## Title page

# Ascan: a novel method for the study of allele specific expression in single individuals.

Federico Zambelli [1,2*], Matteo Chiara [1,2*], Erika Ferrandi [1], Pietro Mandreoli [1,2], Marco Antonio Tangaro [2], Giulio Pavesi [1,2], Graziano Pesole **[2,3]

[1] Department of Biosciences, University of Milan, Milan, Italy

[2] Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, National Research Council, Bari, Italy

[3] Department of Biosciences, Biotechnology and Biopharmaceutics, University of Bari “Aldo Moro”, Bari, Italy

* These authors contributed equally to this work
**To whom correspondence should be addressed.

## Declarations of interest: none

## Abstract

In diploid organisms, two copies of each allele are normally inherited from parents. Paternal and maternal alleles can be regulated and expressed unequally, which is referred to as allele-specific expression (ASE). In this work, we present aScan, a novel method for the identification of ASE from the analysis of matched individual genomic and RNA sequencing data. By performing extensive analyses of both real and simulated data, we demonstrate that aScan can correctly identify ASE with high accuracy and sensitivity in different experimental settings. Additionally, by applying our method to a small cohort of individuals that are not included in publicly available databases of human genetic variation, we outline the value of possible applications of ASE analysis in single individuals for deriving a more accurate annotation of "private" low-frequency genetic variants associated with regulatory effects on transcription. All in all, we believe that aScan will represent a beneficial addition to the set of bioinformatics tools for the analysis of ASE. Finally, while our method was initially conceived for the analysis of RNA-seq data, it can in principle be applied to any quantitative NGS assay for which matched genotypic and expression data are available.

**Availability:** aScan is currently available in the form of an open source standalone software package at: https://github.com/Federico77z/aScan/. aScan version 1.0.3, available at https://github.com/Federico77z/aScan/releases/tag/1.0.3, has been used for all the analyses included in this manuscript. A Docker image of the tool has also been made available at https://github.com/pmandreoli/aScanDocker.

## Keywords

- Allele specific expression
- Transcriptomics

- Bioinformatics
- Next Generation Sequencing
- RNA-Seq expression quantification

## Introduction

A growing body of evidence suggests that -at least in humans and mammals- a substantial proportion of phenotypic diversity is mediated by transcriptional effects [1–4]. In diploid organisms, two copies of each allele are normally inherited from parents. Paternal and maternal alleles can be regulated and expressed unequally, which is referred to as allele-specific expression (ASE) [5]. The most radical form of ASE is genomic imprinting [6], where the allele from one parent is silenced systematically through epigenetic modifications in male and/or female gametes (e.g. DNA methylation, histone modifications) [7]. However, during the past decade several studies have reported widespread differences between the levels of expression of paternal and maternal alleles also at non-imprinted autosomal genes loci, with varying levels of magnitude and across different types of cells and tissues [8,9]. Current data suggest that between 5% [10,11] and 20% [12] of all the heterozygous variants in coding regions of the human genome display some level of ASE, with effects and extents that can vary by cell and tissue type [13–16], developmental stage [17–19] and phenotypic features [20]. For example, levels of ASE are remarkably higher in cancer cells as compared to normal tissues [21–23]. Importantly, ASE has been implicated in the pathogenesis of different types of tumors [24–26], but also with complex phenotypic traits and disorders [27]. Several lines of evidence suggest that patterns of ASE are consistent with Mendelian inheritance [28] and that ASE itself can be attributed to distinct, inheritable, epigenetic marks at homologous genomic regions. Indeed, many genes associated with ASE are located in the proximity of differentially methylated regions (DMRs) [29,30]. Therefore, it follows that a systematic investigation of allele-specific expression can provide a unique perspective for the integration of genetic variation and epigenetic information for the study of transcriptional regulation.

The identification of quantitative trait loci (QTL) associated with different levels of gene expression (eQTLs) [31] is the most common approach for the study of genetic factors that can modulate the expression of a gene, as demonstrated by the extensive catalogues of expression quantitative trait loci (eQTLs) in diverse cell types and conditions that are currently available [32]. Notably, eQTLs are highly enriched in genetic variants associated with complex phenotypic traits and disorders [50], therefore annotation of genetic variants potentially associated with regulatory effects is commonly used in several bioinformatics workflows for the prioritization of genetic variants in clinical genomics investigations [34-36]. That notwithstanding, the approach used for eQTL mapping has some important limitations, among which the most relevant are a limited sensitivity in the detection of effects associated with low-frequency genetic variants and the requirement for a large cohort of individuals to achieve an adequate detection power [37-38]. The study of ASE in single individuals is an alternative and complementary approach that can be employed for the characterization of genetic variants associated with the modulation of gene expression, to address some of the main limitations of eQTL mapping [39]. By this approach, differences in allelic expressions can be compared at different levels of annotation, including genes, transcripts or exons. Since the two alleles under study are compared directly within the same cellular environment (and not across different individuals, tissues or samples), highly specific and accurate insights on the molecular mechanisms that modulate gene expression can be derived. Moreover, since comparisons are based on the genetic profiles of a single person (or specimen), this type of analysis is not restricted to common genetic variants. On the other hand, by definition, outcomes of individual analyses of ASE are specific for the single individual, and can hardly be extended to the study of general patterns of imbalanced allelic expression in larger cohorts or at the population level. Since differences in allelic expression of heterozygous variants can be indicative of different types of molecular mechanisms including (but not limited to) nonsense mediated decay [40], differential binding of transcription factors or epigenetic modifiers [41], alteration of a splice site [42], or a reduced stability of the transcript due to the alteration of the secondary structure of UTRs [43], the study of ASE at level of single individuals may also be relevant in clinical settings, for example for the prioritization and annotation of genetic variants associated with pathological conditions [44].

Although several bioinformatics tools for the analysis of ASE are currently available, at present methods specifically devised for the analysis of ASE at the level of single individuals are scarce, and present some constraints that limit their application in different experimental settings. For example, while several methods based on sophisticated probabilistic models to infer the expression levels of single alleles are currently available, these methods do not provide a statistical framework for the testing of significant differences in alleles expression [45-47]. Conversely, the majority of methods for testing differential ASE are not directly applicable in a wide array of experimental settings, since they require the availability of phased genomic data, or need matched genome sequencing data (from the same individual) to estimate a null model of read ratio distributions [48–51]. In this work we present aScan, a novel and highly accurate method for the identification of ASE from the analysis of matched genome and RNA sequencing data from the same individual. By performing extensive analyses of both real and simulated data, we demonstrate that aScan can identify ASE with high accuracy and sensitivity in different experimental settings. Additionally, we outline a series of relevant considerations for the analysis of ASE at the level of single individuals, with potential implications also for a more accurate annotation of “private” low frequency genetic variants and their associated functional effects in personalized medicine.

All in all, in the light of the results presented in the current study, we believe that aScan will represent a very useful addition to the set of bioinformatics methods for the analysis of ASE.

## Results

## Analysis of simulated data

To evaluate the accuracy of aScan in the identification of biased allelic expression, extensive simulations of genes with ASE were performed, by taking advantage of publicly available gene expression profiles and matched genome sequencing data obtained from 60 distinct individuals [52-53]. Genes were stratified in four distinct classes based on their expression levels: highly expressed (TPM>15), moderately expressed (3<TMP≤15), lowly expressed (1<TPM≤3) and scarcely expressed (0.5<TPM≤1). Different levels of ASE were simulated by selecting 400 distinct genes from every class and by generating different proportions of maternal and paternal allele associated reads with the following ratios: 4:6, 3:7, 2:8, 1:9. A detailed summary

of the total number of polymorphic sites, proportion of heterozygous SNPs, and number of "testable" heterozygous sites associated with exons according to the Refseq annotation of the human genome (release 106) for every individual is reported in Supplementary Table S1.

As outlined in Figure 1A, we observe that aScan attains a very good level of sensitivity in the identification of ASE genes, as - globally - it can detect more than 75% of genes for which biased allele expression was simulated. More importantly, our method displays a very good level of accuracy, with a false discovery rate (FDR) of 1% or below, which is well in line with the theoretical FDR cut-off applied in all our analyses (Figure 1 B). Unsurprisingly- we notice that the ability of aScan to correctly identify genes with ASE is strongly correlated with gene expression levels. Indeed, while a nearly perfect sensitivity is attained for highly expressed genes, only a relatively modest proportion (44.2%) of ASE genes with low expression levels is recovered (Supplementary Figure 1A). Interestingly, we notice that (Supplementary Figure 1B) when only genes where heterozygous sites are covered by at least 10 RNA-seq reads are considered, aScan attains an average sensitivity of 84.6% on our simulated data, and no difference in sensitivity is observed in this case between different classes of gene expression level.  These results suggest that our method displays a good detection power also at relatively low coverage levels, and that the observed reduction of sensitivity for lowly expressed genes is most likely associated with missing or incomplete data (i.e. lack of RNA-seq reads covering heterozygous sites). It is worth noticing how (Figure 1A) aScan displays highly homogeneous levels of sensitivity on our simulated data, irrespective of the magnitude of the simulated differences in allele ratios. Consistent with this observation (Supplementary Figure S2) the corresponding FDRs distributions are highly similar and do not change across the different levels of allele ratios simulated in this study. Similarly, we do not observe a strong dependence between the number of heterozygous sites associated with ASE genes and the detection power of aScan (Supplementary Figure S3), as only a marginal decrease in sensitivity is observed between genes containing a relatively reduced number of heterozygous polymorphic sites (≤3) compared to genes associated with 10 or more heterozygous SNPs. However, a marginal but systematic decrease in sensitivity is observed for lowly expressed genes with a reduced number of polymorphic sites (Supplementary Figure S4 C-D).

A detailed analysis of False Positive calls, that is, genes that were not included in our simulations but were reported to have biased allelic expression according to aScan, displays some striking patterns. Indeed, we observe that the large majority of False Positive calls are associated with extreme biases in allele ratio distributions (Supplementary Figure 5A), and are highly enriched in lowly expressed genes (Supplementary Figure 5B). Consistent with these findings, we observe that the number of False Positive calls is reduced by more than 10-fold if only genes where each heterozygous SNP is covered by 10 or more RNA-seq reads (Supplementary Figure 6). Although these results suggest that a consistent proportion of aScan False Positive calls are probably associated with stochastic effects in the simulation RNA-seq data at lowly expressed genes, in the light of the fact that genes covered by a reduced number of RNA-seq reads are also associated with a reduced sensitivity for the detection of ASE, we conclude that genes, for which the majority of heterozygous sites are not covered by at least 10 reads, should be excluded from this type of analyses.

## Analysis of unphased data

Large scale genome sequencing projects usually apply sophisticated algorithms based on patterns of linkage disequilibrium and/or ad-hoc strategies, such as the sequencing of a relevant number of related individuals, to obtain a precise reconstruction of complete or nearly complete paternal and maternal haplotypes of the subjects under study [53]. However, this approach is not always applicable. The lack of haplotype level annotations can significantly reduce the accuracy of bioinformatics methods for study of ASE [45-51]. To circumvent this limitation, aScan incorporates a simple algorithm for the reconstruction of phased haplotypes based on allele expression data (see Materials and Methods). To evaluate the performances of aScan in the analysis of unphased genomic data, we repeated the analysis of our simulated dataset but by removing the phase information from the VCF files.  As depicted in Figure 2A the levels of sensitivity obtained by aScan on unphased data are substantially equivalent to those derived from the analysis of phased data, and no appreciable reduction or increase in sensitivity is observed. However, a statistically significant (Wilcoxon p-value < 2.2e-16) increase, from 0.01 to 0.02, in the False Discovery rate can be observed (Figure 2B).  Interestingly, while the majority of False Positive calls obtained from phased data were associated with extreme differences in allele ratios at genes with low levels of expression (Supplementary Figure 5A and Supplementary Figure 5B), when unphased genomic data are considered, a significant proportion of the False Positive calls is associated with genes showing only a (predicted) modest shift in allelic expression and relatively high (>10 TPM) expression levels (Supplementary Figure 7A and Supplementary Figure 7B). This suggests that the approach adopted by aScan for the reconstruction of phased haplotypes from unphased genomic data might not be ideal for genes associated with a moderate allelic imbalance and might in turn result in a slightly increased number of False Positive calls. Importantly, (Supplementary Figure 8) we underscore that when a simple threshold of allele ratio imbalance is applied [54-56] and only genes with a ratio of allelic expression of 4:6 or more are considered, False Discovery Rates attained by aScan on unphased data are completely in line with those observed in the analysis of phased VCF files.

## Analysis of individual human samples from 6 tissues

To demonstrate the added value of studies of ASE for the annotation of genetic variants at individual level, we applied aScan for the analysis of gene expression profiles from 6 different tissues (brain, liver, lung, striated muscle, kidney, heart), and matched genotypic data obtained from 3 distinct male individuals of Eastern European ancestry. These data have been previously employed for the study of RNA editing [57] and the correlation of the expression of mitochondrial genes with mitochondrial DNA abundance [58]. For the sake of consistency, the same sample identifiers as provided in [57-58]: S7, S12 and S13, will be used also in this study. As summarized in Table 1 the total number of heterozygous polymorphic sites was highly consistent between individuals ranging from 56004 (S13) to 57657 (S7), and after the application of our strict criteria for the exclusion of highly variable and problematic genomic regions, a total of 51717, 52004 and 49157 heterozygous polymorphic sites were retained respectively for S7, S12 and S13. Total number of mapped reads and estimated expression levels are reported in Supplementary Table S2 and Supplementary Table

S3 respectively, while a summary of the total number of genes tested in each distinct condition is reported in Supplementary Table 4.

According to our analyses, a total of 1286 genes were predicted to display ASE in at least one of the conditions tested. The number of candidate ASE genes, as identified in each organ and individual, are represented in Figure 3.  The number of ASE genes was largely consistent across all the tissues and individuals considered in this study, ranging from 73 (heart S13) to 288 (brain S7). Notably, only 8 genes (ERAP2, GRB10, PEG10, SLC22A18, H19, NTM, RPS9 and LOC100996724 ) display ASE across all the tissues and individuals included in our analyses. Among these, 6 (ERAP2, GRB10, PEG10, SLC22A18, H19 and NTM), are already reported in specialized databases of imprinted human genes [58], suggesting that meta-analyses of ASE in distinct tissues and individuals can represent an effective method for the identification of novel imprinted genes.

While inter-individual comparisons of genes associated with ASE show that only a reduced proportion (12.43%) of these genes is shared between two or more individuals (Figure 5), we observe that patterns of ASE are highly consistent between different tissues of the same individual (Figure 4), and the majority of genes associated with ASE (S12: 89.35%, S13: 90.33%, S7:90.02%) in a specific subject, display unbalanced levels of allelic expression in 2 or more tissues. Conversely, only a relatively limited number of genes (S12: 10.65%, S13: 9.67%, S7:9.98%) display ASE in a single tissue.  This notwithstanding, functional enrichment analyses of ASE genes in each of S12, S13 and S7, as executed by means of the DAVID software [59] (Supplementary Table S5 to S7), did not recover any statistically significant pattern of enrichment.

Publicly available databases of eQTLs and variants associated with regulatory effects are commonly used by different bioinformatics workflows for the prioritization of genetic variants of potential clinical relevance [34-36]. However, the level of accuracy of these information when applied for the annotations of genetic variants in single individuals are not completely clear. To investigate the potential application of ASE analyses in providing a more accurate individual level annotation of genetic variants implicated in the modulation of gene expression, annotations of genetic variants potentially associated with imbalanced allelic expression as derived from the GTEx database were contrasted with those derived from the analysis of ASE patterns in matched tissues of the same subject.

Strikingly, only a relatively limited proportion (24.5%) of the variants of the individuals included in this study are also reported in the GTEx database and the large majority (85.4%) of variants associated with allele specific expression in any of S13, S12 or S7 is not included in GTEx. Interestingly, however we observe that substantially higher proportions of these genetic variants, either associated with ASE (67.2%) or not (71.4%) are reported in comprehensive resources of human genetic variation as for example the Topmed [60], or the GNOMad [61] databases.

A total of 1261 variants observed in any of S7, S12 or S13, were associated with an eQTL in a tissue matching one the six tissues included in this study according to GTEx. However, when individual level expression data were considered, limited evidence of biased allelic expression was observed for these sites, and as outlined in Figure 6A, allele ratio distributions were largely consistent with biallelic expression (median allele ratio 0.523). A striking difference (Figure 6B) was observed when allele ratio distributions associated with these variants were contrasted with equivalent distributions computed from genetic variants associated with ASE according to aScan (Wilcoxon Test p-value < 2.2e-16).

A thorough functional annotation of the variants associated with ASE, that is, variants overlapping the exons of genes displaying ASE, and of linked genomic loci (see below) was carried out to characterize possible underlying molecular mechanisms. Interestingly, Figure 7 shows that a considerable proportion of ASE genes (ranging from 16 to 21% in different individuals) are associated with haplotypes that contain highly deleterious genetic variants, including disruptive splice site variants (1.5% to 1.8%), non-sense variants (5.71% to 9.23%) and/or predicted frameshifts in the CDS sequence (8.97% to 11.11%). Notably, all these haplotypes display reduced levels of allelic expression if compared with the corresponding alternative haplotypes at matched allelic loci, potentially consistent with degradation by non-sense mediated decay and other post-transcriptional regulatory mechanisms. Intriguingly, we underline that a relevant proportion (ranging from 5.21 to 7.22%, Figure 7) of variants associated with ASE identified in this study, are also included in the NHGRI-EBI GWAS catalog [62], one of the most comprehensive resources for the aggregation of human genetic variants associated with phenotypic traits. Additionally, a large number of variants associated with ASE (ranging from 6.43 to 9.13% in different individuals), overlaps functional genomic regulatory elements as reported in the ENSEMBL regulatory build [63], consistent with potential regulatory mechanisms. That notwithstanding, the majority (68%) of polymorphic sites that display unbalanced allelic expression, could not be assigned to a potential “causative” functional annotation according to our analyses. This observation could be explained by the fact that a significant proportion of these variants are not causative of ASE *per se,* but could be in linkage disequilibrium with other distal causative variants.

To investigate this scenario, we performed comparative analyses of genetic variability of core promoters elements, defined as a genomic region encompassing 120 bp from annotated TSSs of gene transcripts, between genes associated, or not associated with ASE. As observed from Figure 8, a limited but statistically significant (Wilcoxon p-value < 2.2e-16) increase of genetic diversity is associated with ASE genes, if compared to genes that show a biallelic pattern of gene expression. This might suggest that at least in part the observed patterns of ASE could be explained again by genetic variants associated with functional regulatory genomic elements *in cis*, as already proposed by other studies [1-4].

## Discussion

The fine regulation of gene expression is a complex process orchestrated by the interplay of genetic and epigenetic factors, external stimuli, but also by post-transcriptional molecular pathways involved in the turn-over and degradation of mature transcripts [64]. In recent years the integration of comprehensive datasets of human genetic variation with assays for the quantification of transcript levels [10,11,32,65], has led to the identification of large sets of human genetic variants associated with differential modulation of gene expression [1-4]. Accordingly, several bioinformatics methods and workflows have been developed [45-51] to integrate different types of genetic and gene expression data and obtain a more detailed picture of the type and extent of molecular signatures associated with the modulation of gene expression.

In this paper we presented aScan, a novel method for the analysis of ASE in single individuals. To demonstrate the application of our method we performed extensive simulations of ASE genes with varying levels of expression and numbers of heterozygous polymorphic sites, by taking advantage of a large publicly available dataset incorporating genetic and gene expression profiles of 60 distinct individuals [52]. We showed that our method can achieve high levels of sensitivity in the detection of ASE genes, with a false discovery rate which is minimal and well in line with the nominal FDR used in all our analyses and especially when only genes with adequate levels of coverage (10 or more RNAseq reads) at heterozygous polymorphic sites are considered. Along with a highly effective statistical framework for the identification of biased allelic expression, aScan incorporates also a simple, but highly effective algorithm, based on the observed allelic expression profiles, for the reconstruction of complete or nearly complete haplotypes from unphased genomic data. This represents a considerable advantage over the majority currently available methods for the study of ASE, which require phased genomic data, and/or the application of dedicated workflows or external tools for the reconstruction of complete or nearly complete paternal and maternal genomic sequences [48-51]. Importantly, highly consistent results have been obtained in the analysis of both phased and unphased data with aScan and especially when, as already suggested also by other studies, a threshold of allelic imbalance ratio was applied [54-56]. Moreover, although aScan has been initially developed for the analysis of RNA sequencing data, the principles used in the implementation of our method and its overall statistical framework are suitable for any quantitative assay based on Next Generation Sequencing technologies, including for example ChIP-seq data for the identification of allele specific transcription factor binding and/or epigenetic modifications, or CAGE-seq data for the study of the differential usage of transcription start sites. A level of flexibility that is not matched by other currently available tools. Finally, the aScan provides detailed reports of ASE patterns at gene, transcript and single marker level, which can facilitate a more detailed annotation and comparison of different functional elements.

Studies on ASE can have important applications in clinical settings, and in particular for the accurate annotation of genetic variants associated with the modulation of gene expression [66-67]. To illustrate the potential benefits of the application of methods for the study of ASE in single individuals in providing a fine grained annotation of genetic variants, we applied our method to the analysis of a small cohort of 3 males of Eastern European ancestry, and matched expression data for 6 different tissues. Several important observations can be derived from this analysis. First, we observe that as expected the usage of large scale datasets and resources for the annotation of eQTL can be sometimes misleading when applied for the annotation of the genetics variants of a specific subject. Indeed, we observe that a significant proportion of rare, and/or population specific variants identified in our subjects is not currently included in the GTEx database, one of the most used and most complete resources for the annotation of human eQTL. The fact that a substantial proportion of these variants was represented in other more comprehensive resources of human genetic variation, might suggest that at present the GTEx database provides a biased sampling of genetic variants from distinct human ethnic/geographic groups. This is a relevant consideration which should be taken into account when annotations derived from GTEx are applied to individuals and human populations whose genetic background is not adequately represented in the database. More importantly we underscore that several genetic variants associated with biased gene expression patterns according to GTEx, were not associated with imbalanced allelic expression in the individuals considered in this study. This finding suggests that, as outlined by previous studies [68,69], detailed analyses of ASE patterns in single individuals have relevant applications in personalised medicine, as these and similar information could be used to improve the functional annotation of genetic variants, including those associated with non-coding functional genetic elements. In turn these augmented, individual specific, levels of annotation could inform variant prioritization strategies, resulting in a more accurate identification of variants and genomic loci potentially associated with a phenotypic condition.

Also, by performing a thorough annotation and comparative analyses of genetic variants and genes associated with ASE, we showed that the application of our method can recapitulate in an accurate manner several of the recent findings on the molecular mechanisms associated with ASE, including: its highly individual specific nature [28]; the association of ASE with an increase in genetic variation at regulatory genomic loci [1-4]; the fact that mechanisms involved in transcripts degradation and stability can effectively impact on measured levels of gene expression [40]. Additionally, the large overlap between genetic variants associated with ASE in our individuals and genomic loci associated with different phenotypic traits according to GWAS studies, and or implicated in the regulation of gene expression levels according to specialized databases of human genomic regulatory elements, are again consistent with previous findings and suggest that ASE could potentially explain a relevant proportion of the phenotypic diversity observed in human populations, but also that a significant proportion of genetic variants possibly linked with ASE are in turn associated with different types of cis-regulatory elements, that can modulate transcriptional regulation and/or the stability of mRNA trascripts. Although we observe also that, according to our functional annotation analyses, genes associated with unbalanced allelic expression in an individual do not necessarily cluster within a particular biological pathway, suggesting that ASE is a widespread phenomenon and that it is not associated with a specific biological process.

All in all, the data presented in the current study provide a first proof of concept of the application of aScan for the detection of allele specific expression and of the potential benefits of this type of analyses for a fine grained annotation of genetic variants at individual level. In the light of the highly individual specific nature of ASE, as outlined also by our results, we suggest that when possible clinical studies should always incorporate individual specific analyses of gene expression patterns, to provide an accurate annotation of genetic variants associated with the modulation of gene expression. Since patterns of ASE are mainly determined by the genetic profile of the individual, these analyses could be executed on different types of tissues/samples, not necessarily related with the pathological condition under study.  For example, ASE measured on circulating RNA (e.g. from a liquid biopsy) may recapitulate the same phenomenon in disease-related cell and tissue types, thus providing relevant clues for understanding disease-related molecular mechanisms. Considering also the constant decrease in the costs of sequencing, we anticipate that, to attain a more accurate “person-specific” annotation and prioritization of genetic variants, novel approaches for the development of precision medicine applications in the forthcoming years will strongly rely on the development and utilization of methods and strategies for the integration of different types of omics experiments. In this respect we believe that by providing an efficient and precise system for the identification of allelic imbalance, at different levels, aScan will provide a highly useful and reliable method.

## Materials and methods

## Algorithm and Implementation

The aScan algorithm requires as input the result of the mapping of RNA-seq reads on the genome in BAM format, a VCF file with the genomic annotation of variants and a

GTF file with a reference transcripts or genes annotation. The current version of aScan takes into consideration only heterozygous single-nucleotide substitutions.

Given a gene (transcript) G with N heterozygous positions, starting from the mapping of the RNA-Seq reads for each position *i* the algorithm computes the nucleotide counts $c_{1i}$ and $c_{2i}$ for the two alleles derived from the sequence reads covering the position. The sum $c_{1i} + c_{2i}$ thus equals the overall read coverage of the position. Then, for each position *i*, the deviation from the theoretical uniform distribution of the two alleles (50-50%) is assessed with a log-likelihood test:

$$\chi_i = 2\left(c_{1i} \ln\ln \frac{c_{1i}}{m} + c_{2i} \ln\ln \frac{c_{2i}}{m}\right)$$

Where m is the expected number of occurrences for each allele. That is, given the null hypothesis that the observed counts result from random sampling from a uniform distribution with equal nucleotide counts $m = (c_{1i} + c_{2i})/2$, the distribution of $\chi_i$ is approximately a chi-squared distribution, with one degree of freedom. That is, it expresses the probability that the observed nucleotide counts are derived from two alleles with equal transcript levels. Similar approaches have already been proposed for the analysis of expression data, for example in microarrays [70] and differential expression of duplicated genes [71].

To assess the allele specificity of the expression of the whole gene (transcript) we compute the sum of the $\chi_i$ values associated with all the N heterozygous positions:

$$\chi(G) = \sum_{j=1}^{N} \chi_i$$

The distribution of *$\chi(G)$* is a chi-squared distribution with N degrees of freedom, summarizing the probability of the whole gene (transcript) not having an allele specific expression, that is, of observing the nucleotide counts by chance in a gene evenly expressed on both alleles. P-values are corrected for multiple testing by applying the

Benjamini Hochberg procedure for the control of False Discovery Rate (FDR). Genes (transcripts) reported having an allele-specific expression are finally those with an overall $\chi(G)$ lower than a given FDR threshold t (0.01 in our experiments).

A further condition can be imposed also on the raw counts for the alleles. Given C1: the sum of nucleotides of the most frequent allele count at each heterozygous position and C2 the sum of nucleotides of the less frequent allele count at each heterozygous position, to be allele specific the gene should have an overall bias towards the most frequent allele, that is, $C1/(C1+C2) > t$ (we used 0.6 in our experiments).

The above calculations do not take into account phasing. In case the latter information is available, then the second condition can be modified accordingly, that is, the two C1 and C2 values can be directly attributed to the two alleles, with once again the condition of having a bias towards one of the two alleles.

The approach just described can be applied not only to whole transcripts or genes, but also to single exons, by considering e.g. only the positions within a cassette exon. In this case the result of the analysis will be the identification of allele specific inclusion/excision of the exon itself.

## Datasets

RNA-seq data of lymphoblastoid cell lines of 60 individuals of CEU (Central European) ancestry were obtained from the Array Express portal [72] under the E-MTAB-197 accession (http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-197). Matched genetic profiles were downloaded from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/. The vcf-subset program, as implemented in the vcf-tools suite [73] was applied to obtain VCF files of single individuals.

RNA-seq data and matched genome sequencing data of three post-mortem healthy individuals of Russian ancestry, as available under the phs000870 accession, were obtained directly from the dbgap database [74]. Annotation of human eQTLs according to the GTEx study [32] was obtained directly from the GTEx portal at: https://storage.googleapis.com/gtex_analysis_v8/multi_tissue_qtl_data/GTEx_Analysis_v8.metasoft.txt.gz.

## Analysis of RNA-seq data

RNA-seq reads were aligned to the reference Refseq (release 106) annotation of the hg19 assembly of the human transcriptome [75], as obtained from http://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh38_latest/refseq_identifiers/GRCh38_latest_genomic.gff.gz, according to the guidelines defined in [56]. Alignments were performed by means of the Bowtie2 [76] program, using the following parameters -D 20 -R 3 -N 1 -L 18 -i S,1,0.50. Gene expression levels were derived, for every individual, by applying the RSEM software [77]. Genes annotated on the mitochondrial genome or on hap chromosomes were not considered, and the respective read counts excluded from downstream normalizations and analyses.

## Simulation of ASE data

For every individual, maternal and paternal haplotypes were reconstructed from the corresponding BAM file by applying the bcf-tools [78] consensus utility to the hg19 reference assembly of the human genome. For every sample, genes were arbitrarily assigned to 1 of 4 possible classes based on their expression levels:

1. highly expressed (TPM>15),
2. moderately expressed (3<TMP≤15),
3. lowly expressed (1<TPM≤3)

4. scarcely expressed (0.5<TPM≤1).

Genomic loci reported in the “black list” of problematic genomic regions, as defined by the Encode project [79] were not considered for the computation of ASE.

To simulate different levels of ASE, 1600 genes, 400 for each of the different classes of expression level as defined above, were selected for every individual. Only genes including at least 1 heterozygous site were considered. For these genes different ratios of allele specific expression (4:6, 3:7, 2:8, 1:9) were simulated by applying the RSEM-simulate-reads utility [77] with RNA-seq models as inferred directly from real data, to the paternal and maternal transcriptomes, as reconstructed from the respective consensus genomic sequence. To avoid confounding effects, equivalent proportions (1:1) of paternal and maternal reads were simulated for all the expressed genes (TPM>0.5) that were selected for the simulation of ASE. Fastq files of simulated paternal and maternal reads were merged into a single fastq file. Finally simulated reads were aligned to the reference Refseq annotation of the hg19 assembly of the human genome, again by means of the RSEM program, using the same procedure as outlined above. The “--output-genome-bam” option was used to obtain genome alignment files of simulated RNA-seq reads in BAM format.

## Filtering of VCF and BAM files

To avoid possible biases in the computation ASE -due to inconsistent mapping of RNA-seq reads- a set of filters was applied to exclude highly variable genomic regions. The genome was segmented in a set of overlapping (by 5bp), genomic windows of 1 read length in size (75 bp) by means of the bedtools makewindows utility [80]. Profiles of SNP and indels density were obtained for every individual, on the genomic intervals defined above, by applying bedtools coverage [80]. Genomic windows containing 3 or more SNPs and/or an indel associated with any other type of genetic variants were filtered from the corresponding VCF file and discarded from subsequent analyses.

BAM files were filtered to retain only reads with a mapping quality of 20 or above by using samtools view [81]. PCR duplicates as identified by applying the SAMBLASTER [82], were removed from subsequent computations

## Execution of aScan and post processing of aScan results

aScan version 1.0.3 as available from https://github.com/Federico77z/aScan/releases/tag/1.0.3 was applied in all our analyses, using the same gtf files that were provided in input to RSEM for the calculation of gene expression (Refseq release 106 annotation of the hg19 assembly of the human genome). A FDR of 0.01 was considered for the identification of genes associated with allele specific expression. To minimize possible false positive calls, in the analysis of real data only genes for which all heterozygous sites were covered by 10 or more RNA-seq reads were considered, additionally a threshold of allelic imbalance ratio of 4:6 or higher was also applied.

Custom Perl scripts were used for the post-processing of the results. The standard libraries of the R programming language were used for graphical representation of the data.

## Functional annotation of genetic variants and ASE genes

Functional enrichment analyses of ASE genes, as identified by aScan were executed by the means of Functional Annotation Clustering utility, with default parameters as available form the DAVID website [59].

Functional annotations of genetic variants were performed by means of the ANNOVAR software [83], based on the Refseq [75] (release 106) annotation of the hg19 reference assembly of the human genome. The following resources were considered for the annotation of allele frequencies in human populations: ExAC [84] (version 1.0 updated 27 February 2017), 1000 Genomes [53] (phase 3), gnomAD [61] (version 2.1, updated 10 December 2018), dbSNP [85] (build 151), Kaviar [86] (version 160204-Public) and TopMed [60] (freeze5, accessed on 28 February 2019, nhlbiwgs.org). The Ensembl regulatory build [63] was used for the annotation of non-coding functional genomic elements. The NHGRI-EBI GWAS catalog [62] was used to obtain a comprehensive catalog of genetic variants reported in GWAS studies.

## Acknowledgments

We thank ELIXIR-Italy for providing the computing and bioinformatics facilities and Maria Rosa Mirizzi for technical assistance.

## Funding

This work was supported by the Italian Ministero dell'Università e Ricerca: PRIN 2017, Consiglio Nazionale delle Ricerche, H2020 projects EOSC-Life (GA 824087), EOSC-Pillar (GA 857650) and ELIXIR Converge (GA 871075), and Elixir-IIB.

## References

1. Jones BL, Swallow DM. The impact of cis-acting polymorphisms on the human phenotype. HUGO J. 2011; 5:13–23
2. Ecker S, Pancaldi V, Valencia A, et al. Epigenetic and Transcriptional Variability Shape Phenotypic Plasticity. BioEssays 2018; 40:1700148
3. Chatterjee S, Ahituv N. Gene Regulatory Elements, Major Drivers of Human Disease. Annu. Rev. Genomics Hum. Genet. 2017; 18:45–63
4. Tomar A, Teperino R. Genetic control of non-genetic inheritance in mammals: state-of-the-art and perspectives. Mamm Genome. 2020;31(5-6):146-156. doi:10.1007/s00335-020-09841-5
5. Knight JC. Allele-specific gene expression uncovered. Trends Genet. 2004; 20:113–116

6. Peters J. The role of genomic imprinting in biology and disease: an expanding view. Nat. Rev. Genet. 2014; 15:517–530
7. Ferguson-Smith AC. Genomic imprinting: the emergence of an epigenetic paradigm. Nat. Rev. Genet. 2011; 12:565–575
8. Kukurba KR, Zhang R, Li X, et al. Allelic Expression of Deleterious Protein-Coding Variants across Human Tissues. PLOS Genet. 2014; 10:e1004304
9. Pirinen M, Lappalainen T, Zaitlen NA, et al. Assessing allele-specific expression across multiple tissues from RNA-seq read data. Bioinformatics 2015; 31:2497–2504
10. Chen J, Rozowsky J, Galeev TR, et al. A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. Nat. Commun. 2016; 7:11101
11. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, et al. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature 2010; 464:773–777
12. Tian L, Khan A, Ning Z, et al. Genome-wide comparison of allele-specific gene expression between African and European populations. Hum. Mol. Genet. 2018; 27:1067–1077
13. Sun M, Zhang J. Allele-specific single-cell RNA sequencing reveals different architectures of intrinsic and extrinsic gene expression noises. Nucleic Acids Res. 2020; 48:533–547
14. Pastinen T. Genome-wide allele-specific analysis: insights into regulatory variation. Nat. Rev. Genet. 2010; 11:533–538
15. Turro E, Su S-Y, Gonçalves et al. Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. Genome Biol. 2011; 12:R13
16. Moyerbrailean GA, Richards AL, Kurtz D, et al. High-throughput allele-specific expression across 250 environmental conditions. Genome Res. 2016; 26:1627–1638
17. Gutierrez-Arcelus M, Baglaenko Y, Arora J, et al. Allele-specific expression changes dynamically during T cell activation in HLA and other autoimmune loci. Nat. Genet. 2020; 52:247–253
18. Huang W-C, Ferris E, Cheng T, et al. Diverse Non-genetic, Allele-Specific Expression Effects Shape Genetic Architecture at the Cellular Level in the Mammalian Brain. Neuron 2017; 93:1094-1109.e7
19. Prendergast JGD, Tong P, Hay DC, et al. A genome-wide screen in human embryonic stem cells reveals novel sites of allele-specific histone modification associated with known disease loci. Epigenetics Chromatin 2012; 5:6
20. Khansefid M, Pryce JE, Bolormaa S, et al. Comparing allele specific expression and local expression quantitative trait loci and the influence of gene expression on complex trait variation in cattle. BMC Genomics 2018; 19:793
21. LaFramboise T, Weir BA, Zhao X, et al. Allele-Specific Amplification in Cancer Revealed by SNP Array Analysis. PLoS Comput. Biol. 2005; 1:
22. Halabi NM, Martinez A, Al-Farsi H, et al. Preferential Allele Expression Analysis Identifies Shared Germline and Somatic Driver Genes in Advanced Ovarian Cancer. PLoS Genet. 2016; 12:e1005755
23. Sandberg R, Ernberg I. Assessment of tumor characteristic gene expression in cell lines using a tissue similarity index (TSI). Proc. Natl. Acad. Sci. 2005; 102:2052–2057
24. Fan J, Lee H-O, Lee S, et al. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. Genome Res. 2018; 28:1217–1227
25. Pinheiro H, Bordeira-Carriço R, Seixas S, et al. Allele-specific CDH1 downregulation and hereditary diffuse gastric cancer. Hum. Mol. Genet. 2010; 19:943–952
26. Iyer SV, Parrales A, Begani P, et al. Allele-specific silencing of mutant p53 attenuates dominant-negative and gain-of-function activities. Oncotarget 2016; 7:5401–5415

27. Crowley JJ, Zhabotynsky V, Sun W, et al. Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. Nat. Genet. 2015; 47:353–360
28. Garg P, Borel C, Sharp AJ. Detection of parent-of-origin specific expression quantitative trait loci by cis-association analysis of gene expression in trios. PloS One 2012; 7:e41695
29. Prickett AR, Barkas N, McCole RB, et al. Genome-wide and parental allele-specific analysis of CTCF and cohesin DNA binding in mouse brain reveals a tissue-specific binding pattern and an association with imprinted differentially methylated regions. Genome Res. 2013; 23:1624–1635
30. Singh P, Cho J, Tsai SY, et al. Coordinated allele-specific histone acetylation at the differentially methylated regions of imprinted genes. Nucleic Acids Res. 2010; 38:7974–7990
31. Nica AC, Dermitzakis ET. Expression quantitative trait loci: present and future. Philos. Trans. R. Soc. B Biol. Sci. 2013; 368:
32. Lonsdale J, Thomas J, Salvatore M, et al. The Genotype-Tissue Expression (GTEx) project. Nat. Genet. 2013; 45:580–585
33. Westra H-J, Franke L. From genome to function by studying eQTLs. Biochim. Biophys. Acta BBA - Mol. Basis Dis. 2014; 1842:1896–1902
34. Ye Y, Zhang Z, Liu Y, Diao L, Han L. A Multi-Omics Perspective of Quantitative Trait Loci in Precision Medicine. Trends Genet. 2020 May;36(5):318-336. doi: 10.1016/j.tig.2020.01.009. Epub 2020 Feb 24. PMID: 32294413
35. Gibson G, Powell JE, Marigorta UM. Expression quantitative trait locus analysis for translational medicine. Genome Med. 2015;7(1):60. Published 2015 Jun 24. doi:10.1186/s13073-015-0186-7
36. Michaelson JJ, Loguercio S, Beyer A. Detection and interpretation of expression quantitative trait loci (eQTL). Methods 2009; 48:265–276
37. Sun W, Hu Y. eQTL Mapping Using RNA-seq Data. Stat. Biosci. 2013; 5:198–219
38. Li H, Deng H. Systems genetics, bioinformatics and eQTL mapping. Genetica 2010; 138:915–924
39. Almlöf JC, Lundmark P, Lundmark A, et al. Powerful Identification of Cis-regulatory SNPs in Human Primary Monocytes Using Allele-Specific Gene Expression. PLOS ONE 2012; 7:e52260
40. Erwood S, Laselva O, Bily TMI, et al. Allele-Specific Prevention of Nonsense-Mediated Decay in Cystic Fibrosis Using Homology-Independent Genome Editing. Mol. Ther. - Methods Clin. Dev. 2020; 17:1118–1128
41. Yang E-W, Bahn JH, Hsiao EY-H, et al. Allele-specific binding of RNA-binding proteins reveals functional genetic variants in the RNA. Nat. Commun. 2019; 10:1338
42. Nembaware V, Lupindo B, Schouest K, et al. Genome-wide survey of allele-specific splicing in humans. BMC Genomics 2008; 9:265
43. Sun W, Gao Q, Schaefke B, et al. Pervasive allele-specific regulation on RNA decay in hybrid mice. Life Sci. Alliance 2018; 1:
44. Fernald GH, Capriotti E, Daneshjou R, et al. Bioinformatics challenges for personalized medicine. Bioinformatics 2011; 27:1741–1748
45. Raghupathy N, Choi K, Vincent MJ, et al. Hierarchical analysis of RNA-seq reads improves the accuracy of allele-specific expression. Bioinformatics 2018; 34:2177–2184
46. Deonovic B, Wang Y, Weirather J, et al. IDP-ASE: haplotyping and quantifying allele-specific expression at the gene and gene isoform level by hybrid sequencing. Nucleic Acids Res. 2017; 45:e32
47. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-

generation DNA sequencing data. Genome Res. 20, 1297–303. doi:10.1101/gr.107524.110
48. Edsgärd D, Iglesias MJ, Reilly S-J, et al. GeneiASE: Detection of condition-dependent and static allele-specific expression from RNA-seq data without haplotype information. Sci. Rep. 2016; 6:21134
49. Skelly DA, Johansson M, Madeoy J, et al. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. Genome Res. 2011; 21:1728–1737
50. Fan J, Hu J, Xue C, et al. ASEP: Gene-based detection of allele-specific expression across individuals in a population by RNA sequencing. PLOS Genet. 2020; 16:e1008786
51. León-Novelo L, Gerken AR, Graze RM, et al. Direct Testing for Allele-Specific Expression Differences Between Conditions. G3 GenesGenomesGenetics 2017; 8:447–460
52. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. Nature 2015; 526:68–74
53. The 1000 Genomes Project Consortium, Delaneau O, Marchini J. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. Nat. Commun. 2014; 5:3934
54. Castel SE, Levy-Moonshine A, Mohammadi P, et al. Tools and best practices for data processing in allelic expression analysis. Genome Biol. 2015; 16:195
55. Fontanillas P, Landry CR, Wittkopp PJ, et al. Key considerations for measuring allelic expression on a genomic scale using high-throughput sequencing. Mol. Ecol. 2010; 19 Suppl 1:212–227
56. Stevenson KR, Coolon JD, Wittkopp PJ. Sources of bias in measures of allele-specific expression derived from RNA-seq data aligned to a single reference genome. BMC Genomics 2013; 14:536
57. Picardi E, Manzari C, Mastropasqua F, et al. Profiling RNA editing in human tissues: towards the inosinome Atlas. Sci. Rep. 2015; 5:14941
58. D’Erchia AM, Atlante A, Gadaleta G, et al. Tissue-specific mtDNA abundance from exome data and its correlation with mitochondrial transcription, mass and respiratory activity. Mitochondrion 2015; 20:13–21
58. Ginjala V. Gene imprinting gateway. Genome Biol. 2001; 2:reports2009
59. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 2009; 4:44–57
60. Taliun D, Harris DN, Kessler MD, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. bioRxiv 2019; 563866
61. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 2020; 581:434–443
62. MacArthur J, Bowler E, Cerezo M, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017; 45:D896–D901
63. Zerbino DR, Wilder SP, Johnson N, et al. The Ensembl Regulatory Build. Genome Biol. 2015; 16:56
64. Lee TI, Young RA. Transcriptional Regulation and its Misregulation in Disease. Cell 2013; 152:1237–1251
65. Knowles DA, Davis JR, Edgington H, et al. Allele-specific expression reveals interactions between genetic variation and environment. Nat. Methods 2017; 14:699–702
66. Hindorff LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. 2009; 106:9362–9367
67. Casamassimi A, Federico A, Rienzo M, et al. Transcriptome Profiling in Human Diseases: New Advances and Perspectives. Int. J. Mol. Sci. 2017; 18

68. Bell CG, Beck S. Advances in the identification and analysis of allele-specific expression. Genome Med. 2009; 1:56
69. Lee C, Kang EY, Gandal MJ, et al. Profiling allele-specific gene expression in brains from individuals with autism spectrum disorder reveals preferential minor allele usage. Nat. Neurosci. 2019; 22:1521–1532
70. Wang S, Ethier S. A generalized likelihood ratio test to identify differentially expressed genes from microarray data. Bioinformatics. 2004;20(1):100-104. doi:10.1093/bioinformatics/btg384
71. Smith RD, Kinser TJ, Smith GDC, Puzey JR. A likelihood ratio test for changes in homeolog expression bias. BMC Bioinformatics. 2019;20(1):149. Published 2019 Mar 20. doi:10.1186/s12859-019-2709-5
72. Parkinson H, Kapushesky M, Shojatalab M, et al. ArrayExpress—a public database of microarray experiments and gene expression profiles. Nucleic Acids Res. 2007; 35:D747–D750
73. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. Bioinformatics 2011; 27:2156–2158
74. Mailman MD, Feolo M, Jin Y, et al. The NCBI dbGaP database of genotypes and phenotypes. Nat. Genet. 2007; 39:1181–1186
75. O'Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016; 44:D733-745
76. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat. Methods 2012; 9:357–359
77. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 2011; 12:323
78. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 2011; 27:2987–2993
79. Amemiya HM, Kundaje A, Boyle AP. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci. Rep. 2019; 9:9354
80. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010; 26:841–842
81. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinforma. Oxf. Engl. 2009; 25:2078–2079
82. Faust GG, Hall IM. SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics 2014; 30:2503–2505
83. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010; 38:e164
84. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 2016; 536:285–291
85. Sherry ST, Ward M-H, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001; 29:308–311
86. Glusman G, Caballero J, Mauldin DE, et al. Kaviar: an accessible system for testing SNV novelty. Bioinformatics 2011; 27:3216–32

## Figures Legends

**Figure 1. Sensitivity and False Discovery rate of aScan on simulated ASE data.** A) Average sensitivity of aScan in the detection of ASE at different ratios of allelic imbalance. Ratios are indicated below each bar. Average sensitivity is represented in the fifth column. A red dotted line is used to indicate the average sensitivity. B) False Discovery rate at different ratios of allelic imbalance. A red dotted line is used to indicate the theoretical FDR (0.01). Ratios are indicated below each bar.

**Figure 2. Sensitivity and False Discovery rate of aScan on unphased data.** A) Average sensitivity at different ratios of allelic imbalance. Ratios are indicated below each bar. Average sensitivity is represented in the fifth column. A red dotted line is used to indicate the average sensitivity. B) False Discovery rate at different ratios of allelic imbalance. A red dotted line is used to indicate the theoretical FDR (0.01). Ratios are indicated below each bar.

**Figure 3. Barplots of ASE genes.** Each panel (A, B and C) displays the number of ASE genes as detected by aScan in each of S12, S13 and S7 respectively, in all the six tissues considered in this study, as well as the total number of ASE genes detected in each individual. Tissues are indicated below each bar.

**Figure 4. Venn diagram of genes associated with ASE.** Venn diagram of genes showing allele specific expression in the 3 individuals considered in this study.

**Figure 5. Intersection of ASE genes.** For every individual (panel A, B and C for S12, S13 and S7 respectively) each barplot indicates the number of genes that are associated with ASE in any number (1 to 6) of tissues of that individual.

**Figure 6. Violin plots of allelic ratio.** A) Violin plots of allelic ratios, as measured as the proportion of RNA-seq reads assigned to each allele, at 1261 heterozygous variants associated with eQTLs according to the GTEx database, but associated with ASE according to aScan. B) Violin plots of allelic ratios of variants associated with ASE according to aScan.

**Figure 7. Pie charts of functional annotation of genetic variants associated with ASE.** Each panel indicates the proportion of SNPs associated with ASE that were annotated with different types of functional annotations according to Annovar, in each of S12 (A), S13 (B), S7(C). Different colors (see the legend) are used to indicate distinct types of annotations.

**Figure 8. Violin plots of proportions of polymorphic sites in core promoters.** Rates of polymorphic sites are displayed on the Y axis. These were calculated as the number of polymorphic sites per 100bp in regions encompassing 120 bp from annotated TSSs. Genes showing ASE (light gray) genes not associated with ASE (noASE, dark gray).

## Supplementary Tables Legends

**Supplementary Table S1. Total number polymorphic sites in 60 CEU individuals.** Tot hom: total number of homozygous polymorphic sites. Total het: total number of heterozygous polymorphic sites. Testable: total number of heterozygous sites that were retained in our analysis after the filtration of problematic and hypervariable genomic regions.

**Supplementary Table S2. Total number of reads.** Total number of Million pairs of reads and mapped pairs of reads for each of the 54 distinct RNA-seq libraries analysed in this study. Subjects and tissues are indicated in the Subject and tissue columns respectively. A progressive identifier (R1 to R3), is used in the Replicate column to delineate technical different replicates of the same sample. Total number of pairs of reads (in millions) are reported in the Tot Reads (M). The Tot Mapped (M) column indicates the total number of reads mapped to gene models according to Refseq (version 106) annotation of hg19 assembly of the human genome.

**Supplementary Table S3. Gene expression levels.** Gene expression levels as obtained from RSEM. Genes are indicated in the rows. Conditions in the columns. Expression levels are reported as transcript per Million. (TPM).

**Supplementary Table S4. Total number genes tested.** For every subject and tissue the table reports the number of genes that were tested for ASE. These include only genes covered by at least 50 reads, and with at least 1 heterozygous polymorphic site covered by at least 10 distinct RNAreads.

**Supplementary Table S5 to S7. Functional enrichment analyses.** Results of functional enrichment analyses of ASE genes as obtained from the DAVID suite. For each individual the top 10 clusters as obtained from the Functional annotation clustering utility are reported.

Results for S12, S13 and S7 are reported in supplementary Tables S5, S6 and S7 respectively.

## Supplementary Figures Legends

**Supplementary Figure S1. Sensitivity of aScan at different expression levels.** A) Violin plots of the sensitivity of aScan on genes with different expression levels. Expression levels are indicated below each violin. Sensitivity is displayed on the Y axis. A dotted red line is used to indicate the average sensitivity. B) Equivalent to A, but considering only genes for which all heterozygous polymorphic sites are covered by 10 or more RNA-seq reads.

**Supplementary Figure S2. FDR distributions.** A) Violin plots of FDR distributions at different expression levels . Expression levels are indicated below each violin. The log10 of the FDR is plotted on the Y axis

**Supplementary Figure S3. Sensitivity of aScan on genes with a different number of polymorphic sites.** Violin plots display levels of sensitivity of aScan in the detection of ASE in genes with a different number of polymorphic sites. Number of polymorphic sites are indicated below each violin. Sensitivity is reported on the Y axis.

**Supplementary Figure S4. Sensitivity of aScan on genes with a different number of polymorphic sites and different expression levels.** Violin plots are used to show levels of sensitivity of aScan in the detection of ASE in genes with a different number of polymorphic sites. Number of polymorphic sites are indicated below each violin. Sensitivity is reported on the Y axis. A) Genes expressed at TPM>10. B) Genes expressed between 15 and 3 TPM. C) Genes expressed between 3 and 1 TPM. D) Genes expressed between 0.5 and 1 TPM.

**Supplementary Figure S5. Ratio of allelic imbalance, and expression levels of False Positive calls.** A) Histogram of ratios of allelic imbalance for False Positive predictions of ASE. B) Violin plot comparing expression distribution of expression levels between all the expressed genes (ALL, light blue) and False Positive predictions of ASE (purple).

**Supplementary Figure S6. Rates of False Positive calls for genes with heterozygous sites covered by at least 10 reads.** A red dotted line is used to indicate the observed False Positive rates when no coverage threshold is applied. Ratios are indicated below each bar.

**Supplementary Figure S7. Ratio of allelic imbalance, and expression levels of False Positive calls on unphased data.** A) Histogram of ratios of allelic imbalance for False Positive predictions of ASE. B) Violin plot comparing expression distribution of expression levels between all the expressed genes (ALL, light blue) and False Positive predictions of ASE (purple).

**Supplementary Figure S8. False Discovery rates on unphased data using a threshold for allelic imbalance.** False Discovery rates at different ratios of allelic imbalance. A red dotted line is used to indicate the theoretical FDR (0.01). Ratios are indicated below each bar. A blue line is used to indicate average FDR levels attained without the application of this imbalance threshold (Figure 2).

## Declarations of interest: none

# CRediT Author Statement

Federico Zambelli: Methodology, Software, Writing- Reviewing and Editing.

Matteo Chiara: Conceptualization, Investigation, Writing- Original draft preparation.

Erika Ferrandi: Investigation, Validation, Writing- Reviewing and Editing.

Pietro Mandreoli: Validation, Software, Data Curation,Writing- Reviewing and Editing.

Marco Antonio Tangaro: Validation,Software,Data Curation,Writing- Reviewing and Editing.

Giulio Pavesi: Methodology, Formal analysis, Supervision, Writing- Reviewing and Editing.

Graziano Pesole: Supervision, Writing- Reviewing and Editing, Funding acquisition

| Subject | Age | Cause of death | Tot SNPs | Tot het SNPs | Tot Testable | Tissues |
|---|---|---|---|---|---|---|
| S7 | 47 | Acute coronary syndrome | 99.606 | 57.657 | 51.717 | brain cortex,<br>liver,<br>abdominal striated muscle,<br>kidney,<br>lung,<br>heart myocardium LV |
| S12 | 54 | Car accident | 98.332 | 56.717 | 52.004 | |
| S13 | 48 | Traumatic asphyxia | 97.388 | 56.004 | 49.157 | |

**Table 1. Salient features of the phs000870 dataset .** Subject: pseudonymised identifier of the subjects. Tot SNPs: total number of single nucleotide polymorphisms in protein coding genes. Tot het SNPs: total number of heterozygous SNPs. TotTestable: total number of SNPs that were retained in our analyses after the filtering of problematic or highly variable genomic regions. Tissues: list of tissues for which RNA sequencing data were available

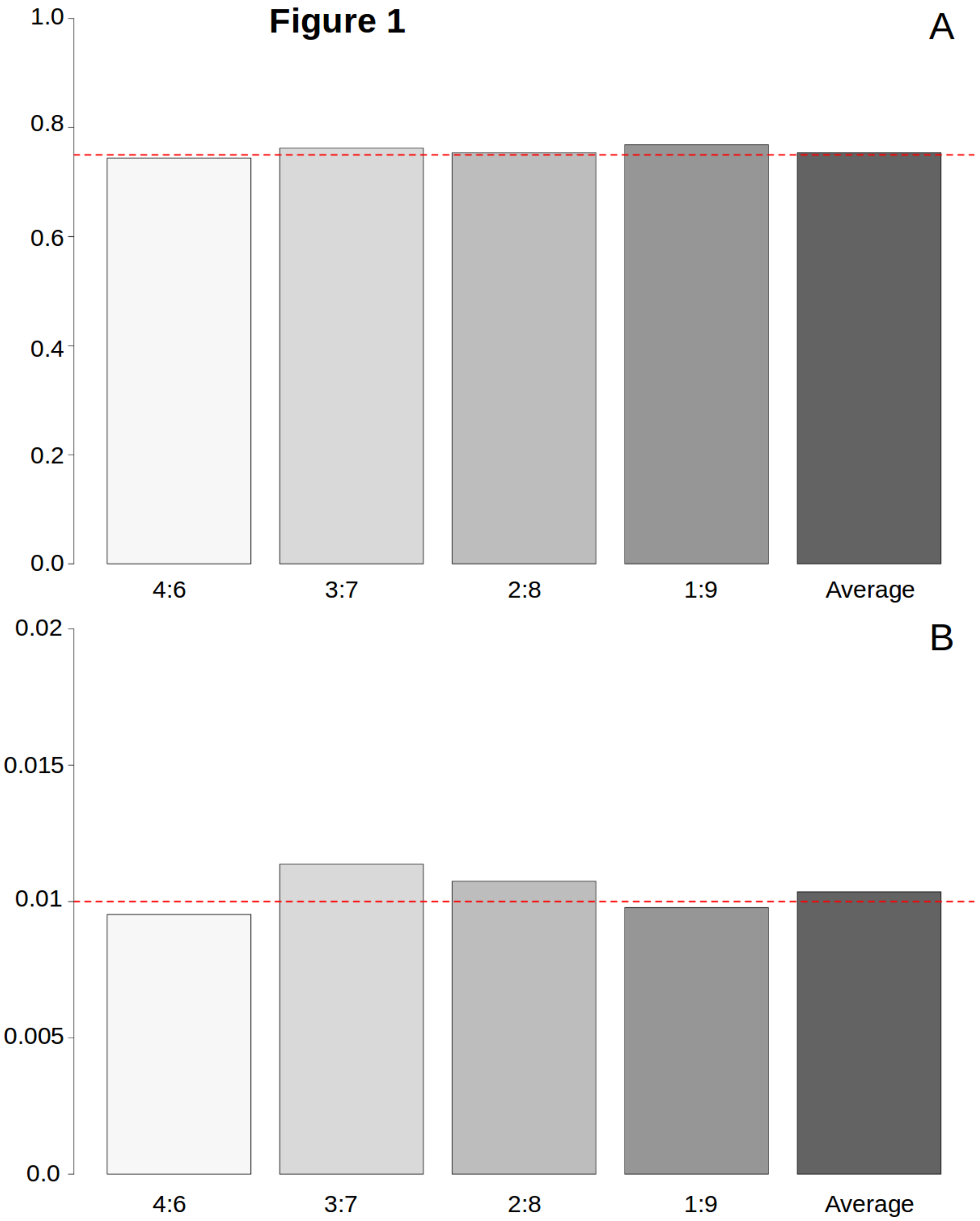

**Figure 1. Sensitivity and False Discovery rate of aScan on simulated ASE data.** A) Average sensitivity of aScan in the detection of ASE at different ratios of allelic imbalance. Ratios are indicated below each bar. Average sensitivity is represented in the fifth column. A red dotted line is used to indicate the average sensitivity. B) False Discovery rate at different ratios of allelic imbalance. A red dotted line is used to indicate the theoretical FDR (0.01). Ratios are indicated below each bar.

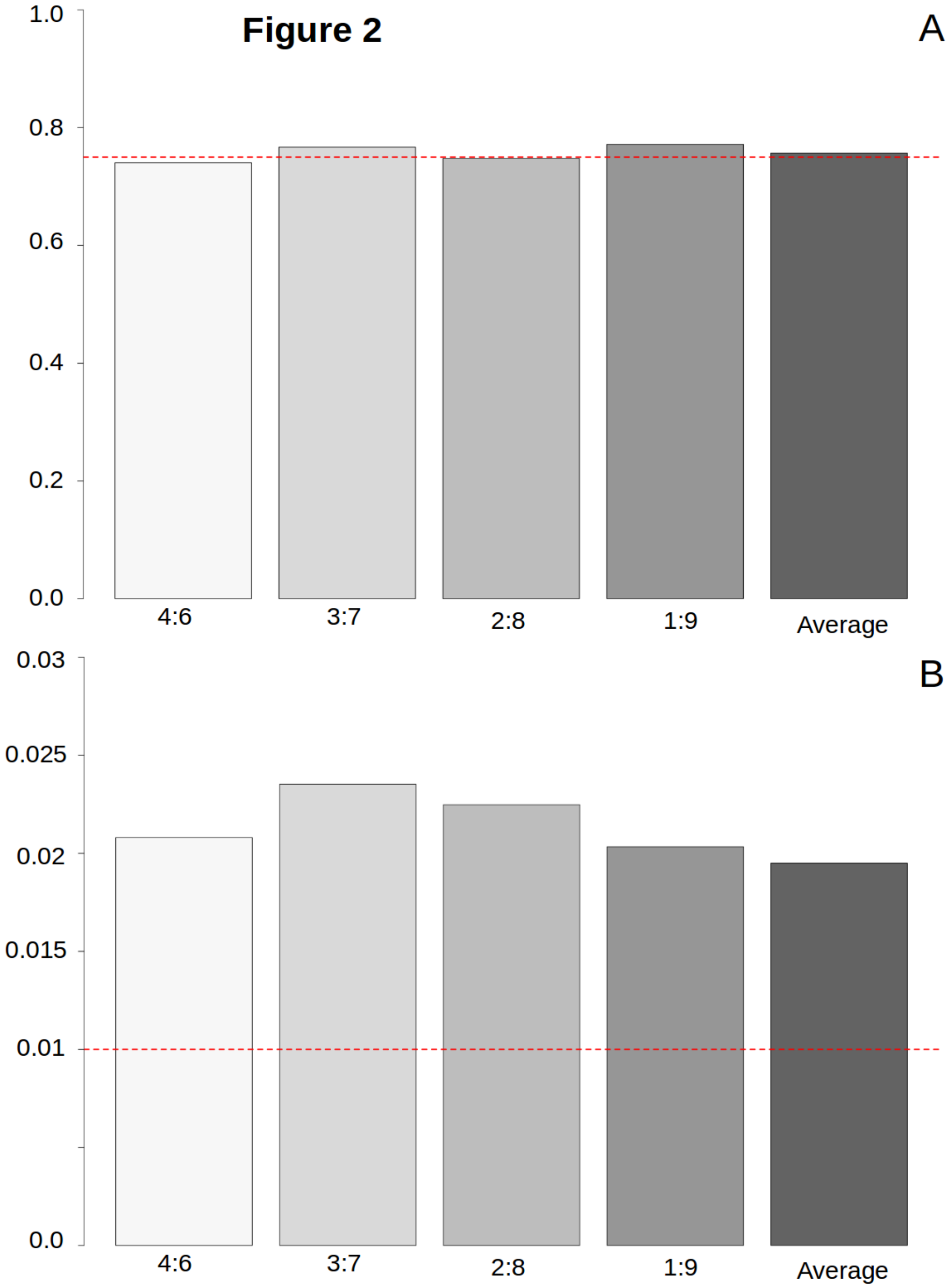

**Figure 2. Sensitivity and False Discovery rate of aScan on unphased data.** A) Average sensitivity at different ratios of allelic imbalance. Ratios are indicated below each bar. Average sensitivity is represented in the fifth column. A red dotted line is used to indicate the average sensitivity. B) False Discovery rate at different ratios of allelic imbalance. A red dotted line is used to indicate the theoretical FDR (0.01). Ratios are indicated below each bar.

# Figure 3

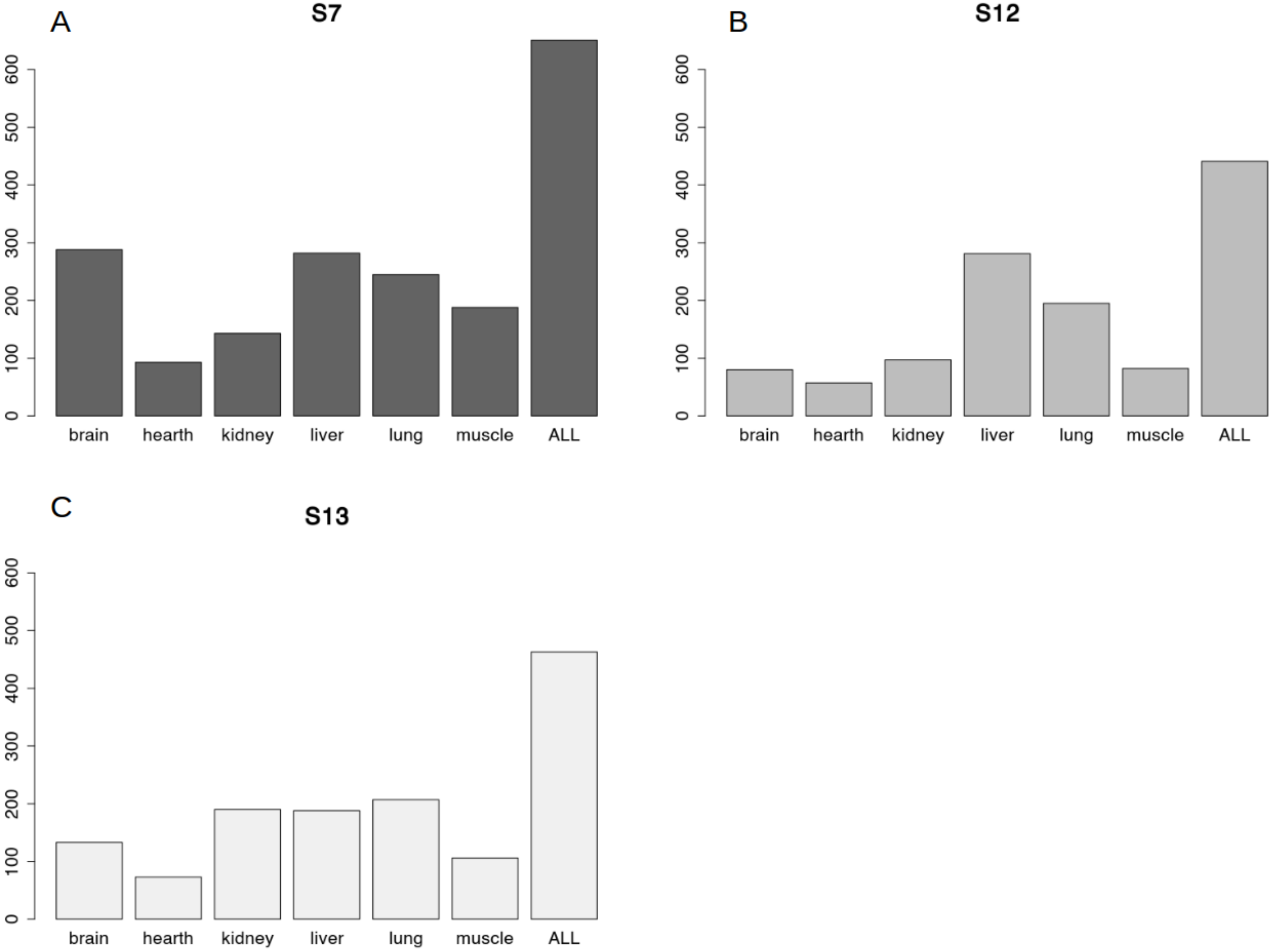

**Figure 3. Barplots of ASE genes.** Each panel (A, B and C) displays the number of ASE genes as detected by aScan in each of S12, S13 and S7 respectively, in all the six tissues considered in this study, as well as the total number of ASE genes detected in each individual. Tissues are indicated below each bar.

Figure 4

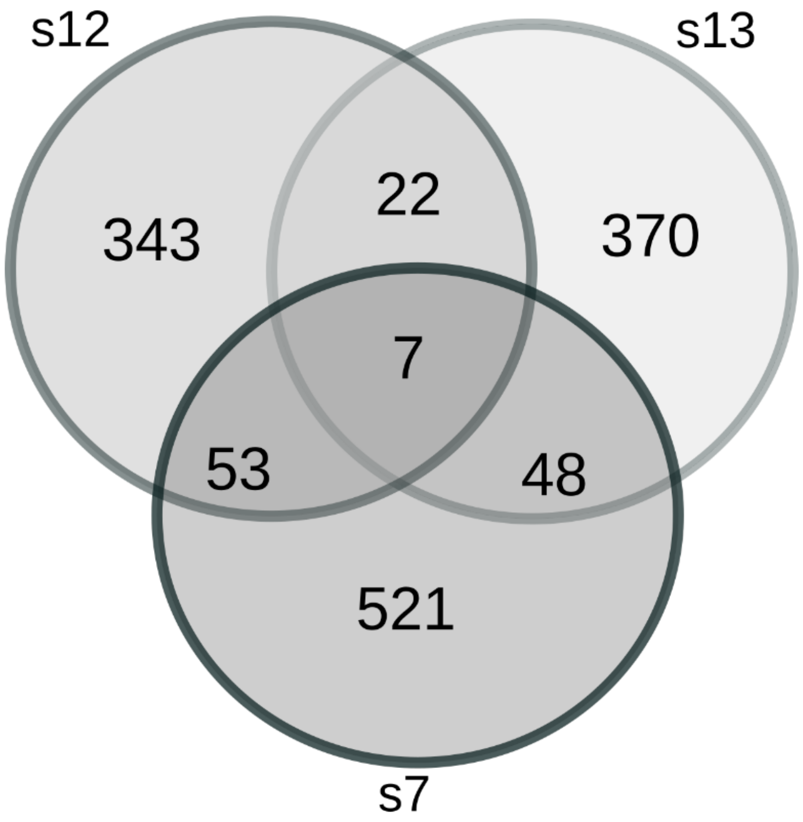

**Figure 4. Venn diagram of genes associated with ASE.** Venn diagram of genes showing allele specific expression in the 3 individuals considered in this study.

# Figure 5

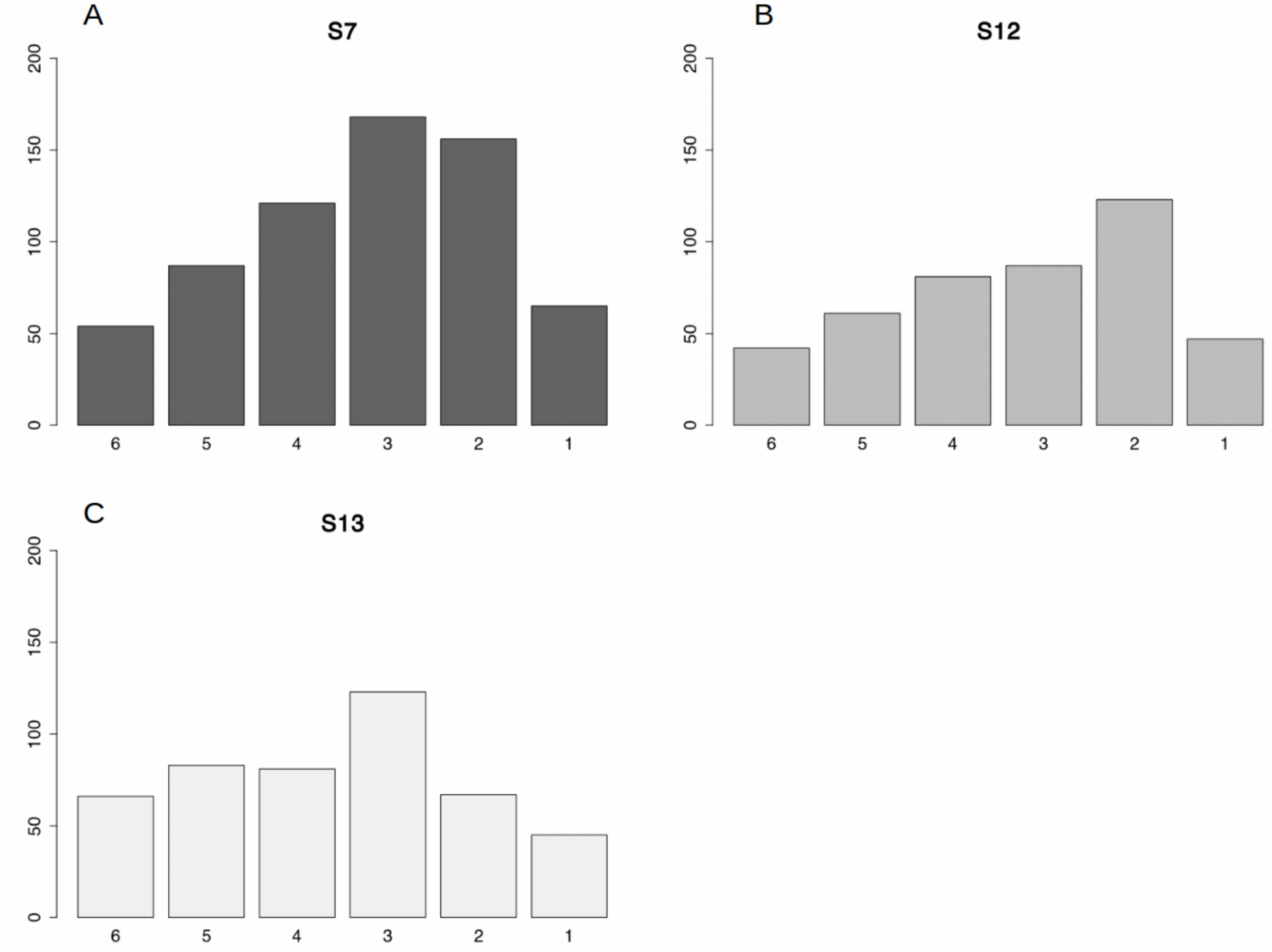

**Figure 5. Intersection of ASE genes.** For every individual (panel A, B and C for S12, S13 and S7 respectively) each barplot indicates the number of genes that are associated with ASE in any number (1 to 6) of tissues of that individual.

# Figure 6

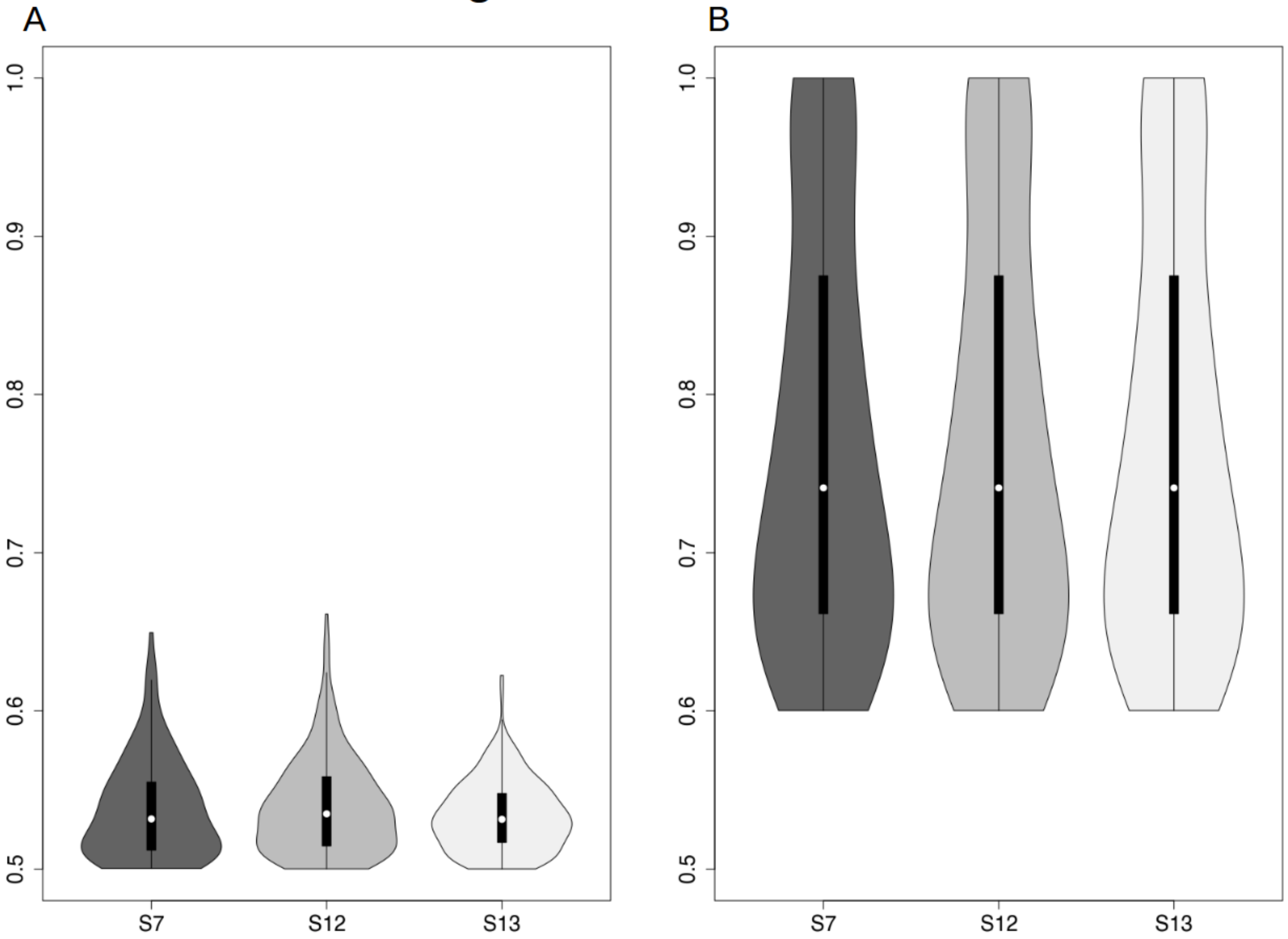

**Figure 6. Violin plots of allelic ratio.** A) Violin plots of allelic ratios, as measured as the proportion of RNA-seq reads assigned to each allele, at 1261 heterozygous variants associated with eQTLs according to the GTEx database, but associated with ASE according to aScan. B) Violin plots of allelic ratios of variants associated with ASE according to aScan.

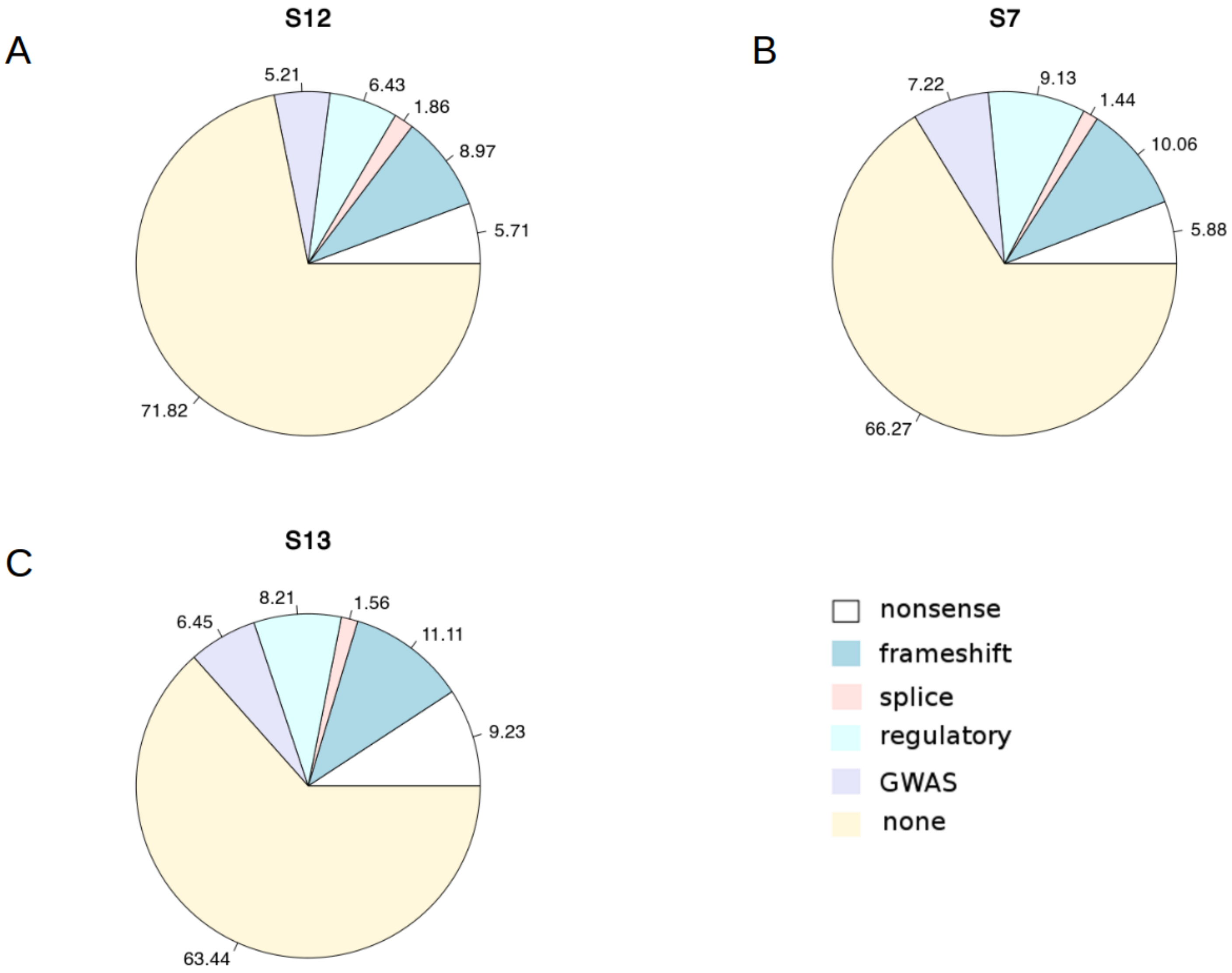

**Figure 7. Pie charts of functional annotation of genetic variants associated with ASE.** Each panel indicates the proportion of SNPs associated with ASE that were annotated with different types of functional annotations according to Annovar, in each of S12 (A), S13 (B), S7(C). Different colors (see the legend) are used to indicate distinct types of annotations.

# Figure 8

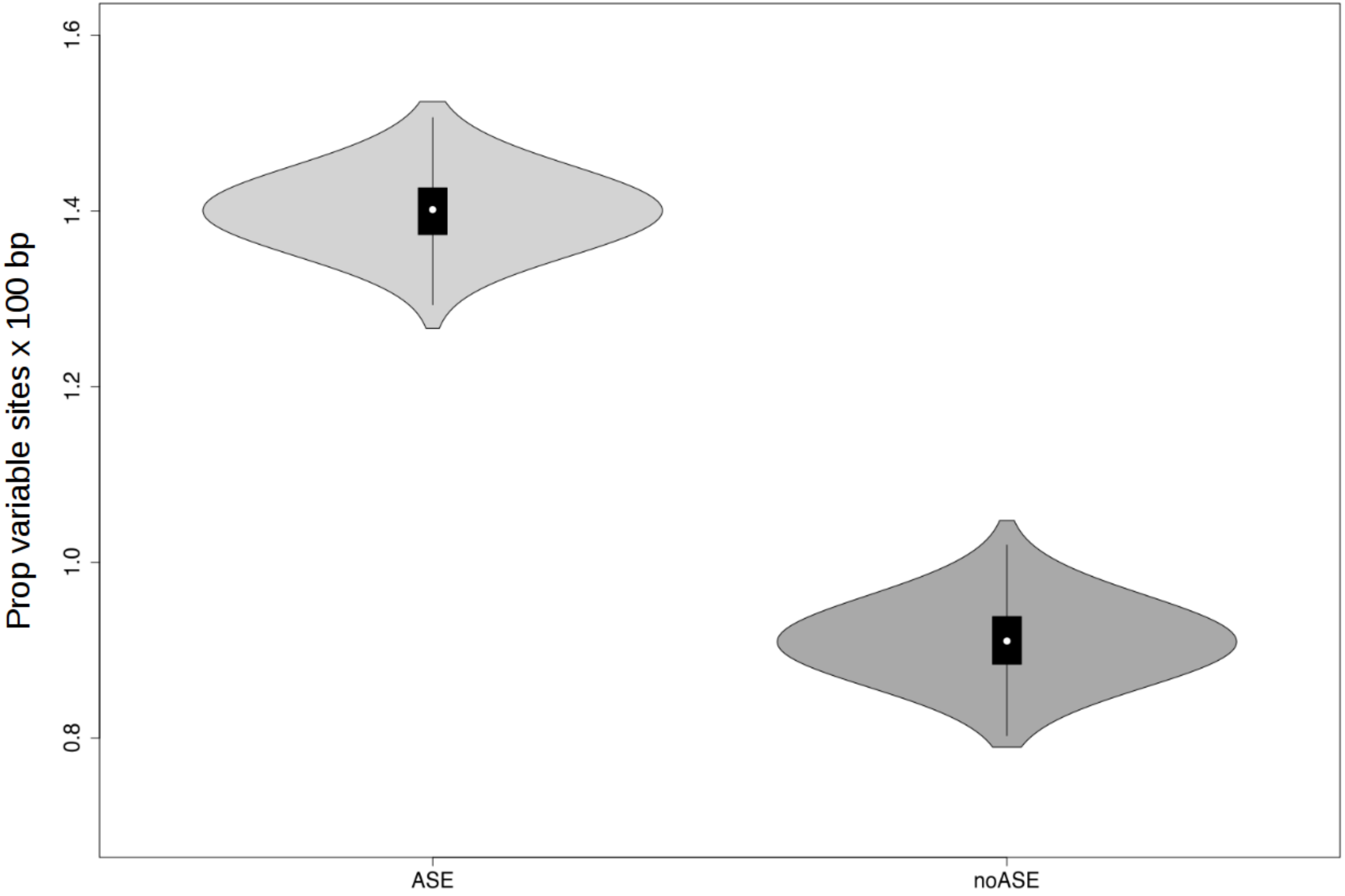

**Figure 8. Violin plots of proportions of polymorphic sites in core promoters.** Rates of polymorphic sites are displayed on the Y axis. These were calculated as the number of polymorphic sites per 100bp in regions encompassing 120 bp from annotated TSSs. Genes showing ASE (light gray) genes not associated with ASE (noASE, dark gray).