide novel insights into the interplay between the regulation of alternative splicing and transcription, and reveal different mechanisms through which sQTLs may impact splicing patterns. Our results confirm that genetic effects on splicing contribute to human complex traits and diseases to an extent comparable with regulatory variation controlling gene expression.

Garrido-Martín, D., Borsari, B., Calvo, M., Reverter, F., Guigó R. (2019) Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. *Submitted*

Garrido-Martín D, Borsari B, Calvo M, Reverter F, Guigó R. Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. Nature communications. 2021;12(1):727–727. DOI: 10.1038/s41467-020-20578-2

[*] Correspondence should be addressed to E-mail: roderic.guigo@crg.eu (Roderic Guigó)

CHAPTER 2

Alternative splicing visualization in large RNA-seq datasets

During our study of genetic effects on alternative splicing (Chapter 1), we faced the problem of visualizing splicing events in a large dataset such as GTEx. Despite the popularity of *sashimi* plots –which display read coverage along a genomic region plus the splice junctions supported by RNA-seq–, current implementations represent each sample on a separate line, while most GTEx tissues have sample sizes greater than 100. Indeed, the increasing availability of large-scale RNA-seq datasets (GTEx, ENCODE, etc.), as well as additional flaws of common tools to generate sashimi plots (annotation-dependence, inefficient implementations, etc.), made us realize the need of a new implementation of the *sashimi* plot, able to deal with a larger number of samples, while offering enhanced visualization. Hence, we developed a command-line tool that presents several advantages over its predecessors: it is annotation-independent, ii) it is a fast, standalone, command line tool, iii) it scales for a large number of samples thanks to several aggregation methods, and iv) it can compress the length of uninformative regions without splicing events.

Garrido-Martín, D., Palumbo, E., Guigó R. and Breschi, A. (2018) ggsashimi: Sashimi plot revised for browser- and annotation-independent splicing visualization. *PLoS computational biology*, 14(8), e1006360.

Garrido-Martín D, Palumbo E, Guigó R, Breschi A. ggsashimi: Sashimi plot revised for browser- and annotation-independent splicing visualization. PLoS computational biology. 2018;14(8):e1006360–e1006360. DOI: 10.1371/journal.pcbi.1006360

CHAPTER 3

A fast non-parametric test of association for multivariate phenotypes

Our work with Anderson test in the context of sQTL mapping (Chapter 1) revealed its large potential to identify genetic effects on multivariate phenotypes. A key feature of this approach is the lack of assumptions on the distribution of the response variables (i.e. the phenotypes). This is likely to result in an increased power to detect genetic associations with many biological traits, which often do not follow known (e.g. Normal, Poisson, etc.) distributions. However, Anderson test relies on permutations to assess significance in complex designs (i.e. more than one predictor in the model). This supposes a major drawback for its usage, especially given the large size and complexity of current datasets, as well as the common presence of confounders. To overcome this limitation, we invested considerable efforts in obtaining the theoretical distribution of the Anderson test statistic. Our result, described in this Chapter, enables to compute asymptotic p values, avoiding the need of permutations and dramatically reducing the computation time. We further illustrate the performance of our method using simulated and real datasets, and present it as a valuable alternative for multivariate GWAS and QTL mapping analyses.

Garrido-Martín, D., Calvo, M., Reverter, F., Guigó R. (2019) A fast non-parametric test of association for multivariate phenotypes. *In preparation*

A fast non-parametric test of association for multivariate phenotypes

Diego Garrido-Martín[1], Miquel Calvo[2], Ferran Reverter[2] and Roderic Guigó*[1,3]

[1]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Catalonia, Spain
[2]Section of Statistics, Faculty of Biology, Universitat de Barcelona (UB), Av. Diagonal 643, Barcelona 08028, Spain
[3]Universitat Pompeu Fabra (UPF), Barcelona, Catalonia, Spain

*Correspondence should be addressed to E-mail: roderic.guigo@crg.eu (Roderic Guigó)

Abstract

The increasing availability of phenotypic data in large cohorts of genotyped individuals requires efficient multivariate methods to identify genetic effects on multiple traits. In this context, Anderson test offers a powerful non-parametric approach. However, it relies on permutations to assess significance in complex designs, which discourages its usage in current large-scale datasets. Here, we derive the limiting distribution of the Anderson test statistic for complex designs and Euclidean distances, providing a framework for the fast computation of asymptotic *p* values. Using a comprehensive set of simulations, we show that the asymptotic test presents controlled type I error rates and high power. We illustrate the applicability of our method by performing context-specific splicing quantitative trait loci mapping across GTEx tissues, and a genome-wide association study of the MRI-derived volumes of hippocampal subfields in the ADNI cohort.

Introduction

In the past years, the availability of human deep phenotype data has dramatically increased[1]. Moreover, recent technological developments have enabled genome-wide profiling of a wide variety of molecular traits, in addition to genotypes[2–4]. However, most genome-wide association studies (GWAS) and quantitative trait loci (QTL) mapping analyses test for association with genetic variants a single trait at a time, even when multiple phenotypes are available[5–8]. In this context, multivariate approaches present several advantages over standard univariate analysis.

Indeed, many human traits share genetic and environmental influences, which may be reflected in their correlation structure[9]. Hence, multivariate analysis offers increased statistical power to detect genetic associations[10,11]. The multivariate setting is particularly suitable to investigate pleiotropy, pervasive in the human genome[12], and it can be advantageous even when only a small subset of the phenotypes are affected by the genetic variants tested[13]. Additionally, it provides a unique framework to study the molecular mechanisms through which genetic variants act, allowing joint analyses across multiple phenotypic layers[14]. Moreover, when the same trait is measured in different conditions (e.g. across tissues or environments), multivariate analyses can be used to characterize context-dependent genetic effects[15,16]. This also applies to longitudinal studies, where a trait is measured over time[17]. Last but not least, as multivariate analyses require fewer individual tests, the multiple testing burden is reduced.

The most widely used approaches for multivariate association testing include methods that build linear combinations of phenotypes, such as principal component analysis (PCA)[18] or canonical correlation analysis (CCA)[19], as well as multivariate linear models, which comprise Multivariate Analysis of Variance (MANOVA)[20], linear mixed models (mvLMMs)[9,21–23] or generalized linear models (mvGLMs)[24]. However, while the former hinder interpretability, the latter often make strong assumptions about the distribution of the dependent variables (e.g. MANOVA and mvLMMs require multivariate normality). Furthermore, although mvLMMs are able to handle relatedness among individuals (i.e. population stratification)[25], fitting them is computationally intensive and may be slow in large datasets, despite continuous implementation enhancements[21,26]. Thus, the development of a fast, non-parametric multivariate alternative, suitable for GWAS and QTL mapping, would be highly valuable.

Anderson[27] introduced a distance approach in order to extend the univariate factorial linear model to several dimensions without requiring a known probability distribution of the dependent variables. The hypothesis of no-effects was tested by a permutation procedure based on a *pseudo-F* statistic, where the sums of squares in ANOVA are replaced by sums of inter-distances between the individuals. This approach was employed to study alternative splicing (AS) across several human populations, using the Hellinger distance between vectors of relative AS isoform abundances as dissimilarity metric[28].

While Anderson's multivariate approach remains conceptually appealing, the increased size and complexity of recent datasets requires a precision on the *p* values that turns the permutational procedure impractical. Anderson and Robinson[29] showed the asymptotic distribution of the numerator of the test statistic in the context of a one-way fixed design. This approach was implemented in `sQTLseekeR` and employed to identify genetic effects on splicing[30]. Here we show the limiting distribution of the Anderson statistic for more complex designs in the Euclidean distance case. Our result also holds after any transformation of the data that preserves the independence of the samples (e.g. square root transformation plus Euclidean distance, equivalent to Hellinger distance between proportions).

In practice, in a typical GWAS setting, e.g. 5 traits measured in 10,000 individuals tested *vs* the genotype plus two additional covariates, our result can offer a 10^6-fold reduction of the computation time of a single test with respect to 10^4 permutations, while achieving *p* values down to 10^{-14}.

Through a comprehensive set of simulations, we evaluated the type I error and power of the asymptotic test in comparison with MANOVA. We developed a procedure to compute asymptotic *p* values, that we implemented in the `mlm` R package, available at `https://github.com/dgarrimar/mlm`. We applied the asymptotic test to real human data in two different scenarios: i) condition-specific splicing QTL mapping across tissues using GTEx data, and ii) a GWAS of the volumes of hippocampal subfields using ADNI data.

Methods

Anderson test statistic

Consider a $n \times q$ matrix of response variables, $\mathbf{Y} = (y_{ij})$, corresponding to n independent observations of a vector of q random variables, and a second matrix $\mathbf{X}$, a $n \times p$ matrix of p predictor variables. Anderson proposed a geometric, permutation-based method in order to study the effects of $\mathbf{X}$[27]. This approach uses a $n \times n$ suitable distance matrix $\mathbf{D}$ between the n individuals based on the $\mathbf{Y}$ outcomes, allowing the computation of a *pseudo-F* statistic. If $\mathbf{D}$ is computed using the Euclidean distance, some properties can be studied in the context of the standard multivariate multiple linear regression (MMR). See Appendix 1 for further details. The aim of MMR is to regress $\mathbf{Y}$ on $\mathbf{X}$ following the model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U} \tag{1}$$

where $\boldsymbol{\beta}$ is a $p \times q$ matrix of parameters and $\mathbf{U}$ a matrix of random errors. MMR generalizes some of the multiple regression results, for instance, the ordinary least squares (OLS) estimation of the $\hat{\beta}$ parameters is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \tag{2}$$

provided that $\mathbf{X}$ has full rank. $\hat{\boldsymbol{\beta}}$ is the solution of the q simultaneous multiple linear regressions on each column of $\mathbf{Y}$, and each column in $\hat{\boldsymbol{\beta}}$ corresponds exactly to the individual multiple regression of the associated column in $\mathbf{Y}$.

If the null hypothesis of interest is $\boldsymbol{\beta} = \mathbf{0}$ (all the coefficients of every variable are null), the Anderson test statistic, in the Euclidean distance case, is equivalent to the following MMR statistic:

$$\tilde{\mathrm{F}} = \frac{\mathrm{tr}(\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}})/\mathrm{rank}(\mathbf{H})}{\mathrm{tr}(\mathbf{Y}^{\mathrm{T}}\mathbf{Y} - \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}})/\mathrm{rank}(\mathbf{I}-\mathbf{H})} = \frac{\mathrm{tr}\left\{\mathbf{Y}^{\mathrm{T}}\mathbf{H}\mathbf{Y}\right\}/\mathrm{rank}(\mathbf{H})}{\mathrm{tr}\left\{\mathbf{Y}^{\mathrm{T}}(\mathbf{I}-\mathbf{H})\mathbf{Y}\right\}/\mathrm{rank}(\mathbf{I}-\mathbf{H})} \tag{3}$$

where $\mathbf{H}$ denotes the usual projection matrix (or *hat* matrix) in linear models, that is:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}} \tag{4}$$

If model (1) includes several predictors (main factors, interactions, continuous covariates), it may be of interest to test the hypothesis $\boldsymbol{\beta}_0 = \mathbf{0}$ about a subset of parameters. In this case the test statistic becomes:

$$\tilde{\mathrm{F}} = \frac{\mathrm{tr}\left\{\mathbf{Y}^{\mathrm{T}}(\mathbf{H}-\mathbf{H}_0)\mathbf{Y}\right\}/\mathrm{rank}(\mathbf{H}-\mathbf{H}_0)}{\mathrm{tr}\left\{\mathbf{Y}^{\mathrm{T}}(\mathbf{I}-\mathbf{H})\mathbf{Y}\right\}/\mathrm{rank}(\mathbf{I}-\mathbf{H})} \tag{5}$$

where $\mathbf{H}_0$ is the hat matrix corresponding to the design matrix $\mathbf{X}_0$, which is $\mathbf{X}$ without the columns associated to the subset of coefficients $\boldsymbol{\beta}_0$.

Null distribution of the test statistic under permutation

The empirical null distribution of the Anderson test statistic ($\tilde{\mathrm{F}}$) can be characterized using permutations, that is, by recomputing $\tilde{\mathrm{F}}$ after random shuffling of the data. Then, *p* values are obtained by comparing the observed value of $\tilde{\mathrm{F}}$ to the distribution of permuted $\tilde{\mathrm{F}}^{\pi}$ values. The only assumption of the permutation test is that the observations are exchangeable under the null hypothesis (H_0). In complex designs, however, it is unclear how to ensure this in order to obtain an exact test (i.e. a test with a type I error rate exactly equal to the significance level selected a priori)[31].

In the case of a model with two main factors (i.e. A and B) and an interaction term (i.e. AB), only under the global null hypothesis observations are exchangeable between the different levels of A and B. However, in the presence of main effects (A or B under the alternative hypothesis, H_1) observations are exchangeable only within levels of other main factors. For example, if B is under H_1, an exact permutation test for A that controls for the effect of B requires permutations to be restricted to the levels of B. In this scenario, unrestricted permutation of raw data yields an approximate test. See ([31]) for a detailed discussion. Notably, there is no exact test for the interaction term controlling for the effect of both main factors, as here the only possible value of the permuted test statistic is the one obtained on the original data.

Computation of asymptotic *p* values

As we describe in this work, the null distribution of the test statistic converges to a weighted sum of independent chi-square variables (see Results). To compute asymptotic *p* values, we can rely on its cumulative density function (CDF). Although such distribution does not have a closed form, it can be approximated with high accuracy, and several approaches are available. We focused on three of these algorithms: Imhof[32], Davies[33] and Farebrother[34], as implemented in the `CompQuadForm` R package[35]. While the first two rely on the numerical inversion of the characteristic function, the third takes advantage of the fact that the CDF can be expressed as an infinite series of central chi-square distributions[35].

To compare their performance, we simulated sets of weights ($\lambda_j \sim U(0, 1)$, with $j \in \{1, \ldots, q\}$), considering different values of q and degrees of freedom for the chi-square distribution. Then, for a range of values of the test statistic we evaluated the obtained *p* values and the computation time. Note that any set of weights can be scaled to obtain values in the interval [0,1], and that scaling both the weights and the test statistic results in identical theoretical *p* values. The typical behaviour of each algorithm is shown in Fig. S1.

Overall, we found almost identical *p* values between the three methods down to a precision of

10^{-10}. However, while Farebrother *p* values decreased monotonically with the value of the test statistic, down to the precision limit ($\approx 10^{-14}$), Imhof and Davies *p* values below 10^{-10} displayed an erratic behaviour, with values ≤ 0. In addition, regarding speed, Farebrother outperformed Imhof and Davies in the majority of scenarios. Hence, we selected the Farebrother method for *p* value calculation. Only when $\lambda_j / \sum\limits_{j=1}^{q} \lambda_j \approx 0$, for one or more j in $\{1, \ldots, q\}$, this approach displayed longer running times, especially for large values of the test statistic. To solve this problem, we dropped the weights for which $\lambda_j / \sum\limits_{j=1}^{q} \lambda_j < t$. We tried several values of t, and found that $t =$ 10^{-3} provides a good balance between speed and accuracy.

`mlm` R package

We have implemented the asymptotic Anderson test in the `mlm` R package, available at `https://github.com/dgarrimar/mlm`. `mlm` enables asymptotic *p* value calculation for the predictors in user-defined MMR models, using the Farebrother method. It allows to select different types of Sums of Squares (I, II or III), as well as logarithm and square root data transformations.

Monte Carlo simulation study

Models

We considered two different MMR models. In (6), the response variables are regressed on two categorical predictors (i.e. factors A and B) and their interaction (AB):

$$\mathbf{Y}_{klm} = \boldsymbol{\mu} + \boldsymbol{\alpha}_k + \boldsymbol{\beta}_l + \boldsymbol{\alpha\beta}_{kl} + \boldsymbol{\epsilon}_{klm} \tag{6}$$

where $\mathbf{Y}_{klm}$ is the q-dimension vector corresponding to the k, l, m sample, $\boldsymbol{\mu}$ the vector of means, $\boldsymbol{\alpha}_k$ the q vector of parameters (one component per response variable) associated to level k of factor A and, similarly, $\boldsymbol{\beta}_l$ is the vector of parameters of level l of factor B, $\boldsymbol{\alpha\beta}_{kl}$ the vector corresponding to level k, l of the interaction AB, and $\boldsymbol{\epsilon}_{klm}$ the vector of random errors.

In (7), the response variables are regressed on a numerical predictor (X) and a factor (A):

$$\mathbf{Y}_{lm} = \boldsymbol{\mu} + \boldsymbol{\alpha}_l + x_{lm}\boldsymbol{\beta} + \boldsymbol{\epsilon}_{lm} \tag{7}$$

where $\mathbf{Y}_{lm}$ is the q-dimension vector corresponding to the m-th sample in the level l of factor A, $\boldsymbol{\mu}$ the vector of intercepts, $\boldsymbol{\alpha}_l$ the q vector of parameters of this level, $\boldsymbol{\beta}$ is the vector of q regression coefficients for covariate X, x_{lm} the observed value (scalar) of the covariate for this sample, and $\boldsymbol{\epsilon}_{lm}$ the vector of random errors.

In both scenarios, we considered balanced and unbalanced designs, e.g. for factor A (analogous for B):

$$n_{A(l)} = \begin{cases} \frac{nu}{(u+a-1)} & l = 1 \\ \frac{n}{(u+a-1)} & 2 \leq l \leq a \end{cases}$$

where $n_{A(l)}$ is the number of samples in level l of factor A, n the total sample size, a the number of levels of factor A, and $u \in \{0.2, 0.5, 1\}$ the degree of imbalance. Values of $n_{A(l)}$ were rounded to integers constrained to $\sum\limits_{l=1}^{a} n_{A(l)} = n$. In practice, $n_{A(l)} \approx un_{A(l)}$, $l \in \{2, \ldots, a\}$. Note that $u = 1$ corresponds to a balanced design.

Data generation

Under the null hypothesis of no association (H_0) between the response variables and the predictors, observations of the former were generated by random sampling from a given multivariate distribution. We considered several distributions, varying the total sample size (n) and the number of response variables (q). In some scenarios, we also considered heteroscedastic situations. Additionally, we simulated the alternative hypothesis (H_1) of one predictor associated to the response variables.

Multivariate normal

We considered first this scenario, as it is assumed in many multivariate linear modeling approaches. Here:

$$\mathbf{Y}_{i\cdot} \sim MVN(\boldsymbol{\mu}, \Sigma)$$

where $\boldsymbol{\mu}$ is the mean and $\Sigma = (\sigma_{jk})$ the covariance matrix of $\mathbf{Y}_{i(l)}$. Under H_0, we set $\boldsymbol{\mu} = \mathbf{0}$. We selected unit variances ($\sigma_{jj} = 1$, $\forall i \in \{1, \ldots, q\}$), and $\sigma_{jk} = c$, where $c \in \{0, 0.2, 0.5, 0.8\}$ is the correlation between any pair of response variables. With these values of c we ensured that Σ is positive definite. We used $\boldsymbol{\mu}$ and Σ as inputs for the `mvrnorm` function in the `MASS` R package[36].

Additionally, we generated observations of the response variables under the alternative hypothesis (H_1) of association with a given predictor X. When X was a factor:

$$\mathbf{Y}_{i(l)\cdot} \sim \begin{cases} MVN(\boldsymbol{\Delta}, \Sigma) & l = 1 \\ MVN(-\boldsymbol{\Delta}, \Sigma) & l = 2 \\ MVN(\mathbf{0}, \Sigma) & l > 2 \end{cases}$$

where $\mathbf{Y}_{i(l)\cdot}$ is any observation of $\mathbf{Y}$ in the level l of factor X, and $\boldsymbol{\Delta} = \Delta\mathbf{1}$, with $\Delta \in \mathbb{R}$. We considered values of Δ ranging from 0 to 0.2 to cover the entire power range in subsequent power analyses. Here two levels change to ensure $\boldsymbol{\mu} = \mathbf{0}$. Observations of **Y** corresponding to the remaining levels of X were simulated under H_0. Note that when X is an interaction term, four levels (rather than two) need to change in opposite directions to ensure $\boldsymbol{\mu} = \mathbf{0}$.

Moreover, when X was a factor, we simulated heteroscedastic conditions as follows:

$$Var(y_{i(l)j}) = \begin{cases} h\sigma_{jj} & l = 1 \\ \sigma_{jj} & l > 1 \end{cases}$$

where σ_{jj} is the variance of any response variable and $h \in \{1, 2, 5, 10\}$ is the degree of heteroscedasticity. Note that $h = 1$ corresponds to a homoscedastic situation.

When X was numerical, we simulated $\mathbf{Y}_{\cdot j}$ so that:

$$Cor(y_{ij}, x_i) = \begin{cases} r & j = 1 \\ 0 & 2 \leq j \leq q \end{cases}$$

while ensuring that $\mathbf{Y}_{i\cdot} \sim MVN(\mathbf{0}, \Sigma)$. We considered values of r ranging from 0 to 0.4 to cover the whole power range in further power analyses.

Vectors of proportions

Our interest in this scenario is related to our previous work with multivariate proportion data for the study of alternative splicing[28,30]. It corresponds to generate points in the $q-1$ simplex. Here:

$$\mathbf{Y}_{i\cdot} \sim S(\boldsymbol{p}, \sigma_g)$$

where $\boldsymbol{p}$ is a given point in the $q-1$ simplex and σ_g is the standard deviation of the generator model. We obtained $\boldsymbol{p}$ so that $p_1 = \frac{L}{(q+L-1)}$ and $p_i = \frac{L}{(q+L-1)}$, $\forall i \in \{2, \ldots, q\}$, with $L \in \{1, 2, \ldots, 10\}$. Note that $L = 1$ corresponds to the center of the simplex, while $L > 1$ to locations that range from the center of the simplex to one of its vertices, $\boldsymbol{e_1} = (1, 0, \ldots, 0)$. To generate observations in the $q-1$ simplex with certain variability around $\boldsymbol{p}$, ensuring that $E(\mathbf{Y}_{i\cdot}) = \boldsymbol{p}$, we implemented an approach that performs random displacements of size δ from $\boldsymbol{p}$ towards the simplex vertices, with $\delta \sim N(0, \sigma_g)$ (see Appendix 2).

Additionally, we generated observations of the response variables under the alternative hypothesis (H_1) of association with a given predictor X. When X was a factor:

$$\mathbf{Y}_{i(l)\cdot} \sim \begin{cases} S(\boldsymbol{p}_{\Delta}, \sigma_g) & l = 1 \\ S(\boldsymbol{p}_{-\Delta}, \sigma_g) & l = 2 \\ S(\boldsymbol{p}, \sigma_g) & l > 2 \end{cases}$$

where $\mathbf{Y}_{i(l)\cdot}$ are the observations of $\mathbf{Y}$ in level l of factor X, and $\boldsymbol{p}_\Delta$ is obtained from $\boldsymbol{p}$ advancing along the geodesic that joins $\boldsymbol{p}$ with the simplex vertex $\boldsymbol{e}_1 = (1, 0 \ldots, 0)$. This displacement depends on a parameter Δ (see Appendix 2). We considered values of Δ ranging from 0 to 0.02, which comprised the whole power range in subsequent power analyses. Here two levels change to ensure $E(\mathbf{Y}_{i\cdot}) = \boldsymbol{p}$. The remaining levels of X were simulated under H_0. Note that when X is an interaction term, four levels (rather than two) need to change in opposite directions to ensure $E(\mathbf{Y}_{i\cdot}) = \boldsymbol{p}$.

Moreover, when X was a factor, we simulated heteroscedastic conditions as follows:

$$\mathbf{Y}_{i(l)\cdot} \sim \begin{cases} S(\boldsymbol{p}, \sigma_g h) & l = 1 \\ S(\boldsymbol{p}, \sigma_g) & l > 1 \end{cases}$$

where $\mathbf{Y}_{i(l)\cdot}$ are the observations of $\mathbf{Y}$ in level l of factor X, σ_g is the standard deviation of the data generator model (see Appendix 2), and $h \in \{1, 2, 5, 10\}$ is the degree of heteroscedasticity. $h = 1$ corresponds to a homoscedastic scenario.

When X was numerical, we simulated $\mathbf{Y}_{\cdot j}$ so that $Cor(y_{i1}, x_i) = r$, with $r \in [0, 0.4]$, as in the multivariate normal scenario. In practice, we first simulated $Cor(y_{i1}, x_i) = 1$ and then added random noise to achieve values of r in the desired range (see Appendix 2).

In this scenario, once $\mathbf{Y}$ was obtained, we applied a square root transformation. This is equivalent to using the Hellinger distance, instead of the Euclidean distance.

Gaussian copula

In this scenario:

$$\mathbf{Y}_{i\cdot} \sim C(\Sigma)$$

where Σ is the correlation matrix of $\mathbf{Y}$. Taking unit variances, we generated several correlation structures as in the MVN scenario. We used the `normalCopula` function from the `copula` R package[37] to generate $\mathbf{Y}$, which was eventually centered. We obtained heteroscedastic conditions as in the MVN scenario. We also simulated the alternative hypothesis (H_1) of association with a given predictor X. When X was a factor:

$$\mathbf{Y}_{i(l)\cdot} = \begin{cases} \mathbf{Y}_{i(l)\cdot} + \boldsymbol{\Delta} & l = 1 \\ \mathbf{Y}_{i(l)\cdot} & l > 1 \end{cases}$$

where $\mathbf{Y}_{i(l)\cdot}$ is any observation of $\mathbf{Y}$ in the level l of factor X, and $\boldsymbol{\Delta} = \Delta\mathbf{1}$, with $\Delta \in \mathbb{R}$. We considered values of Δ ranging from 0 to 0.2, and two levels of X (four when X was an interaction term) changed to ensure $E(\mathbf{Y}_{i\cdot}) = \mathbf{0}$. The remaining levels of X were generated under H_0. When

X was numerical, we obtained $\mathbf{Y}_{\cdot j}$ as in the MVN scenario.

Multinomial

In this scenario:

$$\mathbf{Y}_{i\cdot} \sim MN(N, \boldsymbol{p})$$

where N and $\boldsymbol{p}$ are the number of trials and the vector of event probabilities, respectively. We simulated $N \sim Poisson(\lambda)$, with $\lambda =$ 100, and $\boldsymbol{p}$ as in the multivariate proportion scenario (see above). We used N and $\boldsymbol{p}$ as input for the `rmultinom` function from the `stats` R package[38].

Additionally, we took advantage of the data generation schema developed for the multivariate proportion scenario to obtain observations of the response variables under the alternative hypothesis (H_1) of association with a factor X as follows:

$$\mathbf{Y}_{i(l)\cdot} \sim \begin{cases} MN(N, \boldsymbol{p}_{\Delta}) & l = 1 \\ MN(N, \boldsymbol{p}_{-\Delta}) & l = 2 \\ MN(N, \boldsymbol{p}) & l > 2 \end{cases}$$

where $\mathbf{Y}_{i(l)\cdot}$ are the observations of $\mathbf{Y}$ in level l of factor X, and $\boldsymbol{p}_{\Delta}$ is obtained from $\boldsymbol{p}$ as in the multivariate proportion scenario (see also Appendix 2). Likewise, we considered values of Δ ranging from 0 to 0.02, and two levels of X (four when X was an interaction term) changed to ensure $E(\mathbf{Y}_{i\cdot}) = N\boldsymbol{p}$. The remaining levels of X were simulated under H_0. Once $\mathbf{Y}$ was obtained, we applied a logarithm transformation.

Evaluation of type I error and power

We selected a significance level of $\alpha =$ 0.05. For each combination of conditions, we simulated $m =$ 10,000 sets of response variables ($\mathbf{Y}$). Under H_0, we evaluated the type I error for the association between $\mathbf{Y}$ and the predictor X, for asymptotic Anderson test and MANOVA (Pillai’s trace), as follows:

$$Type\, I\, error = \frac{\sum_{i=1}^{m} \boldsymbol{I}(p \leq \alpha)}{m}$$

where p is the *p* value of the association and $\boldsymbol{I}$ the indicator function. We employed an analogous setting to compute power when simulating under H_1.

Implementation

All the simulations were performed in R v3.5.2[38]. Asymptotic *p* values were computed using `mlm` (`https://github.com/dgarrimar/mlm`) with default parameters. MANOVA *p* values were computed using the `manova` method in `stats`, with default parameters. For parallelization and portability purposes, we embedded the R code in a pipeline built using `nextflow` v0.27.2, a framework for computational workflows[39]. We also used `Docker` container technology (`https://www.docker.com`) to ensure the reproducibility of our results.

Condition-specific splicing QTL mapping

GTEx data

Transcript expression (transcripts per million, TPM) and variant calls were obtained from the V7 release of the Genotype-Tissue Expression (GTEx) Project (dbGaP accession *phs000424.v7.p2*). These correspond to 10,361 samples from 620 deceased donors with both RNA-seq in up to 53 tissues and Whole Genome Sequencing (WGS) data available. Metadata at donor and sample level was also retrieved.

In GTEx, RNA-seq reads are aligned to the human reference genome (build hg19/GRCh37) using `STAR`[40] v2.4.2a, based on the GENCODE v19 annotation (`https://www.gencodegenes.org/human/release_19.html`). Transcript-level quantifications are obtained with `RSEM`[41] v1.2.22. WGS reads are aligned with `BWA-MEM` (`http://bio-bwa.sourceforge.net`) after base quality score recalibration and local realignment at known indels using Picard (`http://picard.sourceforge.net`). Joint variant calling across all samples is performed using `GATK HaplotypeCaller` v3.4 (`https://software.broadinstitute.org/gatk/documentation/tooldocs`). Further details on GTEx data pre-processing and QC pipelines can be found at the GTEx Portal (`https://gtexportal.org`).

Ethnicity- and gender-specific sQTL mapping

For *cis* condition-specific sQTL (cs-sQTL) mapping, we employed a slightly modified version of `sQTLseekeR2`, which implements asymptotic Anderson test to assess the significance of the association between alternative splicing (AS) on one side, and the genotype, the condition of interest, and the interaction between the two on the other. In `sQTLseekeR2` (`https://github.com/dgarrimar/sQTLseekeR2`), AS is modeled as a multivariate outcome, formed by the relative abundances of the alternative transcript isoforms of a gene (*splicing ratios*), after a square root transformation. Adapted `sQTLseekeR2` (nominal pass) was run within a containerized `nextflow` pipeline.

We performed two separate cs-sQTL mapping studies, considering two different conditions: ethnicity (86.7% european american (EA), 11.7% african american (AA) individuals; we discarded other ethnicities accounting altogether for less than 2% of individuals) and gender (63.9% male, 36.1% female individuals). Both genotype and condition were treated as categorical variables. Donor ischemic time, gender and age, as well as the sample RIN (RNA integrity number) and genotyping platform were regressed out from the *splicing ratios* prior to association testing.

We focused on 48 tissues with sample size greater or equal to 70 (for gender-specific sQTL mapping we further discarded 4 tissues from reproductive organs: testis, uterus, ovary and vagina). The *cis* window was defined as the gene body plus 5Kb upstream and downstream the gene boundaries. We considered genes expressed $\geq$ 1 TPM in at least 80% of the samples (samples with lower gene expression were removed from the analysis of the gene), with at least two isoforms and a minimum isoform expression of 0.1 TPM (transcripts with lower expression in all samples were removed). We analyzed only biallelic SNPs and short *indels* (autosomal + X) with MAF $\geq$ 0.01. To ensure reliable results, we required at least 10 samples per observed level of the interaction between the genotype and the context.

In total, 304,101 variants and 12,244 genes were analyzed in our ethnicity-specific sQTL mapping study, while 1,573,134 variants and 14,428 genes were analyzed in our gender-specific sQTL mapping study. As our test statistic is sensitive to the heterogeneity of the *splicing ratios*’ variability between the levels of the interaction term, a permutation-based (10^4 permutations) multivariate homoscedasticity test[42] was also performed for each gene-variant pair (option `--svqtl` in `sQTLseekeR2`). Pairs failing this test after multiple testing correction by `eigenMT` (see below) were not reported as significant cs-sQTLs.

Multiple testing correction

To correct for the fact that multiple variants are tested per gene, we used `eigenMT`[43]. `eigenMT` estimates the effective number of independent tests (M_{eff}) per gene, considering the LD structure among the tested variants. M_{eff} is then used instead of the total number of tests (M) in Bonferroni correction. This allows to compute a gene-level *p* value (corresponding to the smallest –corrected– *p* value per gene). To account for the fact that multiple genes are tested genome-wide, we applied Benjamini-Hochberg false discovery rate (FDR) to gene-level *p* values[43]. We set a FDR threshold of 0.1.

GWAS of the volumes of hippocampal subfields

ADNI data

Volumes of hippocampal subfields (UCSC dataset) and variant calls were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI 1/GO/2, `http://adni.loni.usc.edu`), corresponding to 591 patients with both cross-sectional brain segmentation derived from Magnetic Resonance Imaging (MRI) and WGS data available. Out of them, 555 individuals displayed high-quality (i.e. `Pass`) hippocampal segmentation in both left and right hemispheres. These included 169 cognitive normal (CN), 230 early and 110 late mild cognitive impairment (E/L-MCI), and 46 Alzheimer's disease (AD) individuals. Metadata at patient level was also retrieved.

Within ADNI, cross-sectional brain segmentation (UCSC dataset) was obtained with `FreeSurfer` v5.1 (`http://surfer.nmr.mgh.harvard.edu`), using the 2010 Desikan-Killany atlas, on T1-weighted images acquired at 3 Tesla. As for genotyping, WGS reads were aligned to the human reference genome (build hg19/GRCh37) with `BWA-MEM`, followed by base quality score recalibration, local realignment at known indels and variant calling (`GATK` v3.1). Further information on ADNI MRI image acquisition, genotyping and QC pipelines can be found at `http://adni.loni.usc.edu/`.

Genome-wide association analysis

The multivariate phenotype of interest were the volumes of the following hippocampal subfields: Cornu Ammonis (CA)1, CA2-3, CA4-Dentate Gyrus (DG), fimbria, hippocampal fissure, subiculum, presubiculum and tail. We obtained the total volume of each subfield by summing its corresponding volume in left and right hemispheres. We selected patient's age, gender, years of education, intracranial volume (ICV) and APOE-ε4 allele dosage as relevant covariates. We analyzed only biallelic SNPs and short *indels* (autosomal + X) with MAF $\geq$ 0.05 and at least 5 individuals per observed genotype group.

We used `mlm` (`https://github.com/dgarrimar/mlm`) with default parameters to test for association between genetic variants and the logarithm-transformed volumes of the eight hippocampal subfields. We defined a model that included the selected covariates plus the genotype. Except for gender, all predictors were treated as continuous variables. In total, 5,486,810 variants were tested for association. The analysis was run within a containerized `nextflow` pipeline. We adopted the common $5 \cdot 10^{-8}$ threshold for genome-wide significance. We also applied a permutation-based (10^8 permutations) multivariate homoscedasticity test[42] for each variant. Variants significant at genome-wide level according to this test were excluded from further analyses.

Results

The asymptotic null distribution of the Anderson test statistic

In order to ensure the convergence of the test statistic in (3), it is necessary to impose certain conditions to model (1). For instance, the rows of **U** must be independent with the same $q \times q$ covariance matrix denoted by Σ, thus $cov(\mathbf{U}) = \mathbf{I} \otimes \Sigma$ ($\otimes$ denotes the Kronecker product, **I** the $n \times n$ identity matrix). Theorem 1 in Appendix 1 specifies additional details about the imposed conditions.

Consider the eigenvalue decomposition of $\Sigma = \mathbf{P}\Lambda = \mathrm{diag}(\lambda_j)$. If $\mathrm{vec}(\boldsymbol{\beta}) = \boldsymbol{\beta}_{\mathrm{v}}$ is the vectorized form of the $\boldsymbol{\beta}$ parameter matrix in (1), $\hat{\boldsymbol{\beta}}_{\mathrm{v}}$ the corresponding OLS vector of estimates and $\chi^2_j(p)$ a collection of q independent chi-square variables with p degrees of freedom, then (see Lemma 1 in Appendix 1):

$$(\hat{\boldsymbol{\beta}}_{\mathrm{v}} - \boldsymbol{\beta}_{\mathrm{v}})^{\mathrm{T}}(\mathbf{P}\,\mathbf{P}^{\mathrm{T}}) \otimes (\mathbf{X}^{\mathrm{T}}\mathbf{X})(\hat{\boldsymbol{\beta}}_{\mathrm{v}} - \boldsymbol{\beta}_{\mathrm{v}}) \xrightarrow{d} \sum_{j=1}^{q} \lambda_j \chi^2_j(p) \tag{8}$$

Under the null hypothesis, $\boldsymbol{\beta}_{\mathrm{v}} = \mathbf{0}$, and the limiting distribution in (8) allows to obtain the trace in the numerator of the test statistic in (3):

$$\mathrm{tr}(\mathbf{Y}^{\mathrm{T}}\mathbf{H}\mathbf{Y}) = \hat{\boldsymbol{\beta}}_{\mathrm{v}}^{\mathrm{T}}(\mathbf{P} \otimes (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{\frac{1}{2}})(\mathbf{P}^{\mathrm{T}} \otimes (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{\frac{1}{2}})\hat{\boldsymbol{\beta}}_{\mathrm{v}} \xrightarrow{d} \sum_{j=1}^{q} \lambda_j \chi^2_j(p)$$

Thus, the trace converges to a linear combination of independent chi-square variables where the coefficients are the eigenvalues of the population covariance matrix. In MMR these eigenvalues are estimated by the sample covariance of the residuals, multiplied by $(n-1)/(n-p)$. The denominator in (3) converges in probability to $\sum_{j=1}^{q} \lambda_j$ (see Appendix 1).

If the null hypothesis is about a subset of parameters, Lemma 2 in Appendix 1 proves that the numerator of the test statistic in (5) has the following limiting distribution:

$$\mathrm{tr}\left\{\mathbf{Y}^{\mathrm{T}}(\mathbf{H} - \mathbf{H}_0)\mathbf{Y}\right\} \xrightarrow{d} \sum_{j=1}^{q} \lambda_j \chi^2_j(p - p_0) \tag{9}$$

Again, the trace converges to a linear combination of independent chi-square variables (now with $p - p_0$ degrees of freedom, $p_0 = \mathrm{rank}(\mathbf{H}_0)$), where the coefficients are the eigenvalues of the covariance matrix Σ. These eigenvalues can be estimated in practice by the eigendecomposition of the sample covariance matrix of the residuals of the full model.

The convergence requires mutual independence of the rows of **U**. This condition is guaranteed if the individuals are independently sampled. Therefore, any previous transformation on the rows of **Y** that preserves the independence of the samples has also the limiting distribution described in (9). For instance, this includes the Hellinger distance between proportions, which is the Euclidean distance taking the square root of the values of **Y**.

Comparison between asymptotic and permutational approaches

To evaluate our theoretical results, we first considered the model in (6). The number of levels of A and B selected was 2 and 3, respectively, in a completely crossed, balanced design. We simulated $n = 200$ observations of $q = 3$ response variables. Under the null hypothesis of no association (H_0), $\mathbf{Y}_{i\cdot} \sim MVN(\mathbf{0}, \mathbf{I}_q)$, where $\mathbf{Y}_{i\cdot}$ denotes an observation of the response variables and $\mathbf{I}_q$ the $q \times q$ identity matrix. We generated B under the alternative hypothesis (H_1), so that the observations of $\mathbf{Y}$ in the first level of B had mean $\mathbf{1}$ (see Methods).

We then used our result in (9) to obtain the asymptotic null distribution of the test statistic in (5) for the interaction term ($\tilde{\mathrm{F}}_{AB}$), and compared it with the distribution of permuted $\tilde{\mathrm{F}}^{\pi}_{AB}$ values. Of note, permutations were restricted to occur within the levels of factor B. Provided that A and AB are under H_0, this would correspond to the exact permutation test[31] (see Methods). As shown in Fig. 1, the distribution that we derived matches exactly the one obtained by permutations, even in the upper tail region. We also provide empirical evidence that our theoretical result holds regardless of the distribution of $\mathbf{Y}$ (Fig. S2).

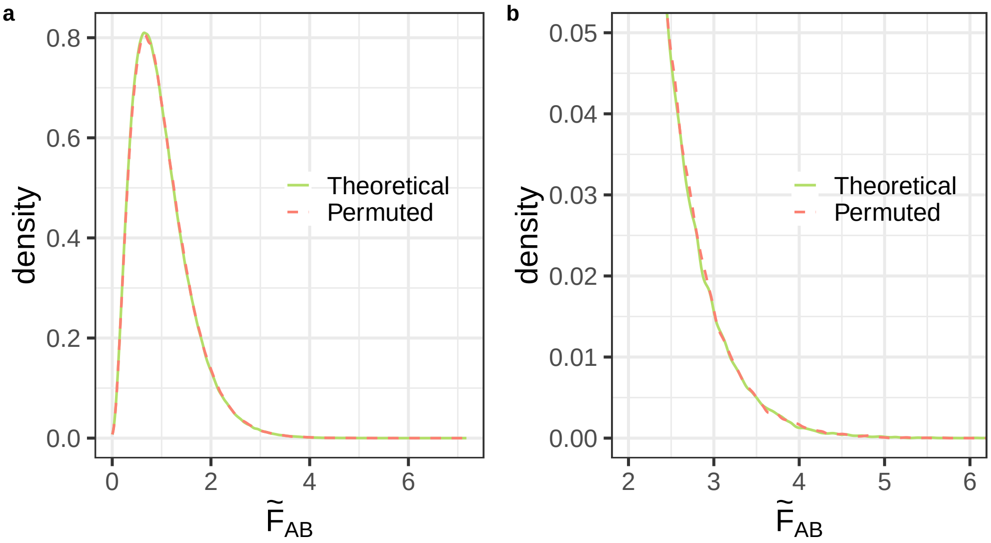

Figure 1. a) Null distribution of the test statistic. Theoretical asymptotic null distribution of the $\tilde{\mathrm{F}}_{AB}$ statistic obtained as proposed in (9), scaled by $\sum_{j=1}^{q} \lambda_j$ (green solid line), compared to the distribution obtained using 10^6 permutations (red dashed line). **b)** Zoom on the upper tail of the distribution.

Recently, McArtor et al.[44] proposed that the coefficients of the linear combination of independent chi-square variables in (9) were indeed the eigenvalues of the centered and squared interdistance matrix. However, these coincide with the eigenvalues of Σ only when there is a single predictor in the model, or under the global null hypothesis (i.e. all the parameters are zero). Indeed, we observed that the distribution suggested by McArtor et al. substantially differs from the one obtained using permutations in the case of partial null hypotheses (i.e. when not all the parameters are 0) (Fig. S3).

To compute *p* values based on the asymptotic null distribution of $\tilde{\mathrm{F}}$ we can rely on its cumulative density function (CDF). Although the CDF of a weighted sum of chi-square random variables does not have a closed form, it can be approximated with high accuracy, and several algorithms are available[32–34]. We compared three of these algorithms, as implemented in the `CompQuadForm` `R` package[35], with the permutation test and selected the Farebrother method, since it performed best in terms of speed and accuracy (see Methods).

Provided that our result in (9) only holds asymptotically, the validity of the proposed asymptotic *p* values will depend on the total sample size (n). In addition, other factors such as the number of response variables (q) or their correlation structure may also have an impact. Thus, we evaluated the relative difference between asymptotic and permutation-based *p* values across a broad range of values of n and q (Fig. 2a). We considered the model depicted in (6), with independent response

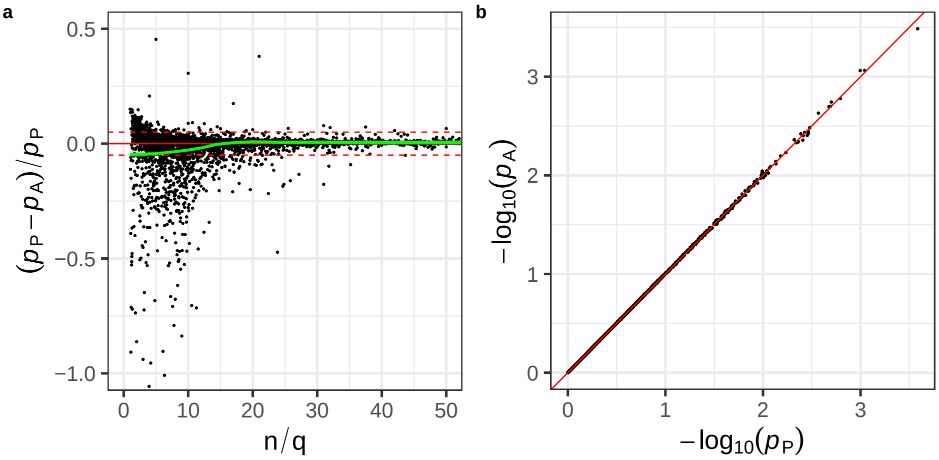

Figure 2. a) Relative bias of asymptotic *p* values *vs* n/q ratio. Relative difference between asymptotic (p_A) and permutation-based (p_P, 10^5 permutations) *p* values for the interaction term (AB) as a function of the ratio between the total sample size and the number of dependent variables (n/q). We considered values of n ranging from 20 to 300, and values of q ranging from 2 to 20. For visualization purposes, we focused on values of $n/q \in$ [0,50] and relative biases $\in$ [-1,0.5]. The horizontal solid red line marks the 0. The horizontal dashed red lines mark the 5% relative bias. A polynomial was fitted to the points using local fitting (LOESS), in order to describe the trend (fit in green, 95% confidence interval in grey). **b)** Comparison of asymptotic and permutation-based *p* values when the asymptotic null holds ($n = 300$, $q = 3$).

variables. Other situations with increasing degrees of correlation (in absolute value) between the response variables would be equivalent to scenarios with fewer independent response variables. When the n/q ratio is small, asymptotic and permutation-based *p* values may differ substantially. As the n/q ratio increases, this bias converges to 0. Overall, when the asymptotic null does not hold, asymptotic *p* values tend to be conservative. As a result, they can still be used without inflated type I error rates (Fig. S5). When the asymptotic null holds (note that this occurs for relatively small values of the n/q ratio, i.e. $n/q \approx$ 20), we observe almost perfect correlation between asymptotic and permutation-based *p* values, even for small *p* values (Fig. 2b). Analogous results were obtained with different distributions of **Y** (Fig. S6).

Simulation study

Type I error

Using the model depicted in (6), we simulated different scenarios varying the distribution and number of response variables. Factor B was simulated under H_1 in all scenarios ($\Delta =$ 0.2 for MVN and C, and $\Delta =$ 0.02 for MN and S, see Methods), while A and AB were simulated under H_0. We considered a relatively large total sample size of $n =$ 500 to ensure that $n/q \geq$ 20. For each combination of conditions, we generated 10,000 datasets, and evaluated the type I error of the test for association with the interaction term (AB) (see Methods). Results are displayed in Table 1. As expected, the distribution of the response variables does not affect the type I error, given that the test does not assume any probabilistic distribution for **Y**. Furthermore, we do not observe any impact of the number of dependent variables (provided that n/q is large enough).

	$q = 2$	$q = 5$	$q = 10$	$q = 15$
MVN	0.049	0.052	0.051	0.049
S	0.051	0.046	0.052	0.048
C	0.052	0.053	0.049	0.048
MN	0.049	0.053	0.053	0.049

Table 1. Type I error rates (AB) for different number (q) and distribution of the dependent variables: MVN = multivariate normal, S = multivariate proportions (simplex), C = gaussian copula, MN = multinomial.

Anderson test, like its parametric counterparts (i.e. ANOVA, MANOVA), is sensitive to differences in multivariate dispersion[27]. To evaluate this, we simulated different degrees of heteroscedasticity (factor B) in balanced and unbalanced designs, with all factors under H_0, and studied the type I error for the interaction term (AB) (see Methods). Results are shown in Table 2.

	$u = 1$	$u = 0.5$	$u = 0.2$
$h = 1$	0.047	0.047	0.047
$h = 2$	0.051	0.111	0.186
$h = 5$	0.068	0.251	0.504
$h = 10$	0.081	0.343	0.683

Table 2. Effects of heteroscedasticity and unbalanced designs on type I error (AB). Here **Y** $\sim MVN(\vec{0}, \mathbf{I}_q)$, $n =$ 500, $q =$ 5. h is the ratio between the variances of the response variables in the first level of factor B and any other level of B. u is the ratio between the sample size of the first level of factor B and any other level. Note that $h = u =$ 1 corresponds to a homoscedastic balanced design (see Methods).

We observed that type I error rates can be substantially inflated when there are differences in the variances of the response variables between the levels of the tested factor. This is particularly problematic in unbalanced designs, while balanced designs are more robust to heteroscedasticity. Overall, the behaviour of Anderson test regarding heteroscedasticity is similar across different distributions of the response variables (Table S1), and analogous to the one displayed by MANOVA (Table S2).

Power

We studied the power of Anderson test using the model depicted in (6). We explored a variety of scenarios, modifying the total sample size, the number of response variables, their distribution and correlation structure. In all cases, factors A and AB were simulated under H_0, while B was simulated under H_1, considering a broad range of Δ values. For each combination of conditions, we simulated 10,000 datasets and evaluated the power of the test for factor B (see Methods).

In Fig. 3, we represent power as a function of Δ, n and q, for different distributions of the response variables. Here we considered uncorrelated variables with unit variance. Overall, we observe a similar behaviour across distributions, and high power to detect small differences (e.g. in a scenario where $\mathbf{Y}_{i\cdot} \sim MVN(\mathbf{0}, \mathbf{I}_q)$, $n =$ 200 and $q =$ 5, a change in the mean of the response variables of $\Delta =$ 0.2 is detected with power 0.96). In addition, power increases with the number of response variables (see also Fig. S4). Note that all the response variables are incremented in Δ (see Methods).

To assess the effect of the correlation structure of the response variables on statistical power, we simulated different scenarios, varying Δ, q and c, where c is the correlation between any two

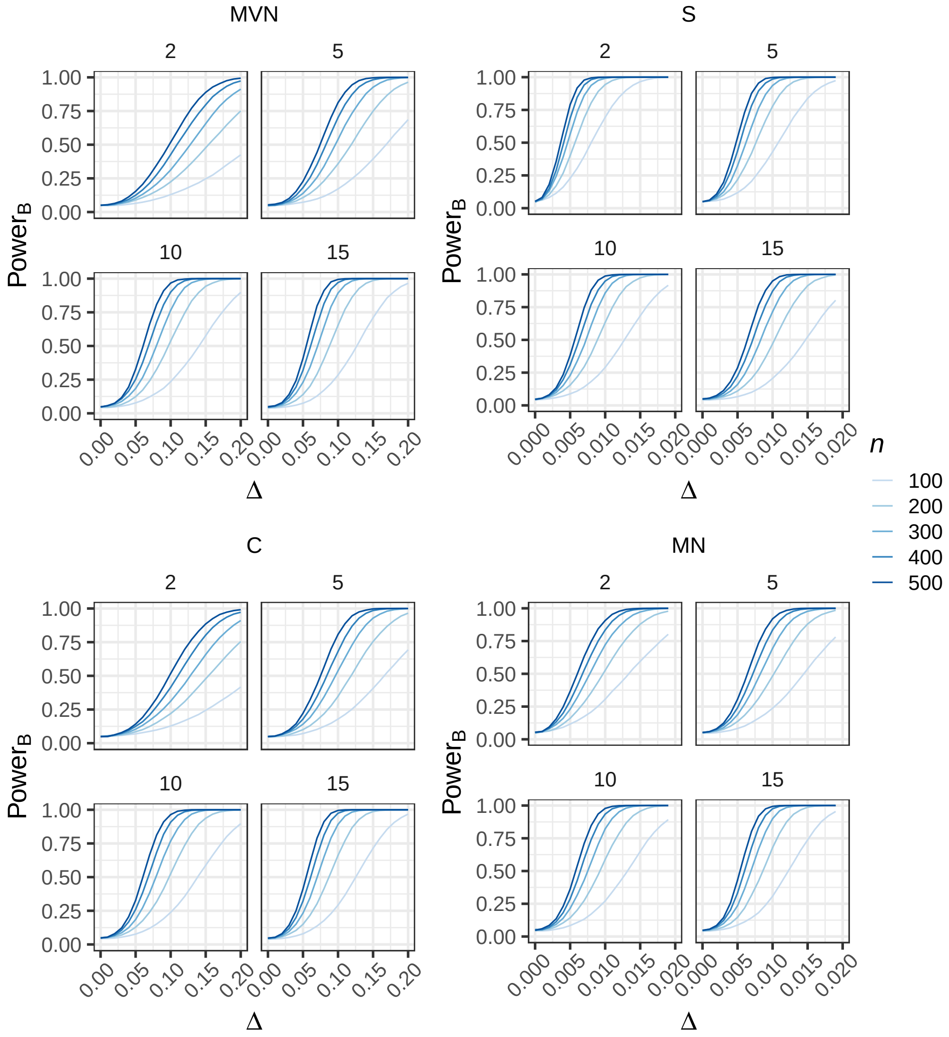

Figure 3. Power curves (factor B) in different scenarios depending on the distribution of Y: MVN = multivariate normal, S = multivariate proportions (simplex), C = gaussian copula, MN = multinomial, . Different values of q (2, 5, 10, 15), n (100, 200, 300, 400) and Δ are evaluated. Note that for MVN and C, $\Delta \in$ [0,0.2], while for S and MN, $\Delta \in$ [0,0.02] (see Methods).

response variables (see Methods). Here $\mathbf{Y}_{i\cdot} \sim MVN(\mathbf{0}, \mathbf{I}_q)$. In addition, n was set to 500 to ensure that the asymptotic null distribution of the test statistic holds. Results are shown in Fig. S7. We observe that power decreases as the correlation between response variables increases. This behaviour is more marked for larger numbers of dependent variables, and holds regardless of the distribution of $\mathbf{Y}$.

When simulating multivariate proportions in the $q-1$ simplex, we were interested in evaluating the effect of the location of the points (L) on statistical power. Hence, we considered several locations, ranging from the center of the simplex to the vicinity of the vertices (see Methods). When $q = 2$, power is identical independently of the location. For $q \geq 3$, we observe increased power in the center of the simplex (Fig. S8).

To evaluate whether the results obtained depend on the model employed, we studied a second model in which the multivariate response was regressed on a numerical covariate (X) and a factor (A), depicted in (7). Here, X follows a $N(0,1)$ distribution $(U(0,1)$ in the case of the S scenario), and A has two levels ($a = 2$, balanced). A was simulated under H_0, whereas X was generated under H_1, controlling its *Pearson* correlation (r) with the first response variable ($\mathbf{Y}_{\cdot 1}$) (see Methods). We performed an identical set of analyses to evaluate power, obtaining analogous results (Fig. S9 to S11). The most relevant difference is that, in this case, H_1 generation involves only one response variable (see Methods), and therefore power decreases with q (e.g. compare Fig. S4 and Fig. S10).

Finally, we compared our test with MANOVA in terms of power across different scenarios. The two approaches display similar power in most scenarios, especially in the case of uncorrelated response variables. When the response variables are correlated, asymptotic Anderson test seems to outperform MANOVA, although this behaviour is reversed in some situations, e.g. when simulating a large number of response variables near the vertices of the simplex (Fig. 4).

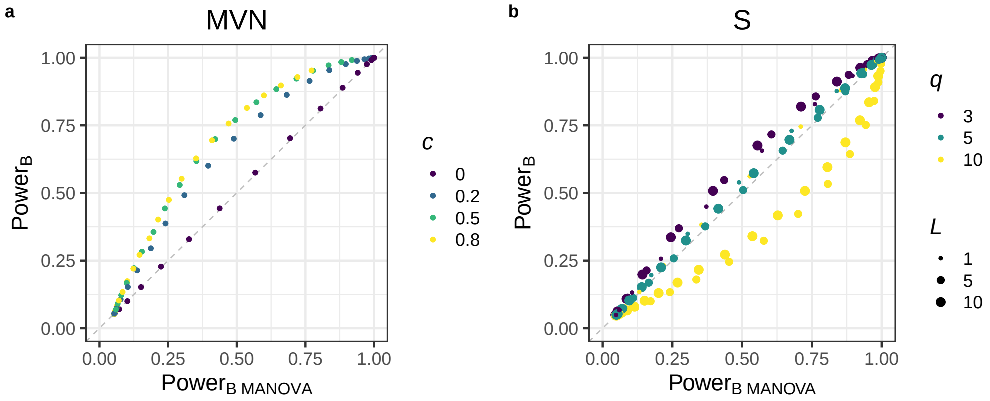

Figure 4. a) Comparison of power (factor B in model (6)) between MANOVA (x-axis) and asymptotic Anderson test (y-axis) for **a)** different values of correlation (c) between any pair of response variables, with $\mathbf{Y}_{i\cdot} \sim MVN(\mathbf{0}, \mathbf{I}_q)$, $q = 3$, $n = 500$ and $\Delta \in [0,0.2]$, and **b)** different values of q and L, with $\mathbf{Y}_{i\cdot} \sim S(q^{-1}\mathbf{1}, \sigma_g)$, $n = 500$ and $\Delta \in [0,0.02]$. The identity line is shown in grey (dashed).

Real data applications

Condition-specific splicing QTL mapping

Alternative splicing (AS) can be treated as a multivariate phenotype, based on the relative abundances of the alternative transcript isoforms generated from a given gene[30]. AS is subject to a tight regulation, often tissue-, cell-type- or condition-specific[45], and its alteration may lead to disease[46]. Here we apply asymptotic Anderson test across a panel of human tissues to identify ethnicity- and gender-specific *cis* genetic effects on AS (i.e. condition-specific splicing quantitative trait loci or cs-sQTLs).

We obtained transcript expression and genotype data from the V7 release of the GTEx Project, and employed a slightly modified version of `sQTLseekeR2` to map *cis* cs-sQTLs in 48 tissues (see Methods). Specifically, we assessed the significance of the interaction between the genotype and the ethnicity of the donor (european american, EA, or african american, AA). The reported ethnicity was confirmed by a PCA of the genotypes (Fig. S12). In a separate analysis, using an analogous approach, we evaluated the genotype-gender interaction. After multiple testing correction, we identified a total of 825 *cis* ethnicity-specific sQTLs affecting 184 genes, and 243 *cis* gender-specific sQTLs affecting 9 genes (Table S3).

These numbers are substantially smaller than the ones reported for regular sQTLs in the same dataset[a]. This is explained, on one side, by the more stringent pre-processing applied here (e.g. at least 10 individuals per level of the interaction are required, see Methods), which resulted in a smaller number of variant-gene pairs tested (see Table S3), but also by the lower statistical power to detect interactions[47]. Overall, gender-specific genetic effects seem less frequent than ethnicity-specific effects. This is consistent with the smaller number of differentially expressed genes between males and females than between EA and AA identified in the GTEx pilot study[48].

As expected, the ratio between the number of genes with ethnicity-specific sQTLs and the number of tested genes grows with the tissue sample size (Fig. S13). Skin (not sun-exposed) and skeletal muscle present the largest values of this ratio. Both tissues are known to have differences between EA and AA individuals[49,50]. Remarkably, skin (sun-exposed), displays a much larger value of this ratio than other tissues with larger sample sizes (e.g. blood, lung, nerve). This is also the case of cultured fibroblasts (skin-derived). Altogether, this suggests that our ethnicity-specific sQTLs are indeed capturing underlying biology.

To illustrate the nature of the cs-sQTLs identified, in Fig. 5 we show the example of the SNP rs28517808, an ethnicity-specific sQTL for the lincRNA SNHG8 in sun-exposed skin. The SNP

[a] See Chapter 1.

affects the relative abundances of the AS isoforms of the target gene (SNHG8) in EA individuals, but not in AA individuals. The exonic structure of the affected isoforms is also displayed. SNHG8 has ethnicity-specific sQTLs in six additional tissues (tibial artery, breast, transformed fibroblasts, esophagus –mucosa and muscularis – and thyroid). This gene has been previously identified as differentially expressed in skin between EA and AA individuals[51], and it has been related to several cancer types[52,53]. In Fig. S14, we also report an example of a gender-specific sQTL that affects the NOSTRIN gene in adipose subcutaneous tissue. NOSTRIN encodes a protein that sequesters endothelial nitric oxid synthase, reducing nitric oxide production and angiogenesis. Based on this, NOSTRIN has been suggested as a potential target to prevent cancer progression and metastasis[54].

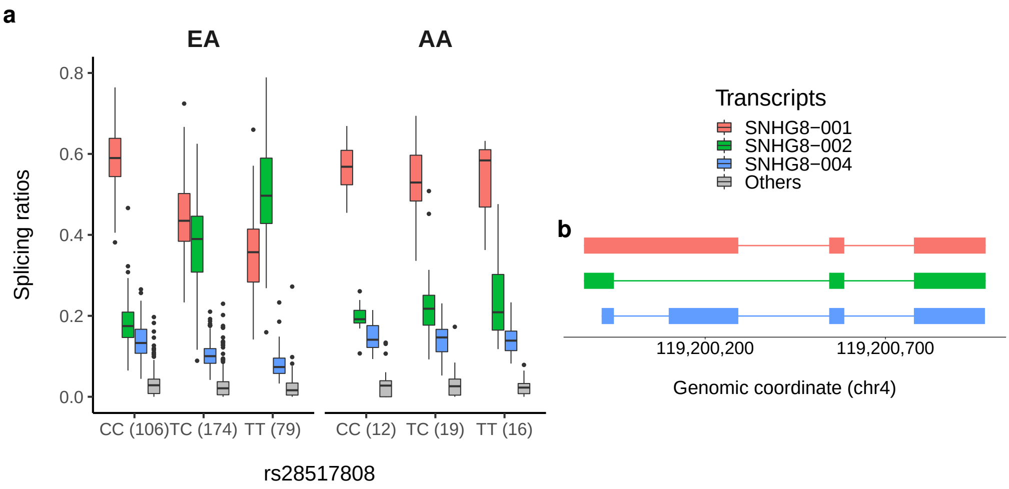

Figure 5. Relative abundances of the three most expressed isoforms in sun-exposed skin from the lincRNA SNHG8 (chr4:119,199,864-119,200,978, forward strand) for each genotype group at the rs28517808 locus (chr4:119,204,466, C/T), in european american (EA) and african american (AA) individuals. The least abundant isoforms are grouped in *Others*. The number of individuals in each genotype group is shown between parentheses. EA individuals that are homozygous for the reference allele (CC) at the SNP locus, express preferentially SNHG8-001 (red). In contrast, EA individuals homozygous for the alternative allele (TT) express preferentially SNHG8-002 (green). EA heterozygous individuals exhibit intermediate abundances. In AA individuals, however, the three isoforms display similar abundances independently of the genotype at rs28517808. **b)** Exonic structure of the isoforms SNHG8-001 (red), SNHG8-002 (green) and SNHG8-004 (blue).

GWAS of hippocampal subfield volumes

The hippocampus is a critical structure for memory, spatial navigation and cognition[55] and it has been related to several major brain disorders, including schizophrenia[56] and Alzheimer's disease[57]. Furhtermore, several GWAS have identified genetic variants associated with whole hippocampal volume[58,59]. However, the hippocampus is a heterogeneous structure, with different subregions that carry out distinct functions[55] and may be differentially affected in a disease context[60]. Here we apply the asymptotic Anderson test to identify genetic variants associated with the volumes of hippocampal subfields in the ADNI cohort (`http://adni.loni.usc.edu`).

Specifically, we obtained the MRI-derived volumes of eight hippocampal subfields (Cornu Ammonis (CA)1, CA2-3, CA4-Dentate Gyrus (DG), fimbria, hippocampal fissure, subiculum, presubiculum and tail) and genotype data from ADNI1/GO/2, corresponding to 555 individuals. Using `mlm`, we tested for association a total of 5,486,810 variants *vs* the logarithm-transformed hippocampal volumes (see Methods).

Despite our small sample size ($n =$ 555) compared to similar GWAS[5,59] ($n >$ 2 · 10^4), we were able to find one significant SNP at genome-wide level (Fig. 6a). This variant (rs34173062, *p* value 1.73 · 10^{-8}) was already found associated with reduced whole hippocampal volume[61] and limbic degeneration[62] in ADNI, and replicated in the UKBB dataset, where it also displayed association with family history of AD[62]. Here, we found an association between its alternative allele and the reduced volume of several hippocampal subfields, particularly marked for CA2_3, CA4_DG, subiculum and presubiculum (Fig. 6b). rs34173062 is a non-synonymous variant in the SHARPIN gene, which encodes a synaptic protein. Although the roles of SHARPIN in AD are still unclear, it has been suggested that it may affect postsynaptic adhesion and scaffolding of the transmitter receptors, altering synaptic stability[62]. Moreover, its impact on inflammation and immune function in the brain has recently been proposed[63]. We also recapitulated the well-known effect of age[57] (*p* value 1.31 · 10^{-14}, Fig. 6c), included as a covariate in the model (see Methods).

We found two additional strongly associated loci, albeit below the genome wide-significance threshold (Fig. 6a). The first comprises variants in the α-amylase gene cluster (e.g. rs79043596, *p* value 8.79 · 10^{-8}), which contains genes recently suggested to play a role in AD[64]. The second includes intronic variants (e.g. rs2060497, *p* value 4.64 · 10^{-7}) in the gene encoding the WWTR1/TAZ transcription factor, one of the main effectors in the Hippo pathway[65], which has been shown to mediate amyloid-β protein precursor (AβPP) signaling[66]. Alternative alleles of both variants relate to reduced volumes of certain hippocampal subregions (Fig. S15).

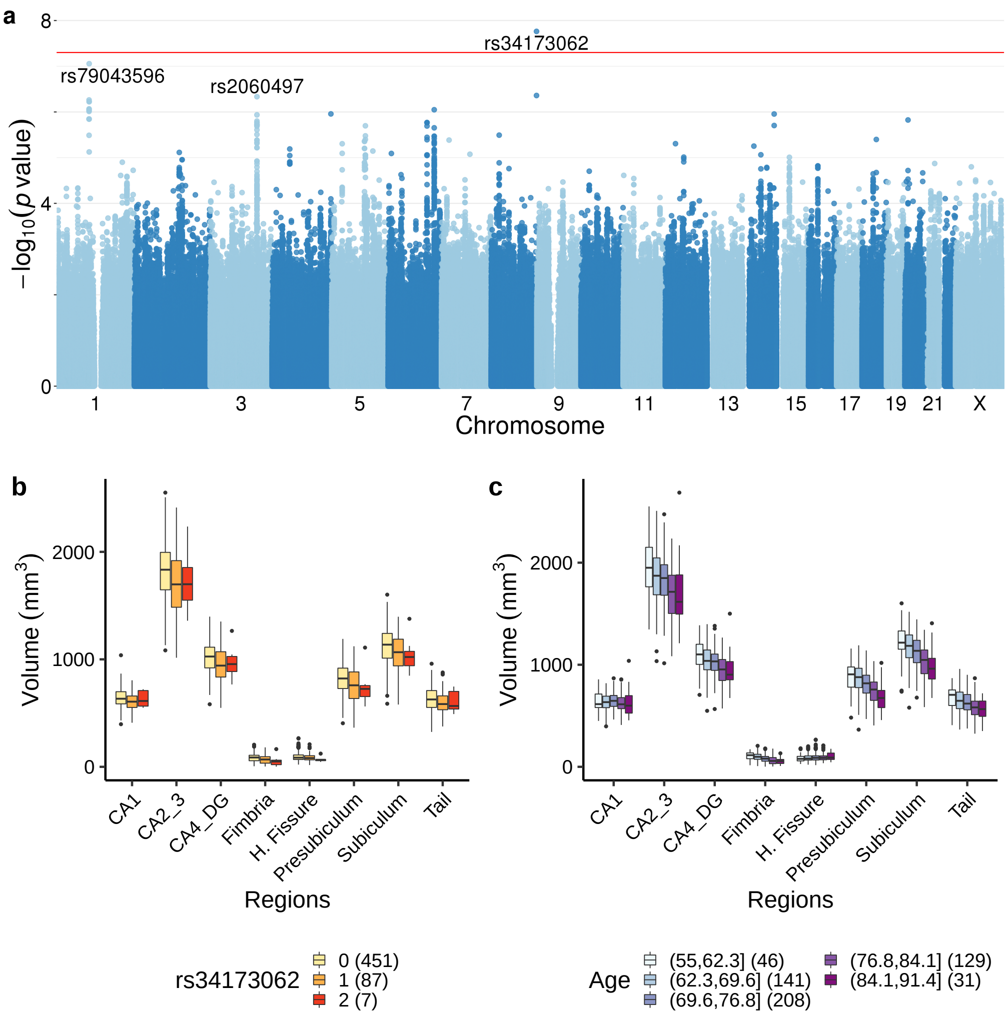

Figure 6. **a)** Manhattan plot showing the $-log_{10}$(p value) of association between genetic variants and the volumes of eight hippocampal subfields. The horizontal red line corresponds to the genome-wide significance threshold selected ($5 \cdot 10^{-8}$). The top associated SNPs are highlighted. **b)** Volumes (mm^3) of eigth hippocampal subfields for each genotype group at the rs34173062 locus (chr8:145158607, G/A). The number of individuals in each genotype group is shown between parentheses. **c)** Volumes (mm^3) of eight hippocampal subfields for different age groups. The number of individuals in each age group is shown between parentheses.

Discussion

In this work, we obtain the limiting distribution of the Anderson test statistic under the null hypothesis for any complex model and Euclidean distance. Our result holds after any transformation that preserves the independence of the observations of the response variables. We provide an efficient approach to compute asymptotic *p* values for the association between any quantitative multivariate phenotype and a set of predictors of interest, that we have implemented in the `mlm` R package (available at `https://github.com/dgarrimar/mlm`). Our comprehensive simulation study, together with our analyses of real datasets, presents the asymptotic Anderson test as a valuable non-parametric approach to identify genetic effects on multi-trait phenotypes in the context of GWAS and QTL mapping.

Our asymptotic approach offers highly accurate *p* values, while dramatically increasing computational efficiency with respect to the permutation test. In addition, it allows to compute *p* values down to a precision limit of 10^{-14}, difficult to reach using permutations given the large size of current biological datasets. Moreover, it is often not trivial to select the permutation schema that ensures an exact test[31].

In extensive simulations, the asymptotic test displayed controlled type I error rates and large power. We also show that our asymptotic result holds for relatively small values of the ratio between the sample size and the number of traits, supporting its applicability to current datasets. In contrast to other approaches for multivariate association testing that require multivariate normality, such as MANOVA[20] or mvLMMs[21–23], our method does not make any assumption regarding the distribution of the traits of interest. This is expected to result in higher power to detect significant associations. Indeed, our simulations showed that asymptotic Anderson test outperformed MANOVA in several contexts, albeit MANOVA still retained larger power in others (see Fig. 4 for an example).

Our approach demonstrated its utility in two real-case scenarios. First, it enabled the identification of context-specific splicing QTLs across GTEx tissues. In particular, we identified over 800 ethnicity-specific sQTLs, especially in tissues with well-characterized differences between individuals of african and european ancestry, such as skin[49]. We also found almost 250 gender-specific sQTLs. The number of genes affected in the two analyses (184 and 9, respectively) suggests that the patterns of splicing regulation may be more conserved between males and females than between EA and AA individuals. Second, our method allowed to perform a multivariate GWAS in the ADNI cohort. The phenotype of interest were the volumes of different hippocampal subfields derived from MRI. Despite our modest sample size, we identified a genome-wide significant SNP (rs34173062), which had been previously associated with reduced whole hippocampal volume, limbic degeneration and Alzheimer’s disease (AD) family history[61,62]. Of note, marked hippocampal atrophy is common in AD[57]. We further identified other strong associations involving loci with potential roles in the physiopathology of AD[62,64,66]. Although the examples described here illustrate the applicability of our method to study the effects of genetic variants, it could be applied to any other predictors of interest.

Nevertheless, our method also presents some limitations. Among them, an increased type I error in heteroscedastic situations. This can be especially problematic in unbalanced designs, which is frequently the case of genetic variants. Heteroscedasticity is a common problem for the majority of linear modelling strategies, although it may be accounted for with GLMs or mixed models. These require either defining *a priori* the variance structure, or inferring it from the actual data. However, the former can be particularly difficult in large and complex biological datasets, while the latter is often highly inefficient. To control for heteroscedasticity, here we employ a non-parametric test which assesses multivariate homogeneity of variances between genotype groups[42]. Unfortunately, this approach is also permutation-based. Hence, further research is needed to characterize the asymptotic distribution of this test statistic.

A second limitation is our assumption of independence regarding the individuals, given that population stratification is often present in human cohorts, and this may lead to false positive associations[25]. Recently, Anderson test was modified to incorporate information on the genetic similarities between pairs of individuals, correcting for population structure in the mixed model way[22]. However this method, implemented in the `GAMMA` software (`http://genetics.cs.ucla.edu/GAMMA`), still relies on permutations for significance assessment. Evaluating whether our asymptotic result can be applied in this context would be a potential avenue for future research. Other multivariate mixed model based alternatives, such as `GEMMA`[21], are currently available, but they can have prohibitive running times with many traits and large samples sizes. For example, a single test using `GEMMA` with 4 traits and n = 5,255 takes about 6.7 minutes[21]. An equivalent test with `mlm` runs in less than 0.04 seconds (10^4 times faster).

References

1. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562,** 203–209 (2018).
2. Consortium, T. E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
3. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).
4. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nature Genetics* **45,** 580–585 (2013).
5. Van der Meer, D. *et al.* Brain scans from 21,297 individuals reveal the genetic architecture of hippocampal subfield volumes. *Molecular Psychiatry,* 1–13 (2018).
6. Natarajan, P. *et al.* Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. *Nature Communications* **9,** 3391 (2018).
7. Li, Y. I. *et al.* RNA splicing is a primary link between genetic variation and disease. *Science* **352,** 600–604 (2016).
8. Consortium, G. Genetic effects on gene expression across human tissues. *Nature* **550,** 204–213 (2017).
9. Korte, A. *et al.* A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature Genetics* **44,** 1066–1071 (2012).
10. Galesloot, T. E., van Steen, K., Kiemeney, L. A. L. M., Janss, L. L. & Vermeulen, S. H. A Comparison of Multivariate Genome-Wide Association Methods. *PLoS ONE* **9,** e95923 (2014).
11. Porter, H. F. & O'Reilly, P. F. Multivariate simulation framework reveals performance of multi-trait GWAS methods. *Scientific Reports* **7,** 38837 (2017).
12. Pickrell, J. K. *et al.* Detection and interpretation of shared genetic influences on 42 human traits. *Nature genetics* **48,** 709–17 (2016).
13. Stephens, M. A Unified Framework for Association Analysis with Multiple Related Phenotypes. *PLoS ONE* **8,** e65245 (2013).
14. Giambartolomei, C. *et al.* A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics* **34,** 2538–2545 (2018).

15. Sul, J. H., Han, B., Ye, C., Choi, T. & Eskin, E. Effectively Identifying eQTLs from Multiple Tissues by Combining Mixed Model and Meta-analytic Approaches. *PLoS Genetics* **9,** e1003491 (2013).

16. Moore, R. *et al.* A linear mixed-model approach to study multivariate gene-environment interactions. *Nature Genetics* **51,** 180–186 (2019).

17. Ning, C. *et al.* Efficient multivariate analysis algorithms for longitudinal genome-wide association studies. *Bioinformatics* (2019).

18. Aschard, H. *et al.* Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *American journal of human genetics* **94,** 662–76 (2014).

19. Ferreira, M. A. R. & Purcell, S. M. A multivariate test of association. *Bioinformatics* **25,** 132–133 (2009).

20. Liu, F. *et al.* A Genome-Wide Association Study Identifies Five Loci Influencing Facial Morphology in Europeans. *PLoS Genetics* **8,** e1002932 (2012).

21. Zhou, X. & Stephens, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods* **11,** 407–409 (2014).

22. Joo, J. W. J. *et al.* Efficient and Accurate Multiple-Phenotype Regression Method for High Dimensional Data Considering Population Structure. *Genetics* **204,** 1379–1390 (2016).

23. Casale, F. P., Rakitsch, B., Lippert, C. & Stegle, O. Efficient set tests for the genetic analysis of correlated traits. *Nature Methods* (2015).

24. Zhang, Y., Zhou, H., Zhou, J. & Sun, W. Regression Models for Multivariate Count Data. *Journal of Computational and Graphical Statistics* (2017).

25. Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics* **11,** 459–463 (2010).

26. Furlotte, N. A. & Eskin, E. Efficient multiple-trait association and estimation of genetic correlation using the matrix-variate linear mixed model. *Genetics* **200,** 59–68 (2015).

27. Anderson, M. A new method for non-parametric multivariate analysis of variance. *Australian Ecology* **26,** 32–46 (2001).

28. Gonzàlez-Porta, M., Calvo, M., Sammeth, M. & Guigó, R. Estimation of alternative splicing variability in human populations. *Genome research* **22,** 528–38 (2012).

29. Anderson, M. J. & Robinson, J. Generalized discriminant analysis based on distances. *Australian & New Zealand Journal of Statistics* **45,** 301–318 (2003).

30. Monlong, J., Calvo, M., Ferreira, P. G. & Guigó, R. Identification of genetic variants associated with alternative splicing using sQTLseekeR. *Nature Communications* **5,** 4698 (2014).

31. Anderson, M. J. Permutation tests for univariate or multivariate analysis of variance and regression. *Canadian Journal of Fisheries and Aquatic Sciences* **58,** 626–639 (2001).

32. Imhof, J. P. Computing the distribution of quadratic forms in normal variables. *Biometrika* **48,** 419–426 (1961).

33. Davies, R. B. Algorithm AS 155: The Distribution of a Linear Combination of χ^2 Random Variables. *Applied Statistics* **29,** 323 (1980).

34. Farebrother, R. W. Algorithm AS 204: The Distribution of a Positive Linear Combination of χ^2 Random Variables. *Applied Statistics* **33,** 332 (1984).

35. Duchesne, P. & Lafaye De Micheaux, P. Computing the distribution of quadratic forms: Further comparisons between the Liu-Tang-Zhang approximation and exact methods. *Computational Statistics and Data Analysis* (2010).

36. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S* (Springer, New York, 2002).

37. Kojadinovic, I. & Yan, J. Modeling Multivariate Distributions with Continuous Margins Using the copula R Package. *Journal of Statistical Software* **34,** 1–20 (2010).

38. R Core Team. *R: A Language and Environment for Statistical Computing* Vienna, Austria, 2018.

39. Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nature Biotechnology* **35,** 316–319 (2017).

40. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).

41. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* **12,** 323 (2011).

42. Anderson, M. J. Distance-based tests for homogeneity of multivariate dispersions. *Biometrics* **62,** 245–253 (2006).

43. Davis, J. R. *et al.* An Efficient Multiple-Testing Adjustment for eQTL Studies that Accounts for Linkage Disequilibrium between Variants. *American Journal of Human Genetics* (2016).

44. McArtor, D. B., Lubke, G. H. & Bergeman, C. S. Extending multivariate distance matrix regression with an effect size measure and the asymptotic null distribution of the test statistic. *Psychometrika* (2017).

45. Fu, X.-D. & Ares, M. Context-dependent control of alternative splicing by RNA-binding proteins. *Nature Reviews Genetics* **15,** 689–701 (2014).

46. Scotti, M. M. & Swanson, M. S. RNA mis-splicing in disease. *Nature Reviews Genetics* **17,** 19–32 (2016).

47. Leon, A. C. & Heo, M. Sample sizes required to detect interactions between two binary fixed-effects in a mixed-effects linear regression model. *Computational Statistics & Data Analysis* **53,** 603–608 (2009).

48. Melé, M. *et al.* Human genomics. The human transcriptome across tissues and individuals. *Science* **348,** 660–5 (2015).

49. Del Bino, S., Duval, C. & Bernerd, F. Clinical and Biological Characterization of Skin Pigmentation Diversity and Its Consequences on UV Impact. *International Journal of Molecular Sciences* **19,** 2668 (2018).

50. Ceaser, T. & Hunter, G. Black and White Race Differences in Aerobic Capacity, Muscle Fiber Type, and Their Influence on Metabolic Processes. *Sports Medicine* **45,** 615–623 (2015).

51. Sanyal, R. D. *et al.* Atopic dermatitis in African American patients is TH2/TH22-skewed with TH1/TH17 attenuation. *Annals of allergy, asthma & immunology : official publication of the American College of Allergy, Asthma, & Immunology* **122,** 99–110 (2019).

52. Chen, C., Zhang, Z., Li, J. & Sun, Y. SNHG8 is identified as a key regulator in non-small-cell lung cancer progression sponging to miR-542-3p by targeting CCND1/CDK6. *OncoTargets and Therapy* **Volume 11,** 6081–6090 (2018).

53. Dong, J. *et al.* lncRNA SNHG8 Promotes the Tumorigenesis and Metastasis by Sponging miR-149-5p and Predicts Tumor Recurrence in Hepatocellular Carcinoma. *Cellular Physiology and Biochemistry* **51,** 2262–2274 (2018).

54. Chakraborty, S. & Ain, R. Nitric-oxide synthase trafficking inducer is a pleiotropic regulator of endothelial cell function and signaling. *The Journal of biological chemistry* **292,** 6600–6620 (2017).

55. Zeidman, P. & Maguire, E. A. Anterior hippocampus: the anatomy of perception, imagination and episodic memory. *Nature Reviews Neuroscience* **17,** 173–182 (2016).

56. Lieberman, J. A. *et al.* Hippocampal dysfunction in the pathophysiology of schizophrenia: a selective review and hypothesis for early detection and intervention. *Molecular Psychiatry* **23,** 1764–1772 (2018).

57. Van de Pol, L. A. *et al.* Hippocampal atrophy in Alzheimer disease: age matters. *Neurology* **66,** 236–8 (2006).

58. Consortium, E. N. I. G. t. M.-A. E. *et al.* Common variants at 12q14 and 12q24 are associated with hippocampal volume. *Nature Genetics* **44,** 545–551 (2012).

59. Hibar, D. P. *et al.* Novel genetic loci associated with hippocampal volume. *Nature Communications* **8,** 13624 (2017).

60. Small, S. A., Schobel, S. A., Buxton, R. B., Witter, M. P. & Barnes, C. A. A pathophysiological framework of hippocampal dysfunction in ageing and disease. *Nature Reviews Neuroscience* **12,** 585–601 (2011).

61. Li, Q. *et al.* Large-scale Feature Selection of Risk Genetic Factors for Alzheimer's Disease via Distributed Group Lasso Regression (2017).

62. Soheili-Nezhad, S. *et al.* A Non-Synonymous SHARPIN Variant is Associated with Limbic Degeneration and Family History of Alzheimer's Disease. *bioRxiv,* 196410 (2019).

63. Asanomi, Y. *et al.* A rare functional variant of SHARPIN attenuates the inflammatory response and associates with increased risk of late-onset Alzheimer's disease. *Molecular Medicine* **25,** 20 (2019).

64. Byman, E., Schultz, N., Fex, M., Wennström, M. & Wennström, M. Brain alpha-amylase: a novel energy regulator important in Alzheimer disease? *Brain Pathology* **28,** 920–932 (2018).

65. Plouffe, S. W., Hong, A. W. & Guan, K.-L. Disease implications of the Hippo/YAP pathway. *Trends in molecular medicine* **21,** 212–22 (2015).

66. Swistowski, A. *et al.* Novel mediators of amyloid precursor protein signaling. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **29,** 15703–12 (2009).

67. Mersmann, O. *microbenchmark: Accurate Timing Functions* 2018.

68. Gower, J. C. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **53,** 325–338 (1966).

69. Bai, Z. D., Rao, C. R. & Wu, Y. *M-Estimation of Multivariate Linear Regression Parameters Under a Convex Discrepancy Function* 1992.

Acknowledgements

We thank Manuel Muñoz, Emilio Palumbo, Natalia Vilor and Juan D. Gispert for useful discussions. The Genotype-Tissue Expression (GTEx) project is supported by the Common Fund of the Office of the Director of the National Institutes of Health (http://commonfund.nih.gov/GTEx). The Alzheimer’s Disease Neuroimaging Initiative (ADNI) is funded by the National Institutes of Health (Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). D.G-M. is supported by a ‘la Caixa’-Severo Ochoa pre-doctoral fellowship. We also acknowledge support of the Spanish Ministry of Economy, Industry and Competitiveness (MEIC) to the EMBL partnership, ‘Centro de Excelencia Severo Ochoa’, the CERCA Programme / Generalitat de Catalunya and the European Regional Development Fund (ERDF).

Author information

Contributions

D.G-M., M.C. and R.G. conceived and designed the study. M.C. worked on the theoretical framework, together with F.R. and D.G-M. D.G-M. implemented the software, performed the simulations and analyzed the data. M.C. and F.R. contributed ideas and statistical advice, helping with the design of the software. D.G-M. and M.C. wrote the original draft. All the authors reviewed the final manuscript.

Competing interests

The authors declare no competing interests.

Supplementary Tables and Figures

	S			C		
	$u = 1$	$u = 0.5$	$u = 0.2$	$u = 1$	$u = 0.5$	$u = 0.2$
$h = 1$	0.045	0.047	0.048	0.053	0.055	0.052
$h = 2$	0.062	0.205	0.402	0.059	0.123	0.201
$h = 5$	0.090	0.385	0.769	0.073	0.258	0.521
$h = 10$	0.096	0.414	0.800	0.085	0.346	0.695

Table S1. Effects of heteroscedasticity and unbalanced designs on type I error (AB), for response variables with different distributions: S = multivariate proportions (simplex), C = gaussian copula. Here n = 500, q = 5. h is the ratio between the variances of the response variables in the first level of factor B and any other level of B. u is the ratio between the sample size of the first level of factor B and any other level. Note that $h = u =$ 1 corresponds to a homoscedastic balanced design, and that in the case of the simplex, heteroskedasticity is introduced at the level of σ_g (see Methods).

MVN	$u = 1$	$u = 0.5$	$u = 0.2$
$h = 1$	0.046	0.048	0.047
$h = 2$	0.053	0.112	0.183
$h = 5$	0.070	0.256	0.503
$h = 10$	0.082	0.351	0.688

Table S2. Effects of heteroscedasticity and unbalanced designs in MANOVA type I error (AB). Here $\mathbf{Y}_{i\cdot} \sim MVN(\mathbf{0}, \mathbf{I}_q)$, $n = 500$, $q = 5$. h is the ratio between the variances of the response variables in the first level of factor B and any other level of B. u is the ratio between the sample size of the first level of factor B and any other level. Note that $h = u = 1$ corresponds to a homoscedastic balanced design (see Methods).

Ethnicity-dependent sQTLs					
Tissue	**Samples**	**Variants**	**Genes**	**cd-sQTLs**	**cd-sGenes**
Adipose - Subcutaneous 🟧	374	95,835	7,450	127	33
Adipose - Visceral (Omentum) 🟧	304	10,461	2,432	1	1
Artery - Tibial 🟥	376	101,410	7,335	108	27
Breast - Mammary Tissue 🟦	244	15,558	3,156	8	2
Cells - Transformed fibroblasts 🟦	296	42,428	5,113	36	14
Colon - Transverse 🟫	238	7,009	1,705	1	1
Esophagus - Mucosa 🟫	348	52,978	6,038	31	12
Esophagus - Muscularis 🟫	325	38,101	5,157	13	5
Lung 🟩	374	61,234	6,636	11	7
Muscle - Skeletal 🟪	478	133,514	7,322	257	52
Nerve - Tibial 🟨	348	78,323	7,036	68	12
Ovary 🟪	120	2,169	764	1	1
Pancreas 🟫	211	8,870	2,089	1	1
Skin - Not Sun Exposed (Suprapubic) 🟦	326	53,572	6,138	96	20
Skin - Sun Exposed (Lower leg) 🟦	406	100,482	7,827	205	41
Thyroid 🟩	388	83,650	7,352	91	22
Whole Blood 🟪	362	39,549	4,093	20	7
			Total (unique)	825	184

Gender-dependent sQTLs					
Tissue	**Samples**	**Variants**	**Genes**	**cd-sQTLs**	**cd-sGenes**
Adipose - Subcutaneous 🟧	385	672,527	10,034	4	1
Breast - Mammary Tissue 🟦	251	526,218	9,936	16	1
Colon - Transverse 🟫	246	503,488	9,788	43	1
Lung 🟩	383	663,111	10,853	128	1
Nerve - Tibial 🟨	361	678,771	10600	23	2
Pancreas 🟫	220	389,911	8,924	10	1
Whole Blood 🟪	369	283,018	6,074	18	1
			Total (unique)	243	9

Table S3. Number of samples, variants and genes tested; cs-sQTLs and cs-sGenes (genes with at least one cs-sQTL) identified across tissues after multiple testing correction (see Methods). Tissues without significant cs-sQTLs are not shown.

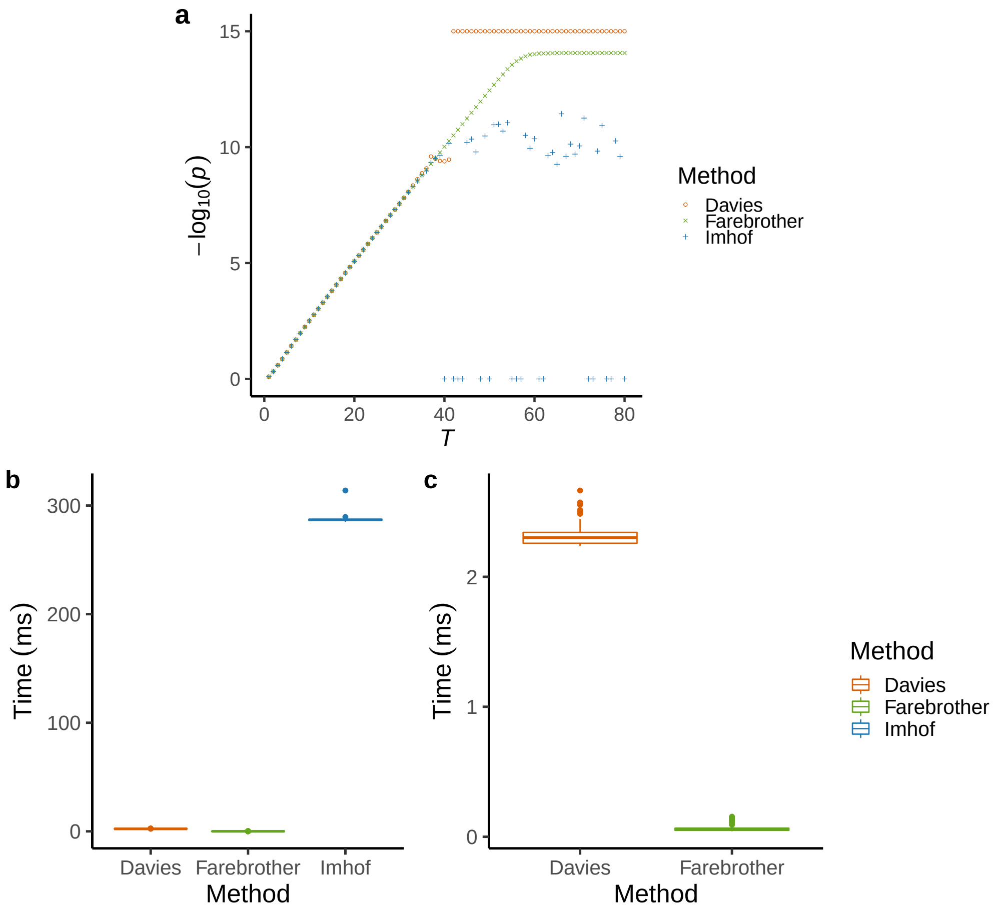

Figure S1. a) Behaviour of the *p* values obtained by Davies, Farebrother and Imhof methods, as implemented in the `CompQuadForm` R package[35], as a function of different values of the test statistic (labelled as T, x-axis). The data shown correspond to a simulation in which $q = 5$ weights are sampled from a uniform distribution ($\lambda_j \sim U(0, 1)$, $j \in \{1, \dots, 5\}$), and chi-square variables have 1 degree of freedom. Farebrother *p* values decreased monotonically with the value of T, down to the precision limit ($\approx 10^{-14}$), Imhof and Davies generated *p* values of 0 (here displayed as $p = 10^{-15}$) or even negative (here displayed as $p = 1$). **b)** Running time of the three methods (in milliseconds), computed across 100 runs with $T = 30$ using the `microbenchmark` R package[67]. **c)** Zoom of b).

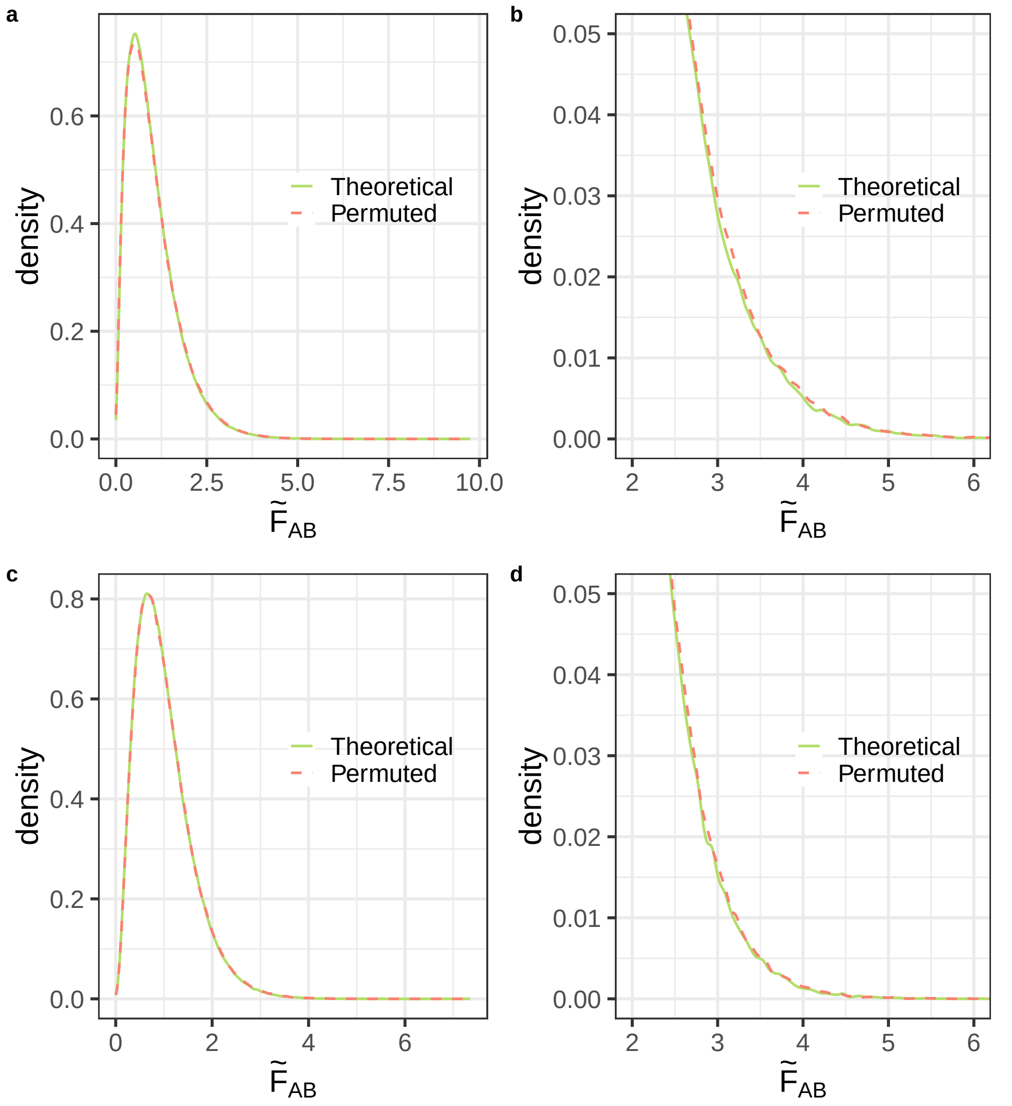

Figure S2. **a)** Theoretical null distribution of the test statistic for the interaction term (green solid line), compared to the distribution obtained using 10^6 permutations (red dashed line), with $\mathbf{Y}_{i\cdot} \sim C(\mathbf{I}_q)$ and $\Delta = 1$. **b)** Zoom of the upper tail of the distribution. **c, d)** Analogous to **a**,**b** with $\mathbf{Y}_{i\cdot} \sim S(\boldsymbol{p}, \sigma_g)$, where $\boldsymbol{p} = q^{-1}\mathbf{1}$. Here $\sigma_g = 0.025$ and $\Delta = 0.2$. In both scenarios we considered model (6) with 2 and 3 levels for A and B, respectively, in a completely crossed, balanced design. We simulated $n = 200$ observations of $q = 3$ response variables and generated B under the H_1 (see Methods).

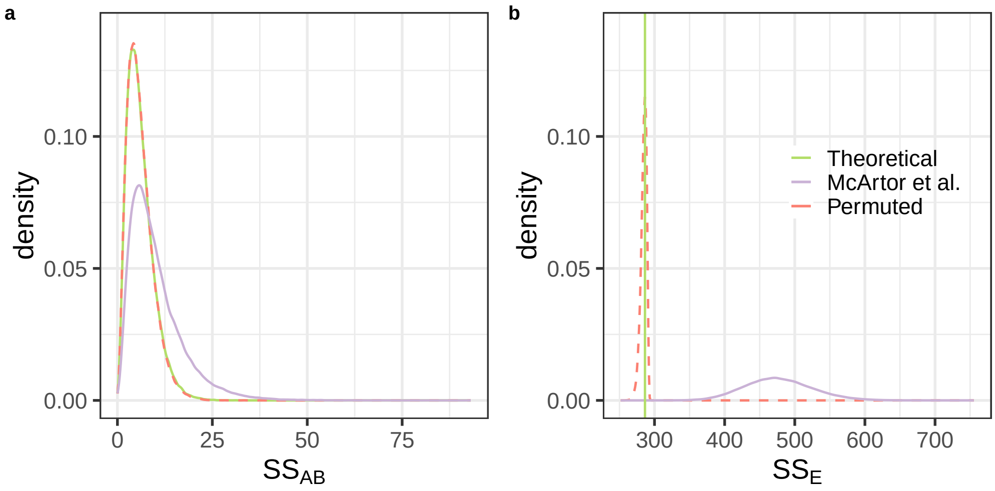

Figure S3. Null distribution of the numerator (**a**) and denominator (**b**) of the test statistic for factor AB. We simulated a scenario analogous to the one employed for Fig. 1. We compared the theoretical distribution derived from (9) (green solid line), to the proposal of McArtor et al.[44] (purple solid line) and the distribution obtained using 10^5 permutations (red dashed line).

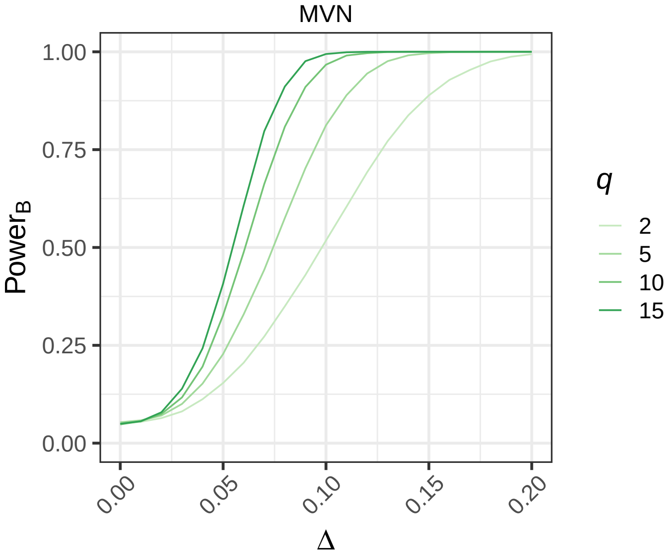

Figure S4. Power curves (factor B) for different values of q. Here, $\mathbf{Y}_{i\cdot} \sim MVN(\mathbf{0}, \mathbf{I}_q)$, $n = 500$ and Δ takes values in the range [0,0.2].

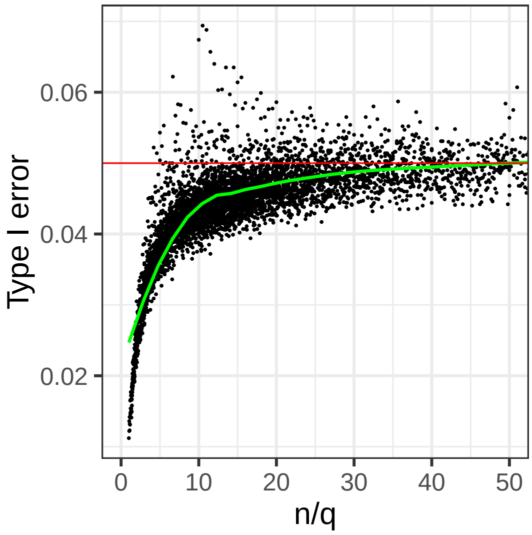

Figure S5. Type I error of asymptotic *p* values for the interaction term (AB) as a function of the n/q ratio. We considered values of n ranging from 20 to 300, and values of q ranging from 2 to 20. For visualization purposes, we focused on the window $n/q \in$ [0,50]. Here we used the model depicted in (6). The horizontal red line marks the 0. A polynomial has been fitted to the points using local fitting (LOESS), in order to describe the trend (fit shown in green, 95% confidence interval in grey).

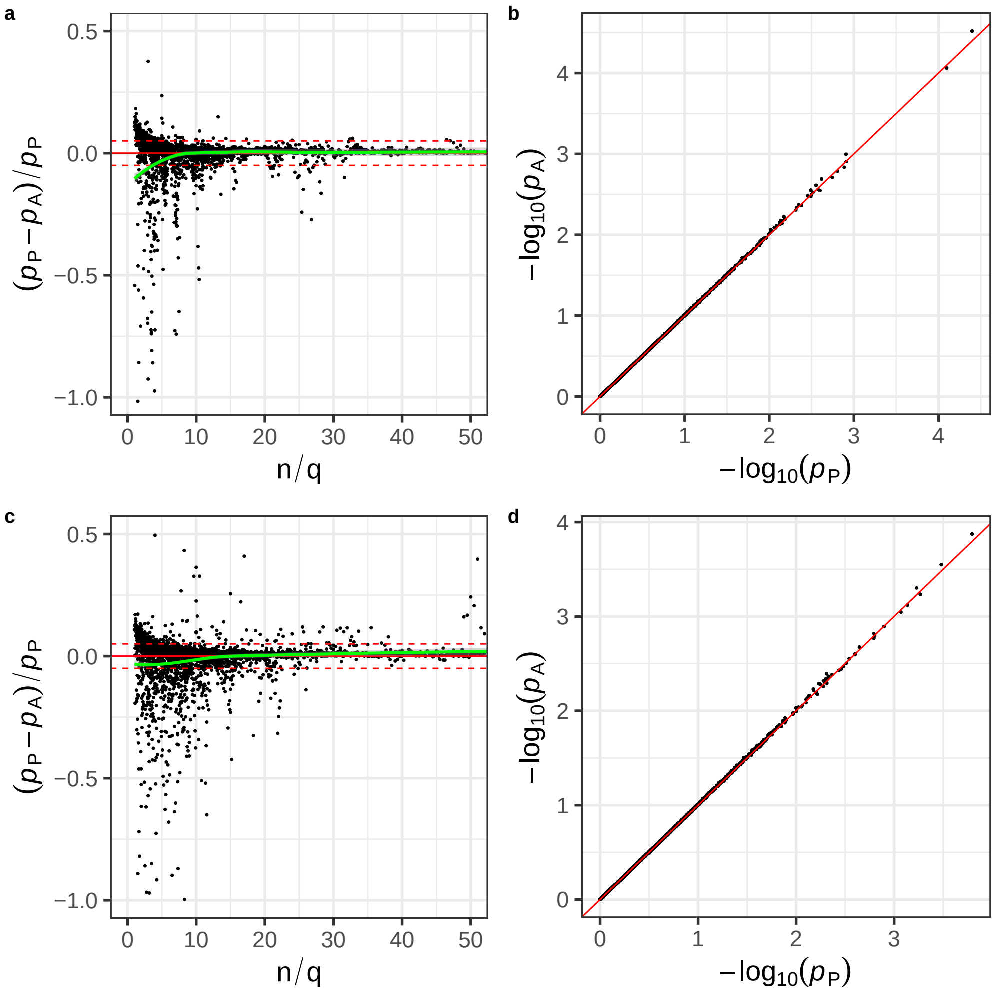

Figure S6. a, c) Relative bias of asymptotic *p* values *vs* n/q ratio, when $\mathbf{Y}_{i\cdot} \sim C(\mathbf{I}_q)$ and $\mathbf{Y}_{i\cdot} \sim S(\mathbf{1/q}, \sigma_g)$, respectively. Relative difference between asymptotic (p_A) and permutation-based (p_P, 10^5 permutations) *p* values for the interaction term (AB) as a function of the ratio between the total sample size and the number of dependent variables (n/q). We considered values of n ranging from 20 to 300, and values of q ranging from 2 to 20. For visualization purposes, we focused on values of $n/q \in$ [0,50] and relative biases $\in$ [-1,0.5]. The horizontal solid red line marks the 0. The horizontal dashed red lines mark the 5% relative bias. A polynomial has been fitted to the points using local fitting (LOESS), in order to describe the trend (fit shown in green, 95% confidence interval in grey). **b, d)** comparison of asymptotic and permutation-based *p* values when the asymptotic null holds ($n = 300$, $q = 3$), when $\mathbf{Y}_{i\cdot} \sim C(\mathbf{I}_q)$ and $\mathbf{Y}_{i\cdot} \sim S(q^{-1}\mathbf{1}, \sigma_g)$, respectively

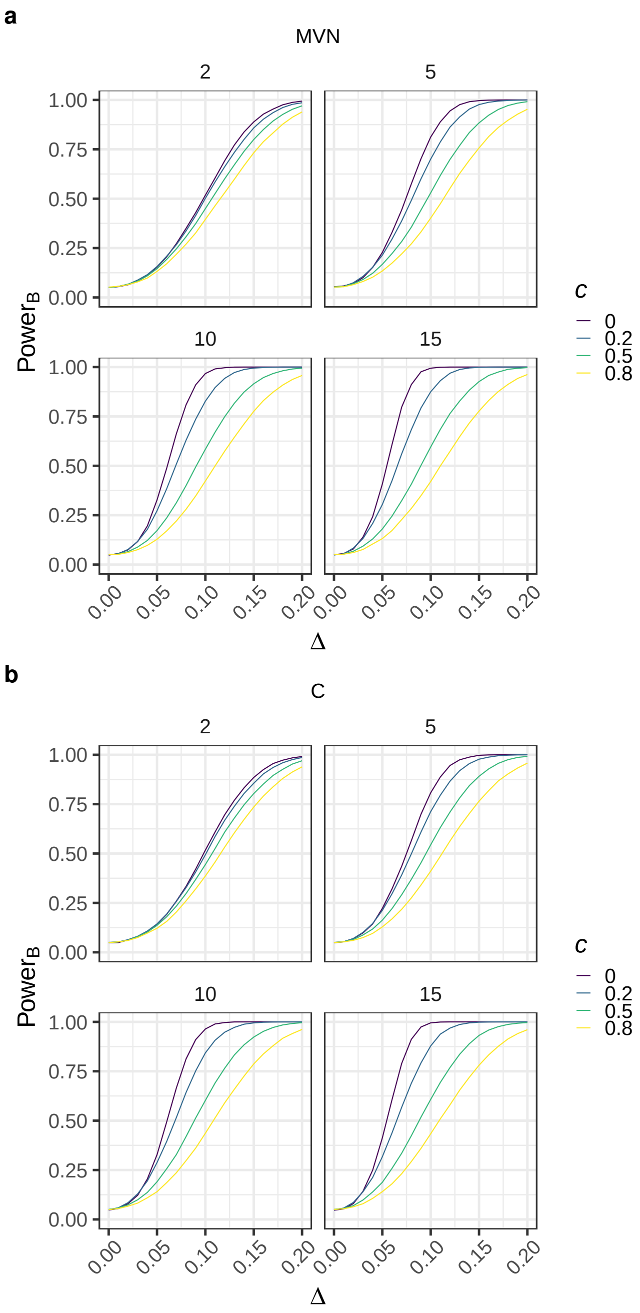

Figure S7. Power curves (factor B) for different values of the number of response variables (q) and the correlation between any pair of response variables (c), with $\Delta \in$ [0,0.2]. **a)** $\mathbf{Y}_{i\cdot} \sim MVN(\mathbf{0}, \mathbf{I}_q)$ and **b)** $\mathbf{Y}_{i\cdot} \sim C(\mathbf{I}_q)$.

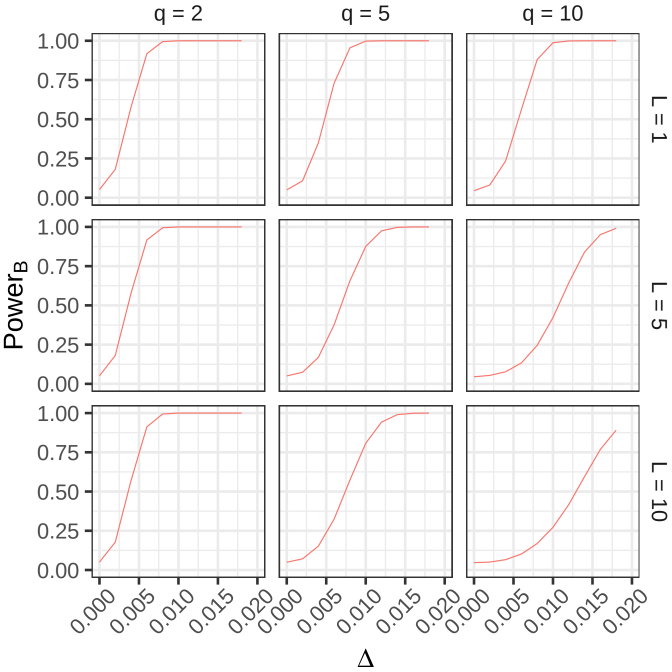

Figure S8. Power curves (factor B) for different values of the number of response variables (q) and the location in the simplex (L). Note that $L = 1$ corresponds to the center of the simplex, while $L = 10$ corresponds to the vicinity of one of the simplex vertices (see Methods).

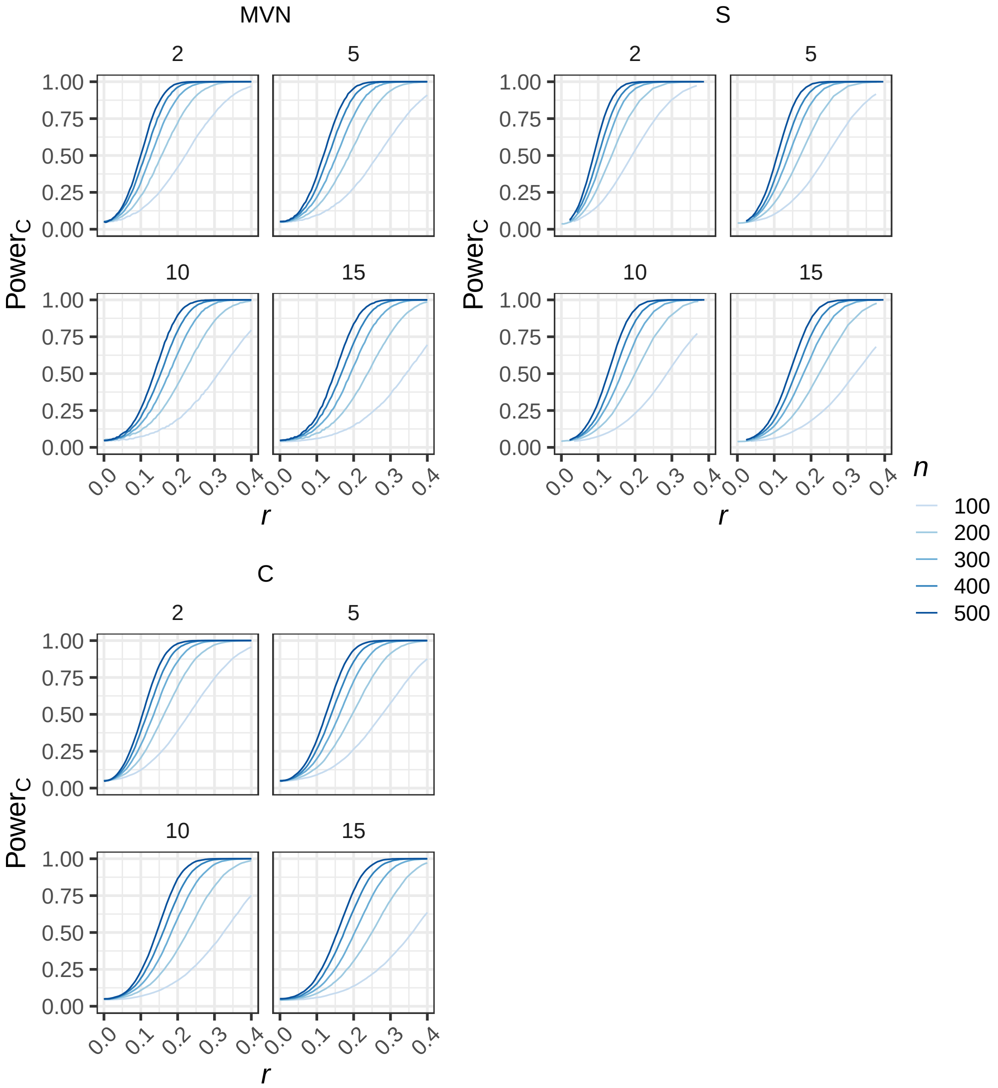

Figure S9. Power curves (covariate C) in different scenarios regarding the distribution of **Y**: MVN = multivariate normal, S = simplex, C = gaussian copula. Different values of n and q are evaluated, with $r \in$ [0,0.4].

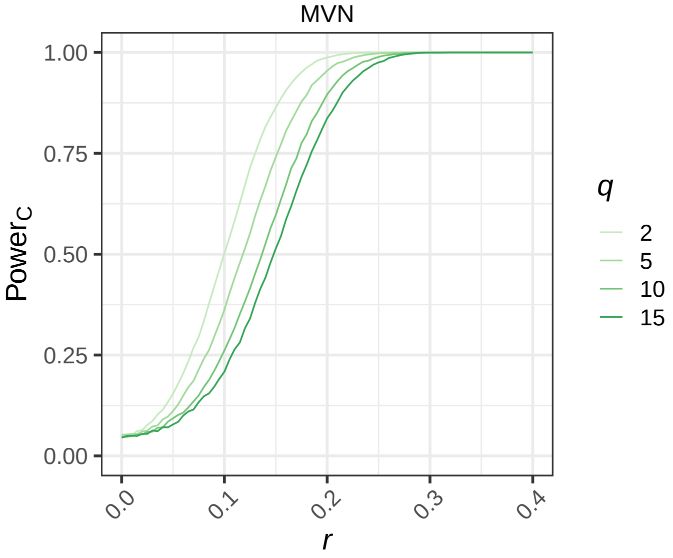

Figure S10. Power curves (covariate C) for different values of q. Here, $\mathbf{Y}_{i\cdot} \sim MVN(\mathbf{0}, \mathbf{I}_q)$, $n = 500$ and r takes values in the range [0,0.4].

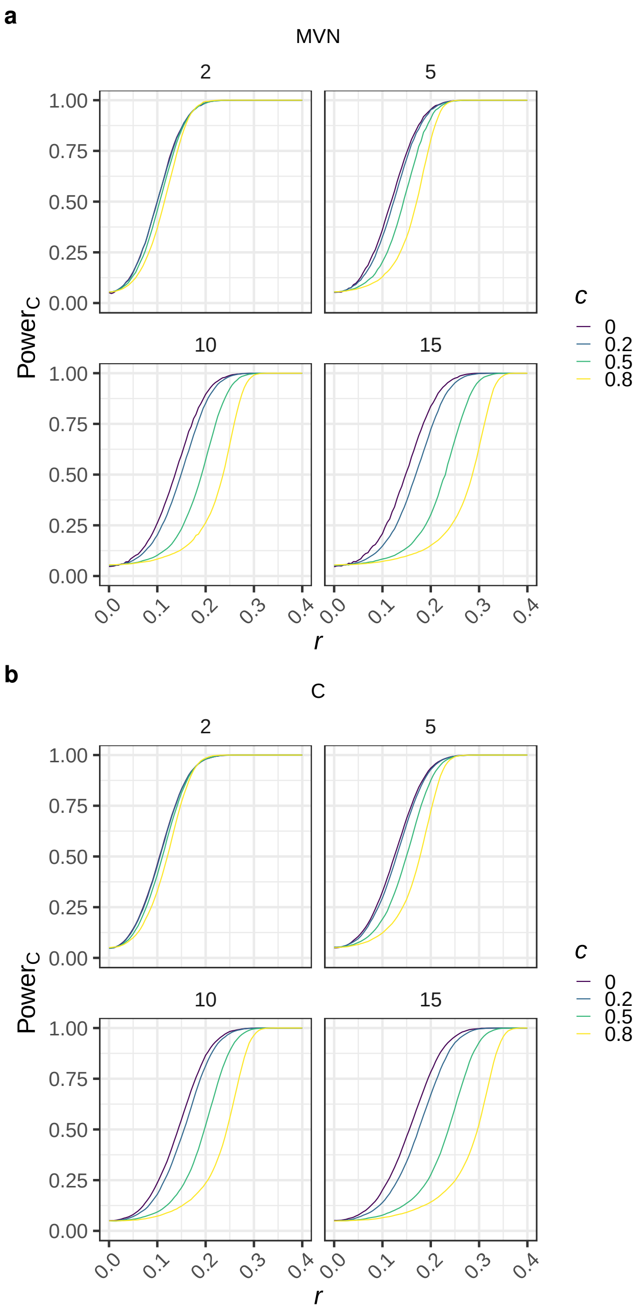

Figure S11. Power curves (covariate C) for different values of the number of response variables (q) and the correlation between any two response variables (c), with $r \in$ [0,0.4]. **a)** $\mathbf{Y}_{i\cdot} \sim MVN(\mathbf{0}, \mathbf{I}_q)$ and **b)** $\mathbf{Y}_{i\cdot} \sim C(\mathbf{I}_q)$.

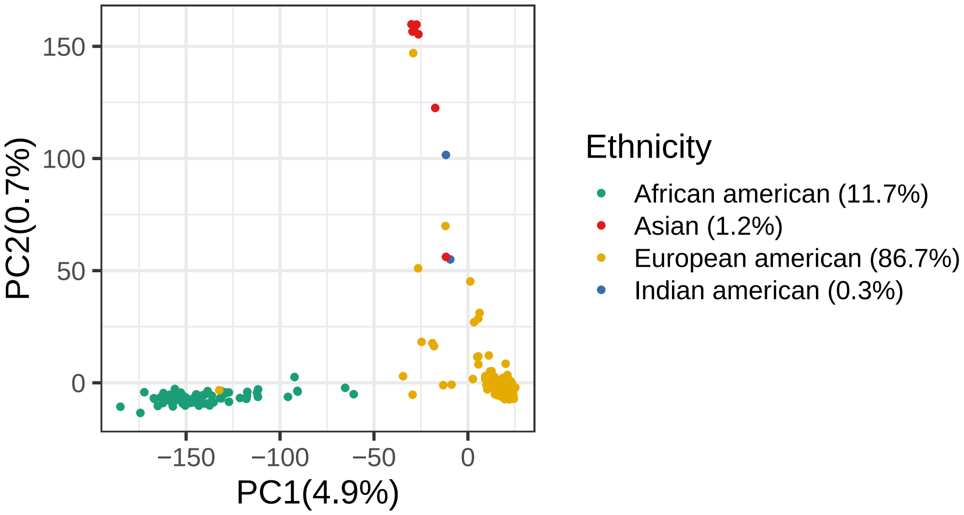

Figure S12. Principal Component Analysis (PCA) based on the genotypes of GTEx individuals, obtained using the `pca` tool in `QTLtools` v1.0 (`https://qtltools.github.io/qtltools`), with parameters `--maf 0.05` and `--distance 50000`. The percentage of variance explained is shown between parentheses. In the legend, the percentage of individuals of each ethnicity is shown between parentheses. Ethnicity, as reported in GTEx by the donor, family/next of kin, or medical record matches the ancestry patterns observed in the PCA. Of note, only european american and african american individuals were selected for cs-sQTL mapping.

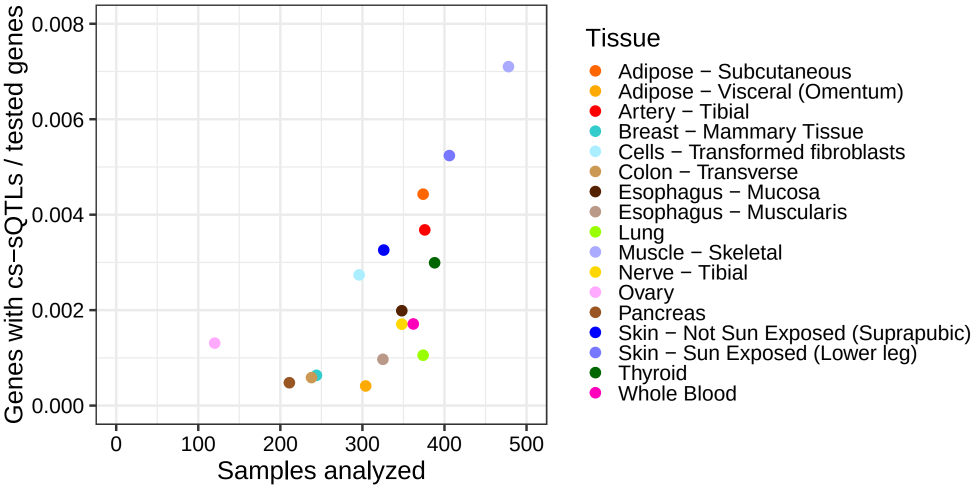

Figure S13. Proportion of genes with cs-sQTLs (over tested genes, y-axis) per tissue with respect to the tissue sample size (x-axis).

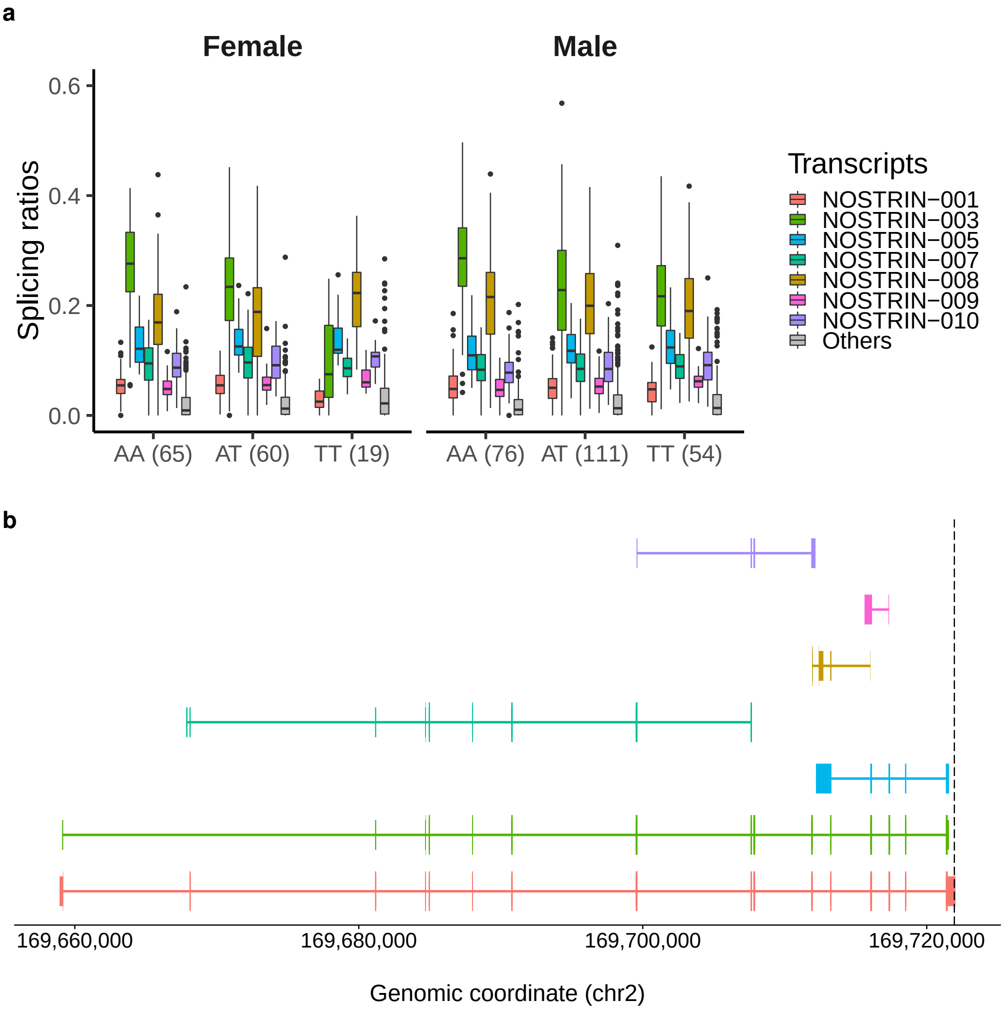

Figure S14. Relative abundances of the most expressed isoforms in adipose subcutaneous tissue from the NOSTRIN gene (chr2:169,643,049-169,722,024, forward strand) for each genotype group at the rs12993143 locus (chr2:169,721,944, A/T), in females and males. The least abundant isoforms are grouped in *Others*. The number of individuals in each genotype group is shown between parentheses. Females homozygous for the reference allele (AA) at the SNP locus, express preferentially NOSTRIN-003 (green). In contrast, females homozygous for the alternative allele (TT) express preferentially NOSTRIN-008 (brown). Female heterozygous individuals exhibit intermediate abundances. In male individuals, however, the three isoforms display similar abundances independently of the genotype at rs12993143. **b)** Exonic structure of the isoforms different isoforms and location of the SNP (dashed line).

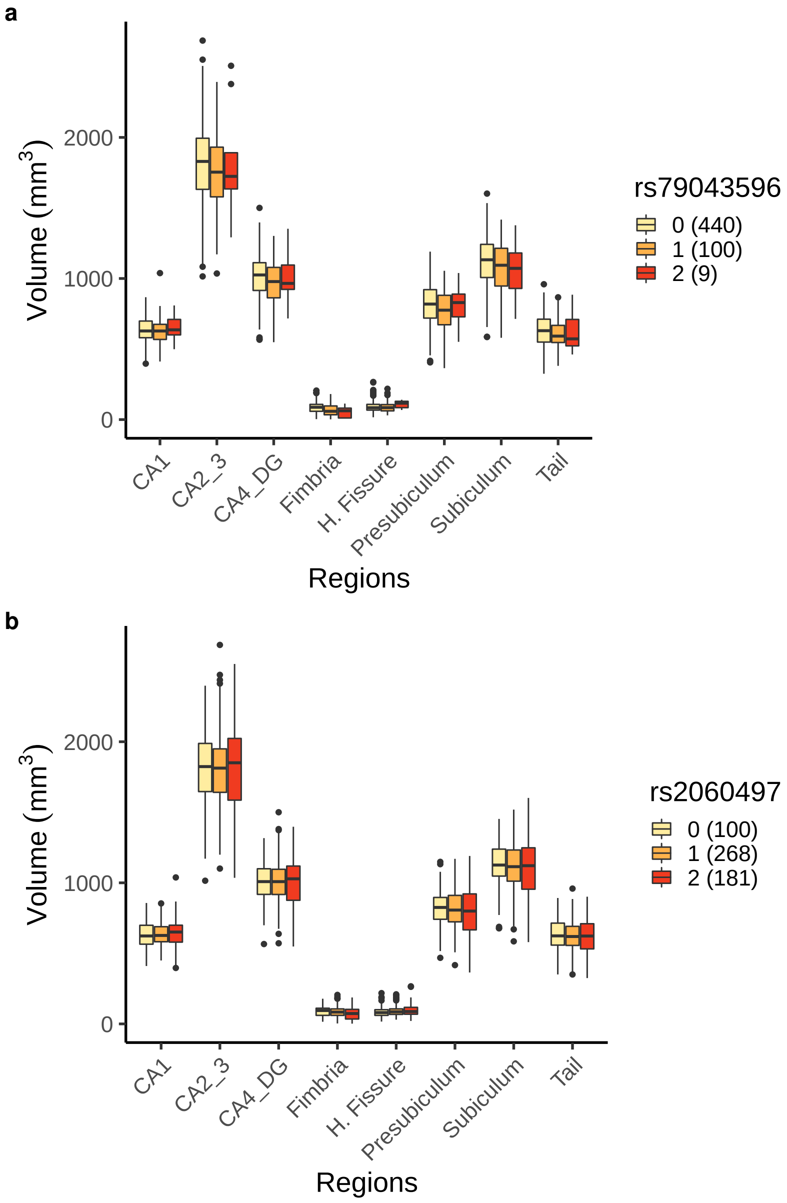

Figure S15. Volumes (mm^3) of 8 hippocampal subfields for each genotype group at **a)** the rs79043596 locus (chr1:104345514, G/C) and **b)** the rs2060497 locus (chr3:149388645, A/C). The number of individuals in each genotype group is shown between parentheses.

Appendices

Appendix 1: mathematical proofs

Anderson denotes by **D** the inter-distances matrix between the samples. If $\mathbf{A} = (a_{ij}) = \left(-\frac{1}{2}d_{ij}^2\right)$, then **G** in (10) is the matrix introduced by Gower[68]

$$\mathbf{G} = \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}\right)\mathbf{A}\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}\right)$$

In this general context, Anderson defines the *pseudo-F* statistic as:

$$\tilde{\mathrm{F}} = \frac{\mathrm{tr}(\mathbf{HGH})/\mathrm{rank}(\mathbf{H})}{\mathrm{tr}((\mathbf{I}-\mathbf{H})\mathbf{G}(\mathbf{I}-\mathbf{H}))/\mathrm{rank}(\mathbf{I}\text{-}\mathbf{H})} \tag{10}$$

Lemma 1. If **Y** is a centered-column matrix, and **D** is computed using the Euclidean distance, the Anderson statistic in (10) can be expressed as:

$$\tilde{\mathrm{F}} = \frac{\mathrm{tr}(\mathbf{Y}^{\mathrm{T}}\mathbf{HY})/\mathrm{rank}(\mathbf{H})}{\mathrm{tr}(\mathbf{Y}^{\mathrm{T}}(\mathbf{I}-\mathbf{H})\mathbf{Y})/\mathrm{rank}(\mathbf{I}\text{-}\mathbf{H})}$$

Proof. If $\mathbf{Y} = (y_{ij})$, any column mean $\overline{\mathbf{y}_{\cdot j}} = 0$. With the Euclidean distance and the definition of **G** it is straightforward to obtain:

$$g_{ij} = -\frac{1}{2}\left(d_{ij}^2 - \overline{d_{\cdot j}^2} - \overline{d_{i\cdot}^2} + \overline{d_{\cdot\cdot}^2}\right) = \sum_{k=1}^{p} y_{ik}y_{jk}$$

Thus, $\mathbf{G} = \mathbf{Y}\ \mathbf{Y}^{\mathrm{T}}$. The result is obtained combining the properties of the trace of the matrix product and the idempotence of the hat matrix **H**. □

Bai et al ([69]) showed general results on the asymptotics of the m-estimation in the multivariate regression field. For our purposes the main result in ([69]) is Theorem 2.4, which in our context can be simplified to:

Theorem 1. *Assume in model (1) the rows of* ***U*** *being iid with the same covariance matrix* Σ*. If* $\Sigma = \boldsymbol{P}\Lambda\boldsymbol{P}^T$, $\Lambda = diag(\lambda_j)$, $vec(\boldsymbol{\beta}) = \boldsymbol{\beta}_v$ *is the vectorized form of the* β *parameter matrix and* $\hat{\boldsymbol{\beta}}_v$ *the corresponding estimates. Under mild regularity conditions of the design matrix, the following result holds:*

$$(\hat{\boldsymbol{\beta}}_v - \boldsymbol{\beta}_v)^T(\boldsymbol{P}\boldsymbol{P}^T) \otimes (\boldsymbol{X}^T\boldsymbol{X})(\hat{\boldsymbol{\beta}}_v - \boldsymbol{\beta}_v) \xrightarrow{d} \sum_{j=1}^{q} \lambda_j \chi_j^2(p)$$

where $\chi_j^2(p)$ *is a collection of* q *independent chi-square variables with* p *degrees of freedom*

Proof. The complete proof can be found in ([69]), however their notation has some differences of the usual MMR notation we have adopted here. To help the reading of ([69]) we translate some symbols and provide details of the most important matrices. First, model in ([69]) is stated as:

$$\mathbf{Y}_i = \mathbf{X}_{i_B}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{E}_i \qquad i = 1, \cdots, n$$

$\mathbf{E}_i$ stands for a i.i.d. q-vector of errors, $\mathbf{X}_{i_B}$ is a $m \times q$ design matrix. Some other noticeable differences between notations are:

1. Bai’s model equates the response of an *individual* i sample, while model (1) stands for the full sample of n individuals.

2. Bai’s m stands for the dimension of $\boldsymbol{\beta}$, then, $m = p \times q$.

Denoting in model (1) row i of $\mathbf{X}$ as $\mathbf{X}_i$ (without B) both design matrices are related by:

$$\mathbf{X}_{i_B}^{\mathrm{T}} = \mathbf{I}_q \otimes \mathbf{X}_i$$

The mild regularity conditions described in the proposition refer specifically to condition (M6) in ([69]), that is, if $\mathbf{S}_n = \mathbf{X}_{1_B}\mathbf{X}_{1_B}^{\mathrm{T}} + \cdots + \mathbf{X}_{n_B}\mathbf{X}_{n_B}^{\mathrm{T}}$ then $\mathbf{S}_n$ must be non-singular for $n \geq n_0$ and

$$d_n^2 = \max_{1 \leq i \leq n} tr(\mathbf{X}_{i_B} \boldsymbol{S}_n^{-1} \mathbf{X}_{i_B}^{\mathrm{T}}) \to 0 \qquad as \quad n \to \infty$$

which can be interpreted as the individual design matrix having a leverage tending to zero, something that seems reasonable in practice. The remaining conditions (M1) to (M5) in ([69]) are trivially satisfied here. Finally, the auxiliary matrices $\mathbf{K}_n$ and $\mathbf{T}_n$ in the proof in ([69]) are, respectively:

$$\mathbf{K}_n = \mathbf{I}_q \otimes (\mathbf{X}^{\mathrm{T}}\mathbf{X})$$

$$\mathbf{T}_n = \Sigma \otimes (\mathbf{X}^{\mathrm{T}}\mathbf{X})$$

□

Consider now a null hypothesis where the parameters from $p_0 + 1$ to p are zero for all the q dimensions, that is, an hypothesis on a subset of the $p \times q$ possible parameters. Under this null hypothesis, for any single column of $\mathbf{Y}$, the corresponding column in matrix β in equation (1) will be multiplied by

$$\mathbf{R} = \left(\mathbf{0}_{(p-p_0)\times p_0}\, , \, \mathbf{I}_{p-p_0}\right)$$

And joining all the dimensions we have the matrix $\mathbf{R}_{\mathrm{v}}$:

$$\mathbf{R}_{\mathrm{v}} = \mathbf{I}_q \otimes \mathbf{R}$$

which allows to express synthetically the hypothesis in the following Lemma, that specifies the limiting distribution of the Anderson’s test for any subset of parameters.

Lemma 2. Assume a null hypothesis where all the parameters from columns $p_0 + 1$ to p in (1) are zero for all the q dimensions, that is:

$$\mathbf{R}_{\mathrm{v}} \boldsymbol{\beta}_{\mathrm{v}} = \mathbf{0}$$

then the numerator of the statistic in (5) converges in law to:

$$\mathrm{tr}\left\{\mathbf{Y}^{\mathrm{T}} \left(\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}} - \mathbf{X}_0(\mathbf{X}_0^{\mathrm{T}}\mathbf{X})_0^{-1}\mathbf{X}_0^{\mathrm{T}}\right) \mathbf{Y}\right\} \xrightarrow{d} \sum_{j=1}^{q} \lambda_j \chi_j^2(p - p_0)$$

Proof. The demonstration has three parts. First, consider the design matrix partitioned in two boxes $\mathbf{X} = (\mathbf{X}_0, \mathbf{X}_1)$ where $\mathbf{X}_0$ corresponds to $\mathbf{X}$ without the columns associated to the subset of coefficients assumed to be zero. Then, prove that $\mathbf{H} - \mathbf{H}_0$ is an idempotent matrix.

Second, obtain the limiting distribution of $\mathbf{R}_{\mathrm{v}}\,(\hat{\boldsymbol{\beta}}_{\mathrm{v}} - \boldsymbol{\beta}_{\mathrm{v}})$ applying Theorem 1 and the *Product Limit Normal Rule*, and then derive the convergence in law of the following expression:

$$\left(\mathbf{I}_q \otimes \left(\mathbf{R}\,(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\,\mathbf{R}^{\mathrm{T}}\right)^{-\frac{1}{2}}\right)\,\mathbf{R}_{\mathrm{v}}\,(\hat{\boldsymbol{\beta}}_{\mathrm{v}} - \boldsymbol{\beta}_{\mathrm{v}}) \xrightarrow{d} N(\mathbf{0},\, \Sigma \otimes \mathbf{I}_{p-p_0})$$

Third, consider again the block partitioning of $\mathbf{X}$ in boxes $\mathbf{X} = (\mathbf{X}_0, \mathbf{X}_1)$ and prove that

$$\begin{aligned} \mathrm{tr}(\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_0^{\mathrm{T}}\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0\hat{\boldsymbol{\beta}}_0) &= \mathbf{y}_{\mathrm{v}}^{\mathrm{T}}\,\left(\mathbf{I}_q \otimes \left(\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}} - \mathbf{X}_0(\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0)^{-1}\mathbf{X}_0^{\mathrm{T}}\right)\right)\,\mathbf{y}_{\mathrm{v}} \\ &= \hat{\boldsymbol{\beta}}_{\mathrm{v}}^{\mathrm{T}}\,\mathbf{R}_{\mathrm{v}}^{\mathrm{T}}\left(\mathbf{I}_q \otimes (\mathbf{R}\,(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\,\mathbf{R}^{\mathrm{T}})^{-\frac{1}{2}}\right)\left(\mathbf{I}_q \otimes (\mathbf{R}\,(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\,\mathbf{R}^{\mathrm{T}})^{-\frac{1}{2}}\right)\mathbf{R}_{\mathrm{v}}\,\hat{\boldsymbol{\beta}}_{\mathrm{v}} \end{aligned}$$

Under the null hypothesis:

$$\left(\mathbf{P} \otimes \left(\mathbf{R}\,(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\,\mathbf{R}^{\mathrm{T}}\right)^{-\frac{1}{2}}\right)\,\mathbf{R}_{\mathrm{v}}\,\hat{\beta}_{\mathrm{v}} \xrightarrow{d} N(\mathbf{0},\, \Lambda \otimes \mathbf{I}_{p-p_0})$$

Because the limit covariance matrix is diagonal, all the $(p - p_0) \times q$ components are independent. The squared elements summed in the numerator correspond to squared univariate normals with zero mean and variance λ_j. Finally, group in $\chi_j^2(p - p_0)$ variables the components with identical eigenvalue. □

Denominators in (3) and (5) are identical to the sum of squared error terms of the simple regressions on each column of $\mathbf{Y}$. A well-known result of the OLS asymptotic properties states each of these terms converges in probability to the corresponding diagonal element of Σ. The assumptions for such convergence are analogous to Bai’s assumptions. Therefore, the denominator of the Anderson’s statistic converges in probability to the trace of Σ, that is, to $\sum_{j=1}^{q} \lambda_j$. Then, applying Lemma 2 and Slutsky’s theorem, the limiting distribution of the statistic is obtained. In practice, probability tails can be computed considering only its numerator (the weights of the linear combination of chi-squares will be λ_j) or the pseudo-F ratio (weights will be $\lambda_j / \sum_{j=1}^{q} \lambda_j$).

Appendix 2: Data generation in the simplex under H_0 and H_1

Given two points (i.e. vectors of proportions) in the $q-1$ simplex, $\boldsymbol{f}_1 = (f_{11}, \ldots, f_{1q})$ and $\boldsymbol{f}_2 = (f_{21}, \ldots, f_{2q})$, a problem of interest is to find the closest point in the simplex to $\boldsymbol{f}_1$, in the direction determined by $\boldsymbol{f}_2$, obtained by adding an amount Δ to f_{11}. We name this point $\boldsymbol{f}'_1$. In our context, we use the Hellinger distance to assess the dissimilarity between vectors of proportions, which is not Euclidean in its natural parametrization. Therefore, to find $\boldsymbol{f}'_1$ we need to respect the geometry induced by the distance.

The geometry of the Hellinger distance in the simplex is easier to visualize if we transform the proportions to their square root, $\boldsymbol{f}_i = (\sqrt{f_{i1}}, \ldots, \sqrt{f_{iq}})$, so that the vectors are located on the surface of the upper octant of a hypersphere of radius 1. In this surface, the shortest path between $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ is the geodesic given by:

$$f'_{1j} = \left\{ \sqrt{f_{1j}}\ \cos(d/2) + \frac{\sqrt{f_{2j}} - \sqrt{f_{1j}}\ \cos(\rho/2)}{\sin(\rho/2)}\ \sin(d/2) \right\}^2 \qquad j \in \{1, \ldots, q\} \tag{11}$$

where d is the distance traveled along the geodesic between $\boldsymbol{f}_1$ and $\boldsymbol{f'_1}$, and ρ the length of the arc of the hypersphere between $\boldsymbol{f_1}$ and $\boldsymbol{f_2}$, $\rho = 2\arccos(\sum_{j=1}^{q} \sqrt{f_{1j} f_{2j}})$.

After some straightforward algebra we can obtain the expression of the components $2 \ldots q$ along the geodesic satisfying $f'_{11} = f_{11} + \Delta$:

$$f'_{1j} = f_{1j} \left(1 - \frac{\Delta}{1 - f_{11}}\right) \quad j \in \{2, \ldots, q\} \tag{12}$$

During data generation in the multivariate proportion scenario, we employed equation (12) to generate $\boldsymbol{p'}$ from $\boldsymbol{p}$, that is, the multivariate mean under H_1 given the multivariate mean under H_0 and different values of Δ. Moreover, to actually generate random observations of vectors of proportions with certain variability, while ensuring that $E(\boldsymbol{f}_i) = \boldsymbol{p}$ (or, equivalently, $E(\boldsymbol{f}_i) = \boldsymbol{p'}$ under H_1), we proceeded as follows:

1. Generate a vector of q step sizes, $\boldsymbol{\delta}$, using any probability distribution, so that that $E(\delta_j) = 0$ and $Var(\delta_j)$ = σ_g^2, for $j \in \{1, \ldots, q\}$. Specifically, we obtained $\delta_j \sim N(0, \sigma_g^2)$.

2. Select randomly the order in which the steps towards the simplex vertices $\{\boldsymbol{e}_1 = (1, 0, 0, \ldots, 0)$, $\boldsymbol{e}_2 = (0, 1, 0, \ldots, 0), \ldots$, $\boldsymbol{e}_q = (0, 0, \ldots, 1)\}$ are taken.

3. If $\boldsymbol{f}_i^{(0)} = \boldsymbol{p}$, for each $j = 1, \ldots q$, $\boldsymbol{f}_i^{(j)}$ is obtained from $\boldsymbol{f}_i^{(j-1)}$ advancing from $\boldsymbol{f}_i^{(j-1)}$ towards the vertex selected in the previous step, using equation (12) with $\Delta = \delta_j$. When $j = q$ the sample generated corresponds to the last step $\boldsymbol{f}_i = \boldsymbol{f}_i^{(q)}$.

Note that σ_g should be small enough so that $\sum\limits_{j=1}^{q} f_{ij} = 1, \forall i \in \{1, \ldots, n\}$. In addition, the variances of the response variables generated, σ^2_{jj}, depend on the parameters σ_g, q, L and the probability distribution selected to generate δ_j. In practice, for a given probability distribution, we selected different values of σ_g to ensure $\overline{\boldsymbol{\sigma_{jj}}} = 0.03$ across different values of q and L.

Note on data generation under H_1 for numerical predictors

Data generation under H_1 for a given predictor X, when it is a factor, is depicted in the main text. When X is numerical, we simulated $f^{(0)}_{ij}$ so that:

$$f^{(0)}_{ij} = \begin{cases} p_j + x_i + \epsilon_i & j = 1 \\ p_j & 2 \leq j \leq q \end{cases}$$

where p_j is the j-th element of the $\boldsymbol{p}$ vector and $\epsilon_i \sim U(-\omega, \omega)$. We then applied the algorithm depicted above. We selected values of ω so that $Cor(f_{i1}, x_i) = r$, with $r \in$ [0, 0.4].

DISCUSSION

This thesis work initially focused on the identification of genetic effects on alternative splicing, by applying a multivariate approach across a large set of human tissues, as depicted in Chapter 1. In addition to the extensive catalogue of splicing QTLs generated, which constitutes itself a resource of interest to the field, we hope that the set of analyses performed helps to advance the understanding of alternative splicing regulation and its contribution to human complex traits and diseases. Moreover, the challenges encountered during our sQTL study motivated the development of new methodological approaches, aimed to handle the increasing size and complexity of current omics datasets. Specifically, in Chapter 2 we described a new command-line tool for splicing event visualization across multiple samples, and in Chapter 3 we presented an efficient non-parametric approach for multivariate association testing in GWAS and QTL mapping. Altogether, this work supposes a valuable contribution to address the fundamental question of how genetics shapes phenotypic traits.

Yet another QTL mapping study?

In the past years, next-generation sequencing technologies have enabled the study of a plethora of molecular phenotypes. Characterizing the effects of human genetic variation throughout the different molecular layers is essential to understand GWAS associations. However, while substantial efforts have been devoted to investigate genetic effects on transcriptional and epigenetic regulation, the interplay between genetic variants and alternative splicing (i.e. sQTLs) has received, in comparison, limited attention.

In light of this, as presented in Chapter 1, we leveraged the GTEx resource (Lonsdale et al., 2013) to generate the most comprehensive set of sQTLs across healthy human tissues reported to date. Our work offers two major advancements with respect to previous multi-tissue sQTL analyses. First, we used a multivariate approach for sQTL mapping (Monlong et al., 2014), which targets global changes in the relative abundances of a gene's transcript isoforms, rather than focusing on specific splicing events. This presents clear advantages over standard univariate methods, given that sQTLs tend to involve simultaneously multiple types of splicing events. Second, we surveyed a much larger number of tissues, showing that genetic effects on splicing are highly shared. This has implications for the study of the effects of disease-causing splicing mutations across different tissues. Furthermore, it allowed to determine which tissues present the most specific patterns of splicing regulation (e.g. brain, muscle, testis). This provides relevant information that may serve as a guide for experimental validations, typically done in cell lines which differ from cells *in vivo*.

The availability of eQTLs in GTEx enabled integrative analyses to explore the regulatory interplay between transcription and splicing. Indeed, we show that a substantial fraction of sQTLs are also eQTLs for the same gene and tissue. This observation, which departs from previous reports (Li et al., 2016), is partially due to the nature of our sQTLs (since they account for changes in transcriptional termini in addition to canonical splicing events), but also highlights the tight association between these two processes in terms of regulation (Naftelberg et al., 2015). Furthermore, we reported many variants affecting the splicing of a gene and the expression of a different one. This suggests that the pleiotropic effect of regulatory variants may be mediated by distinct molecular mechanisms through different genes, uncovering unexpected complexity in the regulatory program encoded by the human genome. Of note, when investigating the completeness of the splicing process, which is related to the degree of coupling between splicing and transcription, sQTLs appeared to play a preferential role in the regulation of post-transcriptional splicing.

An additional landmark of our work is the characterization of the mechanisms through which genetic variants affect splicing patterns, which include, beyond direct impact on donor and acceptor sites, the modification of binding sites of RNA-binding proteins (RBPs). This appears to occur more frequently, although the former often results in stronger effects on splicing. Nevertheless, we expect both mechanisms to cooperate, given that many RBPs are indeed splicing regulators, which tend to bind near splice sites (Fu and Ares, 2014).

Finally, our analyses contributed to shed light upon the relationship between genetic variation, alternative splicing and human phenotypes. We found that sQTLs are enriched in GWAS loci associated with several complex traits and diseases, and most importantly, that they display comparable or even stronger GWAS associations than genetic variants affecting gene expression. In addition, sQTLs altering RBP binding seem to play a particularly relevant role in disease. These observations grant splicing a central role in mediating the impact of genetic variation on human phenotypes, a fact only recently acknowledged (Li et al., 2016).

Our work, however, presents some limitations, which open to potential avenues for future research. An important challenge, common to most splicing analyses, is the correct estimation of the abundances of full-length transcript isoforms from short-read RNA-seq data (Vaquero-Garcia et al., 2016). This might change soon, thanks to continuous developments in the field of long-read RNA sequencing (van Dijk et al., 2018). Still, the higher error rates and especially the lower throughput of these techniques need to be handled (Park et al., 2018).

A second aspect worth mentioning is that here we analyzed the transcriptome of whole tissues. Yet tissues are complex mixtures of several cell types, and given that heterogeneous cellular composition often underlies transcriptional differences (e.g. breast tissue in males and females), cell-type deconvolution would be relevant not only to gain power for standard QTL analyses, but also to identify cell-type specific QTLs (van der Wijst et al., 2018). In this regard, single-cell RNA-seq (Kolodziejczyk et al., 2015), as well as emerging technologies which preserve spatial information about the tissue context or subcellular localization of the RNA (Eng et al., 2019, Rodriques et al., 2019), offer promising results.

A more detailed analysis of the relationship between sQTLs and GWAS loci, as well as the identification of the actual causal variants, could be achieved thanks to statistical co-localization (Hormozdiari et al., 2016, Wen et al., 2017) and fine-mapping approaches (Brown et al., 2017, Hormozdiari et al., 2014), respectively. Nevertheless, this requires further work to adapt the existing methods, or develop new approaches, so that they can be applied within our multivariate, non-parametric framework.

The challenges of Big Data Biology

In recent years, continuous developments of high-throughput sequencing technologies have enabled the genome-wide profiling of a wealth of molecular traits (DNA methylation, chromatin status, transcript expression, etc.), in addition to genotypes (Consortium, 2012, Kundaje et al., 2015, Lonsdale et al., 2013). Moreover, the amount and complexity of this data keeps growing day after day, making of genomics one of the most demanding Big Data domains (Stephens et al., 2015). In parallel, the availability of deep phenotype data has dramatically increased, as a result of considerable efforts to build comprehensive phenotype resources, such as biobanks (Sudlow et al., 2015). Certainly, this data deluge offers countless opportunities for personalized medicine, but also poses several challenges for current computational analyses, including data visualization, integration, computational efficiency and reproducibility, among others. With the aim of overcoming these difficulties, some of them faced during our analysis of genetic effects on alternative splicing across GTEx tissues, we developed new statistical methods and analysis tools.

Efficient visualization of high-dimensional datasets is crucial for exploratory data analysis. Nonetheless, even a relatively simple scenario, such as the visualization of splicing events from RNA-seq, may become intractable with current software when large sample sizes are available. As depicted in Chapter 2, the sashimi plot (Katz et al., 2015) is the standard representation of splicing events. However, currently available implementations display each RNA-seq experiment on a separate line. As a result, visual comparison of more than a few samples becomes unfeasible, and some form of aggregation is required. Moreover, splicing visualization is further hindered by the presence of long intronic regions without splicing events. To exceed these and other limitations, we developed a software to generate sashimi plots, which scales for a large number of samples by multiple aggregation methods (overlay of the signal of different samples, mean, median) and focuses on informative regions, by scaling down genomic segments between splice sites (Chapter 2). This exemplifies the relevance of the development of new visualization tools as the size of genomic datasets keeps increasing.

A second major challenge is the comprehensive analysis and integration of the different layers of information available (Ritchie et al., 2015). Modelling higher-order relationships between molecular traits, as well as between these and the environment, is key in order to understand their interactions and relative contributions to human phenotypes (Civelek and Lusis, 2014). In addition, characterizing the connections between different complex traits and diseases may shed light on their shared risk factors. To our understanding, a first step towards data integration is facilitated by multivariate analyses, which allow joint modelling of multiple traits by taking advantage of their correlated structure (Stephens, 2013). However, despite their undoubted potential and the fact that intrinsically multivariate phenotypes are widespread in Biology (size and connectivity of brain regions, levels of blood lipids, cellular composition of a tissue, expression of genes in the same pathway, composition of the gut microbiota, etc.), these methods are not commonly used. In the context of this thesis, multivariate analysis is employed to identify genetic effects on multiple traits, both in the framework of GWAS and QTL mapping, demonstrating the applicability of this strategy. To cite an example, using a cohort 50 times smaller than the ones employed in similar studies, we were able to identify loci strongly associated with the reduced volume of hippocampal subfields, some of them with potential roles in the physiopathology of Alzheimer's disease. Remarkably, multivariate approaches can also be employed to leverage the information of features that are implicit in the data, and that have been automatically learnt using deep learning (Angermueller et al., 2016). An example is represented by the features that convolutional auto encoder networks can extract from histological images (Ash et al., 2018).

Due to the large scale of current omics datasets, new visualization techniques and integrative approaches also require efficient algorithms to achieve fast computations. We show the importance of this in Chapter 3. The multivariate non-parametric statistical framework provided by Anderson test relied on permutations to assess significance in complex designs, resulting in prohibitive running times with datasets like GTEx. Hence, we took considerable efforts to derive the limiting distribution of the Anderson test statistic, with the aim of computing asymptotic *p* values. Our asymptotic result guarantees high accuracy while reducing dramatically the computation time for large sample sizes. Nevertheless, even highly efficient pipelines often require parallelization strategies able to absorb heavy computational workloads. This applies to the sQTL mapping pipeline presented in Chapter 1, as well as to all the simulations and analyses depicted in Chapter 3.

Another major concern in current large-scale omics analyses is reproducibility. Although in-silico experiments are expected to be reproducible, in practice replicating a typical computational biology pipeline can take months (Garijo et al., 2013). Common sources of computational irreproducibility include the lack of good practices regarding software dependencies and numerical instability. The latter is more difficult to fix, as it arises from small variations across computational platforms, being particularly problematic in high-performance computing (HPC) environments, including the cloud (Di Tommaso et al., 2017). To give an example, even a simple differential expression analysis can result in a distinct number of significant genes when run on different platforms (i.e. Mac OS, Ubuntu, Amazon Linux, etc.). In this thesis work, in order to guarantee the reproducibility of our results while ensuring highly parallel and portable computation, we have taken advantage of Nextflow (`https://www.nextflow.io/`) and Docker (`https://www.docker.com/`) technologies. Specifically, all our main pipelines are written in the Nextflow domain-specific language, and the software dependencies are containerized using Docker.

Finally, we hope that the methodological advancements presented in this thesis contribute to extract valuable information from the highly complex and extensive genetic, molecular and phenotypic data available, supposing a step forward in these exciting times to study Biology.

CONCLUSIONS

The work presented in this thesis addresses the study of genetic effects on alternative splicing across human tissues, and provides novel statistical methods and analysis tools for alternative splicing visualization and multivariate association testing in the framework of GWAS and QTL mapping.

Here is a summary of the main contributions of this thesis:

- We have developed an efficient and reproducible pipeline for the discovery of genetic variants affecting splicing (sQTLs), based on an approach that captures the intrinsically multivariate nature of this phenomenon. We have employed it to analyze the multi-tissue GTEx dataset, generating the most comprehensive catalogue to date of sQTLs in the human genome.

- The analyses of this sQTL catalogue revealed that:

  - Genetic effects on splicing tend to be shared across multiple tissues.
  - Genetic variants often affect both transcription and splicing, but not always of the same target gene.
  - Post-transcriptional splicing is under stronger regulation than co-transcriptional splicing.
  - Genetic variants can affect splicing through i) direct impact on donor and acceptor splice sites and ii) modification of binding sites of RNA-binding proteins (RBPs). While the latter is more common, the former leads to stronger effects.
  - Genetic variants affecting splicing can have a phenotypic impact comparable or even stronger than variants affecting gene expression, with those altering RBP binding playing a prominent role in disease.

- We have developed a command-line tool for alternative splicing visualization across multiple samples. Given a specified genomic region, it generates sashimi plots for individual RNA-seq experiments, as well as aggregated plots for groups of experiments, a feature unique to this software. It is annotation-independent, uses standard bioinformatics file formats and allows the visualization of splicing events even for large genomic regions, by scaling down the genomic segments between splice sites.

- We have extended the statistical framework provided by Anderson test, initially employed for sQTL mapping. Specifically:
  - We have proven that, in the case of complex designs and Euclidean distances, the limiting distribution of the Anderson test statistic is a linear combination of independent chi-square variables, where the coefficients are the eigenvalues of the residual covariance matrix. This result also holds after any transformation that preserves the independence of the observations of the response variables.
  - We developed a fast implementation of the asymptotic test, that allows asymptotic *p* value computation for predictors in user-defined multivariate regression models.
  - In extensive simulations, we showed controlled type I error rates and high power for the asymptotic test.
  - The asymptotic test has proven valuable to identify genetic associations with multivariate phenotypes in the context of genome-wide association studies (GWAS) and QTL mapping analyses.

BIBLIOGRAPHY

Alamancos, G. P., Agirre, E., and Eyras, E. (2014). Methods to study splicing from high-throughput RNA sequencing data. *Methods in Molecular Biology*, 1126(Table 9):357–397.

Alasoo, K., Rodrigues, J., Danesh, J., Freitag, D. F., Paul, D. S., and Gaffney, D. J. (2019). Genetic effects on promoter usage are highly context-specific and contribute to complex traits. *eLife*, 8.

Albert, F. W. and Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, 16(4).

Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. *Genome Research*, 22(10):2008–2017.

Anderson, M. (2001). A new method for non-parametric multivariate analysis of variance. *Australian Ecology*, 26(2001):32–46.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., Li, K., Schwarzfischer, L., et al. (2014). An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–461.

Angermueller, C., Pärnamaa, T., Parts, L., and Stegle, O. (2016). Deep learning for computational biology. *Molecular Systems Biology*, 12(7):878.

Aschard, H., Guillemot, V., Vilhjalmsson, B., Patel, C. J., Skurnik, D., Ye, C. J., Wolpin, B., Kraft, P., and Zaitlen, N. (2017). Covariate selection for association screening in multiphenotype genetic studies. *Nature Genetics*.

Aschard, H., Vilhjálmsson, B. J., Greliche, N., Morange, P.-E., Trégouët, D.-A., and Kraft, P. (2014). Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *American journal of human genetics*, 94(5):662–76.

Aschard, H., Vilhjálmsson, B. J., Joshi, A. D., Price, A. L., and Kraft, P. (2015). Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *American journal of human genetics*, 96(2):329–39.

Ash, J. T., Darnell, G., Munro, D., and Engelhardt, B. E. (2018). Joint analysis of gene expression levels and histological images identifies genes associated with tissue morphology. *bioRxiv*, page 458711.

Banovich, N. E., Lan, X., McVicker, G., van de Geijn, B., Degner, J. F., Blischak, J. D., Roux, J., Pritchard, J. K., and Gilad, Y. (2014). Methylation QTLs Are Associated with Coordinated Changes in Transcription Factor Binding, Histone Modifications, and Gene Expression Levels. *PLoS Genetics*, 10(9):e1004663.

Battle, A., Khan, Z., Wang, S. H., Mitrano, A., Ford, M. J., Pritchard, J. K., and Gilad, Y. (2015). Impact of regulatory variation from RNA to protein. *Science*, 347(6222):664–667.

Battle, A., Mostafavi, S., Zhu, X., Potash, J. B., Weissman, M. M., McCormick, C., Haudenschild, C. D., Beckman, K. B., Shi, J., Mei, R., Urban, A. E., Montgomery, S. B., Levinson, D. F., and Koller, D. (2014). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Research*, 24(1):14–24.

Beasley, T. M., Erickson, S., and Allison, D. B. (2009). Rank-Based Inverse Normal Transformations are Increasingly Used, But are They Merited? *Behavior Genetics*, 39(5):580–595.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Bentley, D. L. (2014). Coupling mRNA processing with transcription in time and space. *Nature Reviews Genetics*, 15(3):163–175.

Bolisetty, M. T., Rajadinakaran, G., and Graveley, B. R. (2015). Determining exon connectivity in complex mRNAs by nanopore sequencing. *Genome Biology*, 16(1):204.

Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*, 169(7):1177–1186.

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525–527.

Brown, A. A., Viñuela, A., Delaneau, O., Spector, T. D., Small, K. S., and Dermitzakis, E. T. (2017). Predicting causal variants affecting expression by using whole-genome sequencing and RNA-seq from multiple human tissues. *Nature Genetics*, 49(12):1747–1751.

Casale, F. P. (2016). *Multivariate linear mixed models for statistical genetics*. PhD thesis, University of Cambridge.

Casale, F. P., Rakitsch, B., Lippert, C., and Stegle, O. (2015). Efficient set tests for the genetic analysis of correlated traits. *Nature Methods*.

Chandramouli, K. and Qian, P.-Y. (2009). Proteomics: challenges, techniques and possibilities to overcome biological sample complexity. *Human genomics and proteomics : HGP*, 2009.

Chen, L., Ge, B., Casale, F. P., Vasquez, L., Kwan, T., Garrido-Martín, D., Watt, S., Yan, Y., Kundu, K., Ecker, S., Datta, A., Richardson, D., Burden, F., Mead, D., Mann, A. L., et al. (2016). Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell*, 167(5):1398–1414.

Chen, M. and Manley, J. L. (2009). Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nature reviews. Molecular cell biology*, 10(11):741–754.

Chu, C.-S., Trapnell, B. C., Curristin, S., Cutting, G. R., and Crystal, R. G. (1993). Genetic basis of variable exon 9 skipping in cystic fibrosis transmembrane conductance regulator mRNA. *Nature Genetics*, 3(2):151–156.

Civelek, M. and Lusis, A. J. (2014). Systems genetics approaches to understand complex traits. *Nature reviews. Genetics*, 15(1):34–48.

Conneely, K. N. and Boehnke, M. (2007). So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests. *American journal of human genetics*, 81(6):1158–68.

Consortium, G. (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–60.

Consortium, G. (2017). Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213.

Consortium, T. B., Anttila, V., Bulik-Sullivan, B., Finucane, H. K., Walters, R. K., Bras, J., Duncan, L., Escott-Price, V., Falcone, G. J., Gormley, P., Malik, R., Patsopoulos, N. A., Ripke, S., Wei, Z., Yu, D., et al. (2018). Analysis of shared heritability in common disorders of the brain. *Science*, 360(6395):eaap8757.

Consortium, T. E. P. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.

Consortium, T. F., the RIKEN PMI, CLST, Forrest, A. R. R., Kawaji, H., Rehli, M., Baillie, J. K., Hoon, M. J. L. d., Haberle, V., Lassmann, T., Kulakovskiy, I. V., Lizio, M., Itoh, M., Andersson, R., Mungall, C. J., et al. (2014). A promoter-level mammalian expression atlas. *Nature*, 507(7493):462–470.

Consortium, T. H. R., McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., Kang, H. M., Fuchsberger, C., Danecek, P., Sharp, K., Luo, Y., Sidore, C., Kwong, A., Timpson, N., et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, 48(10):1279–1283.

Dahl, A., Guillemot, V., Mefford, J., Aschard, H., and Zaitlen, N. (2019). Adjusting for Principal Components of Molecular Phenotypes Induces Replicating False Positives. *Genetics*, 211(4):1179–1189.

Davis, J. R., Fresard, L., Knowles, D. A., Pala, M., Bustamante, C. D., Battle, A., and Montgomery, S. B. (2016). An Efficient Multiple-Testing Adjustment for eQTL Studies that Accounts for Linkage Disequilibrium between Variants. *American Journal of Human Genetics*.

Day, F. R., Loh, P.-R., Scott, R. A., Ong, K. K., and Perry, J. R. B. (2016). A Robust Example of Collider Bias in a Genetic Association Study. *American journal of human genetics*, 98(2):392–3.

de Almeida, S. F., Grosso, A. R., Koch, F., Fenouil, R., Carvalho, S., Andrade, J., Levezinho, H., Gut, M., Eick, D., Gut, I., Andrau, J.-C., Ferrier, P., and Carmo-Fonseca, M. (2011). Splicing enhances recruitment of methyltransferase HYPB/Setd2 and methylation of histone H3 Lys36. *Nature Structural & Molecular Biology*, 18(9):977–983.

de Klerk, E. and 't Hoen, P. A. C. (2015). Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. *Trends in genetics : TIG*, 31(3):128–39.

de la Mata, M. and Kornblihtt, A. R. (2006). RNA polymerase II C-terminal domain mediates regulation of alternative splicing by SRp20. *Nature Structural & Molecular Biology*, 13(11):973–980.

de Lange, K. M., Moutsianas, L., Lee, J. C., Lamb, C. A., Luo, Y., Kennedy, N. A., Jostins, L., Rice, D. L., Gutierrez-Achury, J., Ji, S.-G., Heap, G., Nimmo, E. R., Edwards, C., Henderson, P., Mowat, C., et al. (2017). Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nature Genetics*, 49(2):256–261.

Degner, J. F., Pai, A. A., Pique-Regi, R., Veyrieras, J.-B., Gaffney, D. J., Pickrell, J. K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G. E., Stephens, M., Gilad, Y., and Pritchard, J. K. (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*, 482(7385):390–394.

Delaneau, O., Ongen, H., Brown, A. A., Fort, A., Panousis, N. I., and Dermitzakis, E. T. (2017). A complete tool set for molecular QTL discovery and analysis. *Nature Communications*.

Demenais, F., Margaritte-Jeannin, P., Barnes, K. C., Cookson, W. O. C., Altmüller, J., Ang, W., Barr, R. G., Beaty, T. H., Becker, A. B., Beilby, J., Bisgaard, H., Bjornsdottir, U. S., Bleecker, E., Bønnelykke, K., Boomsma, D. I., et al. (2018). Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nature Genetics*, 50(1):42–53.

Dermitzakis, E. T. (2012). Cellular genomics for complex traits. *Nature Reviews Genetics*, 13(3):215–220.

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., and Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4):316–319.

Ellinghaus, D., Jostins, L., Spain, S. L., Cortes, A., Bethune, J., Han, B., Park, Y. R., Raychaudhuri, S., Pouget, J. G., Hübenthal, M., Folseraas, T., Wang, Y., Esko, T., Metspalu, A., Westra, H.-J., et al. (2016). Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. *Nature Genetics*, 48(5):510–518.

Elliott, L. T., Sharp, K., Alfaro-Almagro, F., Shi, S., Miller, K. L., Douaud, G., Marchini, J., and Smith, S. M. (2018). Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature*, 562(7726):210–216.

Eng, C.-H. L., Lawson, M., Zhu, Q., Dries, R., Koulena, N., Takei, Y., Yun, J., Cronin, C., Karp, C., Yuan, G.-C., and Cai, L. (2019). Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature*, 568(7751):235–239.

Eriksson, M., Brown, W. T., Gordon, L. B., Glynn, M. W., Singer, J., Scott, L., Erdos, M. R., Robbins, C. M., Moses, T. Y., Berglund, P., Dutra, A., Pak, E., Durkin, S., Csoka, A. B., Boehnke, M., et al. (2003). Recurrent de novo point mutations in lamin A cause Hutchinson-Gilford progeria syndrome. *Nature*, 423(6937):293–298.

Feng, Y., Zhang, Y., Ying, C., Wang, D., and Du, C. (2015). Nanopore-based Fourth-generation DNA Sequencing Technology. *Genomics, Proteomics & Bioinformatics*, 13(1):4–16.

Ferreira, M. A. R. and Purcell, S. M. (2009). A multivariate test of association. *Bioinformatics*, 25(1):132–133.

Francesconi, M. and Lehner, B. (2014). The effects of genetic variation on gene expression dynamics during development. *Nature*, 505(7482):208–211.

Fu, X.-D. and Ares, M. (2014). Context-dependent control of alternative splicing by RNA-binding proteins. *Nature Reviews Genetics*, 15(10):689–701.

Furlotte, N. A. and Eskin, E. (2015). Efficient multiple-trait association and estimation of genetic correlation using the matrix-variate linear mixed model. *Genetics*, 200(1):59–68.

Galesloot, T. E., van Steen, K., Kiemeney, L. A. L. M., Janss, L. L., and Vermeulen, S. H. (2014). A Comparison of Multivariate Genome-Wide Association Methods. *PLoS ONE*, 9(4):e95923.

Gamazon, E. R., Segrè, A. V., van de Bunt, M., Wen, X., Xi, H. S., Hormozdiari, F., Ongen, H., Konkashbaev, A., Derks, E. M., Aguet, F., Quan, J., Nicolae, D. L., Eskin, E., Kellis, M., Getz, G., et al. (2018). Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nature Genetics*, 50(7):956–967.

Garieri, M., Delaneau, O., Santoni, F., Fish, R. J., Mull, D., Carninci, P., Dermitzakis, E. T., Antonarakis, S. E., and Fort, A. (2017). The effect of genetic variation on promoter usage and enhancer activity. *Nature Communications*, 8(1):1358.

Garijo, D., Kinnings, S., Xie, L., Xie, L., Zhang, Y., Bourne, P. E., and Gil, Y. (2013). Quantifying Reproducibility in Computational Biology: The Case of the Tuberculosis Drugome. *PLoS ONE*, 8(11):e80278.

Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., and Plagnol, V. (2014). Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genetics*, 10(5):e1004383.

Giambartolomei, C., Zhenli Liu, J., Zhang, W., Hauberg, M., Shi, H., Boocock, J., Pickrell, J., Jaffe, A. E., Pasaniuc, B., and Roussos, P. (2018). A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics*, 34(15):2538–2545.

Goldstein, D. B. (2009). Common Genetic Variation and Human Traits. *New England Journal of Medicine*, 360(17):1696–1698.

Gupta, I., Clauder-Münster, S., Klaus, B., Järvelin, A. I., Aiyar, R. S., Benes, V., Wilkening, S., Huber, W., Pelechano, V., and Steinmetz, L. M. (2014). Alternative polyadenylation diversifies post-transcriptional regulation by selective RNA-protein interactions. *Molecular Systems Biology*, 10(2):719.

Hellquist, A., Zucchelli, M., Kivinen, K., Saarialho-Kere, U., Koskenmies, S., Widen, E., Julkunen, H., Wong, A., Karjalainen-Lindsberg, M.-L., Skoog, T., Vendelin, J., Cunninghame-Graham, D. S., Vyse, T. J., Kere, J., and Lindgren, C. M. (2007). The human GIMAP5 gene has a common polyadenylation polymorphism increasing risk to systemic lupus erythematosus. *Journal of medical genetics*, 44(5):314–21.

Henikoff, S. and Shilatifard, A. (2011). Histone modification: cause or cog? *Trends in genetics : TIG*, 27(10):389–96.

Herzel, L., Ottoz, D. S. M., Alpert, T., and Neugebauer, K. M. (2017). Splicing and transcription touch base: co-transcriptional spliceosome assembly and function. *Nature Reviews Molecular Cell Biology*, 18(10):637–650.

Higgs, D. R., Goodbourn, S. E. Y., Lamb, J., Clegg, J. B., Weatherall, D. J., and Proudfoot, N. J. (1983). α-Thalassaemia caused by a polyadenylation signal mutation. *Nature*, 306(5941):398–400.

Hodges, C., Bintu, L., Lubkowska, L., Kashlev, M., and Bustamante, C. (2009). Nucleosomal fluctuations govern the transcription dynamics of RNA polymerase II. *Science*, 325(5940):626–8.

Hooper, J. E. (2014). A survey of software for genome-wide discovery of differential splicing in RNA-Seq data. *Human Genomics*, 8(1):3.

Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B., and Eskin, E. (2014). Identifying causal variants at loci with multiple signals of association. *Genetics*, 198(2):497–508.

Hormozdiari, F., van de Bunt, M., Segrè, A. V., Li, X., Joo, J. W. J., Bilow, M., Sul, J. H., Sankararaman, S., Pasaniuc, B., and Eskin, E. (2016). Colocalization of GWAS and eQTL Signals Detects Target Genes. *American journal of human genetics*, 99(6):1245–1260.

Huang, Y., Li, W., Yao, X., Lin, Q.-J., Yin, J.-W., Liang, Y., Heiner, M., Tian, B., Hui, J., and Wang, G. (2012). Mediator complex regulates alternative mRNA processing via the MED23 subunit. *Molecular cell*, 45(4):459–69.

Hubner, N., Wallace, C. A., Zimdahl, H., Petretto, E., Schulz, H., Maciver, F., Mueller, M., Hummel, O., Monti, J., Zidek, V., Musilova, A., Kren, V., Causton, H., Game, L., Born, G., et al. (2005). Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nature Genetics*, 37(3):243–253.

Ip, J. Y., Schmidt, D., Pan, Q., Ramani, A. K., Fraser, A. G., Odom, D. T., and Blencowe, B. J. (2011). Global impact of RNA polymerase II elongation inhibition on alternative splicing regulation. *Genome research*, 21(3):390–401.

Jangi, M. and Sharp, P. (2014). Building Robust Transcriptomes with Master Splicing Factors. *Cell*, 159(3):487–498.

Jones, M. (2009). Kumaraswamy’s distribution: A beta-type distribution with some tractability advantages. *Statistical Methodology*, 6(1):70–81.

Joo, J. W. J., Kang, E. Y., Org, E., Furlotte, N., Parks, B., Hormozdiari, F., Lusis, A. J., and Eskin, E. (2016). Efficient and Accurate

Multiple-Phenotype Regression Method for High Dimensional Data Considering Population Structure. *Genetics*, 204(4):1379–1390.

Kahles, A., Lehmann, K.-V., Toussaint, N. C., Hüser, M., Stark, S. G., Sachsenberg, T., Stegle, O., Kohlbacher, O., Sander, C., Cancer Genome Atlas Research Network, S. J., Rätsch, G., Felau, I., Kasapi, M., Ferguson, M. L., Hutter, C. M., et al. (2018). Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. *Cancer cell*, 34(2):211–224.

Kakaradov, B., Xiong, H., Lee, L. J., Jojic, N., and Frey, B. J. (2012). Challenges in estimating percent inclusion of alternatively spliced junctions from RNA-seq data. *BMC Bioinformatics*, 13(Suppl 6):S11.

Katz, Y., Wang, E. T., Airoldi, E. M., and Burge, C. B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature methods*, 7(12):1009–15.

Katz, Y., Wang, E. T., Silterra, J., Schwartz, S., Wong, B., Thorvaldsdóttir, H., Robinson, J. T., Mesirov, J. P., Airoldi, E. M., and Burge, C. B. (2015). Quantitative visualization of alternative exon expression from RNA-seq data. *Bioinformatics*, 31(14):2400–2402.

Kilpinen, H., Waszak, S. M., Gschwind, A. R., Raghav, S. K., Witwicki, R. M., Orioli, A., Migliavacca, E., Wiederkehr, M., Gutierrez-Arcelus, M., Panousis, N. I., Yurovsky, A., Lappalainen, T., Romano-Palumbo, L., Planchon, A., Bielser, D., et al. (2013). Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science*, 342(6159):744–7.

Kim, S., Kim, H., Fong, N., Erickson, B., and Bentley, D. L. (2011). Pre-mRNA splicing is a determinant of histone H3K36 methylation. *Proceedings of the National Academy of Sciences of the United States of America*, 108(33):13564–9.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C., and Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science*, 308(5720):385–9.

Knowles, D. A., Burrows, C. K., Blischak, J. D., Patterson, K. M., Serie, D. J., Norton, N., Ober, C., Pritchard, J. K., and Gilad, Y. (2018). Determining the genetic basis of anthracycline-cardiotoxicity by molecular response QTL mapping in induced cardiomyocytes. *eLife*, 7.

Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., Sasaki, D., Imamura, K., Kai, C., Harbers, M., Hayashizaki, Y., and Carninci, P. (2006). CAGE: cap analysis of gene expression. *Nature Methods*, 3(3):211–222.

Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., and Teichmann, S. A. (2015). The technology and biology of single-cell RNA sequencing. *Molecular cell*, 58(4):610–20.

Kornblihtt, A. R., Schor, I. E., Alló, M., Dujardin, G., Petrillo, E., and Muñoz, M. J. (2013). Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nature Reviews Molecular Cell Biology*, 14(3):153–165.

Korte, A., Vilhjálmsson, B. J., Segura, V., Platt, A., Long, Q., and Nordborg, M. (2012). A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature Genetics*, 44(9):1066–1071.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330.

Kurtovic-Kozaric, A., Przychodzen, B., Singh, J., Konarska, M. M., Clemente, M. J., Otrock, Z. K., Nakashima, M., Hsi, E. D., Yoshida, K., Shiraishi, Y., Chiba, K., Tanaka, H., Miyano, S., Ogawa, S., Boultwood, J., et al. (2015). PRPF8 defects cause missplicing in myeloid malignancies. *Leukemia*, 29(1):126–136.

Lappalainen, T., Sammeth, M., Friedländer, M. R., 't Hoen, P. A. C., Monlong, J., Rivas, M. A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., Barann, M., Wieland, T., Greger, L., van Iterson, M., Almlöf, J., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–11.

Leek, J. T. and Storey, J. D. (2007). Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLoS Genetics*, 3(9):e161.

Li, B. and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*, 12(1):323.

Li, Y. I., Knowles, D. A., Humphrey, J., Barbeira, A. N., Dickinson, S. P., Im, H. K., and Pritchard, J. K. (2018). Annotation-free quantification of RNA splicing using LeafCutter. *Nature Genetics*, 50(1):151–158.

Li, Y. I., van de Geijn, B., Raj, A., Knowles, D. A., Petti, A. A., Golan, D., Gilad, Y., and Pritchard, J. K. (2016). RNA splicing is a primary link between genetic variation and disease. *Science*, 352(6285):600–604.

Li, Z., Chen, J., Yu, H., He, L., Xu, Y., Zhang, D., Yi, Q., Li, C., Li, X., Shen, J., Song, Z., Ji, W., Wang, M., Zhou, J., Chen, B., et al. (2017). Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. *Nature Genetics*, 49(11):1576–1583.

Liu, F., van der Lijn, F., Schurmann, C., Zhu, G., Chakravarty, M. M., Hysi, P. G., Wollstein, A., Lao, O., de Bruijne, M., Ikram, M. A., van der Lugt, A., Rivadeneira, F., Uitterlinden, A. G., Hofman, A., Niessen, W. J., et al. (2012). A Genome-Wide Association Study Identifies Five Loci Influencing Facial Morphology in Europeans. *PLoS Genetics*, 8(9):e1002932.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., et al. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6):580–585.

Loos, R. J. F. and Janssens, A. C. J. W. (2017). Predicting Polygenic Obesity Using Genetic Information. *Cell metabolism*, 25(3):535–543.

Lovci, M. T., Ghanem, D., Marr, H., Arnold, J., Gee, S., Parra, M., Liang, T. Y., Stark, T. J., Gehman, L. T., Hoon, S., Massirer, K. B., Pratt, G. A., Black, D. L., Gray, J. W., Conboy, J. G., et al. (2013). Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nature Structural & Molecular Biology*, 20(12):1434–1442.

Luco, R. F., Pan, Q., Tominaga, K., Blencowe, B. J., Pereira-Smith, O. M., and Misteli, T. (2010). Regulation of alternative splicing by histone modifications. *Science*, 327(5968):996–1000.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., Pendlington, Z. M., Welter, D., Burdett, T., Hindorff, L., Flicek, P., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic acids research*, 45(D1):D896–D901.

Manning, K. S. and Cooper, T. A. (2017). The roles of RNA processing in translating genotype to phenotype. *Nature Reviews Molecular Cell Biology*, 18(2):102–114.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., et al. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–53.

Marigorta, U. M., Rodríguez, J. A., Gibson, G., and Navarro, A. (2018). Replicability and Prediction: Lessons and Challenges from GWAS. *Trends in genetics : TIG*, 34(7):504–517.

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337(6099):1190–5.

McClellan, J. and King, M.-C. (2010). Genetic heterogeneity in human disease. *Cell*, 141(2):210–7.

McVicker, G., van de Geijn, B., Degner, J. F., Cain, C. E., Banovich, N. E., Raj, A., Lewellen, N., Myrthil, M., Gilad, Y., and Pritchard, J. K. (2013). Identification of genetic variants that affect histone modifications in human cells. *Science*, 342(6159):747–9.

Merkin, J., Russell, C., Chen, P., and Burge, C. B. (2012). Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science*, 338(6114):1593–9.

Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Hui, S., Kar, S., Lemaçon, A., Soucy, P., Glubb, D., Rostamianfar, A., Bolla,

M. K., Wang, Q., Tyrer, J., Dicks, E., Lee, A., et al. (2017). Association analysis identifies 65 new breast cancer risk loci. *Nature*, 551(7678):92–94.

Monlong, J., Calvo, M., Ferreira, P. G., and Guigó, R. (2014). Identification of genetic variants associated with alternative splicing using sQTLseekeR. *Nature Communications*, 5(1):4698.

Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J., Guigo, R., and Dermitzakis, E. T. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, 464(7289):773–777.

Moore, R., Casale, F. P., Jan Bonder, M., Horta, D., Franke, L., Barroso, I., and Stegle, O. (2019). A linear mixed-model approach to study multivariate gene-environment interactions. *Nature Genetics*, 51(1):180–186.

Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N. E., Ahfeldt, T., Sachs, K. V., Li, X., Li, H., Kuperwasser, N., Ruda, V. M., Pirruccello, J. P., Muchmore, B., Prokunina-Olsson, L., Hall, J. L., Schadt, E. E., et al. (2010). From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature*, 466(7307):714–719.

Naftelberg, S., Schor, I. E., Ast, G., and Kornblihtt, A. R. (2015). Regulation of Alternative Splicing Through Coupling with Transcription and Chromatin Structure. *Annual review of biochemistry*, 84:165–198.

Natarajan, P., Peloso, G. M., Zekavat, S. M., Montasser, M., Ganna, A., Chaffin, M., Khera, A. V., Zhou, W., Bloom, J. M., Engreitz, J. M., Ernst, J., O’Connell, J. R., Ruotsalainen, S. E., Alver, M., Manichaikul, A., et al. (2018). Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. *Nature Communications*, 9(1):3391.

Ning, C., Wang, D., Zhou, L., Wei, J., Liu, Y., Kang, H., Zhang, S., Zhou, X., Xu, S., and Liu, J.-F. (2019). Efficient multivariate analysis algorithms for longitudinal genome-wide association studies. *Bioinformatics*.

Nowicka, M. and Robinson, M. D. (2016). DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics. *F1000Research*, 5:1356.

Okunola, H. L. and Krainer, A. R. (2009). Cooperative-binding and splicing-repressive properties of hnRNP A1. *Molecular and cellular biology*, 29(20):5620–31.

Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T., and Delaneau, O. (2016). Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics*, 32(10):1479–1485.

Ongen, H. and Dermitzakis, E. T. (2015). Alternative Splicing QTLs in European and African Populations. *American journal of human genetics*, 97(4):567–75.

O'Reilly, P. F., Hoggart, C. J., Pomyen, Y., Calboli, F. C. F., Elliott, P., Jarvelin, M.-R., and Coin, L. J. M. (2012). MultiPhen: Joint Model of Multiple Phenotypes Can Increase Discovery in GWAS. *PLoS ONE*, 7(5):e34861.

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics*, 40(12):1413–5.

Park, E., Pan, Z., Zhang, Z., Lin, L., and Xing, Y. (2018). The Expanding Landscape of Alternative Splicing Variation in Human Populations. *The American Journal of Human Genetics*, 102(1):11–26.

Parkes, M., Cortes, A., van Heel, D. A., and Brown, M. A. (2013). Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nature Reviews Genetics*, 14(9):661–673.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4):417–419.

Patro, R., Mount, S. M., and Kingsford, C. (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature biotechnology*, 32(5).

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, 33(3):290–295.

Pickrell, J. K., Berisa, T., Liu, J. Z., Ségurel, L., Tung, J. Y., and Hinds, D. A. (2016). Detection and interpretation of shared genetic influences on 42 human traits. *Nature genetics*, 48(7):709–17.

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y., and Pritchard, J. K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464(7289):768–772.

Porter, H. F. and O’Reilly, P. F. (2017). Multivariate simulation framework reveals performance of multi-trait GWAS methods. *Scientific Reports*, 7(1):38837.

Pradeepa, M. M., Sutherland, H. G., Ule, J., Grimes, G. R., and Bickmore, W. A. (2012). Psip1/Ledgf p52 binds methylated histone H3K36 and splicing factors and contributes to the regulation of alternative splicing. *PLoS genetics*, 8(5):e1002717.

Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7):459–463.

Raj, T., Li, Y. I., Wong, G., Humphrey, J., Wang, M., Ramdhani, S., Wang, Y.-C., Ng, B., Gupta, I., Haroutunian, V., Schadt, E. E., Young-Pearse, T., Mostafavi, S., Zhang, B., Sklar, P., et al. (2018). Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer’s disease susceptibility. *Nature Genetics*, 50(11):1584–1592.

Reyes, A. and Huber, W. (2018). Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic acids research*, 46(2):582–592.

Rhoads, A. and Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics*, 13(5):278–289.

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., and Kim, D. (2015). Methods of integrating data to uncover genotype-phenotype interactions. *Nature Reviews Genetics*, 16(2):85–97.

Rockman, M. V. and Kruglyak, L. (2006). Genetics of global gene expression. *Nature Reviews Genetics*, 7(11):862–872.

Rodriques, S. G., Stickels, R. R., Goeva, A., Martin, C. A., Murray, E., Vanderburg, C. R., Welch, J., Chen, L. M., Chen, F., and Macosko, E. Z. (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467.

Rosenberg, N. A., Huang, L., Jewett, E. M., Szpiech, Z. A., Jankovic, I., and Boehnke, M. (2010). Genome-wide association studies in diverse populations. *Nature Reviews Genetics*, 11(5):356–366.

Rotival, M., Quach, H., and Quintana-Murci, L. (2019). Defining the genetic and evolutionary architecture of alternative splicing in response to infection. *Nature communications*, 10(1):1671.

Schaid, D. J., Chen, W., and Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, 19(8):491–504.

Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S., and Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. *Genome research*, 22(9):1748–59.

Scotti, M. M. and Swanson, M. S. (2016). RNA mis-splicing in disease. *Nature Reviews Genetics*, 17(1):19–32.

Sekar, A., Bialas, A. R., de Rivera, H., Davis, A., Hammond, T. R., Kamitaki, N., Tooley, K., Presumey, J., Baum, M., Van Doren, V., Genovese, G., Rose, S. A., Handsaker, R. E., Daly, M. J., Carroll, M. C., et al. (2016). Schizophrenia risk from complex variation of complement component 4. *Nature*, 530(7589):177–183.

Shabalin, A. A. (2012). Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. *Bioinformatics*.

Sharma, S., Kohlstaedt, L. A., Damianov, A., Rio, D. C., and Black, D. L. (2008). Polypyrimidine tract binding protein controls the transition from exon definition to an intron defined spliceosome. *Nature Structural & Molecular Biology*, 15(2):183–191.

Sharon, D., Tilgner, H., Grubert, F., and Snyder, M. (2013). A single-molecule long-read survey of the human transcriptome. *Nature Biotechnology*, 31(11):1009–1014.

Shen, S., Park, J. W., Lu, Z.-x., Lin, L., Henry, M. D., Wu, Y. N., Zhou, Q., and Xing, Y. (2014). rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the National Academy of Sciences*, 111(51):E5593–E5601.

Shi, Y. (2017). Mechanistic insights into precursor messenger RNA splicing by the spliceosome. *Nature Reviews Molecular Cell Biology*, 18(11):655–670.

Sivakumaran, S., Agakov, F., Theodoratou, E., Prendergast, J. G., Zgaga, L., Manolio, T., Rudan, I., McKeigue, P., Wilson, J. F., and Campbell, H. (2011). Abundant pleiotropy in human complex diseases and traits. *American journal of human genetics*, 89(5):607–18.

Smemo, S., Tena, J. J., Kim, K.-H., Gamazon, E. R., Sakabe, N. J., Gómez-Marín, C., Aneas, I., Credidio, F. L., Sobreira, D. R., Wasserman, N. F., Lee, J. H., Puviindran, V., Tam, D., Shen, M., Son, J. E., et al. (2014). Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature*, 507(7492):371–375.

Somekh, J., Shen-Orr, S. S., and Kohane, I. S. (2019). Batch correction evaluation framework using a-priori gene-gene associations: applied to the GTEx dataset. *BMC Bioinformatics*, 20(1):268.

Spain, S. L. and Barrett, J. C. (2015). Strategies for fine-mapping complex traits. *Human Molecular Genetics*, 24(R1):R111–R119.

Stacey, S. N., Sulem, P., Jonasdottir, A., Masson, G., Gudmundsson, J., Gudbjartsson, D. F., Magnusson, O. T., Gudjonsson, S. A., Sigurgeirsson, B., Thorisdottir, K., Ragnarsson, R., Benediktsdottir, K. R., Nexø, B. A., Tjønneland, A., Overvad, K., et al. (2011). A germline variant in the TP53 polyadenylation signal confers cancer susceptibility. *Nature Genetics*, 43(11):1098–1103.

Stegle, O., Parts, L., Durbin, R., and Winn, J. (2010). A Bayesian Framework to Account for Complex Non-Genetic Factors in Gene Expression Levels Greatly Increases Power in eQTL Studies. *PLoS Computational Biology*, 6(5):e1000770.

Stein, S., Lu, Z.-x., Bahrami-Samani, E., Park, J. W., and Xing, Y. (2015). Discover hidden splicing variations by mapping personal transcriptomes to personal genomes. *Nucleic Acids Research*, 43(22):10612–10622.

Stephens, M. (2013). A Unified Framework for Association Analysis with Multiple Related Phenotypes. *PLoS ONE*, 8(7):e65245.

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S., and Robinson, G. E. (2015). Big Data: Astronomical or Genomical? *PLOS Biology*, 13(7):e1002195.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, 100(16):9440–5.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., et al. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine*, 12(3):e1001779.

Sul, J. H., Han, B., Ye, C., Choi, T., and Eskin, E. (2013). Effectively Identifying eQTLs from Multiple Tissues by Combining Mixed Model and Meta-analytic Approaches. *PLoS Genetics*, 9(6):e1003491.

Sul, J. H., Raj, T., de Jong, S., de Bakker, P. I. W., Raychaudhuri, S., Ophoff, R. A., Stranger, B. E., Eskin, E., and Han, B. (2015). Accurate and fast multiple-testing correction in eQTL studies. *American journal of human genetics*, 96(6):857–68.

Takata, A., Matsumoto, N., and Kato, T. (2017). Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci. *Nature Communications*, 8:14519.

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, page 1.

Tanackovic, G., Ransijn, A., Thibault, P., Abou Elela, S., Klinck, R., Berson, E. L., Chabot, B., and Rivolta, C. (2011). PRPF mutations are associated with generalized defects in spliceosome formation and pre-mRNA splicing in patients with retinitis pigmentosa. *Human Molecular Genetics*, 20(11):2116–2130.

Thorleifsson, G., Walters, G. B., Gudbjartsson, D. F., Steinthorsdottir, V., Sulem, P., Helgadottir, A., Styrkarsdottir, U., Gretarsdottir, S., Thorlacius, S., Jonsdottir, I., Jonsdottir, T., Olafsdottir, E. J., Olafsdottir, G. H., Jonsson, T., Jonsson, F., et al. (2009). Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nature Genetics*, 41(1):18–24.

Thorvaldsdottir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2):178–192.

Tilgner, H., Knowles, D. G., Johnson, R., Davis, C. A., Chakrabortty, S., Djebali, S., Curado, J., Snyder, M., Gingeras, T. R., and Guigo, R. (2012). Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Research*, 22(9):1616–1625.

Torkamani, A., Wineinger, N. E., and Topol, E. J. (2018). The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, 19(9):581–590.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515.

Trincado, J. L., Entizne, J. C., Hysenaj, G., Singh, B., Skalic, M., Elliott, D. J., and Eyras, E. (2018). SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biology*, 19(1):40.

Ule, J., Stefani, G., Mele, A., Ruggiu, M., Wang, X., Taneri, B., Gaasterland, T., Blencowe, B. J., and Darnell, R. B. (2006). An RNA map predicting Nova-dependent splicing regulation. *Nature*, 444(7119):580–6.

van der Meer, D., Rokicki, J., Kaufmann, T., Córdova-Palomera, A., Moberget, T., Alnæs, D., Bettella, F., Frei, O., Doan, N. T., Sønderby, I. E., Smeland, O. B., Agartz, I., Bertolino, A., Bralten, J., Brandt, C. L., et al. (2018). Brain scans from 21,297 individuals reveal the genetic architecture of hippocampal subfield volumes. *Molecular Psychiatry*, pages 1–13.

van der Sluis, S., Posthuma, D., and Dolan, C. V. (2013). TATES: Efficient Multivariate Genotype-Phenotype Analysis for Genome-Wide Association Studies. *PLoS Genetics*, 9(1):e1003235.

van der Wijst, M. G. P., Brugge, H., de Vries, D. H., Deelen, P., Swertz, M. A., and Franke, L. (2018). Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nature Genetics*, 50(4):493–497.

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., and Thermes, C. (2018). The Third Revolution in Sequencing Technology. *Trends in genetics : TIG*, 34(9):666–681.

Van Nostrand, E. L., Pratt, G. A., Shishkin, A. A., Gelboin-Burkhart, C., Fang, M. Y., Sundararaman, B., Blue, S. M., Nguyen, T. B., Surka, C., Elkins, K., Stanton, R., Rigo, F., Guttman, M., and Yeo, G. W. (2016). Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nature Methods*, 13(6):508–514.

Vaquero-Garcia, J., Barrera, A., Gazzara, M. R., González-Vallinas, J., Lahens, N. F., Hogenesch, J. B., Lynch, K. W., and Barash, Y. (2016). A new view of transcriptome complexity and regulation through the lens of local splicing variations. *eLife*, 5.

Vaquero-Garcia, J., Norton, S., and Barash, Y. (2018). LeafCutter vs. MAJIQ and comparing software in the fast moving field of genomics. *bioRxiv*, page 463927.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics*, 101(1):5–22.

Voineagu, I., Wang, X., Johnston, P., Lowe, J. K., Tian, Y., Horvath, S., Mill, J., Cantor, R. M., Blencowe, B. J., and Geschwind, D. H. (2011). Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature*, 474(7351):380–384.

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., and Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476.

Wang, X., Hou, J., Quedenau, C., and Chen, W. (2016). Pervasive isoform-specific translational regulation via alternative transcription start sites in mammals. *Molecular Systems Biology*, 12(7):875.

Wen, X., Lee, Y., Luca, F., and Pique-Regi, R. (2016). Efficient Integrative Multi-SNP Association Analysis via Deterministic Approximation of Posteriors. *American journal of human genetics*, 98(6):1114–1129.

Wen, X., Pique-Regi, R., and Luca, F. (2017). Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLOS Genetics*, 13(3):e1006646.

Wu, L., Candille, S. I., Choi, Y., Xie, D., Jiang, L., Li-Pook-Than, J., Tang, H., and Snyder, M. (2013). Variation and genetic control of protein abundance in humans. *Nature*, 499(7456):79–82.

Yang, E.-W., Bahn, J. H., Hsiao, E. Y.-H., Tan, B. X., Sun, Y., Fu, T., Zhou, B., Van Nostrand, E. L., Pratt, G. A., Freese, P., Wei, X., Quinones-Valdez, G., Urban, A. E., Graveley, B. R., Burge, C. B., et al. (2019). Allele-specific binding of RNA-binding proteins reveals functional genetic variants in the RNA. *Nature communications*, 10(1):1338.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., and Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7):565–569.

Ye, C. J., Chen, J., Villani, A.-C., Gate, R. E., Subramaniam, M., Bhangale, T., Lee, M. N., Raj, T., Raychowdhury, R., Li, W., Rogel, N., Simmons, S., Imboywa, S. H., Chipendo, P. I., McCabe, C., et al. (2018). Genetic analysis of isoform usage in the human anti-viral response reveals influenza-specific regulation of ERAP2 transcripts under balancing selection. *Genome research*, 28(12):1812–1825.

Yeo, G. W., Coufal, N. G., Liang, T. Y., Peng, G. E., Fu, X.-D., and Gage, F. H. (2009). An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nature structural & molecular biology*, 16(2):130–7.

Zapala, M. A. and Schork, N. J. (2012). Statistical properties of multivariate distance matrix regression for high-dimensional data analysis. *Frontiers in genetics*, 3:190.

Zhang, X., Joehanes, R., Chen, B. H., Huan, T., Ying, S., Munson, P. J., Johnson, A. D., Levy, D., and O'Donnell, C. J. (2015). Identification of common genetic variants controlling transcript isoform variation in human whole blood. *Nature Genetics*, 47(4):345–352.

Zhang, Y., Xu, Z., Shen, X., Pan, W., and Alzheimer's Disease Neuroimaging Initiative, t. A. D. N. (2014). Testing for association with multiple traits in generalized estimation equations, with application to neuroimaging data. *NeuroImage*, 96:309–25.

Zhang, Y., Zhou, H., Zhou, J., and Sun, W. (2017). Regression Models for Multivariate Count Data. *Journal of Computational and Graphical Statistics*.

Zhao, K., Lu, Z. X., Park, J. W., Zhou, Q., and Xing, Y. (2013). GLiMMPS: Obust statistical model for regulatory variation of alternative splicing using RNA-seq data. *Genome Biology*.

Zhou, X. and Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods*, 11(4):407–409.

Zhou, Z. and Fu, X.-D. (2013). Regulation of splicing by SR proteins and SR protein-specific kinases. *Chromosoma*, 122(3):191–207.

Zhu, X., Feng, T., Tayo, B. O., Liang, J., Young, J. H., Franceschini, N., Smith, J. A., Yanek, L. R., Sun, Y. V., Edwards, T. L., Chen, W., Nalls, M., Fox, E., Sale, M., Bottinger, E., et al. (2015). Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *American journal of human genetics*, 96(1):21–36.