A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals

Jieming Chen[1,2], Joel Rozowsky[1,3], Timur R. Galeev[1,3], Arif Harmanci[1,3], Robert Kitchen[1,3], Jason Bedford[1], Alexej Abyzov[1,3,6], Yong Kong[4,5], Lynne Regan[1,2,3], Mark Gerstein*[1,2,3,4]

[1]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA.
[2]Integrated Graduate Program in Physical and Engineering Biology, Yale University, New Haven, CT 06520, USA.
[3]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA.
[4]Department of Computer Science, Yale University, New Haven, CT 06520, USA.
[5]Keck Biotechnology Resource Laboratory, Yale University, New Haven, CT 06511, USA.
[6]Current address: Department of Health Sciences Research, Center for Individualized Medicine, Mayo Clinic, Rochester, MN 55905
*Corresponding author

ABSTRACT

Large-scale sequencing in the 1000-Genomes Project has revealed multitudes of single nucleotide variants. Here, we provide insights into the functional effect of these variants using allele-specific behavior. This can be assessed for an individual by mapping ChIP-seq and RNA-seq reads to a personal genome, and then measuring "allelic imbalances" between the numbers of reads mapped to the paternal and maternal chromosomes. We annotate variants associated with allele-specific binding and expression in 382 individuals by uniformly processing 1,263 functional genomics datasets, developing approaches to reduce the heterogeneity between datasets due to overdispersion and mapping bias. Since many allelic variants are rare, aggregation across multiple individuals is necessary to identify broadly applicable "allelic elements". We also find SNVs for which we can anticipate allelic imbalance from the disruption of a binding motif. Our results serve as an allele-specific annotation for the 1000-Genomes variant catalog and are distributed as an online resource (alleledb.gersteinlab.org).

Deleted: Jason Bedford[1],

Deleted: , Robert Kitchen[1,3], Timur Galeev[1,3]

Deleted: Specifically, we

Deleted: over-dispersion.

Deleted: of the

INTRODUCTION

In recent years, the number of personal genomes has increased dramatically, from single individuals[1,2] to large sequencing projects such as the 1000 Genomes Project[3], UK10K[4] and the Personal Genome Project[5]. These efforts have provided the scientific community with a massive catalog of human genetic variants, most of which are rare.[3] Subsequently, a major challenge is to functionally annotate these variants.

Much of the characterization of variants so far has been focused on those found in the protein-coding regions, but the advent of large-scale functional genomic assays, such as chromatin immunoprecipitation sequencing (ChIP-seq) and RNA sequencing (RNA-seq), has facilitated the annotation of genome-wide variation. This can be accomplished by correlating functional readouts from the assays to genomic variants, particularly in identifying regulatory variants, such as mapping of expression quantitative trait loci (eQTLs)[6–8] and allele-specific[9,10] variants. eQTL mapping assesses the effects of variants on expression profiles across a large population of individuals and is usually used for detection of common regulatory variants. On the other hand, allele-specific approaches assess phenotypic differences directly at heterozygous loci within a single genome. Using each allele in a diploid genome as a perfectly matched control for the other allele, allele-specific variants can be detected even at low population allele frequencies. Therefore, allele-specific approaches are very powerful, in terms of functionally annotating personal genomes, especially for identifying rare cis-regulatory variants on a large scale.

Early high throughput implementations of allele-specific approaches employed microarray technologies, and thus are restricted to a small subset of loci.[11,12] Later studies have used ChIP-seq and RNA-seq experiments for genome-wide measurements of allele-specific variants but have been mostly limited to a single assay with a variety of individuals,[13] or a few individuals with deeply-sequenced and well-annotated genomes.[10,14] For instance, GM12878, a very well-characterized lymphoblastoid cell-line from a Caucasian female, has several RNA-seq datasets and a huge trove of ChIP-seq data for more than 50 transcription factors (TFs) distributed across multiple studies.[15,16] Merging these datasets to create a database is advantageous. The database consolidates a catalog of annotated allele-specific variants in a central repository. Datasets belonging to the same individuals can also be combined to increase statistical power in detection and to have more features for intra- and inter-individual comparisons (such as having more TFs and populations or investigating allele-specific binding and expression coordination).

Deleted: are

Deleted: simply having

Deleted: facilitates

Deleted: across

However, it is not optimal to simply aggregate results from multiple studies, even for the same biological sample. This is because disparate studies might design RNA-seq and ChIP-seq experiments with various goals in mind. Even if allele-specific analyses are conducted, they are often performed with different methods and sets of tools, parameters and variations of the same test (Supplementary Table 1). In addition, each allele-specific analysis is also sensitive to the technical issues associated with variant calling and processing, RNA-seq and ChIP-seq experiments, such as thresholding and read mapping.[17–19] For example, homozygous SNVs incorrectly called as heterozygous will result in reads mapping to one allele (over the other), giving rise to false signals of allelic imbalance. Variants called using shorter reads such as those in RNA-seq datasets can also contain many artefacts. Thus, it is important to have a call set, particularly obtained from whole genome DNA sequencing, such as those from the 1000

Formatted: Font color: Red

Genomes Project. Also, allele-specific SNVs detected in copy number variants have a higher rate of false positives, since copy number changes can easily masquerade as allelic imbalance.

Therefore, the task of merging has to be carried out in a uniform, standardized fashion to yield interpretable results. To this end, we organize and unify datasets from eight different studies into a comprehensive data corpus and repurpose it especially for allele-specific analyses. We first alleviate allelic mapping bias via the use of personal genomes and then the removal of reads that overlap regions exhibiting mapping bias in simulations. We also take into account the overdispersion of each dataset when we harmonize and pool them together and then a ~~second~~ time on the pooled sets, during the detection of allele-specific variants. Overall, we detect more than 7K and 85K single nucleotide variants (SNVs) associated with allele-specific binding (ASB) and expression (ASE) events respectively, over 382 individuals from the 1000 Genomes Project. We are able to present a survey for these allele-specific variants in various general and specific categories of coding and non-coding genomic elements and annotations (e.g. coding regions and enhancers) in a population-aware manner. We identified genomic regions that are enriched or depleted in allelic activity. Finally, using our consolidated data, we investigate the extent of purifying selection in allele-specific SNVs and the inheritance of allele-specific expression and allele-specific binding in two different transcription factors. The variants and annotations are available as an online resource, AlleleDB (http://alleledb.gersteinlab.org/).

RESULTS

AlleleDB Workflow

In general, the AlleleDB workflow uniformly processes two pieces of information from each individual: the DNA sequence, and reads from either the ChIP-seq or RNA-seq experiment to assess SNVs associated with ASB or ASE respectively (Figure 1). **(1)** It starts by first constructing a diploid personal genome for each of the 382 individuals, using DNA variants from the 1000 Genomes Project. **(2)** It then aligns the ChIP-seq or RNA-seq dataset to each of the haploid genomes instead of the human reference genome, and chooses the better uniquely mapped alignment. This reduces reference bias that can potentially result in erroneous read mapping.[14] Because each individual can have multiple ChIP-seq or RNA-seq datasets, the alignment is performed to each personal genome twice. **(2a)** In the first round, the alignment is performed for each of 276 ChIP-seq and 987 RNA-seq datasets to calculate a measure of overdispersion (with respect to an expected binomial distribution), ρ (see 'Methods'). We observe that if there is a greater overdispersion in the empirical allelic ratio (defined as the proportion of reads that map to the reference allele) distribution of a dataset, the binomial test tends to overestimate the number of allele-specific events (Figure 2). There are varying degrees of overdispersion in our datasets, even between biological replicates. In general, RNA-seq datasets are generally more consistent in overdispersion than ChIP-seq datasets. Differing overdispersion in individual datasets poses a challenge later in Step 2b when we pool and merge multiple datasets. In order to harmonize the datasets, we flag and filter datasets that are deemed to be more overdispersed in allelic ratio distributions, leaving us with 186 ChIP-seq and 955 RNA-seq datasets for allele-specific detection (Supplementary Table 2). **(2b)** The second alignment is performed by 'pooling' the 186 ChIP-seq and 955 RNA-seq datasets that have not been filtered in Step 2a. The pooling is performed for each individual and each transcription factor (for ChIP-seq); e.g. CTCF (CCCTC-binding factor) ChIP-seq datasets for NA12878 that

UNC.

Formatted: Font color: Red
Formatted: Font color: Red
Formatted: Font color: Red
Formatted: Font color: Red
Deleted: Briefly, it
Formatted: Font: Bold
Deleted: (1)
Formatted: Font color: Red
Deleted: ; in
Deleted: step
Deleted: , or pool,
Formatted: Font color: Red
Deleted: pooling
Deleted: has

were not filtered were pooled together. An overdispersion parameter is re-calculated for each pooled set. **(3)** The third module filters reads that preferentially map to one allele over the other due to sequence homology (Figure 1), which we term 'ambiguous mapping bias'. This bias occur when reads containing one allele map to multiple locations and are thus removed, not because of worse alignment, but because of ambiguous alignment. For a uniquely mapped read that overlap at least one heterozygous SNV on one parental genome ('original read'), we simulate reads that represent all possible haplotypes of that read, even though we found that most original reads overlap only 1 heterozygous SNV (typically >90%; Supplementary Table 3). We then align the simulated reads to the other parental genome. Original reads with simulated reads that map to multiple locations or do not map back to the same location on the other parental genome are removed. (Figure 1). We subsequently re-align the filtered read pile to the diploid personal genome (see 'Methods'). **(4)** Finally, we obtain allelic counts from the personal genome alignments, and a beta-binomial test is performed using the 'pooled' overdispersion parameter calculated in Step 2b to detect allele-specific SNVs. For ChIP-seq data, the SNVs are further pared down to those within peak regions. We also remove SNVs if they lie in regions predicted to be copy number variants. Please refer to the 'Methods' section for more detailed description.

We build a database, AlleleDB (http://alleledb.gersteinlab.org/), to house the annotations, the allele-specific and accessible SNVs. AlleleDB can be downloaded as flat files or queried and visualized directly as a UCSC track in the UCSC Genome browser[20] as specific genes or genomic locations. This enables cross-referencing of allele-specific variants with other track-based datasets and analyses, and makes it amenable to all functionalities of the UCSC Genome browser. Heterozygous SNVs found in the stipulated query genomic region are color-coded in the displayed track; Figure 4a shows a schematic that illustrates an example of a visualization.

ASB and ASE Inheritance analyses using CEU trio

The CEU trio is a well-studied family and with multiple ChIP-seq studies performed on different TFs. Previous studies have also presented allele-specific inheritance.[10,15,21] Here, after uniformly processing datasets from multiple studies, we are able to analyze and compare the heritability of ASE and ASB across two DNA-binding proteins in a consistent manner (Figure 3; see 'Methods'). For the DNA-binding protein CTCF and PU.1 (also SPI1, or spleen focus forming virus proviral integration proto-oncogene), we observe a high parent-child correlation (Figure 3, Supplementary Table 4), denoting great similarity in allelic directionality (Pearson's correlation, $r \geq 0.77$ in both parent-child plots). We also observe considerable heritability in ASE, but to a lesser degree. In general, the high inheritance of allele-specific SNVs observed in the same allelic direction from parent to child also implies a sequence dependency in allele-specific behavior.

AS variants and enrichment analyses

Using the AlleleDB variants found in the personal genomes of the 2 parents of the trio and 379 unrelated individuals from Phase 1 of the 1000 Genomes Project, we focus on autosomal SNVs and detected 85,742 unique ASE and 7,462 ASB SNVs, representing 16% and 6% of the accessible SNVs respectively (Table 1). 15% of our candidate ASE SNVs and 3% of ASB SNVs are in the coding DNA sequences (CDS); these correspond to log odds ratios of 0.3 (enrichment) and -0.2 (depletion) respectively, when compared to the non-coding regions (Supplementary Figure 1).

Deleted: **(3)** Finally,

Deleted: (see

Deleted:).

Formatted: Font color: Red

Formatted: Font color: Red

Formatted: Font color: Red

Formatted: Font color: Red

Formatted: Font color: Red

Formatted: Font color: Red

Formatted: Font color: Red

Formatted: Font color: Red

Formatted: Font color: Red

Formatted: Font color: Red

Of great interest, is the annotation of these allele-specific SNVs with respect to known genomic elements, both coding and non-coding. We calculate the enrichment of ASB and ASE SNVs in various genomic categories. To do so, we have to define 'accessible' and 'control' SNVs. 'Accessible' SNVs are heterozygous SNVs that have at least the minimum number of reads needed to be statistically detectable for allelic imbalance, which is calculated independently for each dataset (see 'Methods').

We further define the 'control' SNVs as the non-allele-specific subset of accessible SNVs, by excluding ASB or ASE SNVs from each set of accessible SNVs. Thus, the control SNVs are non-allele-specific but are matched in terms of the minimum number of reads to the detected allele-specific SNVs (see 'Methods'). This matching is especially pertinent to our enrichment analyses, since the Fisher's exact test is dependent on the choice of the null expectation (i.e. controls).

To estimate the degree of allele-specificity in both coding and non-coding genomic elements, we calculate the enrichment of allele-specific SNVs by comparing allele-specific SNVs relative to the control SNVs using Fisher's exact tests. The enrichment analyses are performed in two ways: 'expanded' and 'collapsed'. The former counts each occurrence of SNV in a population-aware manner, where each control or allele-specific SNV is counted for each individual at each locus. The latter collapses and counts a control or allele-specific SNV location as a unique SNV as long as it occurs in at least one individual (Figure 4b). Both enrichment analyses are performed in genomic annotations (or categories) with differing granularities, from broad genomic categories to individual binding motifs and genes. Broad genomic categories are grouped based on similar functional context. These include 708 non-coding genomic categories from the ENCODE project[22] (e.g. DNaseI hypersensitivity sites and transcription factor binding motifs) and six gene sets known to be involved in monoallelic expression (MAE)[23,24] (e.g. imprinted genes,[25] olfactory receptor genes[26]).

We further calculate the enrichment of allele-specific SNVs in 19,257 autosomal protein-coding genes from GENCODE[27] in both collapsed and population-aware expanded fashion. The database allows us to visualize allele-specific SNVs across the gene region and over multiple individuals. For example, *SNRPN* and *SNURF* are maternally-imprinted genes, shown to be highly implicated in the Prader-Willi Syndrome, an imprinting disorder.[28] Indeed, they are two of our most highly-ranked allele-specific genes by overall odds ratio (column 'AS.OR' in Supplementary File 2). When *SNURF* is queried in our database, we can see clearly that the allele-specificity is supported not only by evidence from 61 ASE loci across the gene but a number of variants are shown to be also allele-specific over multiple individuals, one variant even up to 169 individuals. The concurrent visualization of ASB and ASE SNVs with respect to genomic elements using the UCSC genome browser is also another advantage of AlleleDB. For example, *ZNF331* (zinc finger protein 331) gene contains a good number of both ASE and ASB loci. It has previously been shown experimentally to be consistently expressed from the paternal allele.[29] Our visualization shows ASB loci from POL2 (RNA polymerase II largest subunit), RPB2 (RNA polymerase II second largest subunit) and MYC (also c-Myc, or v-myc avian myelocytomatosis viral oncogene homolog) of several individuals coinciding near *ZNF331* exons; the former two DNA-binding proteins are components of RNA polymerase II (Figure 4a).

Formatted: Font color: Red

Deleted: [21] (e.g.

Deleted: [22]

Deleted: [23]

Deleted: [24]

Field Code Changed

Deleted: [5]

Deleted: ¶

We further calculate the enrichment of allele-specific SNVs in 19,257 autosomal protein-coding genes from GENCODE[26] in both collapsed and population-aware expanded fashion. The database allows us to visualize allele-specific SNVs across the gene region and over multiple individuals. For example, *SNRPN* and *SNURF* are maternally-imprinted genes, shown to be highly implicated in the Prader-Willi Syndrome, an imprinting disorder.[27] Indeed, they are two of our most highly-ranked allele-specific genes by overall odds ratio (column 'AS.OR' in Supplementary File 2). When *SNURF* is queried in our database, we can see clearly that the allele-specificity is supported not only by evidence from 61 ASE loci across the gene but a number of variants are shown to be also allele-specific over multiple individuals, one variant even up to 169 individuals. The concurrent visualization of ASB and ASE SNVs with respect to genomic elements using the UCSC genome browser is also another advantage of AlleleDB. For example, *ZNF331* (zinc finger protein 331) gene contains a good number of both ASE and ASB loci. It has previously been shown experimentally to be consistently expressed from the paternal allele.[28] Our visualization shows ASB loci from POL2 (RNA polymerase II largest subunit), RPB2 (RNA polymerase II second largest subunit) and MYC (also c-Myc, or v-myc avian myelocytomatosis viral oncogene homolog) of several individuals coinciding near *ZNF331* exons; the former two DNA-binding proteins are components of RNA polymerase II (Figure 4a).¶

Additionally, we extend the enrichment analyses to gene elements, such as introns and promoter regions. Figure 5 (and Supplementary Fig 1) shows the enrichment of allele-specific SNVs in elements closely related to a gene model, namely enhancers, promoters, CDS, introns and untranslated regions (UTR). For SNVs associated with allele-specific binding (ASB), we observe an enrichment in the 5' UTRs. This is in line with an enrichment of ASB SNVs in promoters. Though not significant, this suggests functional roles for these variants found in TF binding motifs or peaks found near transcription start sites to regulate gene expression. We see variable enrichments of ASB SNVs in the peaks of particular TFs such as POL2, SA1 (cohesin subunit) and CTCF in promoter regions, while depletion in others, such as PU.1 (Figure 5, Supplementary File 3). These differences might imply that some TFs are more likely to participate in allele-specific regulation than others. Between the two enrichment analyses, we observe more consistent trends in the odds ratios of ASB SNVs than ASE SNVs. The differences are most likely contributed by the presence of common SNVs that are also behaving consistently (either being allele-specific or non-allele-specific) over multiple individuals.

The population-aware analysis gives additional power to calculate enrichment for very specific genomic annotations, namely specific protein-coding genes, enhancers and transcription factor binding motifs; this is unlike broad genomic categories that span over multiple regions in broad genomic categories. By computing the enrichment analysis in a population-aware fashion, we can also define elements based on evidence supported over multiple individuals. This allows us to quantify allele-specific consistency and enrichment even within smaller and specific protein-coding genes (Supplementary File 2), and enhancers (Supplementary File 7), and differentiate those annotations that are significantly and more consistently enriched to be 'allele-specific', depleted to be 'balanced', or otherwise 'indeterminate'. We provide these lists on the AlleleDB resource (http://alleledb.gersteinlab.org/download/).

Rare variants and purifying selection in AS SNVs

To assess the occurrence of ASB and ASB SNVs in the human population, we consider the population minor allele frequencies (MAF). Table 1 shows the breakdown of the accessible and allele-specific SNVs in six ethnic populations (we combined the results for CHB and JPT) and allele frequencies. Yoruba from Ibadan, Nigeria (YRI) contribute the most to both ASE and ASB variants at each allele frequency category. The number of rare allele-specific SNVs (MAF $\leq$ 5%) is about two folds higher in the YRI than the other European sub-populations of comparable (CEU, FIN) or larger (TSI) population sizes (see 'Methods' for full explanation of population abbreviations). However, the percentage of allele-specific SNVs (in accessible SNVs) remain fairly consistent. In general, rare variants do not form the majority of all the allele-specific variants. For each category of allele frequency, the proportion of allele-specific SNVs detected (with respect to accessible SNVs) is fairly comparable across populations (CEU, FIN, GBR, TSI and YRI), with a slight enrichment of ASB SNVs and slight depletion of ASE SNVs as we go towards lower frequencies.

To examine selective constraints in allele-specific SNVs, we then consider the enrichment of rare variants with MAF $\leq$ 0.5%.[3,30] Figure 6 shows a shift of the allele frequency spectrum towards very low allele frequencies in all allele-specific and non-allele-specific SNVs, peaking at MAF $\leq$ 0.5%. We limit our analyses for ASE SNVs to only those found in CDS regions and ASB SNVs

Formatted: Font color: Red
Formatted: Font color: Red
Formatted: Font color: Red
Formatted: Font color: Red
Formatted: Font color: Red
Formatted: Font color: Red

to only those found within known TF motifs (among the 708 non-coding categories in Supplementary File 1). In general, ASE SNVs are shown to have a greater enrichment of rare variants than ASB SNVs. This is probably due to the background of ASE SNVs being in genes versus ASB SNVs mostly in non-coding regions of the genome. Our results in Figure 6 show a statistically significant lower enrichment of rare variants in ASE SNVs as compared to non-ASE SNVs (Fisher's exact test odds ratio=0.2, p<2.2e-16) but statistically insignificant higher enrichment of rare variants in non-ASB SNVs than ASB SNVs (Fisher's exact test odds ratio=1.4, p=0.08). This observation suggests that ASE variants may be under weaker selection than non-ASE variants.

AS variants in TF binding motifs affecting TF occupancy

A pertinent ASB analysis is to identify ASB SNVs that might cause a TF binding difference. To perform this analysis, we focus on the 328 ASB SNVs found across multiple individuals that reside in the binding motifs of 16 TFs. We consider an allele to be disruptive when it occurs less frequently at a position in the motif. Thus, we compare the difference in occurrence between the reference and the alternate allele of the ASB SNV in the position weight matrix (PWM) of a TF binding motif. For instance, if the alternate allele is disruptive, the reference allele is favored, and the difference in occurrence > 0 (see 'Methods'). We then correlate this with the allelic ratio at the ASB SNV. We expect a TF binding motif that favors the reference allele of an ASB SNV (difference in occurrence > 0) to be associated with more binding to the reference allele (i.e. allelic ratio > 0.5). We find a statistically significant correlation between the difference in occurrence and the allelic ratio for the 328 ASB SNVs (Pearson's correlation = 0.70, p<2.2e-16), showing that there is indeed an overall trend for the favored allele to correspond to increased TF binding. In general, the effects of the SNVs are consistent across individuals in the context of the same motifs. As a resource, we provide the list of ASB SNVs with the frequencies of the occurrence of their reference and alternate alleles found in the various TF motifs and their corresponding allelic ratios (Supplementary File 4).

DISCUSSION

The binomial test is typically used to provide statistical significance for the identification of allele-specific SNVs. However, previous studies have observed a deviation from the binomial distribution in read count distributions in ChIP-seq and RNA-seq datasets, which in turn results in broader allelic ratio distributions, i.e. overdispersed.[6,31–33] The beta-binomial test introduces additional parameters to account for overdispersion. Datasets with low overdispersion give very similar results between binomial and beta-binomial tests (Figure 2A). The binomial test tends to overestimate the number of detected allele-specific SNVs in datasets with higher overdispersion, giving rise to more false positives (Figure 2B). In addition to accounting for the overdispersion in the statistical inference of allele-specific SNVs, we propose the use of the overdispersion parameter, ρ, as a means of quality control for flagging datasets that are very different in the spread of the null allelic ratio distributions. This is because high overdispersion can in fact also serve as a strong indicator for potential issues in the datasets, such as uneven and/or sparse read coverage. Hence, while overdispersion could be a biological consequence of allele-specific behavior, we typically assume that allelic ratios of most loci are balanced. The removal of 'outlier' datasets then facilitates the process of homogenizing and harmonizing the datasets.

Formatted: Font color: Red

Formatted: Font color: Red

Formatted: Font color: Red

Formatted: Font color: Red

Formatted: Font color: Red

Formatted: Font color: Red

Formatted: Font color: Red

Deleted: the

Formatted: Font color: Red

Formatted: Font color: Red

Formatted: Font color: Red

Formatted: Font color: Red

Deleted: [6,30–32] We generally assume that most of the SNVs in autosomes would have more balanced allelic ratios. Hence, while overdispersion could be a biological consequence of allele-specific behavior, high overdispersion in ASE distributions would imply biased autosomal gene expression and might in fact indicate potential issues, e.g. sparse uneven coverage. Since there are multiple datasets for each individual and TF, it would be reasonable to homogenize the separate datasets, so that the resultant pools for each individual and TF can facilitate detection of a more conservative set of allele-specific SNVs for AlleleDB. In addition to accounting for the overdispersion in the statistical inference of allele-specific SNVs, we propose the use of the overdispersion parameter, ρ, as a means to select datasets that are more similar in the spread of the distributions. Datasets with low overdispersion give very similar results between binomial and beta-binomial tests (Figure 2A). The binomial test tends to overestimate the number of detected allele-specific SNVs in datasets with higher overdispersion; it is too relaxed in these cases (Figure 2B). Consequently, we adopt a serial two-step approach of first excluding individual datasets with high overdispersion, and then pooling the datasets (by individual and TF) for allele-specific detection, using the beta-binomial test to account for the degree of overdispersion.

Consequently, we propose the utility of overdispersion as both a means of dataset quality control and allele-specific SNV detection in a beta-binomial test.

Another source of error that we investigated and accounted is allelic mapping bias. This occurs when one allele is preferentially aligned over the other in read alignment, resulting in detection of erroneously imbalanced SNVs. In this study, we have accounted for two types of mapping biases, namely the reference bias, and 'ambiguous mapping bias'.

The reference bias occurs when the read with the reference allele is more favorably mapped, since the read with the alternate allele has already at least one mismatch to begin with. Since reads are typically aligned to the haploid human reference genome in conventional allele-specific analyses, the reference bias has been widely regarded as the main source of allelic mapping bias.[17,19,34] There has been a myriad of strategies developed to alleviate reference bias;[35] we provide some examples in Supplementary Table 1. Alignments to a personal genome has been cited as one of the more rigorous but computationally-intensive approach in reference bias reduction.[14,17,34–36] Here, we have demonstrated the utility of personal genomes in bias reduction for allele-specific SNV detection. Additionally, the personal genome is able to handle various mapping artefacts not easily managed by using only the reference genome. Particularly, with the ability to incorporate larger variants beyond single nucleotide variants (such as indels), the personal genome serves as a more representative genome of the individual, as demonstrated by a much better alignment of unique reads.[14,37]

The second allelic mapping bias stems from loci with sequence homology. We term this 'ambiguous mapping bias', because reads from one allele might align ambiguously to multiple locations, resulting in reads with the other allele being unduly favored (Figure 1).[19,38,35] Several strategies have been implemented in dealing with the ambiguous mapping bias (Supplementary Table 1). To date, the primary approach has been the identification and removal of sites in which >5% of the total number of reads exhibit such bias.[13,34,35,39] In our study, we observe that many detected SNVs remain allele-specific even after removing reads that display such bias, showing that this the site removal strategy can be overly conservative (Supplementary Table 5). Hence, we instead identify and remove reads that give rise to allelic mapping bias, thereby retaining robust allele-specific SNVs. While our manuscript was under revision, we notice that this read removal strategy (instead of sites) has also been employed very recently by van de Geijn *et al.*[38] We also show that ambiguous mapping bias seems to have a greater effect on ChIP-seq than RNA-seq datasets, even after accounting for reference bias by the personal genomes (Supplementary Table 5). Besides allelic differences, ambiguous mapping is also highly dependent on the length of the read, as also shown by Degner *et al.* that the bias decreases with increasing read length.[19] We envision that ambiguous mapping bias will be further alleviated by long read technologies being employed in functional assays.

Despite the implementation of personal genome construction, accounting for ambiguous mapping bias and additional filters to lower the number of false positives, it is important to note that the AS SNVs detected are still not necessarily causal. The resultant allelic difference in gene expression and binding can be due to another undetected causal variant that has a strong linkage disequilibrium with the detected variant or, it could be due to a group of variants that act collectively to give the resultant allelic expression or binding.[40] It could also be a result of other

Deleted: However, even with lower number of false positives, it is important to note that the AS SNVs detected are not necessarily causal. The resultant allelic difference in gene expression and binding can be due to another undetected causal variant that has a strong linkage disequilibrium with the detected variant or, it could be due to a group of variants that act collectively to give the resultant allelic expression or binding.[33] It could also be a result of other epigenetic effects such as genomic imprinting where no variants are causal. Nonetheless, we can still prioritize variants in terms of their potential impact. For example, we provide a more confident set of allele-specific SNVs, since they are found to be in the same allelic direction (reference allele) supported by evidence in at least 3 individuals in AlleleDB (Supplementary File 6). Also, a previous experimental study has shown that there is a greater likelihood of an actual TF binding at their canonical motifs.[34] Hence, we also provide a list of high-impact ASB SNVs that cause a change in the PWMs of the transcription factor binding motifs (Supplementary File 4).¶
¶
So far, allele-specific analyses have usually been more SNV- or gene-centric. However, many diseases have been found to implicate allelic activity in particular genomic regions.[35–37] Our downstream analyses focuses additionally on relating allele-specific activity to known genomic elements and annotations, such as CDS and various non-coding regions. However, such element-centric analyses will not be feasible without a large number of ASE and ASB SNVs. It is important to appreciate that a significant portion of SNVs are rare, thus the abundance of and detection of rare AS variants increase with many genomes. Previous studies mostly focus on a very small number of genomes. Hence, it is difficult to perform AS analyses on rare variants from a single study. Yet, having a large number of rare variants is important, especially in quantifying allelic activity in elements, as this requires aggregating information from multiple SNVs across the a genomic region. Consolidating rare allele-specific SNVs is also helpful in defining SNV sets, which allows us to assign allelic activity scores to genomic regions or multiple variants based on allele-specific activity; this is akin to the idea of burden tests for rare variants in association studies.[38,39] The assignment of allelic activity scores is useful when incorporating into large-scale annotation pipelines.[40]¶

epigenetic effects such as genomic imprinting where no variants are causal.[41] Nonetheless, the computational detection of allele-specific SNVs still allows us to prioritize variants in terms of their potential impact. For example, we provide a more confident set of allele-specific SNVs, since they are found to be in the same allelic direction supported by evidence in at least 3 individuals in AlleleDB (Supplementary File 6). Also, a previous experimental study has shown that there is a greater likelihood of an actual TF binding at their canonical motifs.[42] Hence, we also provide a list of high-impact ASB SNVs that cause a change in the PWMs of the transcription factor binding motifs (Supplementary File 4).

So far, allele-specific analyses have usually been more SNV- or gene-centric. However, many diseases have been found to implicate allelic activity in particular genomic regions.[43–45] Our downstream analyses focuses additionally on relating allele-specific activity to known genomic elements and annotations, such as CDS and various non-coding regions. However, such element-centric analyses will not be feasible without a large number of ASE and ASB SNVs. It is important to appreciate that a significant portion of SNVs are rare, thus the abundance of and detection of rare AS variants increase with many genomes. Previous studies mostly focus on a very small number of genomes. Hence, it is difficult to perform AS analyses on rare variants from a single study. Yet, having a large number of rare variants is important, especially in quantifying allelic activity in elements, as this requires aggregating information from multiple SNVs across the a genomic region. Consolidating rare allele-specific SNVs is also helpful in defining SNV sets, which allows us to assign allelic activity scores to genomic regions or multiple variants based on allele-specific activity; this is akin to the idea of burden tests for rare variants in association studies.[46,47] Such an assignment of allelic activity scores is also useful when incorporating into large-scale annotation pipelines.[48]

We have also adopted two ways to analyze enrichment: an expanded approach that capitalizes on the number of individuals and a collapsed approach that computes enrichment based on unique allele-specific SNVs occurring in at least one individual. An expanded population-aware approach emphasizes the common allele-specific variants found across multiple genomes to determine the allele-specificity of an element. An element is deemed more likely to be allelic if it is supported by more evidence of an allele-specific SNV occurring in multiple individuals. On the other hand, a collapsed approach treats each common and rare variant independently. An element that is deemed more allelic in this case but not in the population-aware enrichment analysis, might mean that there are many more rare variants exhibiting allele-specific behavior. A point to note is that this type of allele-specific elements would not have been picked out with a small number of genomes. Thus, a difference in results from the two analyses of the same element can suggest an interplay between rare and common allele-specific SNVs.

Additionally, we can provide some insights into the coordination of ASB and ASE within that category, by comparing ASB and ASE enrichments within specific genomic regions or broad categories (Figure 5). For example, loci that are associated with monoallelic expression have been shown to be also associated with ASB of various transcription factors, such as imprinted[49,50] and immunoglobulins genes[51]. Also, in Figure 4a, we can visualize, in AlleleDB, specific sub-regions within the *ZNF331* gene where ASB and ASE coordination might occur.

Deleted: on

Deleted: ,

Formatted: Font color: Red

Our current catalog of allele-specific SNVs is detected from lymphoblastoid cell lines (LCLs), which is also the predominant cell-line type in the literature. However, it has already been known that there is considerable variability in regulation of gene expression in different tissues.[52] Data from projects, such as GTEx[39], which has more functional assays and sequencing in other tissues and cell lines can be incorporated to provide a more complete allele-specific analysis. Furthermore, our search for datasets shows a dearth of personal genomes with corresponding ChIP-seq and RNA-seq data in non-European populations. It could be a strong reflection on the lack of large-scale functional genomics assays in specific ethnic groups – a concern echoed previously in population genetics.[53] Since many allele-specific variants have been found to be rare at both the individual and the sub-population level, it is of great interest and importance that more individuals of diverse ancestries be represented.

In conclusion, there is great value in integrating existing data, especially across a large collection of genomes. However, it is essential to harmonize heterogeneous datasets in a uniform fashion. As more diverse and accurate personal genomes with haplotype information[54–56] and their corresponding functional genomics data become available, an approach to detect many allele-specific SNVs for a single personal genome will improve annotations of rare SNVs. AlleleDB is easily scaled to accommodate new individual genomes, tissue and cell types. Additionally, the database allows the visualization of ASB and ASE together conveniently. By building the resource using the individuals and variants from the 1000 Genomes Project, AlleleDB can also serve as an allele-specific annotation of the 1000 Genomes Project variant catalog.

Deleted: utility

Deleted: [47–49] and their corresponding functional genomics data become available, an allele-specific approach to detect many allele-specific SNVs for a single personal genome will increase the number of rare allele-specific SNVs detected.

METHODS

Construction of diploid personal genomes

There is a total of 382 genomes used in this study: 379 unrelated genomes, of low-coverage (average depth of 2.2 to 24.8) from Utah residents in the United States with Northern and Western European ancestry (CEU), Han Chinese from Beijing, China (CHB), Finnish from Finland (FIN), British in England and Scotland (GBR), Japanese from Tokyo, Japan (JPT), Toscani from Italy (TSI), and Yorubans from Ibadan, Nigeria (YRI) and 3 high-coverage genomes from the CEU trio family (average read depth of 30x from Broad Institute's, GATK Best Practices v3; variants are called by UnifiedGenotyper). Each diploid personal genome is constructed from the SNVs and short indels (both autosomal and sex chromosomes) of the corresponding individual found in the 1000 Genomes Project. This is constructed using the tool, *vcf2diploid*.[14] Essentially, each variant (SNV or indel) found in the individual's genome is incorporated into the human reference genome, hg19. Most of the heterozygous variants are phased in the 1000 Genomes Project; those that are not, are randomly phased. As a result, two haploid genomes for each individual are constructed. When this is applied to the family of the CEU trio, for each child's genome, these haploid genomes become the maternal and paternal genomes, since the parental genotypes are known. Subsequently, at a heterozygous locus in the child's genome, if at least one of the parents has a homozygous genotype, the parental allele can be known. However, for each of the genomes of the 379 unrelated individuals and the 2 parents from the CEU trio, the alleles, though phased, are of unknown parental origin.

CNV genotyping is also performed for each genome by CNVnator,[57] which calculates the average read depth within a defined window size, normalized to the genomic average for the

Deleted: ¶
CNV genotyping is also performed for each genome by CNVnator,[50] which calculates the average read depth within a defined window size, normalized to the genomic average for the region of the same length.

region of the same length. For each low coverage genome, a window size of 1000 bp is used, while for the high coverage genomes, a window size of 100 bp is used. SNVs found within genomic regions with a normalized abnormal read depth <0.5 or >1.5 are filtered out, since these would mostly likely give rise to spurious allele-specific detection.

RNA-seq and ChIP-seq datasets

In total, we reprocessed 287 ChIP-seq for 14 individuals and 993 RNA-seq datasets for 382 individuals from eight different studies (Supplementary Table 2).

RNA-seq datasets are obtained from the following: gEUVADIS[13], ENCODE[22], Lalonde *et al.* (2011)[58], Montgomery *et al.* (2010)[59], Pickrell *et al.* (2010)[6], Kilpinen *et al.* (2013)[15] and Kasowski *et al.* (2013)[16].

ChIP-seq datasets are obtained from the following: ENCODE[22], Kilpinen *et al.* (2013)[15], Kasowski *et al.* (2013)[16] and McVicker *et al.* (2013)[60].

Read alignment and estimation of ρ

Reads are aligned against each of the derived haploid genome (maternal/paternal genome for trio) using Bowtie 1.[61] When a read is aligned to the same locus, we only pick the alignment that map better to a haplotype. Otherwise, if a read is tied in alignment to both haplotypes, we keep that read and randomly assign it to either haplotype. No multi-mapping is allowed and only a maximum of 2 mismatches per alignment is permitted. This enables the calculation of the proportion of reads that align to the reference allele, or the allelic ratio, at each heterozygous SNV.

To estimate ρ, we adopt a three-step approach. We first obtain the empirical histogram for the allelic ratios of all heterozygous SNVs with read counts ≥ 6. Next, we calculate the expected null distribution (where there is no allelic imbalance) using the probability density function (pdf) of the beta-binomial distribution using the R package, VGAM[62]:

$$P_{betabin}(X = k|n, a, b) = \binom{n}{k} \frac{B(k + a, n - k + b)}{B(a, b)}$$

where n represents the total number of reads at a particular locus, B(x,y) represents the beta function with variables x and y, a and b represent the shape parameters of the beta distribution. For computational efficiency, if $n \geq 1000$, we set it to a maximum of 1000, but retain the allelic ratio at the SNV. The VGAM beta-binomial routines require the input of the overdispersion parameter, ρ, and probability of success (also the mean of the beta distribution), which we fix at p=0.5 since the null hypothesis assumes no allelic imbalance. We then obtain the expected beta-binomial distributions for ρ=0 to ρ=1 with increment of 0.1, and choose ρ that minimizes the least sum of squared errors (LSSE) between the empirical and the expected distributions. Lastly, to further refine our estimate, we iterate a bisection method to arrive at a LSSE (R pseudo-code available in Supplementary File 5).

After removing 11 ChIP-seq and 6 RNA-seq datasets that have insufficient read alignments, we calculate ρ for each 276 ChIP-seq and 987 RNA-seq individual datasets. For RNA-seq datasets,

Formatted: Font color: Red

Deleted: [21]

Deleted: [51], Montgomery *et al.*

Deleted: [52], Pickrell *et al.*

Deleted: [21], Kilpinen *et al.*

Deleted: [53].

Deleted: [54]

Deleted: discard it.

Deleted: [55]:

Formatted: Font color: Red

we removed 32 datasets with $\rho \geq 0.125$, which is one standard deviation higher than the mean ρ amongst the RNA-seq datasets. For ChIP-seq datasets, because many of the datasets have considerable ρ, we use a less stringent arbitrary threshold of $\rho \geq 0.3$ to remove 90 ChIP-seq datasets. Using the resultant 186 ChIP-seq and 955 RNA-seq datasets, we pool datasets by TF and individual for ChIP-seq and by individual for RNA-seq and re-calculate ρ for each pooled dataset. This final ρ is used in the beta-binomial test for allele-specific SNV detection.

Accounting for ambiguous mapping bias

To account for ambiguous mapping bias, **(1)** we first align the reads to each of the two parental haplotypes of the diploid personal genome of each individual (and each TF for ChIP-seq). **(2)** For each haplotype, we retain only those reads that uniquely mapped to regions with heterozygous SNVs. For a uniquely mapped read ('original read') that overlap at least one heterozygous SNV on one parental genome, we simulate reads that represent all possible haplotypes of that read. For example, for reads that overlap a single heterozygous SNV, we simulate the same reads but with a single allele change at the heterozygous SNV position. If the read overlaps multiple heterozygous SNVs, reads with all possible haplotypes are simulated. Due to computational complexity and higher probability of harboring sequencing errors, we remove reads that overlap >5 heterozygous SNVs. **(3)** We then map these simulated reads to the other parental genome. **(4)** Finally, we identify the original reads (red read in Figure 1) which give rise to the simulated reads (blue read in Figure 1) that align to multiple loci in the other haplotype. Subsequently, the original reads are filtered from the read pool before allele-specific SNV detection with the beta-binomial test. We also exclude original reads in which the simulated reads do not map back to the same location and reads in which the neither alleles of the overlapping SNVs matches the nucleotide on the corresponding read; the latter suggests sequencing errors.

Allele-specific SNV detection

Allele-specific SNV detection is performed on the pooled datasets, as mentioned above. Here, a beta-binomial p-value is derived based on the VGAM R package as described in the previous section. Similarly for computational efficiency, if $n \geq 1000$, we set it to a maximum of 1000, but retain the allelic ratio at the SNV. To correct for multiple hypothesis testing, FDR is calculated. Since statistical inference of allele-specificity of a locus is dependent on the number of reads of the ChIP-seq or RNA-seq dataset, this is performed using an explicit computational simulation.[14] Briefly, for each iteration of the simulation, a mapped read is randomly assigned to either allele at each heterozygous SNV and a beta-binomial test is performed using the estimated ρ. At a given p-value threshold, the FDR can be computed as the ratio of the number of false positives (from the simulation) and the number of observed empirical positives. An FDR cutoff of 5% is used. Furthermore, we allow only significant allele-specific SNVs to have a minimum of 6 reads.

For ChIP-seq data, allele-specific SNVs have to be also within peaks. Peak regions are determined by first performing PeakSeq[63] for each of the 14 personal haploid genomes with ChIP-seq data. Only a single read per strand per position is kept and duplicates removed. The fragment length is set to 200 bps. Peak calling is performed with default parameters and the final peak set for each transcription factor is identified at an FDR of 5%. Finally, the coordinates of the peaks (based on the respective personal haploid genomes) are mapped to the reference

Formatted: Font: Bold

Deleted: performs

Deleted: 10% is used for ChIP-seq data and 5% for RNA-seq data, since the latter is typically of deeper coverage.

Deleted: [56] for each of the 14 personal haploid genomes with ChIP-seq data. Only a single read per strand per position is kept and duplicates removed. The fragment length is set to 200 bps. Peak calling is performed with default parameters and the final peak set for each transcription factor is identified at a false discovery rate of 5%. Finally, the coordinates of the peaks (based on the respective personal haploid genomes) are mapped to the reference genome and then finally being merged between the haploid genomes.

genome and then finally being merged between the two haploid genomes. We also make the uniformly-processed peaks available as a resource on the AlleleDB website.

Allele-specific detection has minimal bias towards sites with lower read depth (Supplementary Figures 2 and 3) and is highly reproducible when we compare between replicates (Supplementary Figures 4 and 5). The detection for all TFs and gene expression of 382 individuals took about 600 days in CPU time (1.6 years), but the pipeline is highly parallelizable, thereby streamlining the process.

AlleleDB

The final data and results are organized into a resource, AlleleDB (http://alleledb.gersteinlab.org/), which conveniently interfaces with the UCSC genome browser for query and visualization. Since many in the scientific community are familiar with the genome browser, we hope that this would increase the accessibility and usability of AlleleDB. The query results are also available for download in BED format, which is compatible with other tools, such as the Integrated Genome Viewer[64]. More in-depth analyses can be performed by downloading the full set of allele-specific results. For ASB, the output will be delineated by the sample ID and the associated TFs; for ASE, the output will be categorized by individual and the associated gene. We also provide the raw counts for each accessible SNV and indicate if it is identified as an allele-specific SNV. AlleleDB also serves as an annotation of allele-specific regulation of the 1000 Genomes Project SNV catalog. All supplementary files and additional auxiliary materials can also be downloaded on AlleleDB via http://alleledb.gersteinlab.org/download/

Allele-specific inheritance analyses

The conventional measure of 'heritability' allows the estimation of (additive) genetic contribution to a certain trait. The population genetics definition of 'heritability' in a parent-offspring setting is described by the slope, β, of a regression ($Y=\beta X + \alpha$), with the dependent variable being the child's trait value (Y) and the independent variable (X) being the average trait values of the father and the mother ('midparent').[65] This is a population-based measure typically performed on a large set of trios for a particular trait (e.g. height) and β is not necessarily bound between 0 and 1.

Given that we have only a single trio, we adapt the typical definition of 'heritability' to quantify allele-specific inheritance for each TF. For each TF and parent-child comparison, we consider ASB SNVs from two scenarios: (1) when an allele-specific SNV is heterozygous in all three individuals but common to the two individuals being compared, and (2) when an allele-specific SNV is heterozygous in two individuals and homozygous (reference or alternate) in the third. We define the allelic ratio as the 'trait', which is a continuous value and computed as the proportion of reads that align to the reference allele with respect to the total number of reads mapped to either allele of a particular site. We perform the analyses separately for father-child and mother-child pair to maximize statistics, since a midparent calculation will require that a SNV is allele-specific in all three individuals (Scenario 1).

Given that Pearson's correlation coefficient, r, always gives a value between 0 and 1, we use r instead of β, as our measure of 'heritability'. We also compute and include β values in

Formatted: Font color: Red

Deleted: 1

Deleted: 2

Formatted: Font color: Red

Deleted: 4

Formatted: Font color: Red

Deleted: 3

Formatted: Font color: Red

Deleted: [57].

Deleted: [58] This is a population-based measure typically performed on a large set of trios for a particular trait (e.g.

Supplementary Table 4. The parent-parent comparison is provided as a source of comparison for two unrelated individuals with shared ancestry. For parent-parent β, the maternal allelic ratio is chosen arbitrarily to be the independent variable.

Genomic annotations

Categories of gene elements from Figure 5 and Supplementary Figure 1, such as promoters, CDS regions and UTRs, and 19,257 autosomal protein-coding gene annotations (HGNC symbols) are obtained from GENCODE version 17.[27] Promoter regions are set as 2.5kbp upstream of all transcripts annotated by GENCODE.

Gene annotations also include 2.5kbp upstream of the start of gene. 708 categories of non-coding annotations are obtained from ENCODE Integrative release,[22] which includes broad categories such as TF binding sites and annotations such as distal binding sites of particular TFs, e.g. ZNF274. The details of TF family classification is first described in Vaquerizas *et* al.[66] and then also in Gerstein *et* al.[67] Note that these TF binding sites are separate from those sites in promoter regions in Figure 5 and Supplementary Figure 1, which are based on the 44 TFs and peaks from the ChIP-seq experiments used in our pipeline.

The olfactory receptor gene list is from the HORDE database[26]; immunoglobulin, T cell receptor and MHC gene lists are from IMGT database[68]. Imprinted genes are merged from the Catalog of Parent-of-origin Effects (http://igc.otago.ac.nz/home.html),[69] the GeneImprint website (http://www.geneimprint.com/) and also Lo *et al.*[12] We performed enrichment analyses on a number of enhancer lists, which are derived using the ChromHMM and Segway algorithms[70,71], and data from distal regulatory modules from Yip *et al.* (2012)[72]. The result for the enhancers in Figure 5 is based on the union of these lists. The lists can be found at http://info.gersteinlab.org/Encode-enhancers. An additional enhancer list for experimentally validated enhancers is obtained from VISTA enhancer browser database[73] (http://enhancer.lbl.gov/). Housekeeping gene list is obtained from Eisenberg and Levanon (2013) (http://www.tau.ac.il/~elieis/HKG/)[74].

Enrichment analyses

Enrichment analyses were performed in two ways: 'collapsed' and 'expanded' (Figure 4b). In both cases, we aggregate ASB and ASE SNVs within a specific genomic element, such as a gene or an enhancer. We then use the Fisher's exact test to calculate the odds ratio and the hypergeometric p value, to test for the enrichment of allele-specific SNVs compared to 'control' SNVs, which are non-allele-specific 'accessible' SNVs.

We define the set of accessible SNVs as all heterozygous SNVs that exceed the minimum number of reads required in order for SNVs to be significantly detectable by the beta-binomial test for each dataset; this includes both allele-specific and non-allele-specific SNVs. This is an additional, more stringent criterion imposed beyond the minimum threshold of 6 reads. Given a fixed FDR cutoff, for a larger dataset, the beta-binomial p-value threshold is typically lower, making the minimum number of reads (N) that will produce the corresponding p-value, larger. This alleviates a bias in the enrichment test for including SNVs that do not have sufficient reads in the first place. Considering an extreme allelic imbalance case where all the reads are found on one allele (all successes or all failures, i.e. allelic ratio is 0 or 1), this minimum N can be

Formatted: Font color: Red

Deleted: 2

Deleted:)[67].¶

obtained from a table of expected two-tailed beta-binomial probability density function, such that accessible SNVs are all SNVs with a minimum number of reads, $n \geq \max(6,N)$. The minimum number of reads thus varies with the pooled size (coverage) of the ChIP-seq or RNA-seq dataset. Thus, the accessible SNVs are dataset-specific; they are determined for each pooled ChIP-seq (grouped by individual and TF, not by study) or RNA-seq dataset (grouped by individual). By considering only the cases with the largest effect size, we underestimate the number of accessible SNVs and this provides a conservative approximation of the statistical significance of the enrichment (or depletion). 'Control' SNVs are subsequently derived from accessible SNVs that are non-allele-specific, i.e. they are the set of accessible SNVs that has excluded the respective ASB or ASE SNVs for each dataset.

In the 'collapsed' enrichment analysis, each control or allele-specific SNV is counted once uniquely, as long as it occurs in at least one individual in AlleleDB. The 'expanded' analysis is performed in a population-aware manner, where each control or allele-specific SNV is counted once for each occurrence in an individual. P-values are Bonferroni-corrected and considered significant if ≤ 0.05.

'Allele-specific' and 'balanced' autosomal protein-coding genes, enhancers and transcription factor binding motifs are defined based on statistically significant (Bonferroni-corrected p value ≤ 0.05) enrichments (odds ratio ≥ 1.5) or depletions (odds ratio < 1.5) respectively, as obtained from the 'expanded' enrichment analysis; the rest of the elements with non-significant odds ratios are considered 'indeterminate'.

Analysis of ASB SNVs found in TF motifs

We obtain a list of all TF motifs and their corresponding position weight matrices (PWMs) from Kheradpour and Kellis[75] (http://compbio.mit.edu/encode-motifs/), using the 2013 version. This set of motifs and PWMs is derived from the ENCODE project and include motifs from TRANSFAC and JASPAR. We then take two approaches to find the effects of ASB SNVs. (1) For all ASB SNV positions in the motifs detected by Kheradpour and Kellis, we obtain the occurrence (frequency) of their reference and alternate allele in the respective PWMs. This first approach is only able to find motif-breaking events that disrupt existing motifs in the reference genome. The PWMs of motifs are defined based on the ENCODE project. (2) Our second approach attempts to include both motif-breaking and motif-gaining events caused by ASB SNVs in AlleleDB. Based on each PWM, we further scan a 59-bp window around the ASB SNV (± 29 bp of the SNV) separately for both the reference and alternate alleles for potential motifs. For each candidate motif, we compute the sequence score using the tool TFM-Pvalue[76], where sequence score is defined by summing up the log likelihoods of each position of the PWM. A motif is identified when the P value on its sequence score $\leq$ 1e-6.

We then merge the results from both approaches. The allelic ratio is defined as before, i.e. the ratio of number of reference reads to the total number of reads, thus when the ratio > 0.5, there are more reads that align to the reference allele, signifying more binding to the motif with the reference allele. We compute the difference in occurrence between the reference and alternate allele (occurrence of reference allele minus occurrence of alternate allele) based on the PWM of the motif, thus a positive value indicates that the reference allele is favored (i.e. less disruptive). The Pearson's correlation is calculated between this difference and the allelic ratio.

Deleted: [68]

ENDNOTES

Acknowledgements

The authors would like to thank Drs. Robert Bjornson and Yao Fu for technical help. We acknowledge support from the NIH and from the AL Williams Professorship funds. This work was supported in part by Yale University Faculty of Arts and Sciences High Performance Computing Center.

Authors' Contributions

JC, JR and MG conceived and designed the resource. JC and JB built the AlleleDB website. JC, JR, TG, AH, RK, YK, AA, LR and MG analyzed and interpreted the data. JC wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

None of the authors have any competing interests.

REFERENCES

1. Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452,** 872–6 (2008).

2. Lupski, J. R. *et al.* Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N. Engl. J. Med.* **362,** 1181–91 (2010).

3. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).

4. Muddyman, D., Smee, C., Griffin, H. & Kaye, J. Implementing a successful data-management framework: the UK10K managed access model. *Genome Med.* **5,** 100 (2013).

5. Church, G. M. The personal genome project. *Mol. Syst. Biol.* **1,** 2005.0030 (2005).

6. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464,** 768–72 (2010).

7. Majewski, J. & Pastinen, T. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends Genet.* **27,** 72–9 (2011).

Deleted: , TG

Formatted: Font: Times New Roman, 12 pt

Formatted: Normal, Don't adjust right indent when grid is defined, No widow/orphan control, Don't adjust space between Latin and Asian text, Don't adjust space between Asian text and numbers

Formatted: Font: Times New Roman, 12 pt

Formatted: Normal, Don't adjust right indent when grid is defined, No widow/orphan control, Don't adjust space between Latin and Asian text, Don't adjust space between Asian text and numbers

Formatted: Font: Times New Roman, 12 pt

Formatted: Normal, Don't adjust right indent when grid is defined, No widow/orphan control, Don't adjust space between Latin and Asian text, Don't adjust space between Asian text and numbers

Formatted: Font: Times New Roman, 12 pt

Formatted: Normal, Don't adjust right indent when grid is defined, No widow/orphan control, Don't adjust space between Latin and Asian text, Don't adjust space between Asian text and numbers

Formatted: Font: Times New Roman, 12 pt

Formatted: Normal, Don't adjust right indent when grid is defined, No widow/orphan control, Don't adjust space between Latin and Asian text, Don't adjust space between Asian text and numbers

Formatted: Font: Times New Roman, 12 pt

Formatted: Normal, Don't adjust right indent when grid is defined, No widow/orphan control, Don't adjust space between Latin and Asian text, Don't adjust space between Asian text and numbers

Formatted: Font: Times New Roman, 12 pt

Formatted: Normal, Don't adjust right indent when grid is defined, No widow/orphan control, Don't adjust space between Latin and Asian text, Don't adjust space between Asian text and numbers

8. Montgomery, S. B., Lappalainen, T., Gutierrez-Arcelus, M. & Dermitzakis, E. T. Rare and common regulatory variation in population-scale sequenced human genomes. *PLoS Genet.* **7,** e1002144 (2011).

9. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489,** 101–8 (2012).

10. McDaniell, R. *et al.* Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* **328,** 235–9 (2010).

11. Yan, H., Yuan, W., Velculescu, V. E., Vogelstein, B. & Kinzler, K. W. Allelic variation in human gene expression. *Science* **297,** 1143 (2002).

12. Lo, H. S. *et al.* Allelic variation in gene expression is common in the human genome. *Genome Res.* **13,** 1855–1862 (2003).

13. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501,** 506–11 (2013).

14. Rozowsky, J. *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol.* **7,** 522 (2011).

15. Kilpinen, H. *et al.* Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* **342,** 744–7 (2013).

16. Kasowski, M. *et al.* Extensive variation in chromatin states across humans. *Science* **342,** 750–2 (2013).

17. Stevenson, K. R., Coolon, J. D. & Wittkopp, P. J. Sources of bias in measures of allele-specific expression derived from RNA-sequence data aligned to a single reference genome. *BMC Genomics* **14,** 536 (2013).

18. Hansen, K. D., Brenner, S. E. & Dudoit, S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.* **38,** e131 (2010).

19. Degner, J. F. *et al.* Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25,** 3207–12 (2009).

20. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12,** 996–1006

(2002).

21. Li, X. *et al*. Transcriptome sequencing of a large human family identifies the impact of rare noncoding variants. *Am. J. Hum. Genet.* **95,** 245–56 (2014).

22. Bernstein, B. E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

23. Goldmit, M. & Bergman, Y. Monoallelic gene expression: a repertoire of recurrent themes. *Immunol. Rev.* **200,** 197–214 (2004).

24. Zakharova, I. S., Shevchenko, A. I. & Zakian, S. M. Monoallelic gene expression in mammals. *Chromosoma* **118,** 279–90 (2009).

25. Morison, I. M., Paton, C. J. & Cleverley, S. D. The imprinted gene and parent-of-origin effect database. *Nucleic Acids Res.* **29,** 275–6 (2001).

26. Olender, T., Nativ, N. & Lancet, D. HORDE: comprehensive resource for olfactory receptor genomics. *Methods Mol. Biol.* **1003,** 23–38 (2013).

27. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22,** 1760–74 (2012).

28. Horsthemke, B. & Buiting, K. Imprinting defects on human chromosome 15. *Cytogenet. Genome Res*. **113,** 292–9 (2006).

29. Pollard, K. S. *et al.* A genome-wide approach to identifying novel-imprinted genes. *Hum. Genet.* **122,** 625–634 (2008).

30. Khurana, E. *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342,** 1235587 (2013).

31. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).

32. Skelly, D. A., Johansson, M., Madeoy, J., Wakefield, J. & Akey, J. M. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome Res.* **21,** 1728–1737 (2011).

Deleted: 21

Formatted: Font: Times New Roman, 12 pt

Formatted: Normal, Don't adjust right indent when grid is defined, No widow/orphan control, Don't adjust space between Latin and Asian text, Don't adjust space between Asian text and numbers

Deleted: 22

Formatted: Normal, Don't adjust right indent when grid is defined, No widow/orphan control, Don't adjust space between Latin and Asian text, Don't adjust space between Asian text and numbers

Formatted: Font: Times New Roman, 12 pt

Deleted: 23

Formatted ...

Formatted: Font: Times New Roman, 12 pt

Deleted: 24

Formatted ...

Formatted: Font: Times New Roman, 12 pt

Deleted: 25

Formatted ...

Formatted: Font: Times New Roman, 12 pt

Deleted: 26

Formatted ...

Formatted: Font: Times New Roman, 12 pt

Deleted: 27

Formatted ...

Formatted: Font: Times New Roman, 12 pt

Deleted: 28

Formatted ...

Formatted: Font: Times New Roman, 12 pt

Deleted: 29

Formatted ...

Formatted: Font: Times New Roman, 12 pt

Deleted: 30

Formatted ...

Formatted: Font: Times New Roman, 12 pt

Deleted: 31

Formatted ...

Formatted: Font: Times New Roman, 12 pt

33. Zhang, S. *et al.* Genome-wide identification of allele-specific effects on gene expression for single and multiple individuals. *Gene* **533,** 366–373 (2014).

34. Panousis, N. I., Gutierrez-Arcelus, M., Dermitzakis, E. T. & Lappalainen, T. Allelic mapping bias in RNA-sequencing is not a major confounder in eQTL studies. *Genome Biol.* **15,** 467 (2014).

35. Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E. & Lappalainen, T. Tools and best practices for data processing in allelic expression analysis. *Genome Biol.* **16,** 195 (2015).

36. Satya, R. V., Zavaljevski, N. & Reifman, J. A new strategy to reduce allelic bias in RNA-Seq readmapping. *Nucleic Acids Res.* **40,** e127 (2012).

37. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526,** 75–81 (2015).

38. van de Geijn, B., McVicker, G., Gilad, Y. & Pritchard, J. K. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* **12,** 1061–3 (2015).

39. GTEx Consortium *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science (80-.).* **348,** 648–660 (2015).

40. Tao, H., Cox, D. R. & Frazer, K. A. Allele-specific KRT1 expression is a complex trait. *PLoS Genet.* **2,** 0848–0858 (2006).

41. Baran, Y. *et al.* The landscape of genomic imprinting across diverse adult human tissues. *Genome Res.* **25,** 927–36 (2015).

42. Cusanovich, D. A., Pavlovic, B., Pritchard, J. K. & Gilad, Y. The Functional Consequences of Variation in Transcription Factor Binding. *PLoS Genet.* **10,** (2014).

43. Amin, A. S. *et al.* Variants in the 3' untranslated region of the KCNQ1-encoded Kv7.1 potassium channel modify disease severity in patients with type 1 long QT syndrome in an allele-specific manner. *Eur. Heart J.* **33,** 714–23 (2012).

Deleted: 32

Formatted: Normal, Don't adjust right indent when grid is defined, No widow/orphan control, Don't adjust space between Latin and Asian text, Don't adjust space between Asian text and numbers

Formatted: Font: Times New Roman, 12 pt

Formatted: Normal, Don't adjust right indent when grid is defined, No widow/orphan control, Don't adjust space between Latin and Asian text, Don't adjust space between Asian text and numbers

Moved (insertion) [1]

Formatted: Font: Times New Roman, 12 pt

Deleted: 33

Formatted: Normal, Don't adjust right indent when grid is defined, No widow/orphan control, Don't adjust space between Latin and Asian text, Don't adjust space between Asian text and numbers

Formatted: Font: Times New Roman, 12 pt

Deleted: 34

Formatted: Normal, Don't adjust right indent when grid is defined, No widow/orphan control, Don't adjust space between Latin and Asian text, Don't adjust space between Asian text and numbers

Formatted: Font: Times New Roman, 12 pt

Deleted: 35

Formatted: Normal, Don't adjust right indent when grid is defined, No widow/orphan control, Don't adjust space between Latin and Asian text, Don't adjust space between Asian text and numbers

Formatted: Font: Times New Roman, 12 pt

Deleted: 36

44. Anjos, S. M., Shao, W., Marchand, L. & Polychronakos, C. Allelic effects on gene regulation at the autoimmunity-predisposing CTLA4 locus: a re-evaluation of the 3' +6230G>A polymorphism. *Genes Immun.* **6,** 305–11 (2005).

45. Valle, L. *et al.* Germline allele-specific expression of TGFBR1 confers an increased risk of colorectal cancer. *Science* **321,** 1361–5 (2008).

46. Price, A. L. *et al.* Pooled Association Tests for Rare Variants in Exon-Resequencing Studies. *Am. J. Hum. Genet.* **86,** 832–838 (2010).

47. Han, F. & Pan, W. A data-adaptive sum test for disease association with multiple common or rare variants. *Hum. Hered.* **70,** 42–54 (2010).

48. Fu, Y. *et al.* FunSeq2 : a framework for prioritizing noncoding regulatory variants in cancer. (2014). doi:10.1186/s13059-014-0480-5

49. Boodhoo, A. *et al.* A promoter polymorphism in the central MHC gene, IKBL, influences the binding of transcription factors USF1 and E47 on disease-associated haplotypes. *Gene Expr.* **12,** 1–11 (2004).

50. Kim, J. Do *et al.* Identification of clustered YY1 binding sites in imprinting control regions. *Genome Res.* **16,** 901–911 (2006).

51. Chaumeil, J. & Skok, J. A. The role of CTCF in regulating V(D)J recombination. *Current Opinion in Immunology* **24,** 153–159 (2012).

52. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45,** 580–5 (2013).

53. Bustamante, C. D., Burchard, E. G. & De la Vega, F. M. Genomics for the world. *Nature* **475,** 163–5 (2011).

54. Kitzman, J. O. *et al.* Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat. Biotechnol.* **29,** 59–63 (2011).

55. Peters, B. A. *et al.* Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* **487,** 190–5 (2012).

56. Fan, H. C., Wang, J., Potanina, A. & Quake, S. R. Whole-genome molecular haplotyping

Formatted ...
Formatted ...
Deleted: 37
Formatted ...
Formatted ...
Deleted: 38
Formatted ...
Formatted ...
Deleted: 39
Formatted ...
Formatted ...
Deleted: 40
Formatted ...
Formatted ...
Deleted: 41
Formatted ...
Formatted ...
Deleted: 42
Formatted ...
Formatted ...
Deleted: 43
Formatted ...
Formatted ...
Deleted: 44
Formatted ...
Formatted ...
Deleted: 45.
Moved up [1]: . GTEx Consortium *et al.* The Genotype-
Formatted ...
Deleted: 46
Formatted ...
Formatted ...
Deleted: 47
Formatted ...
Formatted ...
Deleted: 48
Formatted ...
Formatted ...
Deleted: 49
Formatted ...
Formatted ...

of single cells. *Nat. Biotechnol.* **29,** 51–7 (2011).

57. Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **21,** 974–84 (2011).

58. Lalonde, E. *et al.* RNA sequencing reveals the role of splicing polymorphisms in regulating human gene expression. *Genome Res.* **21,** 545–54 (2011).

59. Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464,** 773–7 (2010).

60. McVicker, G. *et al.* Identification of genetic variants that affect histone modifications in human cells. *Science* **342,** 747–9 (2013).

61. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).

62. Yee, T. VGAM: Vector Generalized Linear and Additive Models. (2014). at <http://cran.r-project.org/package=VGAM>

63. Rozowsky, J. *et al.* PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.* **27,** 66–75 (2009).

64. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29,** 24–6 (2011).

65. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era--concepts and misconceptions. *Nat. Rev. Genet.* **9,** 255–66 (2008).

66. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* **10,** 252–263 (2009).

67. Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489,** 91–100 (2012).

68. Lefranc, M.-P. *et al.* IMGT-Choreography for immunogenetics and immunoinformatics. *In Silico Biol.* **5,** 45–60 (2005).

Deleted: 50
Formatted ...
Formatted: Font: Times New Roman, 12 pt
Deleted: 51
Formatted ...
Formatted: Font: Times New Roman, 12 pt
Deleted: 52
Formatted ...
Formatted: Font: Times New Roman, 12 pt
Deleted: 53
Formatted ...
Formatted: Font: Times New Roman, 12 pt
Deleted: 54
Formatted ...
Formatted: Font: Times New Roman, 12 pt
Deleted: 55
Formatted ...
Formatted: Font: Times New Roman, 12 pt
Deleted: 56
Formatted ...
Formatted: Font: Times New Roman, 12 pt
Deleted: 57
Formatted ...
Formatted: Font: Times New Roman, 12 pt
Deleted: 58
Formatted ...
Formatted: Font: Times New Roman, 12 pt
Deleted: 59
Formatted ...
Formatted: Font: Times New Roman, 12 pt
Deleted: 60
Formatted ...
Formatted: Font: Times New Roman, 12 pt
Deleted: 61
Formatted ...
Formatted: Font: Times New Roman, 12 pt

69. Morison, I. M., Ramsay, J. P. & Spencer, H. G. A census of mammalian imprinting. *Trends Genet.* **21,** 457–65 (2005).

70. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9,** 215–6 (2012).

71. Hoffman, M. M. *et al.* Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.* **41,** 827–41 (2013).

72. Yip, K. Y. *et al.* Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol.* **13,** R48 (2012).

73. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res.* **35,** D88–92 (2007).

74. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends Genet.* **29,** 569–74 (2013).

75. Kheradpour, P. & Kellis, M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.* **42,** 2976–87 (2014).

76. Touzet, H. & Varré, J.-S. Efficient and accurate P-value computation for Position Weight Matrices. *Algorithms Mol. Biol.* **2,** 15 (2007).

FIGURE LEGENDS

Figure 1. Workflow for uniform processing of data from 382 individuals and construction of AlleleDB. For each of the 382 individuals, (1) a diploid personal genome is first constructed using the variants from the 1000 Genomes Project. Next, reads from individual (2a) and pooled (2b) ChIP-seq or RNA-seq datasets are mapped onto each of the haploid genome of the diploid genome. In (2a), overdispersion (OD) is measured for each dataset and used to segregate highly overdispersed datasets. (2b) The resultant datasets are pooled and the overdispersion parameter is estimated based on the pooled datasets. (3) Reads that give rise to simulated ambiguous-mapping reads are removed. (4) From the filtered read pile, the numbers of reads that map to either allele is being compared to determine if a heterozygous SNV is allele-specific. A statistical significance is computed (after multiple hypothesis test correction) based on the beta-binomial test using the 'pooled' overdispersion parameter in Step 2b to account for overdispersion. All the candidate allele-specific variants are then deposited in AlleleDB database. Additional

Deleted: 62
Formatted: Normal, Don't adjust right indent when grid is defined, No widow/orphan control, Don't adjust space between Latin and Asian text, Don't adjust space between Asian text and numbers
Formatted: Font: Times New Roman, 12 pt
Deleted: 63
Formatted: Normal, Don't adjust right indent when grid is defined, No widow/orphan control, Don't adjust space between Latin and Asian text, Don't adjust space between Asian text and numbers
Formatted: Font: Times New Roman, 12 pt
Deleted: 64
Formatted: Normal, Don't adjust right indent when grid is defined, No widow/orphan control, Don't adjust space between Latin and Asian text, Don't adjust space between Asian text and numbers
Formatted: Font: Times New Roman, 12 pt
Deleted: 65
Formatted: Normal, Don't adjust right indent when grid is defined, No widow/orphan control, Don't adjust space between Latin and Asian text, Don't adjust space between Asian text and numbers
Formatted: Font: Times New Roman, 12 pt
Deleted: 66
Formatted: Normal, Don't adjust right indent when grid is defined, No widow/orphan control, Don't adjust space between Latin and Asian text, Don't adjust space between Asian text and numbers
Formatted: Font: Times New Roman, 12 pt
Deleted: 67
Formatted ...
Formatted: Font: Times New Roman, 12 pt
Deleted: 68
Formatted ...
Formatted: Font: Times New Roman, 12 pt
Deleted: 69
Formatted ...
Formatted: Font: Times New Roman, 12 pt
Deleted:
Deleted: To determine if a heterozygous SNV is allele- ...

information, such as raw read counts of both accessible non-allele-specific and allele-specific variants, can be downloaded for further analyses.

Figure 2. Comparing the effects of the binomial and beta-binomial tests in datasets with low and intermediate level of overdispersion. The grey bars in each plot represent the empirical allelic ratio distribution. For each panel 2A and 2B, the red and blue lines on the left plots represent the null (expected) allelic ratio distributions associated with the binomial and beta-binomial tests respectively. The red and blue bars on the right plots represent the number of allele-specific (AS) SNVs detected each the binomial and beta-binomial tests respectively. Figure 2A shows the plots for one of the RNA-seq datasets for the individual HG00096. It has a low overdispersion parameter, ρ=0.0205. The empirical distribution does not have heavy tails and the binomial and beta-binomial tests give very similar results. This differs from Figure 2B, which shows the plots for one of the RNA-seq datasets for the individual NA11894. Overdispersion is higher at ρ=0.1234, and the beta-binomial null distribution provides a better fit to the empirical allelic ratio distribution than the binomial distribution. The empirical distribution (grey bars) also show heavier tails, signifying more SNVs with allelic imbalance.

Figure 3. Inheritance of allele-specific behavior. The left panel shows plots for the TF CTCF (top row) and ASE (bottom row) being examined for inheritance in the CEU trio (Father: NA12891, blue; Mother: NA12892, red; Child: NA12878, green). Each point on the plot represents the allelic ratio of a common ASB SNV between the parent (x-axis) and the child (y-axis), by computing the proportion of reads mapping to the reference allele at that SNV. High Pearson's correlations, r, observed in both parent-child comparisons for CTCF ($r \geq 0.77$) signify strong heritability in allele-specific behavior. ASE also shows considerably strong evidence of heritability but has comparatively lower r values. The table at the top right panel presents the r values for ASB in two TFs and ASE in our analyses.

Figure 4. (a) ASB and ASE SNVs in allele-specific gene ZNF331 (chromosome 19, position 54,041,333-54,083,523). From AlleleDB, we can observe the ASB SNVs (filled red bars with the name of the transcription factor (TF) above the bars) and ASE SNVs (filled black bars) found in each individual (row) and genomic positions (columns) along the ZNF331 gene. We can see that many of these SNVs are sparsely distributed across a single individual. By collapsing or combining information from multiple individuals, we can identify genomic regions or elements that are enriched for allele-specific activity. Unfilled black and red bars denote control SNVs are heterozygous SNVs that have enough reads to be tested but are non-allele-specific. **(b) Two approaches for enrichment analyses are performed for each genomic element.** (1) The 'expanded' enrichment is performed in a population-aware fashion, in which each occurrence of allele-specific or control non-allele-specific SNV in each individual is counted. (2) The 'collapsed' enrichment conflates all occurrences over multiple individuals into a single unique SNV position as long as an allele-specific or accessible non-allele-specific SNV occurs in least one individual.

Figure 5. The 'expanded' enrichment analysis is population-aware and shows that some genomic regions are more inclined to allele-specific regulation. We map variants associated with allele-specific binding (ASB; green) and expression (ASE; blue) to various categories of genomic annotations, such as coding DNA sequences (CDS), untranslated regions (UTRs),

Deleted: , while

Deleted: distribution using

Deleted: empirical and expected distributions

Deleted: individual

Deleted: empirical and expected distributions

Deleted: individual

enhancer and promoter regions, to survey the human genome for regions more enriched in allelic behavior. Using the control non-allele-specific SNVs as the expectation, we compute the log odds ratio for ASB and ASE SNVs separately, via Fisher's exact tests. The number of asterisks depicts the degree of significance (Bonferroni-corrected): *, $p<0.05$; **, $p<0.01$; ***, $p<0.001$. For each transcription factor (TF) in AlleleDB, we also calculate the log odds ratio of ASB SNVs in promoters, providing a proxy of allele-specific regulatory role for each available TF. Genes known to be mono-allelically expressed such as imprinted and MHC genes (CDS regions) are highly enriched for both ASB and ASE SNVs. The actual log odds ratio of ASB SNVs in imprinted genes, both ASB and ASE SNVs in immunoglobulin genes and ASE SNVs for 3'UTR, MHC and olfactory receptor genes are indicated on the bars.

Figure 6. A considerable fraction of allele-specific variants are rare but do not form the majority. A lower proportion of allele-specific SNVs than non-allele-specific SNVs are rare, suggesting less selective constraints in allele-specific SNVs. The minor allele frequency (MAF) spectra of ASB (green filled circle), control non-ASB SNVs (green open circle), ASE (blue filled circle) and control non-ASE SNVs (blue open circle) are plotted at a bin size of 100. The peaks are in the bin for MAF $\leq 0.5\%$. The inset zooms in on the histogram at MAF $\leq 2.5\%$. The proportion of rare variants in descending order: ASE- > ASE+ > ASB+ > ASB-. Comparing ASE+ to ASE- gives an odds ratio of 0.2 (Bonferroni-corrected hypergeometric $p < 2.2e\text{-}16$), while comparing ASB+ to ASB-, gives an odds ratio of 1.4 ($p=0.08$), signifying statistically significant depletion of ASE SNVs but statistically insignificant enrichment of ASB SNVs relative to the respective non-allele-specific control SNVs. Statistically significant depletion in ASE suggests that ASE SNVs are under less purifying selection.

TABLE LEGEND

Table 1. Breakdown of SNVs in each ethnic population. Heterozygous (HET), accessible (ACC) and ASE SNVs are in Table 1A and ASB SNVs are in Table 1B for 381 unrelated individuals (exclude NA12878). Table 1C shows the same HET, ACC and both ASE and ASB SNVs detected in a single individual, NA12878, who is also part of the trio family. For each of the last 3 columns, each category of HET, ACC and allele-specific SNVs is further stratified by the population minor allele frequencies: common (MAF > 0.05), rare (MAF ≤ 0.01) and very rare (MAF ≤ 0.005). The number of allele-specific SNVs is given as a percentage of the ACC SNVs. Table 1 also provides the number of individuals from each ethnic population with RNA-seq and ChIP-seq data available for the ASE and ASB analyses respectively.

SUPPLEMENTARY MATERIALS

Supplementary Figure 1

This figure shows the results for the 'collapsed' enrichment analysis. We map variants associated with allele-specific binding (ASB; green) and expression (ASE; blue) to various categories of genomic annotations, such as coding DNA sequences (CDS), untranslated regions (UTRs), enhancer and promoter regions, to survey the human genome for regions more enriched in allelic behavior. Using the control non-allele-specific SNVs as the expectation, we compute the log odds ratio for ASB and ASE SNVs separately, via Fisher's exact tests. The number of asterisks depicts the degree of significance (Bonferroni-corrected): *, $p<0.05$; **, $p<0.01$; ***, $p<0.001$. For each transcription factor (TF) in AlleleDB, we also calculate the log odds ratio of ASB SNVs in promoters, providing a proxy of allele-specific regulatory role for each available TF.

Genes known to be mono-allelically expressed such as imprinted and MHC genes (CDS regions) are highly enriched for both ASB and ASE SNVs. The actual log odds ratio of ASB SNVs in imprinted genes, both ASB and ASE SNVs in immunoglobulin genes and ASE SNVs for MHC genes are indicated on the bars. Between the two enrichment analyses, we observe consistent trends in the odds ratios of ASB SNVs and ASE SNVs across the MAE gene sets, except for the T cell receptors. The category is enriched in ASE SNVs when we collapsed the SNV count but, interestingly, depleted when we expand the enrichment analysis in a population-aware fashion (Figure 5). This suggests that the allele-specific expression in certain T cell receptors is not consistently observed in all individuals. Also, there is a consistent depletion in ASE SNVs for the constitutively expressed housekeeping genes, implying that most housekeeping genes give a more balanced (biallelic) expression (Figure 5).

Supplementary Figure 2
This figure shows the percentage of (a) ASB and (b) ASE SNVs (opaque bars with black boundaries) when compared to the accessible SNVs (ACC; transparent bars with no boundaries) as a function of read depth, for 379 unrelated individuals (trio excluded). Here, we display >90% of ASB and ASE SNVs, by not showing those with extreme read depths. Despite the bias in SNV counts towards low read depth, the percentages of our ASB and ASE SNVs that are called are relatively consistent across all read depths (% ASB or ASE; indicated by circles).

Supplementary Figure 3
This figure shows the number of accessible (transparent-colored bars) and ASE SNVs (opaque-colored bars with black boundaries) per individual, grouped and colored by population: CEU (blue), CHB (orange), FIN (magenta), GBR (red), JPT (yellow), TSI (grey) and YRI (green). The CEU trio are represented by the three spikes at the far left. In general, the YRI have more accessible and ASE sites, probably because they have higher number of heterozygous SNVs in their genomes. The number of ASE sites in addition to the proportion with regards to their accessible sites per individual are relatively consistent.

Supplementary Figure 4
This figure shows the replication of AS calls at increasing read depths. We randomly subsampled subsets of various read coverage from a pooled RNA-seq dataset of NA12878 – 100M, 200M, 300M, 400M and 490M ('M' denotes 'million of reads') – such that each smaller pool of reads is a direct subset of the larger sets, with 490M denoting the entire set of reads. For instance, 100M is a subset of all the other sets. We then ran the AlleleDB pipeline. We show that >77% ASE sites are consistent in at least 2 subsets, with very small number of sites unique to each set.

Supplementary Figure 5
This figure shows that the replication of AS calls between technical replicates. We randomly sampled two subsets of 245M ('M' denotes 'million of reads') from a pooled RNA-seq dataset of NA12878, without replacement, i.e. these two sets are mutually exclusive. We then run the AlleleDB pipeline. The Venn diagram shows that the calls between the replicates are very comparable (>70% overlap), demonstrating that our calls reproduce very well.

Supplementary Table 1

Moved (insertion) [2]

Moved down [3]: This figure shows that the replication of AS calls between technical replicates. We randomly sampled two subsets of 245M ('M' denotes 'million of reads') from a pooled RNA-seq dataset of NA12878, without replacement, i.e. these two sets are mutually exclusive. We then run the AlleleDB pipeline. The Venn diagram shows that the calls between the replicates are very comparable (>70% overlap), demonstrating that our calls reproduce very well.¶
¶

Moved (insertion) [3]

Moved up [2]: This figure shows the number of accessible (transparent-colored bars) and ASE SNVs (opaque-colored bars with black boundaries) per individual, grouped and colored by population: CEU (blue), CHB (orange), FIN (magenta), GBR (red), JPT (yellow), TSI (grey) and YRI (green). The CEU trio are represented by the three spikes at the far left. In general, the YRI have more accessible and ASE sites, probably because they have higher number of heterozygous SNVs in their genomes. The number of ASE sites in addition to the proportion with regards to their accessible sites per individual are relatively consistent.¶

This table shows the inconsistencies of the eight studies performing allele-specific analyses using different tools and parameters, e.g. read mapping with a range of read aligners, alignment to different reference genomes and variations of statistical tests in detecting the allele-specific variants. We uniformly processed the tools and parameters in AlleleDB.

Supplementary Table 2
This table shows the number of individual datasets being flagged and segregated due to insufficient reads and due to having an "overdispersed" allelic ratio distribution.

*We define an "overdispersed" ChIP-seq dataset as those with $\rho \geq 0.3$, while an "overdispersed" RNA-seq dataset is defined more strictly by $\rho \geq 0.125$, which is one standard deviation more than the mean overdispersion in the RNA-seq datasets in our processing.

Supplementary Table 3
This table shows the number of uniquely mapped maternal (column 2) and paternal (column 3) reads that overlap a certain number of heterozygous SNVs (column 1) from an example dataset from NA12878 CTCF ChIP-seq assay. ~97% of reads that map uniquely to the maternal or paternal haplotype overlap only 1 heterozygous SNV. On average, we find that >90% of uniquely mapped reads that overlap any heterozygous SNVs at all, overlap only 1 heterozygous SNV.

Supplementary Table 4
This table shows the slope and Pearson's correlation results for two DNA-binding proteins, PU.1 and CTCF, and ASE for parent-child and parent-parent comparisons.

Supplementary Table 5
This table summarizes the results in examining the effects of accounting for ambiguous mapping bias via the removal of sites (column 3) and reads (column 4) using four datasets. We chose two ChIP-seq and two RNA-seq datasets from NA12878. We find that removal of sites often filters SNVs that might be still allele-specific even after removing reads that show ambiguous mapping bias (AMB), indicating that site removal can be over-conservative and read removal is able to retain AS SNVs that are still allele-specific. Also, in our study, we find that AMB seems to have a greater effect on ChIP-seq datasets. Between 10-21% of the detected AS SNVs are removed in ChIP-seq compared to 1-4% in RNA-seq datasets, depending on which bias removal strategy was adopted.

Supplementary File 1
This Excel file contains results from our 'collapsed' and 'expanded' enrichment analyses for 708 categories from ENCODE, including the Fisher's exact test odds ratios, p-values (original and Bonferroni-corrected), the number of allele-specific SNVs and accessible non-allele-specific (control) SNVs found in each category. The results for five gene element categories from GENCODE and 16 enhancer categories are also included. 'NA' is marked in categories where odds ratio cannot be calculated due to insufficient numbers in non-allele-specific SNVs. These are tabulated for ASB, ASE and allele-specific SNVs; the latter is the results for the combined number of ASB and ASE SNVs.

Supplementary File 2
This Excel file contains results from our 'collapsed' and 'expanded' enrichment analyses for the 19,257 autosomal protein-coding genes (HGNC symbols) from GENCODE, including the Fisher's exact test odds ratios, p-values (original and Bonferroni-corrected), the number of allele-specific SNVs and accessible non-allele-specific (control) SNVs found in the gene region and the promoter region (upstream 2500bp). The results for housekeeping genes and 4 monoallelically-expressed gene categories are also included. 'NA' is marked in categories where odds ratio cannot be calculated due to insufficient numbers in non-allele-specific SNVs. These are tabulated for ASB, ASE and allele-specific SNVs; the latter is the combined number of ASB and ASE SNVs. Based on results in AS, we define enhancer regions that are "allele-specific" (Bonferroni p value ≤ 0.05, odds ratio ≥ 1.5), "balanced" (Bonferroni p value ≤ 0.05, odds ratio < 1.5) and "indeterminate".

Supplementary File 3
This Excel file contains the ASB 'collapsed' and 'expanded' enrichment analyses in promoter regions for 44 TFs used in our database, including the Fisher's exact test odds ratios, p-values (original, Bonferroni-corrected), the number of ASB SNVs, accessible non-allele-specific (control) SNVs both found and not found in the gene region. ASB SNVs for each TF are contributed by different individuals. If either of the parents in the CEU trio is involved, ASB SNVs for NA12878 are not included. Those TFs with only ASB SNVs from NA12878 are annotated '1' under the column 'NA12878 only'. 'NA' is marked in categories where odds ratio cannot be calculated due to insufficient numbers in any of the last three columns.

Supplementary File 4
This Excel file contains the ASB SNVs that reside in TF motifs described in Kheradpour and Kellis[75]. Under the column 'motif', the information is delimited by "#" in this order: motif identifier (as defined in Kheradpour and Kellis), start position of motif (0-based), end position of motif (1-based), strand and position of SNV in motif. Allelic ratios at each SNV position are defined above, i.e. ratio of number of reference reads to number of alternate reads.

Deleted: [68].

Supplementary File 5
This Word file contains the R pseudocode for the bisection method that is used to estimate the overdispersion parameter.

Supplementary File 6
This Excel file contains sets of more confident ASB and ASE SNVs. For the more confident 2,394 ASE SNVs, they are identified because ≥ 38 individuals (column 'indCount' ≥ 38) possess each of them. At the same time, for each of the SNV, the allele that has more reads for each individual (columns 'winningAllele' and 'alleleCounts') are consistently found in 80% of the individuals (column 'freq' ≥ 0.8) that possess this ASE SNV. The more confident 183 ASB SNVs are defined by having ≥ 3 individuals possessing that ASB SNV, regardless of the identities of TFs (columns ind_TF and indCount ≥ 3). Also, the allele that has more reads for each ind_TF (columns 'winningAllele' and 'alleleCounts') are found in 80% of ind_TF (column 'freq' ≥ 0.8).

Supplementary File 7

This zip file contains a tab-delimited file that shows the results from our 'expanded' enrichment analysis for 882 experimentally-determined VISTA[73] enhancers and 410,486 enhancer regions from the union of lists by Ernst and Kellis (2012)[70], Hoffman *et. al*. (2013)[71]), and data from distal regulatory modules from Yip *et al*. (2012)[72]. The results include the number of allele-specific SNVs and accessible non-allele-specific (control) SNVs. 'NA' is marked in categories where odds ratio cannot be calculated due to insufficient numbers in non-allele-specific SNVs. These are tabulated for ASB, ASE and AS SNVs; the latter is the combined number of ASB and ASE SNVs. Based on results in AS, we define enhancer regions that are "allele-specific" (Bonferroni p value ≤ 0.05, odds ratio ≥ 1.5), "balanced" (Bonferroni p value ≤ 0.05, odds ratio < 1.5) and "indeterminate".

Deleted: [66] enhancers and 410,486 enhancer regions from the union of lists by Ernst and Kellis (2012)[63], Hoffman *et. al*. (2013)[64]), and data from distal regulatory modules from Yip *et al*. (2012)[65].