Perspectives on allele-specific expression

Perspectives on Allele-Specific Expression

Siobhan Cleary,[1] and Cathal Seoighe[1]

[1]School of Mathematics, Statistics and Applied Mathematics, National University of Ireland, Galway, Ireland; email:cathal.seoighe@nuigalway.ie

Xxxx. Xxx. Xxx. Xxx. YYYY. AA:1–24

https://doi.org/10.1146/((please add article doi))

Keywords

allelic imbalance, allele specific expression

Abstract

Diploidy has profound implications for population genetics and susceptibility to genetic diseases. Although two copies are present for most genes in the human genome, they are not necessarily both active or active at the same level in a given individual. Genomic imprinting, resulting in exclusive or biased expression in favour of the allele of paternal or maternal origin, is now believed to affect hundreds of human genes. A far greater number of genes display unequal expression of gene copies due to cis-acting genetic variants that perturb gene expression. The availability of data generated by RNA sequencing applied to large numbers of individuals and tissue types has generated unprecedented opportunities to assess the contribution of genetic variation to allelic imbalance in gene expression. Here we review the insights gained through the analysis of these data about the extent of the genetic contribution to allelic expression imbalance, the tools and statistical models for gene expression imbalance, and what the results obtained reveal about the contribution of genetic variants that alter gene expression to complex human diseases and phenotypes.

Contents

1. Introduction

Allelic imbalance arises when there is a difference in the states or activities of the alleles of a locus in a diploid (or higher ploidy) organism. Much of the research on allelic imbalance has focused on differences in messenger RNA (mRNA) abundance, which we will refer to as allelic expression imbalance. Imbalance in mRNA abundance between alleles has been referred to as allele-specific expression (ASE) (1). This term is often used to refer to gene expression imbalance, without regard to whether the difference in expression is due to genetic variants or epigenetic effects, such as imprinting or random monoallelic expression (2, 3). However, as it is suggestive of an effect that arises from the allele itself, we propose that the term allele-specific expression should be reserved for imbalance with a genetic origin and adopt that usage here. We use the term allelic expression imbalance, when the cause of the differences in expression between alleles is not specified. Similarly, we use allelic imbalance to refer to any differences between alleles in chromatin state, expression level or relative isoform abundance and allele-specific imbalance when these differences are genetic in origin.

Terminology as used in this review

1.1. Allelic Imbalance

Difference in chromatin state or mRNA or protein levels between the alleles at a locus.

1.2. Allelic Expression Imbalance

Allelic Imbalance that results in differences in mRNA levels, from complete imbalance with exclusive expression of one allele to subtle differences in expression between alleles.

1.3. Allelic-Specific Imbalance

Difference in chromatin state or mRNA or protein levels between the alleles at a heterozygous locus that is due to the genetic differences between the alleles.

1.4. Allele-Specific Expression

Allelic expression imbalance that is genetic in origin, i.e. a difference in expression levels between alleles due to differences between the allele sequences.

ASE has a close association with expression quantitative trait loci (eQTLs), which are genetic loci with an effect on gene expression. eQTLs can act in cis, affecting the expression of a gene on the same chromosome and typically located close to the locus, or in trans, in which case the eQTL and the affected gene may be unlinked. When it acts in cis, an eQTL typically (though not inevitably (4, 5)) results in imbalance between the alternative alleles in heterozygotes. Consequently, allelic imbalance is often used to support the identification of cis-eQTLs. The high power of ASE for the detection of regulatory variants, particularly rare variants (6), and the contribution of regulatory variants to variation in human phenotypes and complex disease susceptibility have led to increasing interest in the phenomenon. This in turn has driven a proliferation of methods to detect ASE itself and to leverage ASE to infer regulatory variants and their effects on gene expression.

Here we review recent methodological developments and results of analyses of ASE, primarily focusing on human data. We begin with an overview of the types of allelic imbalance and review the mechanisms through which other types of imbalance can lead to ASE. We will discuss the range of statistical models and computational pipelines that have been developed to identify allelic imbalance from high throughput sequencing data and to leverage ASE to identify eQTLs and to prioritize genomic variants that may be causal for human diseases. We review estimates of the prevalence of ASE in human samples and discuss the implications of ASE for penetrance of genetic diseases and its potential for discovering the causal variants underlying the phenotypic associations identified in genome-wide association studies.

2. Mechanisms of Allelic Imbalance and ASE

Genetic variants can have an impact on chromatin structure (7, 8), on gene transcription (9, 7, 10), and post-transcriptional processes, including mRNA splicing (11, 12, 13),microRNA (miRNA) binding (14) and mRNA translation (15, 16, 17) (**Fig. 1**). In many cases, these variants can affect the expression level of the linked allele, leading to ASE, as well as leading to other measurable forms of allele-specific imbalance. For example, genetic variants that alter transcription factor binding sites can lead to imbalance in the transcription factor binding, and can result in ASE by altering the rate of initiation of transcription. Similarly, allele-specific imbalance in mRNA splicing, which itself can impact on gene function and

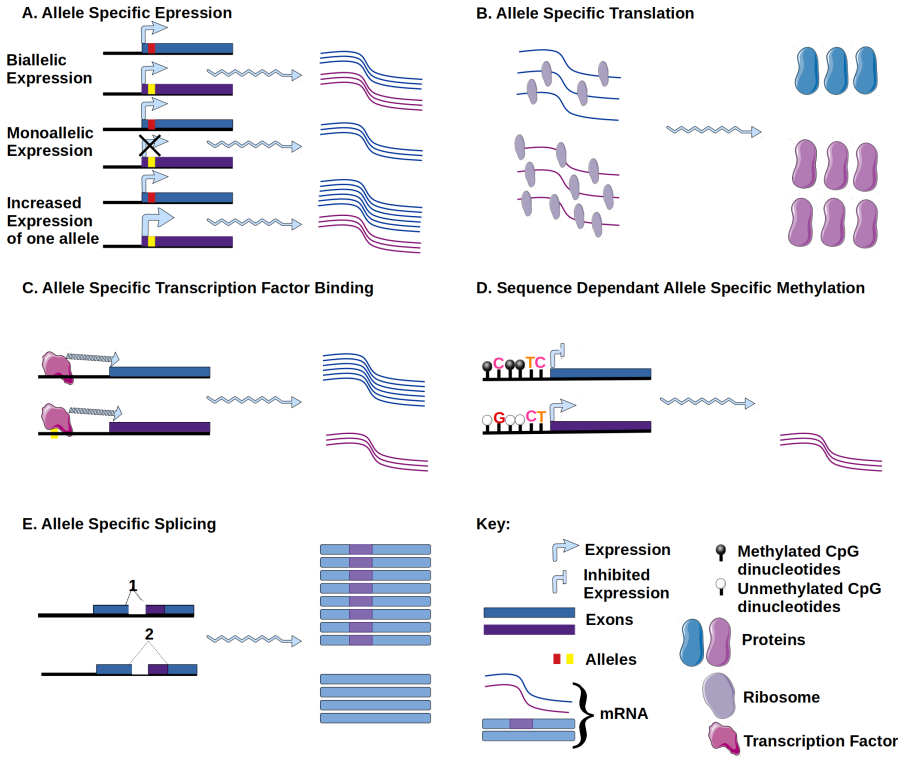

Figure 1

Types of allelic imbalance. A) Allelic expression imbalance. Three cases are shown: equal expression of both alleles (top); exclusive expression of one allele (middle); higher expression of one allele (bottom). B) Allelic imbalance in translation. Genetic variants can alter the rate of mRNA translation, resulting in different levels of ribosome occupancy between alleles. C) Imbalance in transcription factor binding. In the example shown a sequence variant reduces transcription factor binding affinity resulting in allele-specific expression. D) DNA methylation imbalance: methylation inhibiting the expression of one allele. If the difference in methylation results from cis-acting genetic variants it can lead to allele-specific expression E) Allele-Specific Splicing: A variant that alters splicing results in different isoforms from the two alleles.

reveal splicing regulatory variants, can lead to ASE, by altering the frame of translation and inducing nonsense-mediated decay. Below, we consider some of the main mechanisms leading to allele-specific imbalance, highlighting the potential of some of these to give rise to ASE.

2.1. Transcriptional regulation

Heterozygous single nucleotide polymorphisms (SNPs) in non-coding regions may be causal for ASE by affecting transcription factor binding affinity or causing differences in chromatin state between alternative alleles with a downstream effect on the rate of transcription. Chromatin modifications also play a key role in allelic imbalance of non-genetic origin (including genomic imprinting and X chromosome inactivation). Interplay can also occur

between genetic and epigenetic effects, through allele-specific DNA methylation (18).

2.1.1. Transcription factor binding and chromatin accessibility. The alleles of a heterozygous SNP can have different affinities for a transcription factor resulting in allelic imbalance in transcription factor occupancy (10) and distinct rates of transcription for each allele (19)(**Fig. 1C**). Analysis of allele-specific transcription factor binding has played an important role in understanding how non-coding DNA can affect gene expression and contribute to disease phenotypes. In order to dissect fully the implications of altered binding, the causal gene regulatory variant, the transcription factor that binds to it and the target gene should all be identified (20). Epigenetic marks, such as DNA methylation and histone modifications, can be inherited across cell generations, giving rise to distinct populations of cells expressing the same allele (7, 21). Imprinting and X chromosome inactivation are due to epigenetic effects rather than genetic differences between alleles. Epigenetic differences, caused by genetic differences between alleles can also result in allele-specific expression, via a process termed sequence dependent allele-specific methylation (22, 23)(**Fig. 1**D). This can affect non-imprinted, autosomal genes in a tissue and individual specific manner. Approximately 5% of CpG sites show evidence of substantial (>30%) imbalance in DNA methylation (24). Histone modifications are more complex than CpG methylation with hundreds of different types of histone tail modifications possible, including acetylation, methylation, phosphorylation, ubiquitination, sumoylation and ADP-ribosylation of distinct amino acids in the tails of the H3 and H4 histones (7). The effects of genetic variants on gene expression are, in many cases, mediated by their impact on chromatin modifications (25). Differences in chromatin accessibility can also result from allele-specific transcription factor binding and may make a substantial contribution to complex diseases (8).

2.2. Post-transcriptional mechanisms

2.2.1. Nonsense mediated decay and alternative splicing. Nonsense mediated decay (NMD) is a key cellular quality control mechanism that results in the elimination of mRNAs carrying premature termination codons (PTCs) that might result in malformed proteins (26). This process takes place in the cytoplasm and is associated with the termination of translation and mRNA degradation (27). NMD also plays a role in controlling mRNA expression level, contributing to the regulation of a large number of human genes (28). A heterozygous SNP at which one of the alleles results in a PTC can result in degradation of mRNA derived from that allele, resulting in ASE (29). Genetic variants can affect mRNA splicing by altering splicing signals in the transcript. Such mutations can occur within or close to splice donor or acceptor sites, around the branch point or in exonic or intronic enhancer or suppressor sites (13). Common effects on splicing include exon skipping, intron retention, alternate $3'$ or $5'$ exon ends and mutually exclusive exons (30). Because they act in cis, transcribed splicing mutations typically result in allele-specific splicing (12, 13). When a mutation that alters mRNA splicing introduces an in-frame stop codon (e.g. by skipping an exon within the coding region that is not a multiple of three nucleotides in length) it can trigger nonsense-mediated decay (30) targeted towards the affected allele. This results in a lower abundance of the mature mRNA for the allele causing mis-splicing than from the wild-type allele and consequently results in ASE. Even when NMD is not triggered, differences between the protein isoforms resulting from genetic variants that affect splicing

can have important functional consequences (31).

2.2.2. Variants affecting mRNA binding sites. RNA-binding proteins (RBPs) play a role in post-transcriptional gene regulation by binding to RNA in a sequence specific manner, modulating the fate of the bound RNA. Genetic variants on the mRNA can disrupt the interaction of RBPs with the mRNA, resulting in allelic imbalance in RNA binding and, potentially, ASE or allelic variation in mRNA localization or translation (32). Application of a method developed to detect allelic imbalance in RNA binding to enhanced crosslinking and immunoprecipitation sequencing (eCLIP-Seq) data from ENCODE revealed genomic variants that alter mRNA splicing as well as gene expression level (33, 34), illustrating the potential of allele-specific RNA binding to cause ASE. miRNAs and long non-coding RNAs (lncRNA) contribute to post-transcriptional regulation of gene expression. These non-coding RNAs can themselves display allele-specific imbalance in their expression, as well as inducing ASE in the genes they regulate (35, 36). Compared to protein-coding mRNAs, lncRNAs show greater levels of allelic imbalance in their expression (37). The interaction of miRNAs with their target mRNAs can be affected by SNPs within sites in the mRNA that are complementary to the miRNA (14) and again this is likely to result in allele-specific expression.

2.3. Translational mechanisms

Genomic variants that create or disrupt upstream initiation codons in the 5′ UTR, alter mRNA secondary structure, affect the translation start site or nearby sequence motifs or create novel mRNA isoforms can impact mRNA translation (15). Because these variants all act in cis they can result in allelic imbalance in the rate of mRNA translation, as well as changing the resulting protein product in some cases. Treating the ratio of ribosome-associated and non ribosome-associated RNA as a quantitative trait, Li *et al.* (16) identified SNPs inferred to have a significant association with mRNA translation rate in human lymphoblastoid cell lines. A more recent study (38) used translating ribosome affinity purification (TRAP) to identify genetic variants associated with alterations in ribosome occupancy and found evidence that sequence variants in upstream open reading frames (uORFs), miRNA binding sites and poly-A signals led to variation in translation efficiency in 1-2% of transcripts in mouse astrocytes. Allelic variation in the rate of translation can create imbalance in protein abundance derived from the two alleles, even in the absence of ASE at the mRNA level. Imbalance in RNA-editing between alleles can also result in differences at the protein level, in this case unexpected variation in the amino acid sequence. A genome-wide search for allele-specific RNA editing revealed examples of synonymous SNPs, resulting in nearby nonsynonymous changes caused by RNA editing (17).

3. Allelic Imbalance Analysis

3.1. Experimental design considerations and computational pipelines

Analysis of allelic imbalance and ASE from high throughput sequencing data typically involves generation of counts of sequence reads mapped to each allele. Generating this data involves multiple steps, as detailed below, each of which is associated with potential biases and confounding factors. Several efficient and scalable pipelines are available for these tasks, such as AlleleWorkbench (39), WASP (40), CloudASM (41) and ALEA (42).

The discussion below will focus primarily on the inference of allelic expression imbalance from RNA-Sequencing (RNA-Seq) data, but similar approaches can be adapted for other types of allelic imbalance.

3.1.1. Sequencing. In order to have sufficient power to discriminate between the expression levels of alternative alleles, analysis of allelic imbalance requires higher coverage than is generated in a typical RNA-Seq experiment focusing on total expression analysis (4). A threshold of 30 reads spanning the location of interest is often applied to infer allelic imbalance at individual heterozygous sites (43, 44). This can limit the number of genes with sufficient coverage to detect allelic imbalance. For example, with a median of 55 million mapped reads per sample in the Geuvadis study there was a median of only 3,000 genes that met this threshold (45). This is affected by sample heterozygosity as well as sequencing depth and does not consider the possibility of mapping to haplotypes, rather than individual heterozygous SNPs. The Genotype-Tissue Expression (GTEx) project performed RNA sequencing to a median depth of 83 million reads (46) and, using haplotype based methods, far higher proportions of genes could be tested for imbalance (47). In the case of bulk RNA-Seq experiments only allelic imbalance that is mitotically stable i.e. all daughter cells from the original cell share the pattern of expression of one allele, can be detected. Single-cell RNA-Seq can provide information on dynamic imbalance that changes over time due, for example, to bursts of transcriptional activity (48, 49). This can result in patterns of allelic imbalance that are not stable over successive cell generations (50).

3.1.2. Alignment and removal of PCR duplicates. One of the first steps in software pipelines for the analysis of allelic imbalance is to align the sequence reads to a reference genome or transcriptome. Errors in the alignment, or mapping, can have a substantial impact on the results obtained (45). Mapping errors (mapping a read to the wrong location or failure to map a read) can occur with greater frequency for reads containing the alternative than the reference allele at heterozygous SNPs (51), leading to false-positive signals of allelic imbalance. A number of strategies have been proposed to mitigate sequence alignment biases. These include the use of a masked reference (52), personalised diploid genomes (53) or transcriptomes (54), haplotype genomes for alignment (55), the use of SNP-tolerant mappers such as GSNAP (56),STAR-WASP (57), ASE-lux (58) and SNP-omatic (59) and methods that use remapping strategies such as WASP (40). Methods that align sequence reads to a diploid transcriptome that includes genetic variants have been reported to result in improved estimation of ASE (54).

Inference of imbalance at the gene level provides an incomplete picture, as there may be different degrees of imbalance between splice isoforms, with functional implications. It has also been reported that the inference of ASE at the gene level is biased when splice isoforms are ignored (60). For the task of inferring expression imbalance at the isoform level (61, 60), sequence reads must be mapped to both the isoform and allele from which they are derived. There can be ambiguity about both the transcript isoform and allele to which the read maps, particularly as many reads do not overlap heterozygous SNPs or transcript regions that distinguish between different isoforms. This has been addressed by applying a weighted allocation of reads using a hierarchical expectation-maximization (EM) strategy, which was reported to lead to an improvement in the inference of allelic imbalance in general, including at the splice isoform level (54).

The polymerase chain reaction (PCR) amplification step in the preparation of sequenc-

ing libraries can result in the same cDNA fragment being sequenced more than once. This results in sequence reads with identical mapping coordinates. Although it is straightforward to identify these duplicate reads and remove them, this is generally not recommended for RNA-Seq data due to loss of information for highly expressed genes. However, statistical tests of allelic imbalance are often not robust to the presence of duplicate reads and therefore potential PCR duplicates should be removed prior to analysis of allelic imbalance (45). Many tools for removing duplicates retain the read with the best mapping score, but for analysis of allelic imbalance it is essential to use tools,such as WASP (40), that select the retained reads at random, to avoid mapping bias in favour of the reference allele.

3.1.3. Genotyping and haplotype phasing. Generation of allele-specific read counts requires at least one heterozygous SNP within the targeted feature (gene, transcription factor binding site etc). Heterozygous SNPs can be identified separately using genotyping arrays or genomic DNA sequencing. Alternatively, the heterozygous SNPs can be inferred from the reads that map to the feature of interest. In the case of allele-specific expression, for example, genotype can be inferred from the RNA-Seq reads. However, this carries the risk that features that show extreme imbalance can be mistakenly called as homozygous, leading to false negatives in the inference of allelic imbalance. Conversely, sequencing errors, transcription errors or even rare somatic mutations that result in a site that is homozygous in the germline incorrectly being called heterozygous can lead to false positive inference of allelic imbalance. Errors may also occur when genotyping is performed on genomic DNA. In this case, homozygous sites incorrectly called heterozygous can lead to false positive inference of allelic imbalance (40). More recent methods for the analysis of allelic imbalance take account of uncertainty in genotyping (62, 40, 63).

Accurate SNP phase data supports the inference of allelic imbalance, by allowing reads to be mapped to haplotypes spanning multiple heterozygous SNPs. The information contained in the sequence reads can be used for this purpose, with the higher accuracy obtained when long read data are available (61). Haplotypes inferred from population phasing can be combined with the information contained in RNA-Seq reads spanning heterozygous SNPs to improve accuracy (64). However, this tends to be accurate for common variants but uncertain for rare variants.

Allele-specific read counts are the required input for many ASE tools (65, 66, 67, 68, 62). These can be determined for heterozygous SNPs using tools such as ASEReadCounter (45). However, mapping reads to haplotypes rather than individual heterozygous SNPs provides greater power for ASE analysis (64). Haplotype-specific expression levels can be estimated from RNA-Seq data using phASER (64) and haplotypes obtained from the RNA-Seq reads can be integrated with population-level phasing using phASER-pop (64) to extend haplotypes to putative regulatory variants in untranscribed regions (**Fig. 2C**). Some tools such as IDP-ASE (61) and BYASE (60) perform haplotyping as part of ASE estimation. For tools, such as EAGLE (69), that take read counts as input it is possible to supply gene level haplotypic counts instead of heterozygous SNP counts (70) as phaser generates one count per gene (64).

3.1.4. Considerations for analysis of ASE in cancer. Somatic copy number alteration (SCNA) can be a confounding factor in analysing allelic imbalance in cancer studies, leading to false positives for ASE (71). A recent pan cancer study revealed that SCNAs accounted for 84.3% of the observed allelic imbalance (72). Some studies address this by filtering

positions that overlap with copy number variation (53, 73). Methods have been developed to take account of copy number variation and tumor purity when assessing allelic imbalance of somatic mutations (74). Due to the presence of high frequency somatic mutations and copy number alterations in cancer, genotyping is usually based on the normal sample. Comparison of the cancer and normal sample can then reveal the allele that is retained in cancer in the case of loss of heterozygosity, which can be informative about the process leading to cancer development (75). Alternatively, ASE can be estimated for tumor and normal samples separately and the proportions of SNPs showing ASE can be compared between the two groups (76). Other studies have compared the variant allele frequency of heterozygous SNPs in whole exome sequences and transcriptome sequences (77, 78, 79) or used the allelic ratios in genomic DNA to correct for the effects of copy number variants (67).

3.2. Statistical Methods

A wide range of statistical models have been developed for the analysis of ASE. Broadly, they can be characterized by whether the goal is to detect allelic imbalance within individual samples or to combine data across multiple samples, either to characterize ASE or to use it to help estimate the effects of putative regulatory variants (**Fig. 2**). For the former goal the simplest method is to treat the number of reads mapping to the reference (or alternative) allele as a binomial random variable. Several Bayesian methods (61, 80, 81) have also been proposed to analyze ASE within individuals. Methods focused on estimating ASE can be differentiated based on whether they are applied on a gene by gene basis in individual samples, as is the case with the binomial test and also some more specialist methods (82, 61), or whether they attempt to learn model parameters by considering multiple genes simultaneously (e.g. (80, 81)). It is worth noting that when applied to single samples none of these methods can confirm ASE, as defined here, as it is not normally possible to distinguish whether or not the observed imbalance has a genetic origin. Methods have been developed to infer a genetic origin for the imbalance by relating expression imbalance to genotype across multiple individuals (83). Of particular note has been the development of models designed to learn about the effects of regulatory variants by combining ASE with variation in gene expression levels across individuals (40, 63, 84). Building on these, recent work has leveraged ASE to estimate the expected variance in gene expression for human genes, with important implications for understanding genetic disease mechanisms (85). Though the focus in this section is on methods for analysis of allele specific expression, similar methods can be applied to other types of allelic variation, such as allele-specific chromatin modifications.

3.2.1. The binomial test and its limitations. Some early studies of allelic imbalance were based on microarray data and adapted methods from gene expression analysis and genotyping to compare the expression of alternative alleles (9, 86). Contemporary studies, using sequencing, generate counts of alleles mapping to reference and alternative alleles (**Fig. 2A**). These counts were initially compared using a binomial test (45, 87, 43, 44). Applied to individual heterozygous SNPs, it is straightforward to evaluate a null hypothesis that a randomly sampled sequence read has the same probability of being generated from the reference or alternative allele. This null hypothesis can be modified to account for mapping bias in favour of the reference allele (51) by setting a slightly higher probability of

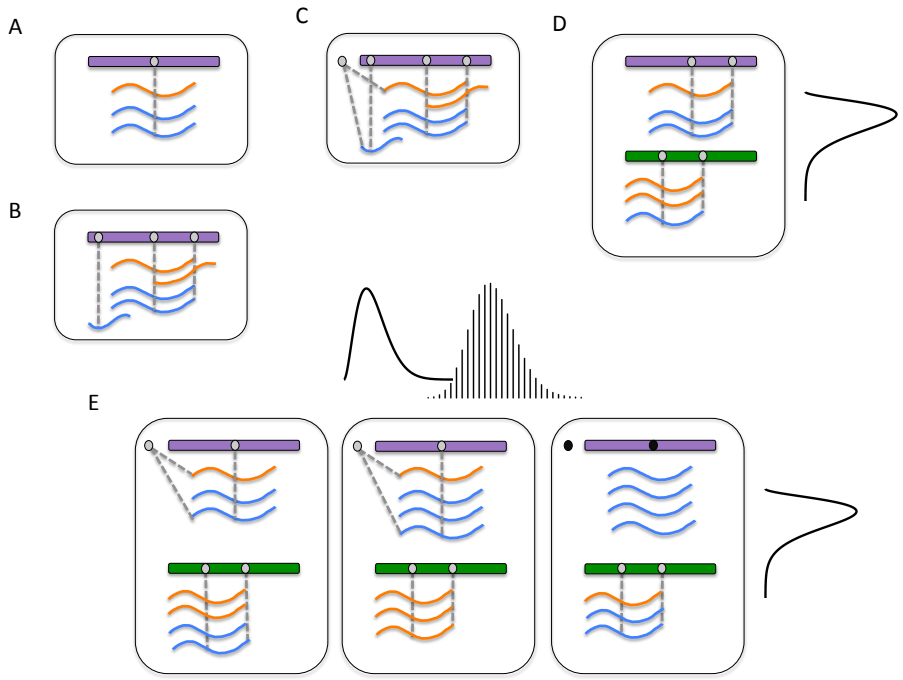

Figure 2

Illustration of the types of statistical models used in the analysis of allelic expression imbalance. Boxes represent individuals. Filled grey circles represent heterozygous SNPs and black circles represent homozygous SNPs. Sequence reads mapped to alleles of A) a single heterozygous SNP or B) haplotypes spanning multiple expressed heterozygous SNPs can be tested for unequal representation of the two alleles. C) Haplotypes can be extended to putative regulatory SNPs when population-based phasing is taken into account. If data from multiple individuals are available this allows the extent and direction of expression imbalance to be tested for correlation with the allele at the putative regulatory SNP. D) Statistical modeling can be used to learn the parameters of distributions describing the variation of ASE across genes within a single sample. E) Models can be constructed to combine evidence from ASE in heterozygous individuals with evidence from variation in gene expression level across individuals to detect regulatory variants. These models include distributions describing allelic expression ratios across SNPs in the same gene and across different genes as well as distributions for total expression level of the gene in different individuals.

a read being generated from the reference allele, under the null hypothesis of no imbalance (43). Further improvements in power can be obtained by mapping reads to phased haplotypes rather than to individual heterozygous sites and information within RNA-Seq reads, including allelic imbalance, can be leveraged to obtain phased information even for rare variants(64, 61, 88). Statistical models have also been developed for joint inference of heterozygous SNPs and detection of ASE from RNA-Seq reads (62). In common with many other methods to infer ASE (e.g. (67)), the latter method uses a likelihood ratio test to evaluate a null hypothesis corresponding to equal representation of alternative alleles, while accounting for uncertainty in the inferred genotypes.

Inference of allelic imbalance using the binomial test, and its variants, has several major caveats. Allele-specific count data tends to be overdispersed, relative to the binomial distribution, meaning that the variance in the count of reads mapping to an allele is higher than expected for a binomial random variable (45, 40). This overdispersion is likely to have both biological causes, reflecting a high prevalence of true allelic imbalance, as well as technical causes. It is possible to treat the number of reads derived from one of the alleles or haplotypes as a beta-binomial (or a binomial-logit-normal (85)) instead of a binomial random variable (45, 40, 82). The beta-binomial is a two-parameter distribution that arises when the parameter of a binomial random variable is itself a beta-distributed random variable. It can be parameterized with a mean and an overdispersion parameter (82), with the latter controlling the extent of the increase in variance relative to the binomial parameter. However, if the overdispersion is primarily biological in origin, reflecting a high frequency of allelic imbalance, including an overdispersion parameter estimated from the data in the null hypothesis may result in a reduction in power to detect ASE.

One of the technical sources of overdispersion is the presence of duplicate reads, but this can be addressed by removal of duplicates as discussed previously, or through the use of molecular barcodes (89). A lack of reproducibility of allelic imbalance results between technical replicates has recently been reported and interpreted to suggest that other steps in library preparation may be more important sources of bias than PCR amplification for allelic expression analysis (90). This lack of reproducibility is in contrast to earlier results, obtained from technical replicates in the Geuvadis study, which suggested that the variance across technical replicates was similar to its expectation under the binomial distribution following implementation of quality control steps (45). A key shortcoming of hypothesis testing for allelic imbalance is that it places the emphasis on evaluating a null hypothesis, which may be unrealistic and sensitive to sequencing depth, rather than on estimating the extent of the imbalance between alleles. Lastly, methods to detect allelic imbalance in single individuals cannot easily distinguish between genetic and epigenetic causes and therefore cannot be used to infer ASE (which as used here implies a genetic origin). Despite the above potential limitations the binomial test remains in use for detecting allelic expression imbalance (47), perhaps due to the ease of interpretation and use.

3.2.2. Bayesian models for allelic imbalance. Several Bayesian methods have been developed for the analysis of allelic imbalance. Considering data from just a single gene and a single individual, but multiple SNPs, IDP-ASE (61) simultaneously performs haplotype reconstruction and inference of allelic expression imbalance from RNA-Seq data. Taking a flat prior it samples from the joint posterior probability of the reconstructed haplotypes and the probability that a random read is derived from one or other of the haplotypes in an individual. Skelly et al. (80) developed a hierarchical Bayesian model for allelic imbalance that considers data from multiple genes simultaneously (**Fig. 2D**). This was first used with RNA-Seq data derived from crosses of *Saccharomyces cerevisiae* strains and data from a single human cell line (80). The study also included genomic data, which allowed technical artifacts, such as mapping bias, to be taken into account. The model for the RNA-Seq data consisted of a mixture prior with a component corresponding to allelic imbalance genes and another for non-allelic imbalance genes, for which the allele-specific read counts have the same distribution as in the genomic data. For the imbalance component, allele-specific read counts in a given gene were modeled using a beta-binomial, parameterized with the expected value and overdispersion. Across all genes, both the expected value and overdispersion were

themselves beta-distributed, with independent parameters, allowing for genes with variable or relatively constant allelic imbalance across heterozygous SNPs. Markov Chain Monte Carlo (MCMC) was used to obtain samples from the joint posterior distribution of the proportion of genes with imbalanced expression, expected value and overdispersion of the imbalance for each gene as well as parameters describing how these vary across the genes with allelic imbalance. An advantage of this Bayesian approach is the capacity to make inferences about the overall proportion of genes affected by allelic imbalance and the effect size distribution across these genes. A Bayesian implementation of a mixed effects binomial regression model was used by the same group to combine information across individuals and across tissues to estimate ASE associated with Neanderthal introgression (81). The parameter of the binomial distribution describing the number of non-reference reads was modeled as a sum of a fixed intercept term (corresponding to the ASE effect) and random effects for tissue and individual. Recently, Dong et al. developed a Bayesian model, together with a Python library (60) to estimate gene and isoform level expression imbalance for any ploidy >1. The authors claim that their method compares favourably to existing methods and gives consistent results across technical replicates. To the best of our knowledge, however, no independent benchmarking has been carried out to evaluate the performance of these methods.

3.2.3. Combining ASE and expression level to estimate regulatory effects of genetic variants. When data from multiple individuals are available, it can become possible to infer a genetic cause for the observed allelic imbalance (i.e. ASE, as defined in this review), by identifying an association between the imbalance and nearby putative regulatory variants (83). This goal is enhanced when haplotypes can be extended beyond the transcribed region to encompass putatively causal variants in regulatory regions (47). In addition to its high power to detect regulatory variants even at low frequencies (91, 6), linking allelic imbalance with known cis-eQTLs provides a confirmation that an eQTL that colocalizes with the affected gene acts in cis (65). A regulatory variant acting in cis has the capacity to both alter the expression level of the gene across samples and alter the relative expression levels of alleles in individual samples in which the variant is heterozygous. In the absence of negative feedback loops that may buffer the effect of regulatory variants in some cases (5), there is a straightforward relationship between the effect of a regulatory variant on gene expression levels across samples and the extent of the associated allelic imbalance. Mohammadi et al. (92) defined the allelic fold change (aFC) as the expression of an alternative relative to a reference allele in a heterozygous individual. In a linear model relating gene expression across individuals to genotype (encoded as 0, 1, 2 for reference homozygotes, heterozygotes and alternative allele homozygotes, respectively) the aFC corresponds to $\beta_1/\beta_0 + 1$, where β_0 is the intercept and β_1 is the effect on gene expression per copy of the alternative allele.

Instead of estimating ASE and eQTL effects separately, if the goal is to use ASE to understand the impact of regulatory variants on expression, it is preferable to model the gene expression level and allelic imbalance effects of regulatory variants jointly (**Fig. 2E**). In this way ASE complements and greatly enhances the power to identify cis-acting eQTLs. Several similar statistical models have been applied for this purpose (40, 63, 84). The model underlying the combined haplotype test (CHT) in WASP (40) as well as those of RASQUAL (63) and TReCASE (84) all involve a likelihood function based on the joint probability of the number of reads mapping to a gene of interest across individuals and the reads mapping to specific haplotypes within individuals. All three use likelihood ratio

tests to test for a genetic effect on expression. Read counts across individuals are modeled as a negative binomial (or negative beta binomial in the case of CHT) and allele-specific read counts are modeled using the beta binomial. The negative binomial was used in early methods for RNA-Seq analysis as an alternative to the Poisson distribution, in order to allow for the excess variance observed in read counts across samples, in a similar way in which the beta-binomial is used as an overdispersed alternative to the binomial distribution, as discussed above. There are some differences in the details of the methods, in particular in the way overdispersion is modeled. For example, CHT uses combined gene and sample level overdispersion parameters for the read depth, while RASQUAL uses a single gene-specific parameter to model overdispersion in both the total read counts and the allele-specific counts. Other differences in the modeling choices, which include the handling of uncertainty in genotyping and phasing, are conveniently summarized in supplementary table 2 of Kumasaka *et al.* (63). These authors reported higher power for their RASQUAL method, compared to CHT and TReCASE(63); however, again no comprehensive and impartial evaluation of the performance of models and tools for these and related tasks has yet been carried out, to the best of our knowledge.

3.2.4. The expected variance in gene expression. The objective of most of the methods discussed above is to estimate the relative expression of different gene alleles. Mohammadi *et al.* (85) recently developed a method with the goal instead of exploiting ASE in order to estimate the expected variance in gene expression, V_G, associated with the set of all regulatory variants acting in cis. The model assumes heterozygous transcribed SNPs in imperfect linkage with an unobserved biallelic regulatory variant. It also allows for the existence of a large number of other cis-acting regulatory variants with smaller effects and invokes the central limit theorem to derive a binomial-logit-normal distribution for the count of reads derived from the reference allele of the transcribed variant. This distribution is similar to the beta-binomial random variable that has been used frequently to model ASE, as discussed above. The genetic variation in gene expression estimated in this way is of fundamental scientific interest, providing information about the selective constraints acting on gene expression, relevant to understanding how gene expression evolves over time, but it also provides a means to prioritize candidate disease associated genes and mutations. The method, referred to as analysis of expression variation (ANEVA), can be used to enhance the detection of expression outliers (6), by highlighting expression outliers in genes with normally constrained gene expression. The combination of ANEVA with a dosage outlier test (referred to as ANEVA-DOT), was applied to patients with Mendelian muscle dystrophy and myopathy (MDM) and was shown to have high power to recover known causal variants as well as suggesting novel potential causal variants, one of which was confirmed in the study (85). In general, this statistical model for ASE and others inspired by it may become an important part of the arsenal of analytical tools for the critical problem of diagnosing causal variants for genetic diseases.

3.2.5. Single cell data and cell type specific imbalance. Bulk RNA-Seq data is derived from samples that generally consist of multiple different cell types, mixed in varying proportions. Genes can be regulated differently in the constituent cell types, resulting in variability in the extent of imbalance that depends on cell type composition. This can be addressed by performing analysis of allelic imbalance in single-cell RNA-Seq data, after first classifying cells by cell type (93). However, it remains expensive and challenging to generate appro-

priate single cell data and large scale surveys of gene expression across tissues in multiple individuals have to-date involved only bulk RNA sequencing. Gene expression deconvolution methods can be used to estimate cell type proportions in bulk RNA-Seq data (94), particularly when gene expression within the constituent cell types is available from relevant single-cell experiments. This can then be linked with imbalance analysis to infer the cell types that are affected by allelic imbalance. The BSCET method makes use of statistical interaction between the extent of allelic imbalance and cell type proportions across samples to infer the cell types affected by allelic imbalance in bulk RNA-Seq data (95).

4. Prevalence of Allele Specific Expression

Several studies have reported the frequency with which allelic expression imbalance is observed (43, 47, 9, 96, 97, 44, 81, 80). As discussed above, there are multiple genetic and epigenetic mechanisms that can lead to allelic expression imbalance; however, most allelic expression imbalance is reported to arise from genetic variation (43). Therefore, estimates of the overall prevalence of allelic expression imbalance provide an indication of ASE prevalence. There are at least two different quantities that can be considered. The first is the frequency with which the alleles are imbalanced within an individual. This has been estimated by testing heterozygous SNPs for evidence of imbalance (43). However, rejection of the simple null hypothesis of equal expression of two alleles does not guarantee that the imbalance is biologically meaningful. Any sequence heterogeneity between the alleles may have some effect on gene regulation and rejection of the null hypothesis may then become a question of the precision of the measurement, which tends to be greater for more highly expressed genes. Methods that consider all genes simultaneously and estimate the proportion of imbalanced genes and the effect size distribution are, therefore, preferable (80). A second measure of prevalence of allelic imbalance that has been reported is the proportion of genes that show imbalance in at least some subset of individuals, when data from a cohort of individuals is analyzed. Given a large enough sample of individuals, high sequencing depth and samples from sufficient tissues, this proportion is likely to approach one, and it therefore requires thresholds on the strength of imbalance and the proportion of individuals displaying imbalance in a particular tissue type (47) to be meaningful.

4.1. Divergent reports of ASE frequency

In 2002, Yan et al. (98) developed an experimental method to assess differences in expression between alleles of heterozygous SNPs and applied the method to data from 13 genes in 96 individuals from the Centre d'Étude du Polymorphisme Humain (CEPH) pedigrees. For six of these genes, there was evidence of allelic imbalance, and this imbalance followed a pattern consistent with Mendelian inheritance. This was followed in 2003 by an estimate of the prevalence of ASE in human using microarrays (9). Of 602 genes that could be tested, 54% showed evidence of allelic expression imbalance. Using reciprocal crosses of two mouse subspecies and a method based on consistent rejection of the null hypothesis of balanced allelic expression (p-value < 0.05) across replicates, Pinter et al. (96) estimated that 20% of mouse genes show evidence of allelic expression imbalance in any given tissue. The majority of the imbalance resulted from genetic effects rather than imprinting or random monoallelic expression. By crossing inbred mice from three subspecies and applying a slightly different method that also focused on rejection of the null hypothesis of balanced expression, Crowley

et al. (97) reported that over 80% of genes showed evidence of allelic imbalance. Using Bayesian modeling Skelly also estimated a high frequency (80%) of ASE in a hybrid of two diverse Saccharomyces cerevisiae strains (80). Applying the same method to a single human cell line, they estimated a frequency of approximately 20% of allelic imbalance (80). Studies that have investigated allelic imbalance in humans rely on standing genetic variation, rather than crosses of divergent strains and the prevalence of ASE may therefore depend on the heterozygosity of the individual. Data from human lymphoblastoid cell lines, generated by the Geuvadis consortium (43), suggested that 6.5% of human genes show evidence of ASE, again using a binomial test (with a significance level of 0.005). A similar frequency of ASE (390 out of 6385 sites interrogated, or 6.1%) was reported by the pilot study of GTEx (44), using the same p-value threshold. This was reduced to 2.3%, when reads were downsampled to achieve a common sequence depth of 30 reads. This decrease by nearly a factor of three illustrates that the reported frequency of ASE based on statistical hypothesis tests is not a reliable indicator of the underlying prevalence. Estimation of the prevalence requires parameterized models, such as those described earlier, that can provide estimates of the proportion of genes affected within or across individuals and the distribution of the effect size. If the estimate of 20% for the weight of the allelic imbalance component in the model of Skelly *et al.* (80) referred to above is reasonable, this suggests that locus-specific tests may fail to detect a substantial proportion of ASE. This may be due to limitations in sequencing depth and insufficient power to detect weaker ASE effects.

Approximately 25% of heterozygous SNPs that tag an introgressed haplotype from Neanderthals showed evidence of ASE (81). In some sense, this resembles a natural experiment analogous to the reciprocal crosses that were used to estimate ASE prevalence in mouse (97, 96), except that the crossed populations are outbred and the data are collected many generations after the hybridizations, so that the introgressed segments may have been affected by evolutionary selection. Interestingly, there was no significant difference in the prevalence of ASE between heterozygous SNPs that tagged a Neanderthal allele compared to other heterozygous SNPs matched for minor allele frequency. This is surprising, given that the Neanderthal alleles should be associated with more divergent regulatory regions, creating more opportunities for allelic imbalance. The lack of a difference was interpreted as evidence of post-introgression purifying selection acting on variants that affect gene regulation (81). However, it is worth noting that the comparison involves Neanderthal haplotypes that are at low frequency in modern humans, potentially due to the relatively small contribution of the Neanderthal introgression and modern human haplotypes at comparable frequencies, some of which will have been suppressed by purifying selection in modern humans. Although no differences are reported in ASE prevalence between introgressed and non-introgressed haplotypes, a cross-tissue analysis suggested lower relative expression of Neanderthal haplotypes in brain and testis, compared to other tissues (81).

4.2. Survey of ASE across tissues and over time

Generation of RNA-Seq data from over 838 individuals across 49 human tissues by the GTEx consortium (46) has provided a real opportunity to gain insights into the prevalence and patterns of ASE. Analysis of the most recent release of GTex suggested that a very high proportion of genes show evidence of ASE in at least some of the samples (47). Among protein-coding genes, 53% showed evidence of strong ASE (at least two fold difference in expression between the alleles) in at least 50 individuals in at least one of the 49 tissues,

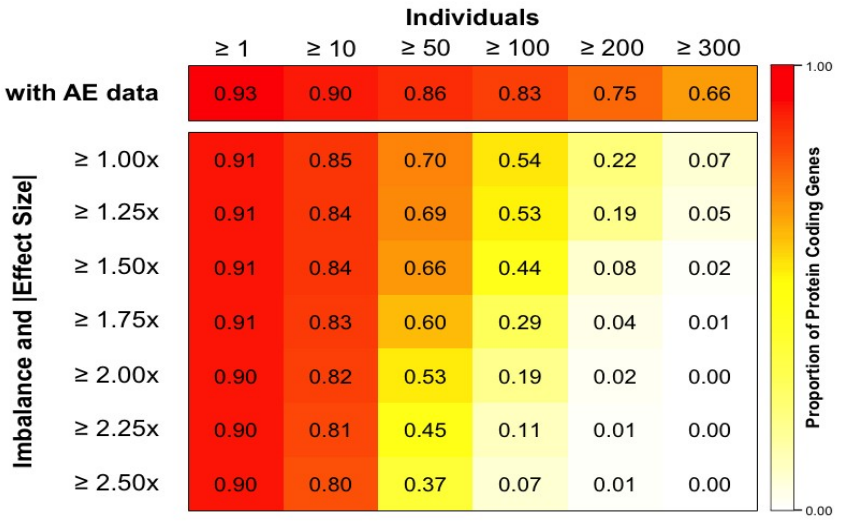

Figure 3

Supplementary Figure 5 from (47). The top row gives the proportion of protein-coding genes with allelic imbalance data in at least the number of individuals shown in the column for at least one GTEx tissue. The remaining rows show the proportion of protein-coding genes with statistically significant allelic imbalance (binomial test FDR ¡ 0.05) in at least the number of individuals shown in the column in at least one tissue, as a function of the minimum effect size (expression ratio between the alleles) given in the row

.

(**Fig. 3**). Given the mean number of samples per tissue (311) this corresponds to strong imbalance in a substantial fraction of the samples. Note that these results show that most genes can be affected by ASE, but does not translate easily into an estimate of the probability that a given gene will show expression imbalance in a given sample.

Analysis of the prevalence of ASE across samples suggested some differences across GTEx tissues, with testis having the largest number of genes with detected imbalance, though this appeared to have been driven largely by the number of expressed genes (47). An earlier analysis of whole-blood RNA-Seq data from 65 individuals at age 70 and at age 80 from the Prospective Investigation of the Vasculature in Uppsala Seniors (PIVUS) cohort (99) suggested a small (2.7%) but statistically significant increase in the prevalence of ASE with age (100), though there were examples of genes for which ASE tended to decrease as well as increase with age. Many of the genes that showed changes in ASE over time were associated with the immune response and suggested to be involved in the aging process (100). Changes in ASE with age suggest that it may be valuable to evaluate the frequency and effect size distribution of ASE across genes at a sample level. This is likely to reflect sequence heterozygosity, but given the relationship with age, may also have associations with phenotype or disease risk.

4.3. Caveats

Recent results suggest that ASE is affected by technical artifacts arising most likely during the preparation of sequencing libraries (90). Mendelevich et al. simulated replicate RNA-Seq datasets and found that the differences in allelic imbalance between technical replicates were greater than expected from the simulations. They used this difference to calculate an overdispersion factor, which was found to be relatively stable for a given sample. This was then used as a correction factor for the inference of imbalance, resulting in substantial reductions in false positive rates. The absence of technical replicates makes it difficult to assess the potential of false positives to contribute to the high rates of allelic imbalance reported by the GTEx consortium. However, it is difficult to envision how such technical artifacts could result in strong and consistent signals of ASE across such high proportions of individuals. Encouragingly, the allelic fold changes reported by GTEx from ASE were also highly consistent (Spearman rho = 0.83) with fold changes estimated orthogonally on the basis of eQTL analysis. Interestingly, excluding individuals who were heterozygous for a known eQTL led to a relatively small drop (median of 7.5%) in the number of genes with evidence of ASE in at least one sample of a given tissue (47). It is therefore possible that some of the ASE that is not supported by eQTLs is artifactual; however, it seems more likely that this result points to a large number of low-frequency eQTLs that have not been discovered.

Although the focus of this review is primarily on ASE, it is interesting to note the recent controversy over the prevalence of imbalance resulting from random monoallelic expression (RMAE). By applying single cell sequencing to mouse fibroblasts and human CD8+ T cells Reinius *et al.* (50) found only a small proportion (<1%) of generally low expression genes were affected by clonal RMAE, while a much larger number of genes displayed evidence of dynamic imbalance that was not stable across cell divisions. This contrasted with prior studies (101, 102) that had suggested thousands of genes display RMAE and generated some debate in the literature (103, 104), with the difference in views coming down primarily to questions of semantics and thresholds applied in the analysis. Reinius and Sandberg (104) argued that reporting the number of genes that are affected by clonal RMAE in any cell clone overstates the effect if only a small proportion of genes are affected in an individual cell clone. To some extent, there is a similar danger inherent in estimating the prevalence of ASE of genetic origin by reporting genes that display ASE in any samples, when a large number of samples are analyzed, or by reporting ASE that affects a substantial proportion of the individuals in at least one tissue (47). Frequent ASE in a gene across individuals does appear to reflect the genetic variability in the expression level of the gene, and can be informative about the phenotypic consequences of regulatory variants (85). Although more vulnerable to sample level artifacts, estimates of the extent of allelic imbalance within an individual may also be informative as a measure of expression heterozygosity in that individual.

5. Allelic Imbalance and Disease

5.1. Contribution of allelic imbalance to disease

ASE has the potential to contribute to disease when expression imbalance favours the disease allele and the functional allele is lowly expressed (52, 105, 106, 107, 75, 108). A number of studies have reported a tendency for higher expression of the minor allele in

cases, across several diseases, including autism spectrum disorder, cardiac disease and Zellweger spectrum disorder (52, 105, 106, 107, 109). This is consistent with a higher impact of disease-causing minor variants in individuals with ASE favouring their expression. It may provide a mechanism for disease alleles that would otherwise be recessive to make a greater contribution to phenotype in heterozygous individuals (107, 109). Conversely, allelic imbalance can in some cases function to compensate for high-impact autosomal dominant disease variants when the wild-type allele is more highly expressed (108).

Loss of function (LOF) mutations in tumor suppressor genes that might otherwise be recessive can contribute to cancer development when the more highly expressed allele is affected (78). Allele-specific loss of heterozygosity across cancer samples (where the same allele is consistently lost or downregulated) has been used to identify germline variants contributing to cancer risk, highlighting likely risk variants in genes involved in DNA repair (75). ASE may also result in one somatic mutation being sufficient to prevent the expression of a tumour suppressor gene at the level required to suppress tumorigenesis (98), increasing the risk of cancer development. Consequently, at the population level, ASE of cancer associated genes may be a risk factor for cancer development, as well as for other genetic diseases (109). This is further supported by the observation that LOF mutations in tumour suppressor genes are common in human populations (76). Allelic imbalance is also associated with oncogenes in cancer, with a tendency for the allele carrying the oncogenic somatic mutations to be over-expressed (74). Cancer driver mutations that result in constitutive activation of genes that contribute to cancer development (e.g. TERT (110)) can lead directly to expression imbalance, if activation of one allele is sufficient to confer the corresponding oncogenic property.

5.2. Use of allelic imbalance to infer causal mechanisms of disease-associated loci

The analysis of allelic expression imbalance complements eQTL analysis by providing information on what is happening at the individual level. This can help to understand the impact of cis-regulatory variation and interpretation of the effects of rare variants (111). The sharing of environmental and technical factors between different alleles within the same sample is a particular advantage of ASE for understanding the function of cis-factors (85). Combining eQTL and ASE to fine map functional genetic variants identified fewer but more accurate causal variants that were enriched for active regions in the genome (112). The analysis of allelic imbalance has been used to assess the function of putative disease-causing regulatory variants identified through GWAS studies (11, 112, 19). Specific forms of allelic imbalance, including imbalance in DNA methylation (105, 113), NMD (106, 72) or splicing (11) can reveal the mechanisms leading to disease, providing a means to explore the impact of putative disease-causing variants identified through GWAS.

6. Conclusions

Allelic imbalance is common in humans at all levels, from chromatin state to mRNA expression level and splicing and the rate of protein translation. Evidence from large-scale studies suggests that most allelic imbalance in gene expression is genetic in origin. Genetic imbalance in gene expression, referred to here as ASE, has implications for disease risk and the severity of the phenotypic impacts of disease-causing coding sequence variants. The cells of

diploid individuals provide a read-out on the impacts of regulatory variants, enabling the functional consequences of putative regulatory variants to be explored, even in the case of rare variants. A large number of computational methods and statistical models have been developed to assess the information provided by allelic imbalance, both within individuals and across individuals, with the latter methods often enabling ASE to be combined with inter-individual variation in expression level to assess the impact of regulatory variants. Further independent benchmarking of these methods will help to guide optimal analysis. With the increasing power to detect genetic loci with subtle effects on human phenotype variation in human populations, there is an ever-increasing demand for methods that can accurately infer the effects of genomic variants at the molecular level. This is coupled with a need for improved methods to assess the likely consequences of de novo variants with a suspected involvement in disease. Leveraging ASE to help assess the variance in gene expression in the normal population in order to prioritize variants that have a large effect on gene expression, relative to the variance in the population is a particularly promising development and one that underscores the power of ASE to derive disease-relevant insights.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

CS and SC are funded by Science Foundation Ireland grant number 16/IA/4612.

LITERATURE CITED

1. Knight JC. 2004. Allele-specific gene expression uncovered. *Trends in Genetics* 20:113–116
2. Buckland PR. 2004. Allele-specific gene expression differences in humans. *Human molecular genetics* 13:R255–R260
3. Gregg C. 2014. Known unknowns for allele-specific expression and genomic imprinting effects. *F1000Prime Reports* 6
4. Pastinen T. 2010. Genome-wide allele-specific analysis: insights into regulatory variation. *Nature Reviews Genetics* 11:533–538
5. Bader DM, Wilkening S, Lin G, Tekkedil MM, Dietrich K, et al. 2015. Negative feedback buffers effects of regulatory variants. *Molecular systems biology* 11:785
6. Li X, Kim Y, Tsang EK, Davis JR, Damani FN, et al. 2017. The impact of rare variation on gene expression across tissues. *Nature* 550:239–243
7. Lee MP. 2012. Allele-specific gene expression and epigenetic modifications and their application to understanding inheritance and cancer. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* 1819:739–742
8. Zhang S, Zhang H, Zhou Y, Qiao M, Zhao S, et al. 2020. Allele-specific open chromatin in human ipsc neurons elucidates functional disease variants. *Science* 369:561–565
9. Lo HS, Wang Z, Hu Y, Yang HH, Gere S, et al. 2003. Allelic variation in gene expression is common in the human genome. *Genome research* 13:1855–1862
10. Reddy TE, Gertz J, Pauli F, Kucera KS, Varley KE, et al. 2012. Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome research* 22:860–869

11. Amoah K, Hsiao YHE, Bahn JH, Sun Y, Burghard C, et al. 2020. Allele-specific alternative splicing in human tissues. *bioRxiv*
12. Nembaware V, Wolfe KH, Bettoni F, Kelso J, Seoighe C. 2004. Allele-specific transcript isoforms in human. *FEBS letters* 577:233–238
13. Nembaware V, Lupindo B, Schouest K, Spillane C, Scheffler K, Seoighe C. 2008. Genome-wide survey of allele-specific splicing in humans. *BMC genomics* 9:265
14. Kim J, Bartel DP. 2009. Allelic imbalance sequencing reveals that single-nucleotide polymorphisms frequently alter microrna-directed repression. *Nature biotechnology* 27:472–477
15. Robert F, Pelletier J. 2018. Exploring the impact of single-nucleotide polymorphisms on translation. *Frontiers in genetics* 9:507
16. Li Q, Makri A, Lu Y, Marchand L, Grabs R, et al. 2013. Genome-wide search for exonic variants affecting translational efficiency. *Nature communications* 4:1–6
17. Zhou ZY, Hu Y, Li A, Li YJ, Zhao H, et al. 2018. Genome wide analyses uncover allele-specific rna editing in human and mouse. *Nucleic acids research* 46:8888–8897
18. Do C, Shearer A, Suzuki M, Terry MB, Gelernter J, et al. 2017. Genetic–epigenetic interactions in cis: a major focus in the post-gwas era. *Genome biology* 18:120
19. Cavalli M, Pan G, Nord H, Arzt EW, Wallerman O, Wadelius C. 2016. Allele-specific transcription factor binding in liver and cervix cells unveils many likely drivers of gwas signals. *Genomics* 107:248–254
20. Cavalli M, Pan G, Nord H, Arzt EW, Wallerman O, Wadelius C. 2019. Allele specific chromatin signals, 3d interactions, and motif predictions for immune and b cell related diseases. *Scientific reports* 9:1–14
21. Wang H, Lou D, Wang Z. 2019. Crosstalk of genetic variants, allele-specific dna methylation, and environmental factors for complex disease risk. *Frontiers in genetics* 9:695
22. Yang HH, Hu N, Wang C, Ding T, Dunn BK, et al. 2010. Influence of genetic background and tissue types on global dna methylation patterns. *PLoS One* 5:e9355
23. Orjuela S, Machlab D, Menigatti M, Marra G, Robinson MD. 2020. Damefinder: A method to detect differential allele-specific methylation. *Epigenetics & Chromatin* 13:1–19
24. Onuchic V, Lurie E, Carrero I, Pawliczek P, Patel RY, et al. 2018. Allele-specific epigenome maps reveal sequence-dependent stochastic switching at regulatory loci. *Science* 361
25. Ng B, White CC, Klein HU, Sieberts SK, McCabe C, et al. 2017. An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. *Nat Neurosci* 20:1418–1426
26. Hug N, Longman D, Cáceres JF. 2016. Mechanism and regulation of the nonsense-mediated decay pathway. *Nucleic acids research* 44:1483–1495
27. Kervestin S, Jacobson A. 2012. Nmd: a multifaceted response to premature translational termination. *Nature reviews Molecular cell biology* 13:700–712
28. Alonso CR. 2005. Nonsense-mediated rna decay: A molecular system micromanaging individual gene activities and suppressing genomic noise. *Bioessays* 27:463–466
29. Montgomery SB, Lappalainen T, Gutierrez-Arcelus M, Dermitzakis ET. 2011. Rare and common regulatory variation in population-scale sequenced human genomes. *PLoS Genet* 7:e1002144
30. Park E, Pan Z, Zhang Z, Lin L, Xing Y. 2018. The expanding landscape of alternative splicing variation in human populations. *The American Journal of Human Genetics* 102:11–26
31. Li YI, Van De Geijn B, Raj A, Knowles DA, Petti AA, et al. 2016. Rna splicing is a primary link between genetic variation and disease. *Science* 352:600–604
32. Sheinberger J, Hochberg H, Lavi E, Kanter I, Avivi S, et al. 2017. Cd-tagging-ms2: detecting allelic expression of endogenous mrnas and their protein products in single cells. *Biology Methods and Protocols* 2:bpx004
33. Yang EW, Bahn JH, Hsiao EYH, Tan BX, Sun Y, et al. 2019. Allele-specific binding of rna-binding proteins reveals functional genetic variants in the rna. *Nature communications* 10:1–15

34. Bahrami-Samani E, Xing Y. 2019. Discovery of allele-specific protein-rna interactions in human transcriptomes. *The American Journal of Human Genetics* 104:492–502
35. Messemaker TC, van Leeuwen SM, van den Berg PR, EJ't Jong A, Palstra RJ, et al. 2018. Allele-specific repression of sox2 through the long non-coding rna sox2ot. *Scientific reports* 8:1–13
36. Võsa U, Esko T, Kasela S, Annilo T. 2015. Altered gene expression associated with microrna binding site polymorphisms. *PloS one* 10:e0141351
37. Johnsson PA, Hartmanis L, Ziegenhain C, Hendriks GJ, Hagemann-Jensen M, et al. 2020. Deducing transcriptional kinetics and molecular functions of long non-coding rnas using allele-sensitive single-cell rna-sequencing. *bioRxiv*
38. Liu Y, Fischer AD, Pierre CLS, Macias-Velasco JF, Lawson HA, Dougherty JD. 2020. Trap-based allelic translation efficiency imbalance analysis to identify genetic regulation of ribosome occupancy in specific cell types in vivo. *bioRxiv*
39. Soderlund CA, Nelson WM, Goff SA. 2014. Allele workbench: transcriptome pipeline and interactive graphics for allele-specific expression. *PloS one* 9:e115740
40. Van De Geijn B, McVicker G, Gilad Y, Pritchard JK. 2015. Wasp: allele-specific software for robust molecular quantitative trait locus discovery. *Nature methods* 12:1061–1063
41. Dumont EL, Tycko B, Do C. 2020. Cloudasm: an ultra-efficient cloud-based pipeline for mapping allele-specific dna methylation. *Bioinformatics* 36:3558–3560
42. Younesy H, Möller T, Heravi-Moussavi A, Cheng JB, Costello JF, et al. 2014. Alea: a toolbox for allele-specific epigenomics analysis. *Bioinformatics* 30:1172–1174
43. Lappalainen T, Sammeth M, Friedländer MR, Ac't Hoen P, Monlong J, et al. 2013. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501:506–511
44. Consortium G, et al. 2015. The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans. *Science* 348:648–660
45. Castel SE, Levy-Moonshine A, Mohammadi P, Banks E, Lappalainen T. 2015. Tools and best practices for data processing in allelic expression analysis. *Genome biology* 16:195
46. Consortium G, et al. 2020. The gtex consortium atlas of genetic regulatory effects across human tissues. *Science* 369:1318–1330
47. Castel SE, Aguet F, Mohammadi P, Ardlie KG, Lappalainen T. 2020. A vast resource of allelic expression data spanning human tissues. *Genome Biology* 21:1–12
48. Benitez JA, Cheng S, Deng Q. 2017. Revealing allele-specific gene expression by single-cell transcriptomics. *The international journal of biochemistry & cell biology* 90:155–160
49. Tunnacliffe E, Chubb JR. 2020. What is a transcriptional burst? *Trends in Genetics* 36:288–297
50. Reinius B, Mold JE, Ramsköld D, Deng Q, Johnsson P, et al. 2016. Analysis of allelic expression patterns in clonal somatic cells by single-cell rna–seq. *Nature genetics* 48:1430
51. Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, et al. 2009. Effect of read-mapping biases on detecting allele-specific expression from rna-sequencing data. *Bioinformatics* 25:3207–3212
52. Lee C, Kang EY, Gandal MJ, Eskin E, Geschwind DH. 2019. Profiling allele-specific gene expression in brains from individuals with autism spectrum disorder reveals preferential minor allele usage. *Nature neuroscience* 22:1521–1532
53. Rozowsky J, Abyzov A, Wang J, Alves P, Raha D, et al. 2011. Alleleseq: analysis of allele-specific expression and binding in a network framework. *Molecular systems biology* 7:522
54. Raghupathy N, Choi K, Vincent MJ, Beane GL, Sheppard KS, et al. 2018. Hierarchical analysis of rna-seq reads improves the accuracy of allele-specific expression. *Bioinformatics* 34:2177–2184
55. Wood DL, Nones K, Steptoe A, Christ A, Harliwong I, et al. 2015. Recommendations for accurate resolution of gene and isoform allele-specific expression in rna-seq data. *PloS one* 10:e0126911

56. Pandey RV, Franssen SU, Futschik A, Schlötterer C. 2013. Allelic imbalance metre (a llim), a new tool for measuring allele-specific gene expression with rna-seq data. *Molecular ecology resources* 13:740–745
57. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, et al. 2013. Star: ultrafast universal rna-seq aligner. *Bioinformatics* 29:15–21
58. Miao Z, Alvarez M, Pajukanta P, Ko A. 2018. Aselux: an ultra-fast and accurate allelic reads counter. *Bioinformatics* 34:1313–1320
59. Manske HM, Kwiatkowski DP. 2009. Snp-o-matic. *Bioinformatics* 25:2434–2435
60. Dong L, Wang J, Wang G. 2020. Byase: A python library for estimating gene and isoform level allele-specific expression. *Bioinformatics*
61. Deonovic B, Wang Y, Weirather J, Wang XJ, Au KF. 2017. Idp-ase: haplotyping and quantifying allele-specific expression at the gene and gene isoform level by hybrid sequencing. *Nucleic acids research* 45:e32–e32
62. Harvey CT, Moyerbrailean GA, Davis GO, Wen X, Luca F, Pique-Regi R. 2015. Quasar: quantitative allele-specific analysis of reads. *Bioinformatics* 31:1235–1242
63. Kumasaka N, Knights AJ, Gaffney DJ. 2016. Fine-mapping cellular qtls with rasqual and atac-seq. *Nature genetics* 48:206–213
64. Castel SE, Mohammadi P, Chung WK, Shen Y, Lappalainen T. 2016. Rare variant phasing and haplotypic expression from rna sequencing with phaser. *Nature communications* 7:1–6
65. Fan J, Hu J, Xue C, Zhang H, Susztak K, et al. 2020. Asep: Gene-based detection of allele-specific expression across individuals in a population by rna sequencing. *PLoS Genetics* 16:e1008786
66. Xie J, Ji T, Ferreira MA, Li Y, Patel BN, Rivera RM. 2019. Modeling allele-specific expression at the gene and snp levels simultaneously by a bayesian logistic mixed regression model. *BMC bioinformatics* 20:530
67. Liu Z, Gui T, Wang Z, Li H, Fu Y, et al. 2016. cisase: a likelihood-based method for detecting putative cis-regulated allele-specific expression in rna sequencing data. *Bioinformatics* 32:3291–3297
68. Edsgärd D, Iglesias MJ, Reilly SJ, Hamsten A, Tornvall P, et al. 2016. Geneiase: Detection of condition-dependent and static allele-specific expression from rna-seq data without haplotype information. *Scientific reports* 6:21134
69. Knowles DA, Davis JR, Edgington H, Raj A, Favé MJ, et al. 2017. Allele-specific expression reveals interactions between genetic variation and environment. *Nature Methods* 14:699–702
70. Oliva M, Muñoz-Aguirre M, Kim-Hellmuth S, Wucher V, Gewirtz AD, et al. 2020. The impact of sex on gene expression across human tissues. *Science* 369
71. de Santiago I, Liu W, Yuan K, O'Reilly M, Chilamakuri CSR, et al. 2017. Baalchip: Bayesian analysis of allele-specific transcription factor binding in cancer genomes. *Genome biology* 18:39
72. Calabrese C, Davidson NR, Demircioğlu D, Fonseca NA, He Y, et al. 2020. Genomic basis for rna alterations in cancer. *Nature* 578:129–136
73. Li G, Bahn JH, Lee JH, Peng G, Chen Z, et al. 2012. Identification of allele-specific alternative mrna processing via transcriptome sequencing. *Nucleic acids research* 40:e104–e104
74. Bielski CM, Donoghue MT, Gadiya M, Hanrahan AJ, Won HH, et al. 2018. Widespread selection for oncogenic mutant allele imbalance in cancer. *Cancer Cell* 34:852–862
75. Luft J, Young RS, Meynert AM, Taylor MS. 2020. Detecting oncogenic selection through biased allele retention in the cancer genome atlas. *BioRxiv*
76. Clayton EA, Khalid S, Ban D, Wang L, Jordan IK, McDonald JF. 2020. Tumor suppressor genes and allele-specific expression: mechanisms and significance. *Oncotarget* 11:462
77. Batcha AM, Bamopoulos SA, Kerbs P, Kumar A, Jurinovic V, et al. 2019. Allelic imbalance of recurrently mutated genes in acute myeloid leukaemia. *Scientific reports* 9:1–11
78. Rhee JK, Lee S, Park WY, Kim YH, Kim TM. 2017. Allelic imbalance of somatic mutations in cancer genomes and transcriptomes. *Scientific reports* 7:1–10

79. Halabi NM, Martinez A, Al-Farsi H, Mery E, Puydenus L, et al. 2016. Preferential allele expression analysis identifies shared germline and somatic driver genes in advanced ovarian cancer. *PLoS genetics* 12:e1005755
80. Skelly DA, Johansson M, Madeoy J, Wakefield J, Akey JM. 2011. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from rna-seq data. *Genome research* 21:1728–1737
81. McCoy RC, Wakefield J, Akey JM. 2017. Impacts of neanderthal-introgressed sequences on the landscape of human gene expression. *Cell* 168:916–927
82. Mayba O, Gilbert HN, Liu J, Haverty PM, Jhunjhunwala S, et al. 2014. Mbased: allele-specific expression detection in cancer tissues and cell lines. *Genome biology* 15:405
83. Lefebvre JF, Vello E, Ge B, Montgomery SB, Dermitzakis ET, et al. 2012. Genotype-based test in mapping cis-regulatory variants from allele-specific expression data. *PLoS One* 7:e38667
84. Sun W. 2012. A statistical framework for eqtl mapping using rna-seq data. *Biometrics* 68:1–11
85. Mohammadi P, Castel SE, Cummings BB, Einson J, Sousa C, et al. 2019. Genetic regulatory variation in populations informs transcriptome analysis in rare disease. *Science* 366:351–356
86. Pant PK, Tao H, Beilharz EJ, Ballinger DG, Cox DR, Frazer KA. 2006. Analysis of allelic differential expression in human white blood cells. *Genome research* 16:331–339
87. AC't Hoen P, Friedländer MR, Almlöf J, Sammeth M, Pulyakhina I, et al. 2013. Reproducibility of high-throughput mrna and small rna sequencing across laboratories. *Nature biotechnology* 31:1015–1022
88. Berger E, Yorukoglu D, Zhang L, Nyquist SK, Shalek AK, et al. 2020. Improved haplotype inference by exploiting long-range linking and allelic imbalance in rna-seq datasets. *Nature Communications* 11:1–9
89. Marx V. 2017. How to deduplicate pcr. *Nature methods* 14:473–476
90. Mendelevich A, Vinogradova S, Gupta S, Mironov AA, Sunyaev S, Gimelbrant AA. 2020. Unexpected variability of allelic imbalance estimates from rna sequencing. *bioRxiv*
91. Almlöf JC, Lundmark P, Lundmark A, Ge B, Maouche S, et al. 2012. Powerful identification of cis-regulatory snps in human primary monocytes using allele-specific gene expression. *PloS one* 7:e52260
92. Mohammadi P, Castel SE, Brown AA, Lappalainen T. 2017. Quantifying the regulatory effect size of cis-acting genetic variation using allelic fold change. *Genome research* 27:1872–1884
93. Choi K, Raghupathy N, Churchill GA. 2019. A bayesian mixture model for the analysis of allelic expression in single cells. *Nature communications* 10:1–11
94. Shen-Orr SS, Gaujoux R. 2013. Computational deconvolution: extracting cell type-specific information from heterogeneous samples. *Current opinion in immunology* 25:571–578
95. Fan J, Wang X, Xiao R, Li M. 2020. Detecting cell-type-specific allelic expression imbalance by integrative analysis of bulk and single-cell rna sequencing data. *bioRxiv*
96. Pinter SF, Colognori D, Beliveau BJ, Sadreyev RI, Payer B, et al. 2015. Allelic imbalance is a prevalent and tissue-specific feature of the mouse transcriptome. *Genetics* 200:537–549
97. Crowley JJ, Zhabotynsky V, Sun W, Huang S, Pakatci IK, et al. 2015. Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. *Nature genetics* 47:353–360
98. Yan H, Yuan W, Velculescu VE, Vogelstein B, Kinzler KW, et al. 2002. Allelic variation in human gene expression. *Science* 297:1143–1143
99. Lind L, Fors N, Hall J, Marttala K, Stenborg A. 2005. A comparison of three different methods to evaluate endothelium-dependent vasodilation in the elderly: the prospective investigation of the vasculature in uppsala seniors (pivus) study. *Arteriosclerosis, thrombosis, and vascular biology* 25:2368–2375
100. Balliu B, Durrant M, de Goede O, Abell N, Li X, et al. 2019. Genetic regulation of gene expression and splicing during a 10-year period of human aging. *Genome biology* 20:230
101. Gimelbrant A, Hutchinson JN, Thompson BR, Chess A. 2007. Widespread monoallelic expres-

sion on human autosomes. *Science* 318:1136–1140
102. Savova V, Chun S, Sohail M, McCole RB, Witwicki R, et al. 2016. Genes with monoallelic expression contribute disproportionately to genetic diversity in humans. *Nature genetics* 48:231–237
103. Vigneau S, Vinogradova S, Savova V, Gimelbrant A. 2018. High prevalence of clonal monoallelic expression. *Nature genetics* 50:1198–1199
104. Reinius B, Sandberg R. 2018. Reply to 'high prevalence of clonal monoallelic expression'. *Nature genetics* 50:1199–1200
105. Izzi B, Pistoni M, Cludts K, Akkor P, Lambrechts D, et al. 2016. Allele-specific dna methylation reinforces pear1 enhancer activity. *Blood* 128:1003–1012
106. McKean DM, Homsy J, Wakimoto H, Patel N, Gorham J, et al. 2016. Loss of rna expression and allele-specific expression associated with congenital heart disease. *Nature communications* 7:12824
107. Falkenberg KD, Braverman NE, Moser AB, Steinberg SJ, Klouwer FC, et al. 2017. Allelic expression imbalance promoting a mutant pex6 allele causes zellweger spectrum disorder. *The American Journal of Human Genetics* 101:965–976
108. de Klein N, van Dijk F, Deelen P, Urzua CG, Claringbould A, et al. 2020. Imbalanced expression for predicted high-impact, autosomal-dominant variants in a cohort of 3,818 healthy samples. *bioRxiv*
109. Castel SE, Cervera A, Mohammadi P, Aguet F, Reverter F, et al. 2018. Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. *Nature genetics* 50:1327–1334
110. Huang F, Bielski C, Rinne M, Hahn W, Sellers W, et al. 2015. Tert promoter mutations and monoallelic activation of tert in cancer. *Oncogenesis* 4:e176–e176
111. Consortium G, et al. 2017. Genetic effects on gene expression across human tissues. *Nature* 550:204–213
112. Zou J, Hormozdiari F, Jew B, Castel SE, Lappalainen T, et al. 2019. Leveraging allelic imbalance to refine fine-mapping for eqtl studies. *PLoS genetics* 15:e1008481
113. Chiba H, Kakuta Y, Kinouchi Y, Kawai Y, Watanabe K, et al. 2018. Allele-specific dna methylation of disease susceptibility genes in japanese patients with inflammatory bowel disease. *PloS one* 13:e0194036