# Allele-specific binding and expression variation in 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[1,3,4]**

[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

Gene expression is a complex trait that underlies many cellular phenotypes, thus it has many implications in understanding disease and trait etiology. Major advances in genome-wide functional assays such as ChIP-seq and RNA-seq have enabled the interrogation of genetic factors that might be causing changes in gene expression. Many studies have provided ChIP-seq and/or RNA-seq data of variable number of individuals from different ethnic populations, using various tools, parameters and algorithms. Here, we pool DNA sequences, RNA-seq and ChIP-seq data of 383 individuals from these disparate data sources and put them through a standardized pipeline to provide a resource and systematic survey for variants that assess allele-specific behavior, particularly in allele-specific binding (ASB) of transcription factors and allele-specific expression (ASE) of genes. We apply our *AlleleSeq* pipeline, which first uses the genomic variants from the 1000 Genomes Project to construct a diploid personal genome for each of the 383 individuals. Personal genome construction reduces reference and read-depth biases, which can increase false positives in the ensuing detection. We then directly integrate the individuals' corresponding ChIP-seq and RNA-seq datasets obtained from eight sources (notably the ENCODE and gEUVADIS projects), for allele-specific variant detection. Subsequently, we investigate the enrichment of ASB and ASE variants in more than 100 categories of annotated genomic elements, such as transcription factor binding sites, enhancers and promoters. We also perform population-based analyses of these individuals, showing intra- and inter-population differences. Further, using a family trio, we examine the parent-child inheritance of allele-specific behavior in binding and show that ASB is inherited among some transcription factors but not in others. The results are consolidated as a resource, AlleleDB, and serves as an allele-specific annotation of variants found in the 1000 Genomes Project catalog.

Deleted: AlleleDB:
Deleted: 380
Deleted: [7]
Deleted: Kong[1]
Deleted: ,[6]
Deleted: [3]Molecular
Deleted: Department
Deleted: [5]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA.
[6]Keck
Deleted: [7]Department
Deleted: AlleleDB: Allele-specific binding and expression variation in 380 individuals¶
¶
Jieming Chen[1,2], Joel Rozowsky[1,3], Jason Bedford[1], Arif Harmanci[1,3], Alexei Abyzov[1,3], Yong Kong[1], Robert Kitchen[1,3], Lynne Regan[1,2,3], Mark Gerstein[1,3,4]¶
¶
[1]Program in Computational Biology and Bioinformatics,¶
[2]Integrated Graduate Program in Physical and Engineering Biology,¶
[3]Molecular Biophysics and Biochemistry Department,¶
[4]Department of Computer Science,¶
Yale University, New Haven, CT 06520, USA¶
¶
¶
Formatted: Underline
Deleted: It is known to be highly variable among individuals.
Deleted: 380
Deleted: one particular class of cis-regulatory variants. These variants potentially modulate
Deleted: 380
Deleted: types/
Deleted: Section Break (Next Page)

## Introduction

The combined efforts of large sequencing projects such as the 1000 Genomes Project has provided the scientific community with a massive catalog of human genetic variants across multiple individuals from the human population.[cite] A substantial number of variants has been found to be very rare and also postulated to possess regulatory functions, especially in influencing gene expression. [cite GWAS, Encode].

Much of the characterization of cis- and trans-acting regulatory variation on gene expression has focused on the mapping of expression quantitative trait loci (eQTLs).[cite] eQTLs are detected based on correlations between the genotypes of a large group of individuals and their corresponding gene expression profiles. A more direct assessment of variant in association with cis-acting regulation can be provided by allele-specific (AS) approaches.[cite old and new papers] These approaches generally compare phenotypic differences between the alleles at heterozygous loci, so that each allele at these positions in a diploid genome acts as a within-sample control for the other allele. There are several immediate advantages for AS approaches with respect to eQTL mapping. First, for eQTLs to be detected, it has to exist in significant allele frequencies relative to the population and effect sizes.[cite] On the other hand, allele-specific behavior can be detected across any heterozygous site, regardless of its allele frequency, so very rare variants can be possibly detected. This is extremely important in light of the vast number of rare variants found to be present in the human population by the 1000 Genomes Project. Second, eQTL mapping is contingent on population size for sufficient statistics, while the allele-specific approach works for a single sample. This makes it an attractive strategy for biological samples such as primary cells and tissues that are difficult to obtain in large numbers. Third, in eQTL mapping, correlation is drawn between total expression measured between samples and their genotypes, that is, allele-insensitive. As such, trans-acting factors such as negative feedback mechanism that sought to reduce total expression variance across samples with different genotypes will not be detected. However, in an AS approach, a heterozygous site can be directly associated with a differential readout by comparing between the two different alleles (within one individual). Such a within-sample control in an AS approach also eliminates normalization issues across multiple assays, since factors that phenotypic differences between samples due to various environmental conditions are being controlled for.

Despite all the advantages, the utility of allele-specific approaches is heavily dependent on high-throughput technology. Early implementations of AS approaches employed microarray technologies, and thus are restricted to a subset of loci.[cite 2002 Yan] Recent progress in large-scale functional genomics assays, such as ChIP-seq and RNA-seq experiments, have facilitated genome-wide detection of candidate variants associated with allele-specific binding (ASB) and expression (ASE) events. However, previous studies have mostly been limited to very small sample sizes of single cell types.[cite] Nonetheless, they already demonstrated that allele-specific variants are widespread and can be tissue-specific.[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 samples and experimental assays (e.g. number of transcription factors for ChIP-seq experiments on a single sample), 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 distributed in at least 10 separate studies.[cite, ENCODE, kasowski] Aggregation of these datasets has obvious advantages in analyses, be it increasing statistical power or simply having more transcription factors for inter-sample comparisons. A huge caveat is that these studies are typically designed for various purposes, resulting in disparate sets of computational tools, strategies and threshold parameters used in the processing of data in

**Formatted:** Underline

**Formatted:** Font: Bold, Underline

**Deleted:** Until recently, interrogation of these variants has focused largely on 1.5% of the genome – the protein-coding regions. [cite] However, a

**Deleted:** mainly located in the remaining 98.5% of the genome - non-coding genome,

**Deleted:** are

**Deleted:** The functional annotation and interpretation of these variants become a huge challenge.

**Deleted:** variations influencing

**Deleted:** been

**Deleted:** ) mapping.[

**Deleted:** In this approach,

**Deleted:** the

**Deleted:** approaches. Early use of

**Deleted:** (AS) approaches

**Deleted:** found in the entire genome

**Deleted:** The advent of genome-wide

**Deleted:** genomic

**Deleted:** particularly

**Deleted:** facilitates

**Deleted:** Detection of allele-specific events is accomplished by directly comparing the read counts from these assays at heterozygous variants, so that each allele in a diploid genome acts as a within-sample control for the other allele. There are several immediate advantages: (1) this is a direct association between a heterozygous site and the observed differential phenotype; (2) the detection of rare variants is possible; (3) this eliminates normalization issues across multiple assays. ¶
¶
Here, we detect

**Deleted:** associated with allele-specific binding and expression from the

**Deleted:** lines of ~380

the respective studies. These portend that simple pooling of results from multiple studies may not be optimal, even for the same biological sample. Moreover, suitable datasets are scattered in the literature. Thus, the tasks of searching and then merging have 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 single comprehensive data corpus and repurpose it specifically for allele-specific analyses. In total, we reprocess 142 ChIP-seq and 475 RNA-seq datasets of 383 individuals in our uniform pipeline (Figure 1). Coupled with the construction of 383 personal genomes using variants from the 1000 Genomes Project, we use the tool that we previously developed, AlleleSeq,[cite] to detect single nucleotide variants (SNVs) associated with ASB and ASE events. We construct a database to house all the personal genomes and detected AS SNVs. Finally, using our consolidated data, we are able to present a systematic survey of these detected allele-specific SNVs in 382 unrelated individuals of seven ethnicities in various categories of genomic elements and to investigate the inheritance of allele-specific binding in eight different transcription factors in a Caucasian trio family.

## Results

### Candidate ASE and ASB variants in 383 individuals

There are several layers of information with respect to an individual that needs to be integrated in order to more accurately detect allele-specific 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 allele-specific expression (ASE) or binding (ASB) (Figure 1). Here, we implement a uniform pipeline (see Methods) to combine personal genomic, transcriptomic and binding data and to standardize our detection of potential 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. Table 1 shows the breakdown of the AS SNVs in various sub-populations: 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 Nigeria (YRI). YRI contributes the most to both ASE and ASB variants. Interestingly, while very rare AS SNVs comprise a substantial proportion in all populations, it is about two folds higher in the YRI (~48% ASE SNVs and ~34% ASB SNVs with MAF $\leq$ 5%) than the other European sub-populations of comparable (CEU, FIN) or larger (TSI) population sizes. Common SNVs (MAF > 5%) constitute the majority of the ASE and ASB variants in all populations. This is especially useful in functional annotation of variants via aggregation analyses [cite]. On average, in each person, ~0.1% of heterozygous sites (~2,000 SNVs) potentially tags for ASE and ~0.5% of heterozygous sites (~10,000 SNVs) for ASB (Supp Table). However, the number of AS variants detected is correlated, to some degree, with the pooled size of the functional genomic datasets available (ChIP-seq and RNA-seq) and the number of transcription factors being included for ASB, for each individual (do a correlation?).

### AlleleDB, a resource for allele-specific behavior in the human genome

We build a database, AlleleDB (http://alleledb.gersteinlab.org/), to house these putative allele-specific SNVs. AlleleDB can be queried directly, in terms of gene or genomic locations and it directly processes and displays the results as a UCSC track in the UCSC Genome browser (Figure 1). [cite] All heterozygous SNVs found in the stipulated query, including those not deemed as allele-specific but are 'accessible' (see Methods) by the pipeline, are color-coded (AS SNVs are red, others are black) in the displayed track. 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. Since many in the scientific community are familiar with the genome browser, we hope that this would increase the

**Deleted:** of various ethnicities. Together with their corresponding ChIP-seq and RNA-seq datasets from various sources [cite],

**Deleted:** Subsequently, we provide

**Deleted:** "

**Deleted:** " (AS)

**Deleted:** the context of a trio family and

**Deleted:** . Finally, we construct a database to house all the personal genomes and detected AS SNVs and provide it as a public online resource – AlleleDB. This also serves as the allele-specific annotation of variants found in the 1000 Genomes Project catalog.

**Deleted:** ¶

**Formatted:** Underline

**Formatted:** Font: Bold, Underline

**Deleted:** A wealth of data can be amassed from multiple sources. However, they are initially used for various purposes, and are subjected to different alignment and detection algorithms, and reference genomes, in their original publications.

**Deleted:** enforce

**Deleted:** integrate

**Deleted:** Table 1 shows the data provenance of the 380 unrelated individuals and 1 trio available in our AlleleDB.

**Moved (insertion) [1]**

accessibility and usage 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 [cite]. 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.

**Enrichment analyses**

Of great interest, is the annotation of these allele-specific SNVs in the human genome with respect to known genomic elements, both coding and non-coding. Only ~56% of our candidate ASE SNVs and ~6% of ASB SNVs are found in CDS. Using the low-coverage personal genomes of 382 unrelated individuals from Phase 1 of the 1000 Genomes Project and the 2 parents of the trio, we further investigate the enrichment (or depletion) of these AS SNVs in 954 categories of genomic elements, including gene annotations from GENCODE, and transcription binding motifs from ENCODE.[cite, Methods] The comparisons are performed against a 'matched' set of control SNVs, which we termed 'accessible' SNVs (see Methods). Accessible SNVs are heterozygous and possess at least the minimum number of reads that is detectable statistically to be allele-specific, but are not identified by AlleleSeq. Subsequently, we use the Fisher's exact test to estimate the odds ratios and p values of finding AS variants in these regions, relative to the expected odds provided by the control SNVs.

For categories more closely related to a gene structure, namely enhancers, promoters, coding DNA sequences (CDS), introns and untranslated regions (UTR) (Figure 2). In general, both categories of AS SNVs are more likely found in the 3' and 5' UTRs, strongly indicating allele-specific regulatory roles in these regions. [any lit evidence? For regulation? Regulatory role allelespecific] 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.[cite] For SNVs associated with allele-specific binding (ASB), we also observe an enrichment in the promoters. This is expected, since many TF binding motifs can be found near transcription start sites in the promoter regions to regulate gene expression. However, we observe variable enrichments of ASB SNVs in particular categories of binding sites for transcription factor families in promoter regions such as xx, xx and xx (Figure 2, Supp fig). These differences suggest that some TFs are more likely to participate in allele-specific regulation than others. Surprisingly, an enrichment of ASE, as well as, ASB SNVs is observed in CDS. Several studies have found that many transcription factors (TFs) bind in the protein-coding regions, for instance to regulate codon usage. [cite] More ASB SNVs found in these regions might suggest an allele-specific mechanism to such regulatory roles of the TFs. However, more experimental characterization would be required to determine if such allele-specific or differential binding (evidenced by ChIP-seq experiments) do exist and if so, whether it leads to any phenotypic differences at all. A recent paper found that many TFs do not elicit any change in gene expression when knocked out [cite 2014 paper by pritchard and gilad]

We also compute the enrichment of AS SNVs in various gene categories. Some of them have been known to be involved in monoallelic expression, namely (1) imprinted genes [cite], genes known to undergo allelic exclusion: (2) olfactory receptor genes [cite], (3) immunoglobulin, (4) genes associated with T cell receptors and the major histocompatibility complex [cite] and (5) a list of genes found to experience random monoallelic expression found in a study by Gimelbrant *et al*. [cite]. Expectantly, most of these have been found to be significantly enriched in ASE SNVs (except for olfactory receptors), especially when compared to the constitutively expressed housekeeping genes (Figure 2). For ASB SNVs, there is a depletion found in the olfactory genes but enrichment in imprinted and MHC genes, possibly indicating differences in allelic exclusion mechanisms and the role of transcription factors in allele-specific

**Moved (insertion) [2]**

**Deleted:** First, we

**Deleted:** from various sources, such as

**Deleted:** Using the low-coverage personal genomes of 380 unrelated individuals from Phase 1 of the 1000 Genomes Project, and the 2 parents of the trio, the

**Deleted:**

**Moved (insertion) [3]**

**Deleted:** we found a significant

**Deleted:** both the UTR regions but a depletion in the CDS. The slight enrichment of ASE SNVs in 3' UTR more

**Formatted:** Font color: Auto

**Formatted:** Font color: Auto

**Deleted:** and UTR regions

**Formatted:** Font color: Red

**Deleted:** none in introns

**Deleted:** also

**Deleted:**

**Deleted:** ,[

**Deleted:** ]

**Deleted:** (Supp Fig). Expectantly, these are found to be highly enriched for ASE SNVs. However, when we test for enrichment on

**Deleted:** (RMAE)

**Deleted:** ], we did not find any enrichment

**Deleted:** ASE SNVs

**Deleted:** was observed only

**Deleted:** . In addition, we examine AS SNVs in the context

**Deleted:** loss-

regulation of these genes [can we find any articles?]. Interestingly, a statistically significant enrichment is also observed in the housekeeping genes.

**Minor Allele Frequency (MAF) analyses**

To examine the occurrence of ASE and ASB SNVs in the human population, we plot the distributions of allele frequency of ASB and ASE SNVs, based on the population minor allele frequencies (MAF) from Phase 1 of the 1000 Genomes Project. In general, rare variants do not form the majority of all the AS variants. Nonetheless, we observe a skew towards very low allele frequencies, peaking at MAF < 0.5% in AS SNVs, compared to other categories of MAF (Figure 3). However, such 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 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.

**ASB Inheritance analyses using CEU trio**

The CEU trio is a well-studied family and particularly, many ChIP-seq experiments were performed on different transcription factors (TFs). This presents an opportunity to investigate the inheritance of allele-specific behavior using data from more TFs. While previous studies have also observed strong inheritance, the datasets are usually limited to a few TFs [cite McDaniel Kilpinen]. Using high-coverage genomes from the CEU family trio, we investigate the inheritance of allele-specific binding events in eight DNA-binding proteins (Figure 4 and Supp fig). Each plot compares two individuals in the trio. Each point in the plot is obtained by calculating the ratio of the number of reads that map to the reference allele and the total number of reads mapping to the site. High correlation between 2 individuals, i.e. significantly more points in the B and C quadrants, denotes great similarity in allelic directionality. Since we only look at SNVs in which at least 2 individuals are heterozygous, we are able to color each SNV accordingly; when 3 individuals are all heterozygous at that SNV (black), when 2 individuals are heterozygous and the third is homozygous for the reference allele (red) and when 2 heterozygous and the third is homozygous for the derived allele. There is no apparent clustering, signifying that there is no cryptic dependence on the genotype of the third individual. We observe high correlation between parent-and-child-only transmission (NA12878 versus NA12891 or NA12892) for CTCF (hypergeometric p=4.3e-8 and p=3.2e-8), PU.1 (p=xx) and POL2 (p=xx), implying AS inheritance. However, MYC (hypergeometric p=0.61 and p=0.4), PAX5, RPB2 and SA1 exhibit statistically insignificant enrichment of points in quadrants B and C. This shows that allele-specific behavior might not necessarily be apparent in some TFs. Also, the heritability of AS SNVs in the same allelic direction implies a sequence dependency in allele-specific behavior.

**Discussion**

It has been known that there is considerable inter-individual variability in gene regulation. [cite] Genetic variants associated with allele-specific regulation constitute a portion of cis-regulatory variation. As such, a systematic characterization of variants associated with both ASB and ASE on a large sample size is valuable in the understanding of gene regulation variability caused by allele-specific events. Previous studies have focused mainly on ASE when exploring allele-specific behavior, [cite a few papers] and investigating ASB has been limited to a few TFs, such as CTCF and Pol2. Here, we are able to re-use existing large ChIP-seq and RNA-seq datasets (that have been used for other purposes) solely for allele-specific analyses, without having to generate new ones.

A part of our strategy is the selection of individuals that are found in the 1000 Genomes Project. We distinguish samples by their ancestry and found that there are only 1 individual each for CHB and JPT. It

**Deleted:** -function variants, which are premature stop codons and splice sites. There is an evident overrepresentation of ASE SNVs in loss-of-function variants than by chance (odds ratio=1.5, p=3E-7), as noted as well in a previous study [cite]. This is also observed in the ASB SNVs but it is not

**Deleted:** (odds ratio, OR=1.7, p=0.06).

**Deleted:** Combining both

**Deleted:** in the 1KG

**Deleted:** (AF),

**Deleted:** AF<

**Deleted:** 005

**Deleted:** allele-specific

**Deleted:** +); comparing

**Deleted:** with

**Deleted:** OR

**Deleted:** xx with a

**Deleted:** -value=xx,

**Deleted:** and ASE+,

**Deleted:** OR xx

**Deleted:** xx).

**Deleted:** Using the

**Deleted:** individuals of one family of trio from Utah with northwestern Caucasian ancestry (CEU),

**Deleted:** ). This is performed

**Deleted:** mapped

**Deleted:** ,

**Formatted:** Font color: Red

**Formatted:** Font color: Red

**Deleted:** in Figure

**Deleted:** Spearman correlation, rho=xx,

**Formatted:** Font color: Auto

**Deleted:** xx

**Deleted:** rho=xx,

**Deleted:** rho=xx,

**Deleted:** MYC,

**Deleted:** low correlations (rho=xx, p=xx).

**Deleted:** ¶

**Formatted:** Font: Not Bold

**Formatted:** Underline

**Deleted:** Previous studies have explored allele- ...

**Deleted:** Using detected AS SNVs and a contro ...

**Deleted:** non-AS SNVs, we set out to survey A ...

**Deleted:** integrative nature of allele-specific (A ...

**Deleted:** issues associated with

could be that we did not scour the literature hard enough for ChIP-seq and RNA-seq data on non-Caucasian and non-African samples, or it could reflect the lack of research, in general, in specific ethnic groups. Since many AS variants have been found to be rare, it is of great interest that more samples of diverse ancestries be represented in a dataset.

Enrichment tests are an important part of our analyses. They are highly dependent on the choice of the null expectation (controls). We intentionally choose a set of control SNVs matched by both heterozygosity and statistical accessibility. Our enrichment analyses also emphasize on relating allele-specific activity to known genomic elements, such as CDS and various non-coding regions. Together, these aid in the enduring effort in characterization of genomic variants on two levels: firstly, at the single nucleotide level, our detected AS SNVs can serve as an annotation to the 1000 Genomes Project variant catalog in terms of allele-specific cis-regulation; secondly, by associating AS SNVs with elements such as promoters and enhancers, we might be able to define genomic regions, with more allele-specific activity.

In our analyses, enrichment tests are also performed on rare variants, defined by minor allele frequencies in the human population (from the 1000 Genomes Project). This has been shown to be a considerable indicator for negative selection, where conserved, and probably functional, regions are shown to have a very high fraction of rare variants. [cite] Our results show lower enrichment of rare variants in AS SNVs than non-AS SNVs. This suggests that, as a whole, they are under less selective constraints than non-AS SNVs. This was also noted in previous studies using only a high-coverage single individual [cite]. A weaker selection may also account for more toleration to varying gene expression profiles across individuals. In addition, a substantial number of rare SNVs (MAF $\leq$ 0.5%) among the AS SNVs (ASB and ASE) denotes that these SNVs seldom overlap across individuals. As we increase sample size, the number of rare and private SNVs increases dramatically as well. While there is utility of common SNVs in identifying single nucleotide regulatory sites via aggregation analyses, the existence of so many rare variants recurring in the same genomic region might imply that a more element- or region-based approach might be appropriate to better investigate allele-specific behavior.

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 of heterogeneity in various standards and parameters etc. However, there are still several concerns. First, the integrative nature of allele-specific approach means that AS variant detection is highly sensitive to sequencing errors in heterozygous SNV calling, as well as biases and issues associated with ChIP-seq and RNA-seq analyses. Second, a certain property of the dataset that was not discussed because of its extant uniformity in the current collection, is tissue specificity. Here, our AS SNVs are all detected in lymphoblastoid cell lines (LCLs). Most genomic sequences and functional genomic datasets in current 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. [cite] More extensive projects are already underway to involve functional assays and sequencing in other tissues and cell lines, such as GTex [cite] and ENCODE [cite]. Further, more accurate allelic information is also being achieved with the advent of phasing information in next-generation sequencing, [cite phased HeLa, Shendure, Church, Illumina]. In light of these datasets and new technologies, AlleleDB is intended as a scalable public resource, so that as technology evolves and more experimental data becomes available, new samples (of potentially varied ancestries), tissue and cell types can be similarly processed and the detected allele-specific SNVs can be included in this central repository. 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**

**Deleted:** To alleviate these issues in enrichment tests, we

**Deleted:** (please refer to the Methods section).

**Moved up [3]:** [any lit evidence? For regulation?

**Deleted:** Generally, within the context of a gene structure, the UTRs and promoters (for ASB only) seem to demonstrate an enrichment of allele-specific regulation, while intronic regions seems to exhibi ...

**Deleted:** Regulatory role allelespecific] A cave ...

**Deleted:** , when compared to non-AS variants ...

**Deleted:** such as the CDS

**Deleted:** AS regions

**Deleted:** regions

**Deleted:** our

**Deleted:** majority of rare SNVs (allele frequency <

**Formatted:** Font color: Red

**Deleted:** also

**Formatted:** Font color: Red

**Deleted:** This might account for

**Formatted:** Font color: Red

**Deleted:** wide range

**Formatted:** Font color: Red

**Deleted:** inter-individual variability in gene ...

**Deleted:** are typically located within the same ...

**Formatted:** Font color: Red

**Deleted:** detected

**Formatted:** Font color: Red

**Deleted:** but their recurrence

**Formatted:** Font color: Red

**Deleted:** Also, we examine the sharing of rare ...

**Deleted:** Given a family of trio, we can explore ...

**Deleted:** strong heritability of ASB events in a ...

**Deleted:** ), which most

**Deleted:** . It has

**Deleted:** . By consolidating data and results fro ...

**Deleted:** are being performed

**Deleted:** in projects

**Deleted:** ] and

**Deleted:** ¶ ...

**Deleted:** (http://alleledb.gersteinlab.org/), to ho ...

**Deleted:** queried directly, in terms of gene or ...

**Moved up [1]:** This enables cross-referencing ...

**Deleted:** We also provide the raw counts for ea ...

**Moved up [2]:** AlleleDB also serves as an ...

**Construction of diploid personal genomes**

There are 380 low-coverage (average depth of 2.2 to 24.8) unrelated genomes and 3 high-coverage genomes from the CEPH 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*, which is part of the AlleleSeq suite. [cite] 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 a family of trio (CEU), 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. These personal genomes are provided in AlleleDB.

CNV genotyping is also performed for each genome by CNVnator.[cite] This is achieved by calculating the average read depth within a window, normalized to the genomic average for the region of the same length. For each low-coverage genome, a bin size of 1000 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 detection. Read depths for each heterozygous SNV in each genome are also provided in AlleleDB.

**Deleted:** Complete genomics data for YRI trio¶
¶

**Allele-specific SNV detection**

AS SNV detection is 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 highly dependent on the size 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 are 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; we then determine the peaks by performing PeakSeq using the control and unmapped reads provided 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 varies with the pooled size (coverage) of the ChIP-seq or RNA-seq dataset. For a larger dataset, the binomial p-value threshold is lower, making the minimum number of reads (n) that will produce the corresponding p-value, larger. We can obtain this from a table of a two-tailed binomial probability density function computed for each n

and each number of successes (Supp Figure). Enrichment analyses are performed using the Fisher's exact test. P-values are considered significant if < 0.05 (denoted by asterisks in the figures).

SNV/Gene lists data provenance: (1) GENCODE; (2) ENCODE; (3) genes for random monoallelic expression from Gimelbrant *et. al.* [cite] (4) olfactory receptor gene list from [cite]; (5) immunoglobulin, T cell receptor and MHC gene list from [cite]; (6) loss-of-function variants from 1000 Genomes Project as annotated by Variation Annotation Tool [cite]. Enhancer lists are also obtained from data at http://www.ebi.ac.uk/~swilder/Superclustering/concordances4/ [Hoffman et al, NAR, 2013; Hoffman et al, Nature Methods, 2013; and Ernst and Kellis, Nature Methods, 2012] and http://encodenets.gersteinlab.org/metatracks/ (described in Yip et al, Genome Biol, 2012). Promoter regions are set as 2kbp upstream of all transcripts annotated by GENCODE. Transcription factor motifs and peaks are obtained from TRANSFAC.[cite]

**AS inheritance analyses**

We compute the proportion of reads that align to the reference allele with respect to the total number of reads mapped to 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, and (2) when an AS SNV is heterozygous in two individuals and homozygous (reference or alternate) in the third. Each SNV is colored black when 3 individuals are all heterozygous at that SNV, red when 2 individuals are heterozygous and the third is homozygous for the reference allele and blue when 2 heterozygous and the third is homozygous for the derived allele. Allele-specific inheritance is then determined by Spearman correlation of the proportion of reads between the 2 individuals.

**Acknowledgements**

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