Uniform Survey of Allele-Specific Binding and Expression Across 383 Individuals

Jieming Chen[1,2], Joel Rozowsky[1,3], Jason Bedford[1], Arif Harmanci[1,3], Alexei Abyzov[1,3,6], Yong Kong[4,5], Robert Kitchen[1,3], 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, Mayo Clinic, Rochester, MN 55905

Abstract

Large-scale sequencing of personal genomes has revealed a large number of genomic variants, creating significant challenges regarding their functional annotation. We focus on variants associated with allele-specific behavior, where allelic imbalance can be directly detected using functional assays. Overall, we find 169,235 allele-specific binding and 144,083 allele-specific expression SNVs across 383 personal genomes, representing 41% and 22% of heterozygous sites that are accessible to ASB and ASE detection respectively. Through comparison of allelic with non-allelic sites, we identify genomic annotations that are significantly enriched in allele-specific SNVs, such as the expression of PTPRG and SNURFand the binding sites of transcription factors BCLAF1 and E2F1. We also observe that allele-specific SNVs tend to be in regions under less purifying selection. These variants and their annotations are offered as a community resource via AlleleDB (http://alleledb.gersteinlab.org/).

Formatted: Font: 12 pt

Deleted: As the cost of sequencing gets cheaper, more personal genomes are being sequenced via large genome sequencing projects and clinical sequencing. These efforts have found a large number of genomic variants. An increasingly important challenge is to functionally annotate variants in personal genomes on a large scale, especially those regulatory variants located in the non-coding genome. This can be done by overlapping readouts from functional genomic assays such as RNA-seq and ChIP-seq. We focus on a specific class of regulatory variation that is made up of variants associated with allele-specific behavior, which is exhibited when a differential phenotypic effect, such as binding or transcription, is observed between the two alleles in a diploid genome. This effect can be reflected by an allelic imbalance in the signals from the functional genomic assays. Previous allele-specific analyses have based mainly on either a few deeply-sequenced, well-annotated genomes or a single assay over the genomes of a variety of individuals. Here, we endeavor to combine data across multiple studies to detect allele-specific variants. Unfortunately, allele-specific variants detection is sensitive to various technical issues such as heterozygous variant detection and read mapping. Hence, simply pooling the results of these disparate studies is not optimal. Moreover, different studies use various tools, parameters and thresholds. Therefore, it is imperative to subject each dataset to uniform processing. To this end, we pool DNA sequences, RNA-seq and ChIP-seq data from separate data sources. This amounted to 383 individuals from seven ancestries and we put them through a standardized processing pipeline. We consolidate the results in a database, AlleleDB. We are able to annotate ~168K putative variants that assess allele-specific binding (ASB) of DNA-binding factors and ~143K variants in allele-specific expression (ASE) of genes, using ChIP-seq and RNA-seq data respectively. Across the seven populations, the number of ASB and ASE SNVs vary, with the Africans possessing the most number of AS variants in general. Within a high coverage European personal genome, there is an average of ~1 and 2% SNVs of ASB and ASB SNVs among heterozygous SNVs. Additionally, using data from only unrelated individuals and ~1000 categories of genomic annotations, we are able to define ~800 categories of non-coding genomic regions and ~800 protein-coding genes that are significantly enriched or depleted of ASB and ASE variants respectively, thereby ascertaining the susceptibility of allele-specific regulation to various elements of the human genome. These data and analyses will aid in the functional annotation of regulatory variation in personal genomes.¶

Formatted: Font: 12 pt

Introduction

In recent years, the number of personal genomes has increased dramatically, from single individuals[1–3] to large sequencing projects such as the 1000 Genomes Project[4], UK10K[5] and the Personal Genome Project[6]. These efforts have provided the scientific community with a massive catalog of human genetic variants, most of which are rare.[4] Subsequently, a major challenge is to functionally annotate all of these variants.

Much of the characterization of variants so far has been focused on those found mainly 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 some form of functional readouts from the assays to genomic variants, particularly in identifying regulatory variants, such as mapping of expression quantitative trait loci (eQTLs)[7–9] and allele-specific (AS)[10–12] 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, AS 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, AS variants can be detected regardless of their allele frequencies. Therefore, AS approaches are very useful, in terms of functionally annotating personal genomes, for identifying cis-regulatory variants on a large scale.

Early high throughput implementations of AS approaches employed microarray technologies, and thus are restricted to a subset of loci.[13–15] Later studies have used ChIP-seq and RNA-seq experiments for genome-wide scans of AS variants but have been mostly limited to a single assay with a variety of individuals,[16] or a few individuals with deeply-sequenced and well-annotated genomes.[11,12] 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 of more than 50 transcription factors (TFs) distributed in more than 1 studies.[17–19] Merging these datasets is advantageous, be it increasing statistical power or simply having more features for more intra- and inter-individual comparisons (such as TFs and populations).

AS variant detection is extremely sensitive to the technical issues of variant calling and RNA-seq and ChIP-seq experiments, such as heterozygous variant calling and read mapping.[20–23] Moreover, studies with the appropriate datasets are typically designed for various purposes, resulting in disparate sets of computational tools, strategies and threshold parameters used in the processing of data in each respective study. These reasons portend that a simple pooling of results from multiple studies may not be optimal even for the same biological sample. The task of merging has to be carried out in a uniform and meaningful manner to yield interpretable results. To this end, we organize and unify datasets from eight different studies into a comprehensive data corpus and repurpose it specifically for allele-specific analyses. We detect more than 169K and 144K single nucleotide variants (SNVs) associated with allele-specific binding (ASB) and expression (ASE) events respectively. We are able to present a systematic survey of these detected AS SNVs in various categories of coding and non-coding genomic annotations. The variants and annotations are available in a resource, AlleleDB (http://alleledb.gersteinlab.org/). Finally, using our consolidated data, we investigate the extent

Formatted: Font: 12 pt

Deleted: In recent years, the number of personal genomes has increased dramatically, from single individuals [cite Watson, venter] to large sequencing projects such as the 1000 Genomes Project,[cite] UK10K [cite] and the Personal Genome Project [cite]. These efforts have provided the scientific community with a massive catalog of human genetic variants, most of them rare.[cite] Subsequently, one of the major challenges is to functionally annotate all of these variants. ¶
¶
Much of the characterization of variants so far has been focused on those found mainly in the protein-coding regions, but the advent of large-scale functional genomic assays, such as ChIP-seq and RNA-seq, has facilitated the annotation of genome-wide variation. This can be accomplished by correlating some form of functional readouts from the assays to genomic variants, particularly in identifying regulatory variants, such as mapping of expression quantitative trait loci (eQTLs) and allele-specific (AS) variants.[cite] eQTLs are detected by assessing the effects of variants on expression profiles across a large population of individuals. A sizeable cohort is required in order to achieve statistical power to detect variants of low frequencies, which places a constraint in the annotation of variants in personal genomes. On the other hand, allele-specific approaches assess phenotypic differences at heterozygous loci within a single genome, so that each allele at these positions in a diploid genome acts as a perfectly matched control for the other allele. [cite old and new papers] As such, they can detect AS variants regardless of their allele frequencies. Therefore, it is very useful to identify AS variants on a large scale, in terms of functionally annotating personal genomes. ¶
¶
Early high throughput implementations of AS approaches employed microarray technologies, and thus are restricted to a subset of loci.[cite 2002 Yan] Later studies have used ChIP-seq and RNA-seq experiments for genome-wide scans of AS variants but have been mostly limited to a few individuals with deeply-sequenced and well-annotated genomes,[cite] or a single assay with a variety of individuals.[cite] Consequently, there is a need to garner more RNA-seq and ChIP-seq data for more extensive allele-specific analyses. While a straightforward strategy is to increase the number of individuals and experimental assays (e.g. number of transcription factors for ChIP-seq experiments on a single individual), this requires large amounts of resources. A less expensive alternative is to tap into the wealth of existing ChIP-seq and RNA-seq experimental data, by pooling already available datasets from numerous studies. 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 of more than 50 transcription factors (TFs) distributed in more than 1 studies.[cite, ENCODE, kasowski, Kilpinen] Merging of these datasets has obvious advantages in analyses, be it increasing statistical power or simply having more features for inter-individual comparisons (such as TFs and populations). ¶
¶ ...

of purifying selection in allele-specific SNVs and the inheritance of allele-specific binding in eight different transcription factors.

Results

AlleleDB, a resource for allele-specific behavior genome annotation

There are two layers of information with respect to an individual that needs to be integrated in order to more accurately detect AS SNVs: (1) the DNA sequence of the individual, and (2) reads from either the RNA-seq or ChIP-seq experiment to look for SNVs associated with ASB or ASE. Here, we implement a uniform pipeline to combine personal genomic, transcriptomic and binding data and to standardize our detection of potential AS SNVs (Figure 1). First, we construct a diploid personal genome for each of the 383 individuals, using variants from the 1000 Genomes Project. Next, we pool the reads from each individual's ChIP-seq or RNA-seq and align them to each of the haploid genome. In total, we reprocess 142 ChIP-seq and 475 RNA-seq datasets for 383 individuals. Lastly, the AS SNVs are detected based on allelic imbalance of reads between the two haplotypes at heterozygous loci. For ChIP-seq data, the SNVs are additionally pared down to those within peak regions (see Methods).

We further define sets of 'control' SNVs. This is especially pertinent to our enrichment analyses, since the results are dependent on the choice of the null expectation (controls). The control SNVs are not allele-specific and are derived from a set of 'accessible' SNVs, which are heterozygous SNVs and possess at least the minimum number of reads to be statistically detectable for allelic imbalance. The accessible SNVs are determined for each ChIP-seq (grouped by individual and TF, not by study) or RNA-seq dataset (Table 1). In other words, these controls match the AS SNVs by statistical accessibility and being heterozygous.

By comparing AS SNVs relative to the control SNVs in each genomic annotation (see methods), we investigate the enrichment (or depletion) of AS SNVs in 20,144 protein-coding genes from GENCODE (version 17)[24] and 952 categories of non-coding genomic elements, including DNaseI hypersensitivity sites and transcription factor binding motifs from ENCODE Integrative release.[17] This provides a systematic survey of ASB and ASE with respect to various functional annotations in the human genome.

We build a database, AlleleDB (http://alleledb.gersteinlab.org/), to house the annotations, and the candidate AS and accessible SNVs. AlleleDB can be downloaded as flat files or queried and visualized directly as a UCSC track in the UCSC Genome browser[25] as specific genes or genomic locations. This enables cross-referencing of AS 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 (AS SNVs are red, accessible SNVs are black) in the displayed track.

Enrichment analyses

Of great interest, is the annotation of these allele-specific SNVs with respect to known genomic elements, both coding and non-coding. Using the AlleleDB variants found in the personal genomes of the 2 parents of the trio and 380 unrelated individuals from Phase 1 of the 1000

Formatted: Font: 12 pt

Deleted: allele-specific

Formatted: Font: 12 pt

Deleted: allele-specific expression (ASE)

Formatted: Font: 12 pt

Deleted: binding (ASB) (Figure 1).

Formatted: Font: 12 pt

Deleted: (see Methods)

Formatted: Font: 12 pt

Deleted: allele-specific SNVs. Eventually, our pipeline detected a total of 143,316 unique ASE SNVs for 382 unrelated individuals and 168,539 unique ASB SNVs from a collective ChIP-seq dataset of 19 transcription factors for 18 unrelated individuals. We also define a set of control SNVs, for each TF dataset (for ASB) and each individual expression dataset (for ASE). This is especially imperative in our enrichment analyses, which are highly dependent on the choice null expectation (controls). We first define a set of 'accessible' SNVs, which are heterozygous and possess at least the minimum number of reads that is determined separately for each ChIP-seq (grouped by TF, not by study) or RNA-seq dataset (Table 1). The control ...

Formatted: Font: 12 pt, Font color: Auto

Formatted: Font: 12 pt

Formatted: Font: 12 pt

Formatted ...

Deleted: allele-specific

Formatted ...

Formatted ...

Deleted: , in terms of gene or genomic locations,

Formatted ...

Deleted: (Figure 1). [cite]

Formatted ...

Deleted: allele-specific

Formatted ...

Deleted: All heterozygous

Formatted ...

Deleted: , including accessible SNVs,

Formatted ...

Deleted: others

Formatted ...

Formatted: Font: 12 pt

Formatted ...

Formatted: Font: 12 pt

Formatted ...

Deleted: Only

Genomes Project, we focus on autosomal SNVs and found that ~56% of our candidate ASE SNVs and ~6% of ASB SNVs are in coding DNA sequences (CDS). Overall, we detected 144,083 ASB and 169,235 ASE SNVs, representing 22% and 41% of the accessible SNVs respectively (Table 1). Further, for ASB SNVs, we observed statistical significance ($p < 0.05$) for 787 non-coding categories and 15 protein-coding genes and for ASE SNVs, 598 non-coding categories and 831 genes, with varying degree of enrichment and depletion of AS SNVs (Supp file). Table 2 shows the top 10 genes and non-coding regions enriched in AS SNVs.

Figure 2 shows the enrichment of AS SNVs to provide a survey of AS regulation in elements closely related to a gene model, namely enhancers, promoters, CDS, introns and untranslated regions (UTR). In general, both categories of AS SNVs are more likely found in the 5' and 3' UTRs, suggesting allele-specific regulatory roles in these regions. On the other hand, intronic regions seem to exhibit a dearth of allele-specific regulation. For SNVs associated with allele-specific expression (ASE), a greater enrichment in 3' UTR than 5' UTR regions might be, in part, a result of known RNA-seq bias.[26,27] For SNVs associated with allele-specific binding (ASB), we also observe an enrichment in the promoters, hinting at functional roles in these variants found in TF binding motifs or peaks found near transcription start sites in the promoter regions to regulate gene expression. However, we see variable enrichments of ASB SNVs of particular TFs in promoter regions such as RPB2 and SA1, while depletion in others, such as PU.1 and POL2 (Figure 2, Supp file). These differences imply that some TFs are more likely to participate in allele-specific regulation than others. Enrichments of ASE, as well as, ASB SNVs are both observed in CDS. It is likely that the enrichment of ASB SNVs is due predominantly to a small set of CDS regions, in light that there are only 15 protein-coding genes with statistically significant enrichment of ASB SNVs. Nonetheless, an enrichment of ASB SNVs might suggest an allele-specific mechanism in the regulatory roles of some of the TFs that bind to these regions.

We also compute the enrichment of AS SNVs in various gene categories. Some of them have been known to be involved in monoallelic expression (MAE), namely imprinted genes,[28] and three sets of genes known to undergo allelic exclusion: olfactory receptor genes,[29] immunoglobulin,[30] genes associated with T cell receptors and the major histocompatibility complex.[31] Monoallelic exclusion is a process exhibiting monoallelic expression, whereby one allele is being expressed while the other is silenced or repressed.[32,33] A list of genes found to experience random monoallelic expression (RME) in a study by Gimelbrant *et al* (2007) is also included.[34] As expected, most of the MAE gene sets have been found to be significantly enriched in both ASB and ASE SNVs, with the exception of the olfactory receptor and RME genes. Interestingly, while a statistically significant enrichment of ASB SNVs is observed in the constitutively expressed housekeeping genes, there is no enrichment in ASE SNVs (Figure 2).

Rare variants and purifying selection in AS SNVs

To assess the occurrence of ASB and ASB SNVs in the human population, we consider the minor allele frequencies (MAF). Table 1 shows the breakdown of the accessible and AS SNVs in seven ethnic populations and allele frequencies. Yoruba from Ibadan, Nigeria (YRI) contributes the most to both ASE and ASB variants at each allele frequency category. The number of very rare AS SNVs (MAF ≤ 0.5%) is about two folds higher in the YRI (~48% ASE SNVs and ~34% ASB SNVs with MAF ≤ 5%) than the other European sub-populations of comparable (CEU, FIN)

Formatted ...

Deleted: found

Deleted: Using the AlleleDB variants found in (...

Formatted ...

Formatted ...

Deleted: ≤

Formatted ...

Deleted: . ¶ ...

Formatted ...

Formatted ...

Deleted: more

Formatted ...

Deleted: structure

Formatted ...

Deleted: [any lit evidence? For regulation? ...

Formatted ...

Deleted: .[cite]

Formatted ...

Deleted: observe

Formatted ...

Deleted: fig

Formatted ...

Deleted: Several studies have found

Formatted ...

Deleted: many TFs bind in

Formatted ...

Deleted: regions, for instance to regulate codon ...

Formatted ...

Deleted: to such

Formatted ...

Deleted: .

Formatted ...

Deleted: We also compute the enrichment of AS ...

Formatted ...

Deleted: examine

Formatted ...

Deleted: ASE

Formatted ...

Deleted: population

Formatted ...

Deleted:) from Phase 1 of the 1000 Genomes ...

Formatted ...

Deleted: YRI (

Formatted ...

Deleted: While

Formatted ...

Deleted: with

Formatted ...

Deleted: 005 comprise a considerable proportio ...

Formatted ...

or larger (TSI) population sizes. In general, rare variants do not form the majority of all the AS variants. Nonetheless, we observe a shift towards very low allele frequencies in AS SNVs, peaking at MAF ≤ 0.5% (Figure 3).

To examine selective constraints in AS SNVs, we consider the enrichment of rare variants with population minor allele frequencies (MAF) ≤ 0.5%.[4,35] Our results show lower enrichment of rare variants in AS SNVs when compared to non-AS SNVs. This posits that, as a whole, AS SNVs are under less selective constraints than non-AS SNVs. Our population study is similar to previous studies that use only a single high-coverage individual.[35,36] Such weaker selection may be a result of accommodating varying degrees of gene expression across individuals.

ASB Inheritance analyses using CEU trio

The CEU trio is a well-studied family and particularly, many ChIP-seq studies were performed on different TFs. Previous studies have presented AS inheritance in a few TFs as a case-study.[11,19] Here, we provide a more comprehensive and statistical investigation of the heritability of ASB (Figure 4 and Supp file). For the DNA-binding protein CTCF, we observe a high parent-child correlation, i.e. significantly more points in the B and C quadrants (red quadrants on each plot in Figure 4) compared to the A and D quadrants (grey quadrants in Figure 4), denoting great similarity in allelic directionality (bonferroni-corrected binomial p=1.2e-46 and p=4.2e-53). The inheritance of AS SNVs in the same allelic direction from parent to child implies a sequence dependency in allele-specific behavior. While there is also a high correlation between the unrelated parents, the number of common allelic SNVs in both parents is substantially lower. We interpret this as a combined effect of the sequence heritability of AS behavior and genetic similarity within the same population. Besides CTCF, PU.1, SA1 and POL2 also show AS inheritance (Supp fig). On the contrary, MYC (binomial p=8.2e-5 and p=1.1e-7), PAX5 and RPB2 exhibit enrichment of points in quadrants B and C with very much lower statistical significance (Supp fig), indicating that AS inheritance is not as apparent in some TFs – inheritance of AS behavior may not be a universal phenomenon.

Discussion

Research on regulatory variants has so far focused mainly on eQTL mapping of common variants. AS analyses can provide a complementary approach to detect regulatory variants. Firstly, we found a substantial number of very rare AS SNVs with MAF ≤ 0.5%. This group of SNVs is harder to access by eQTL mapping and the number is expected to increase with more personal genomes. Secondly, in eQTL mapping, correlation is drawn between total expression measured between individuals in a population and their genotypes. This is allele-insensitive as the total expression across a locus is measured. As such, effects from trans-factors such as negative feedback mechanism that sought to reduce total expression variance across individual genomes with different genotypes will not be detected. However, in an AS approach, even if the total expression is the same across genotypes, difference in allelic expression can still be detected. Such a within-individual control in an AS approach also eliminates normalization issues across multiple assays. Thirdly, eQTL mapping is contingent on population size for sufficient statistics, while the AS approach can detect AS effects *en masse* within a single

Deleted: ¶
¶

Formatted: Font: 12 pt, No underline, Font color: Auto

Deleted: skew

Formatted: Font: 12 pt, No underline, Font color: Auto

Deleted: %, compared to other MAF categories

Formatted ...

Deleted: However, such an enrichment of very rare SNVs is exhibited more in non-allele-specific SNVs (ASE-, ASB-) than in the corresponding allele-specific SNVs (ASE+, ASB+). Comparing ASE+ to ASE- gives an odds ratio of 0.67 (hypergeometric p < 2.2e-16), while comparing ...

Formatted ...

Moved (insertion) [1]

Formatted ...

Deleted: Common SNVs (MAF > 5%) constitu ...

Formatted ...

Deleted: has also

Formatted ...

Deleted: and consequently

Formatted ...

Deleted: approaches provide several advantages ...

Formatted ...

Deleted: be

Formatted ...

Deleted: to

Formatted ...

Deleted: . First, for eQTLs to be detected, it has ...

Formatted ...

Deleted: , that

Formatted ...

Deleted: ,

Formatted ...

Deleted: .

Formatted ...

Deleted: between alleles

Formatted ...

Deleted: , since factors that phenotypic differen ...

Formatted ...

Deleted: allele-specific

Formatted ...

Deleted: works for

Formatted ...

individual's genome. This makes it an attractive strategy for biological samples such as primary cells and tissues that are difficult to obtain in large numbers.

An AS approach is able to detect many AS SNVs for a single personal genome. But as the number of personal genomes increases, the number of private or rare variants accumulates and many of them might be involved in regulation. Thus, it is important to capitalize on existing personal genomes, ChIP-seq and RNA-seq datasets, motivating the development of a pipeline that can uniformly process a large number of personal genomic data for AS detection.

Our search for datasets shows a dearth of personal genomes with ChIP-seq and RNA-seq data in non-European and non-African 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 and is recently being increasingly addressed[37] Also, since many AS 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.

Our analyses place an emphasis on relating allele-specific activity to known genomic annotations, such as CDS and various non-coding regions, and many diseases have been found to implicate ASE in particular genomic regions.[38–40] Therefore, our analyses can help to characterize genomic variants on two levels: firstly, at the single nucleotide level, where our detected AS SNVs can serve as an annotation to variant catalogs (e.g. 1000 Genomes Project) in terms of allele-specific cis-regulation; secondly, by associating AS SNVs with a genomic annotation, we might be able to define categories of genomic regions more attuned to allele-specific activity. Additionally, a comparison between ASB and ASE SNVs in the same category of genomic region can provide some insights to the contribution of ASB by TFs in the ASE of genes. For example, the high enrichment of AS SNVs in most loci associated with monoallelic expression can imply coordination of ASB events with ASE. The exceptions are the groups of RME and olfactory receptor genes, where another mechanism (besides ASB) might be causing ASE in these genes. This can help to prioritize downstream experimental characterization to determine if such allele-specific binding (evidenced by ChIP-seq experiments) do exist and if so, whether it leads to any phenotypic differences.[41]

The final data and results are organized into a resource, AlleleDB, 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 the BED format, which is compatible with other tools, such as the Integrated Genome Viewer[42]. More in-depth analyses can be performed by downloading the full set of AS 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 AlleleSeq identified it as an AS SNV. AlleleDB also serves as an annotation of allele-specific regulation of the 1000 Genomes Project SNV catalog, for use by the scientific community especially for research in gene expression.

Finally, we have shown that there is great value and utility in pooling data, and it has to be processed in a uniform fashion to eliminate issues of heterogeneity in various standards and

Formatted: Font: 12 pt

Deleted: Even though AS variant detection works very well for a personal genome, the enrichment of rare variants in the human variant catalog and considerable unexplained inter-individual variability in gene regulation indicate that it is still valuable to gather more personal genomes and their corresponding functional datasets to discover more private SNVs that might be involved in regulation. More personal genomes also allow more comprehensive inter-individual comparisons. Previous ChIP-seq and RNA-seq studies have focused mainly on a few genomes for either ASE or ASB analyses. [cite a few papers] Also, studies investigating ASB has been limited to a few DNA-binding proteins, such as CTCF and POL2. As such, the main motivation is to develop a scalable uniform pipeline to capitalize on existing ChIP-seq and RNA-seq datasets that might have been used for other research objectives and repurpose them for allele-specific analyses, without having to generate new ones. ¶
¶
A

Formatted: Font: 12 pt, No underline, Font color: Auto

Deleted: .[cite Bustamente review on research diversity]

Formatted: Font: 12 pt, No underline, Font color: Auto

Formatted: Font: 12 pt

Moved up [1]: This posits that, as a whole, AS SNVs are under less selective constraints than non-AS SNVs.

Deleted: Our analyses place an emphasis on relating allele-specific activity to known genomic elements, such as CDS and various non-coding regions. Together, these aid in the enduring effort ...

Formatted: Font: 12 pt, No underline, Font color: Auto

Deleted: This observation was also noted in previous studies using only a high-coverage single individual [cite encodenets, alleleseq]. A weaker ...

Formatted: Font: 12 pt, No underline, Font color: Auto

Deleted: centralized in

Formatted: Font: 12 pt, No underline, Font color: Auto

Deleted: plugs into

Formatted: Font: 12 pt, No underline, Font color: Auto

Deleted: [cite].

Formatted: Font: 12 pt, No underline, Font color: Auto

Formatted: Font: 12 pt

Deleted: Finally, we have shown that there is great value and utility in pooling of data and it has to be processed in a uniform fashion to eliminate issues ...

parameters etc. However, there are still several concerns. First, our current catalog of AS SNVs is detected from lymphoblastoid cell lines (LCLs) and most genomic sequences and functional genomic datasets in the literature are predominantly derived from LCLs. However, it has already been known that there is considerable variability in regulation of gene expression in different tissues.[43] More extensive projects, such as GTex[43] and ENCODE[36], are already underway to involve more functional assays and sequencing in other tissues and cell lines. Further, more accurate allelic information is also being achieved with the advent of longer reads to help in haplotype reconstruction and phasing in next-generation sequencing.[44–46] As technology evolves and more personal genomes and functional genomics data become available, AlleleDB is intended as a scalable resource to accommodate new individual genomes (of potentially diverse ancestries), tissue and cell types. Such should be especially valuable, not only for researchers interested in allele-specific regulation but also for the scientific community at large.

Materials and Methods

Genomic annotation

Categories of genomic regions, such as CDS regions and UTRs, and protein-coding gene annotations are obtained from GENCODE version 17.[24] 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. 952 categories of non-coding annotations are obtained from ENCODE Integrative release,[17] which includes broad categories such as TF binding sites and more specific annotations such as distal binding sites of particular TFs, e.g. ZNF274. Genes for random monoallelic expression are from Gimelbrant *et. al*. (2007) [34] The olfactory receptor gene list is from the HORDE database[29]; immunoglobulin, T cell receptor and MHC gene lists are from IMGT database[30]. We performed enrichment analyses on a number of enhancer lists, which are derived from data in VISTA enhancer browser database[47], Ernst and Kellis (2012)[48] and Hoffman et. al. (2013)[49]. They can be found at the following URLs:

1) http://enhancer.lbl.gov/
2) http://www.ebi.ac.uk/~swilder/Superclustering/concordances4/
3) http://encodenets.gersteinlab.org/metatracks/.

Construction of diploid personal genomes

There are a total of 383 genomes used in this study: 380 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*,[12] 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 CEU

Formatted: Font: 12 pt, No underline, Font color: Auto

Formatted: Font: 12 pt

Formatted: Font: 12 pt, No underline, Font color: Auto

Formatted: Font: 12 pt, No underline, Font color: Auto

Deleted: . [cite]

Formatted: Font: 12 pt, No underline, Font color: Auto

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 380 unrelated individuals, the alleles, though phased, are of unknown parental origin.

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. 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 AS detection.

RNA-seq and ChIP-seq datasets

RNA-seq datasets are obtained from the following sources: gEUVADIS[16], ENCODE[17], Lalonde *et al.* (2011)[51], Montgomery *et al.* (2010)[52], Pickrell *et al.* (2010)[7], Kilpinen *et al.* (2013)[19] and Kasowski *et al.* (2013)[18].

ChIP-seq datasets are obtained from the following sources: ENCODE[17], McVicker *et al.* (2013)[53], Kilpinen *et al.* (2013)[19] and Kasowski *et al.* (2013)[18].

Allele-specific SNV detection

AS SNV detection is generally performed by AlleleSeq.[12] For each ChIP-seq or RNA-seq dataset, reads are aligned against each of the derived haploid genome (maternal/paternal genome for trio) using Bowtie 1.[54] No multi-mapping is allowed and only a maximum of 2 mismatches per alignment is permitted. Sets of mapped reads from various datasets are merged into a single set for allele counting at each heterozygous locus. Here, a binomial p-value is derived by assuming a null probability of 0.5 sampling each allele. 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.[12] Briefly, for each iteration of the simulation, a mapped read is randomly assigned to either allele at each heterozygous SNV and performs a binomial test. 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 positives. An FDR cutoff of 10% is used for ChIP-seq data and 5% for RNA-seq data, since the latter is typically of deeper coverage. Furthermore, we allow only significant AS SNVs to have a minimum of 6 reads. For ChIP-seq data, AS SNVs have to be also within peaks. Peak regions are provided as per those called from each publication of origin, except for the dataset from McVicker *et al.* (2013), in which there are no peak calls. In the latter case, we determine the peaks by performing PeakSeq[55] using the unmapped control reads provided by McVicker et al. (2013) via personal communication with the author [cite, parameters? Arif?].

Enrichment analyses

Accessible SNVs, in addition to being heterozygous, also exceed the minimum number of reads detectable statistically by the binomial test. This is an additional criterion imposed, besides the minimum threshold of 6 reads used in the AlleleSeq pipeline. The minimum number of reads

Formatted: Font: 12 pt

Deleted: CNV genotyping is also performed for each genome by CNVnator,[cite] which calculates the average read depth within a defined window size, normalized to the genomic average for the 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 AS detection.¶
¶

Formatted: Font: 12 pt, No underline, Font color: Auto

Formatted: Font: 12 pt

Deleted: AS SNV detection is generally performed by AlleleSeq. For each ChIP-seq or RNA-seq dataset, reads are aligned against each of the derived haploid genome (maternal/paternal genome for trio) using Bowtie 1.[cite] No multi-mapping is allowed and only a maximum of 2 mismatches per alignment is permitted. Sets of mapped reads from various datasets are merged into a single set for allele counting at each heterozygous locus. Here, a binomial p-value is derived by assuming a null probability of 0.5 sampling each allele. To correct for multiple hypothesis testing, AlleleSeq calculates FDR. 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, as described in the original AlleleSeq publication [cite]. Briefly, for each iteration of the simulation, AlleleSeq randomly assigns a mapped read to either allele at each heterozygous SNV and performs a binomial test. 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 positives. An FDR cutoff of 10% is used for ChIP-seq data and 5% for RNA-seq data, since the latter is typically of deeper coverage. Furthermore, we allow only significant AS SNVs to have a minimum of 6 reads. For ChIP-seq data, AS SNVs have to be also within peaks. Peak regions are provided as per those called from each publication of origin, except for the dataset from McVicker *et al*. [cite], in which there are no peak calls. In the latter case, we determine the peaks by performing PeakSeq [cite] using the unmapped control reads provided via personal communication with the author [cite, parameters? Arif?]. ¶
¶

Formatted: Font: 12 pt, No underline, Font color: Auto

Formatted: Font: 12 pt

Formatted: Font: 12 pt, No underline, Font color: Auto

varies with the pooled size (coverage) of the ChIP-seq or RNA-seq dataset. Given a fixed FDR cutoff, for a larger dataset, the 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), this minimum N can be obtained from a table of expected two-tailed binomial probability density function, such that accessible SNVs are all SNVs with number of reads, n = max(6,N). The control (non-AS) ASB or ASE SNVs are accessible SNVs excluding the respective ASB or ASE SNVs. Enrichment analyses are performed using the Fisher's exact test. P-values are Bonferroni-corrected and considered significant if < 0.05.

AS inheritance analyses

We compute the allelic ratio 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, for each pair of individuals in the trio family, i.e. parent-child and parent-parent. Since AS events can only be detected at heterozygous sites, we consider two scenarios: (1) when an AS SNV is heterozygous in all three individuals but common to the two individuals being compared, and (2) when an AS SNV is heterozygous in two individuals and homozygous (reference or alternate) in the third. P-values are generated by a binomial test of quadrants B and C against a random null distribution (probability = 0.5). The p-values are also Bonferroni-corrected and considered significant if < 0.05.

Acknowledgements

The authors would like to thank Dr. Rob Bjornson for technical help.

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. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5,** e254 (2007).
4. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
5. Muddyman, D., Smee, C., Griffin, H. & Kaye, J. Implementing a successful data-management framework: the UK10K managed access model. *Genome Med.* **5,** 100 (2013).
6. Church, G. M. The personal genome project. *Mol. Syst. Biol.* **1,** 2005.0030 (2005).
7. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464,** 768–72 (2010).
8. Majewski, J. & Pastinen, T. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends Genet.* **27,** 72–9 (2011).

Deleted: (Supp Figure),

Formatted: Font: 12 pt, No underline, Font color: Auto

Formatted: Font: 12 pt, No underline, Font color: Auto

Formatted: Font: 12 pt

Formatted: Font: 12 pt, No underline, Font color: Auto

Formatted: Font: 12 pt

Formatted: Font: 12 pt, No underline, Font color: Auto

Deleted:

Formatted: Font: 12 pt, No underline, Font color: Auto

Deleted:

Formatted: Font: 12 pt, No underline, Font color: Auto

Deleted: then ranked to determine a 'degree' of allele-specific inheritance

Formatted: Font: 12 pt, No underline, Font color: Auto

Formatted: Font: 12 pt

Deleted: ¶

Formatted: Font: 12 pt

Formatted: Font: 12 pt, No underline, Font color: Auto

Formatted: Font: 12 pt

Formatted: Font: 12 pt, No underline, Font color: Auto

Formatted: Font: 12 pt

9. 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).
10. Birney, E., Lieb, J. D., Furey, T. S., Crawford, G. E. & Iyer, V. R. Allele-specific and heritable chromatin signatures in humans. *Hum. Mol. Genet.* **19,** R204–9 (2010).
11. McDaniell, R. *et al.* Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* **328,** 235–9 (2010).
12. Rozowsky, J. *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol.* **7,** 522 (2011).
13. Yan, H., Yuan, W., Velculescu, V. E., Vogelstein, B. & Kinzler, K. W. Allelic variation in human gene expression. *Science* **297,** 1143 (2002).
14. Ge, B. *et al.* Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. *Nat. Genet.* **41,** 1216–22 (2009).
15. Lo, H. S. *et al.* Allelic variation in gene expression is common in the human genome. *Genome Res.* **13,** 1855–62 (2003).
16. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501,** 506–11 (2013).
17. Bernstein, B. E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
18. Kasowski, M. *et al.* Extensive variation in chromatin states across humans. *Science* **342,** 750–2 (2013).
19. Kilpinen, H. *et al.* Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* **342,** 744–7 (2013).
20. Harismendy, O. *et al.* Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.* **10,** R32 (2009).
21. 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).
22. Hansen, K. D., Brenner, S. E. & Dudoit, S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.* **38,** e131 (2010).
23. Degner, J. F. *et al.* Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25,** 3207–12 (2009).
24. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22,** 1760–74 (2012).
25. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).
26. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10,** 57–63 (2009).
27. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320,** 1344–9 (2008).
28. 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).
29. Olender, T., Nativ, N. & Lancet, D. HORDE: comprehensive resource for olfactory receptor genomics. *Methods Mol. Biol.* **1003,** 23–38 (2013).
30. Lefranc, M.-P. *et al.* IMGT-Choreography for immunogenetics and immunoinformatics. *In Silico Biol.* **5,** 45–60 (2005).

31. Complete sequence and gene map of a human major histocompatibility complex. The MHC sequencing consortium. *Nature* **401,** 921–3 (1999).

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

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

34. Gimelbrant, A., Hutchinson, J. N., Thompson, B. R. & Chess, A. Widespread monoallelic expression on human autosomes. *Science* **318,** 1136–40 (2007).

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

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

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

38. 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).

39. 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).

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

41. Cusanovich, D. A., Pavlovic, B., Pritchard, J. K. & Gilad, Y. The functional consequences of variation in transcription factor binding. *PLoS Genet.* **10,** e1004226 (2014).

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

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

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

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

46. Fan, H. C., Wang, J., Potanina, A. & Quake, S. R. Whole-genome molecular haplotyping of single cells. *Nat. Biotechnol.* **29,** 51–7 (2011).

47. 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).

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

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

50. 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).

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

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

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

54. 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).

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

Figure caption

Figure 1. Uniform processing of data from 343 individuals and construction of AlleleDB. For each of the 383 individuals, a diploid personal genome is first constructed using the variants from the 1000 Genomes Project. Next, reads from ChIP-seq or RNA-seq data are mapped onto each of the haploid genome of the diploid genome. At each heterozygous SNV, a comparison is made between the number of reads that map to either allele, and a statistical significance (after multiple hypothesis test correction) is computed to determine if a SNV is allele-specific (AS). All the candidate AS variants are then deposited in AlleleDB database. Additional information, such as raw read counts of both accessible non-AS and AS variants, can be downloaded for further analyses.

Figure 2. Some genomic regions are more susceptible 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), enhancer and promoter regions, to survey the human genome for regions more enriched in allelic behavior. Using the accessible non-AS SNVs as the expectation, we compute the log odds ratio of ASB and ASE SNVs individually, via Fisher's exact tests. The number of asterisks depicts the degree of significance: *, 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 monoallelically expressed such as imprinted and MHC genes (CDS regions) are highly enriched for both ASB and ASE SNVs. The actual log odds ratio of T cell receptor genes for ASE SNVs and MHC genes for ASB and ASE SNVs are indicated on the bars.

Figure 3. A considerable fraction of AS variants are rare but do not form the majority. Lesser proportion of AS SNVs than non-AS SNVs are rare, suggesting less selective constraints in AS SNVs. The minor allele frequency (MAF) spectra of ASB (green filled circle), accessible non-ASB SNVs (green open circle), ASE (blue filled circle) and accessible non-ASE SNVs (blue open circle) are plotted at a bin size of 100. The peaks are in the bin for MAF < 0.5%.The inset zooms in on the histogram at MAF < 3%. Comparing ASE+ to ASE- gives an odds ratio of 0.67 (hypergeometric p < 2.2e-16), while comparing ASB+ to ASB-, gives an odds ratio of 0.96 (p=0.0021), signifying statistically significant depletion of AS variants relative to non-AS variants in both cases. This depletion suggests that AS SNVs are under less purifying selection.

Figure 4. Inheritance of allele-specific binding events is evident in some TFs but not so apparent in others. The top panel shows the legend for each plot. At the lower panel, the TFs

Formatted: Font: 12 pt, Font color: Auto

Formatted: Font: 12 pt

Formatted: Font: 12 pt, No underline, Font color: Auto

Deleted: chromosome

Formatted: Font: 12 pt, No underline, Font color: Auto

Formatted: Font: 12 pt, No underline, Font color: Auto

Formatted: Font: 12 pt, No underline, Font color: Auto

Deleted: . The database can also be queried and visualized directly as a track by the UCSC Genome Browser

Formatted: Font: 12 pt, No underline, Font color: Auto

Formatted: Font: 12 pt

Formatted: Font: 12 pt, No underline, Font color: Auto

Deleted: can

Formatted: Font: 12 pt, No underline, Font color: Auto

Deleted: genomic

Formatted: Font: 12 pt, No underline, Font color: Auto

Formatted: Font: 12 pt, No underline, Font color: Auto

Deleted: susceptibility to

Formatted: Font: 12 pt, No underline, Font color: Auto

Deleted: can

Formatted: Font: 12 pt, No underline, Font color: Auto

Deleted: olfactory receptor

Formatted: Font: 12 pt, No underline, Font color: Auto

Deleted: (>1.5); the

Formatted: Font: 12 pt, No underline, Font color: Auto

Formatted: Font: 12 pt

Formatted: Font: 12 pt, No underline, Font color: Auto

Deleted: The inset zoomed in on the histogram at MAF < 3%.

Formatted: Font: 12 pt

Formatted: Font: 12 pt, No underline, Font color: Auto

Deleted: Here

Formatted ...

CTCF (top row) and MYC (bottom row) are being examined for inheritance. For each TF, three plots compare two individuals in the CEU trio (Father: NA12891, Mother: NA12892, Daughter: NA12878), with the identity of the individual on the x-axis denoted by green and that on the y-axis by blue. Each point on the plot represents the allelic ratio of a common ASB SNV between the two individuals, by computing the proportion of reads mapping to the reference allele at that SNV, i.e. SNVs in the red quadrants (quadrants B and C in legend) signify that the allelic behavior is in the same direction in both individuals. The significance is statistically evaluated by the Bonferroni-corrected p value of a binomial test (under each plot). In CTCF (top row), there is an enrichment of points in quadrants B and C (red quadrants) versus A and D (grey quadrants) in parent-child comparisons (first 2 columns), with very significant p values. This signifies that inheritance of ASB is evident in CTCF. For parent-parent transmission (third column), both parents belong to the same ancestry, thus we expect ASB SNVs to be similar (B+C quadrants than in A+C quadrants), with a p-value lower than those of parent-child comparisons. They are unrelated, so there is also a lower number of common ASB SNVs between the parents. However, MYC (bottom row) shows the trend to a much lesser degree, with smaller number of overlap between parent and child and less deviation between quadrants B+C and A+D, as suggested by the lower significance of the p-values. For MYC, AS inheritance does not seem apparent.

Table 1.

Table 1 shows the breakdown of SNVs in each ethnic population: heterozygous (HET), accessible (ACC) and ASE SNVs in Table 1A and ASB SNVs in Table 1B. For each of the last 3 columns, each category of HET, ACC and AS SNVs is further stratified by the minor allele frequencies: common (MAF > 0.05), rare (MAF ≤ 0.01) and very rare (MAF ≤ 0.005). The number of AS 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.

Table 2.

Table 2 shows the top 10 protein-coding genes and non-coding categories, with the enrichment odds ratio and p-values (original and Bonferroni-corrected). The p-values are considered significant if < 0.05.

Supplementary Figures

Supplementary Figure 1

This figure shows the legend as per Figure 4 in the upper panel and the binomial test results for all eight DNA-binding proteins. CTCF, PU.1, SA1 and POL2 exhibit AS inheritance but MYC, RPB2 and PAX5 do not seem to have very apparent AS inheritance.

Supplementary Figures

Supplementary File 1

This file includes the Fisher's exact test odds ratios and p-values (original, Bonferroni-corrected, FDR-corrected) of a total of 973 categories: 952 non-coding categories from ENCODE and 20

Deleted: compares

Formatted: Font: 12 pt, No underline, Font color: Auto

Deleted: blue

Formatted: Font: 12 pt, No underline, Font color: Auto

Deleted: red

Formatted: Font: 12 pt, No underline, Font color: Auto

Formatted: Font: 12 pt, No underline, Font color: Auto

Deleted:

Formatted: Font: 12 pt, No underline, Font color: Auto

Deleted:

Formatted: Font: 12 pt, No underline, Font color: Auto

Formatted: Font: 12 pt

Formatted: Font: 12 pt, No underline, Font color: Auto

Formatted: Font: 12 pt

Formatted: Font: 12 pt, No underline, Font color: Auto

Formatted: Font: 12 pt

Formatted: Font: 12 pt, No underline, Font color: Auto

non-coding categories from GENCODE and enhancers (see Methods). The results for CDS regions and 6 monoallelically-expressed gene categories are also included.

Supplementary File 2

This file includes the Fisher's exact test odds ratios and p-values (original, Bonferroni-corrected, FDR-corrected) of a total of 20,144 protein-coding genes from GENCODE (See Methods).

Formatted: Font: 12 pt, Bold