Allele-specific binding and expression variation in 383 individuals
Processing and detecting allele-specific binding and expression variation in 383 individuals
Allele-specific binding and expression variant detection in 383 individuals
Survey of allele-specific binding and expression variation in 383 individuals
Survey of allele-specific variation in 383 individuals
Survey of allelic 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[1,2,3,4]

[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

Allele-specific behavior is exhibited when there is a differential phenotypic effect between the two alleles in a diploid genome. Genomic variants associated with allele-specific events constitute an important class of regulatory variation, thus it is extremely useful to annotate them on a large scale. To detect them, we can overlap genomic variants with regions of allelic imbalance identified by large-scale functional genomic assays, such as ChIP-seq and RNA-seq. Previous studies on allele-specific analyses are based 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 aggregate allele-specific variants across these multiple studies. In doing so, we gain not only statistical power, but also a wider range of analyses that can be performed. Unfortunately, because of the integrative nature of allele-specific variant detection, it is sensitive to various technical issues such as heterozygous variant detection and read mapping. Therefore, simply pooling the results of these disparate studies is not desirable. Moreover, different studies use various tools, parameters and thresholds. Henceforth, it is imperative to subject each dataset to uniform processing. To this end, we pool DNA sequences, RNA-seq and ChIP-seq data of 383 individuals from these separate data sources and put them through a standardized processing pipeline. 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. This allows us to define ~100 genomic regions that are enriched or depleted in these allelic variants, thereby ascertaining parts of the genome that might be more susceptible to functional changes due to sequences. We also perform population-based analyses of these individuals, showing intra- and inter-population differences. All the results are consolidated in a resource, AlleleDB.

Deleted: Gene expression
Deleted: complex trait that underlies many cellular phenotypes
Deleted: has many implications in understanding disease and trait etiology. Major advances in genome-wide
Deleted: have enabled the interrogation of genetic factors that might be causing changes in gene expression. Many
Deleted: have provided ChIP-seq and/
Deleted: RNA-seq data of variable number
Deleted: from
Deleted: ethnic populations, using
Deleted: algorithms. Here, we
Deleted: disparate
Deleted: to provide a resource and systematic survey for
Deleted: behavior, particularly in allele-specific
Deleted: transcription
Deleted: . 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.
Deleted: 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
Deleted: as
Deleted: , and serves as an allele-specific annotation of variants found in the 1000 Genomes Project catalog
Formatted: Font: Bold
Formatted: Font: Bold, Underline
Deleted: ¶
--Page Break--
¶

Introduction

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 moderately rare (~ 58% are of population allele frequency < 0.5%).[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 huge cohort is required in order to achieve statistical power to detect variants of low frequencies, thus is constrained in their ability to detect very rare variants. 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. Henceforth, 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 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 (TFs) 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 TFs for more inter-sample comparisons.

Unfortunately, because several layers of data are integrated in AS variant detection, it is extremely sensitive to the technical details of issues such as heterozygous variant calling and read mapping. 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 the respective studies. These portend that a simple pooling of results from multiple studies may not be optimal, even for the same biological sample. Further, 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 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 detect more than 168K and 143K single nucleotide variants (SNVs) associated with ASB and ASE events respectively. We construct a database to house all the personal genomes and detected AS SNVs.

Deleted: The combined efforts of

Deleted: has

Deleted: across multiple individuals from the human

Formatted: Font color: Red

Deleted: .[cite] A substantial number of

Deleted: has been found to be very rare and also postulated to possess regulatory functions, especially in influencing gene expression. [cite GWAS, Encode]

Deleted: 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.

Moved down [1]: 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 ...

Deleted: This is extremely important in light of the vast number of rare variants found to be present in ...

Deleted: Despite all the advantages, the utility of allele-specific approaches is heavily dependent on ...

Deleted: Recent progress in large-scale functional genomics assays, such as

Deleted: , have facilitated

Deleted: detection

Deleted: candidate

Deleted: associated with allele-specific binding (ASB) and expression (ASE) events. However, ...

Deleted: been

Deleted: very small sample sizes of

Deleted: cell types.[cite] Nonetheless, they already demonstrated that allele-specific variants are ...

Deleted: ¶
¶

Deleted: transcription factors

Deleted: A huge caveat

Deleted: that these

Deleted: Moreover

Deleted: single

Deleted: use the tool that we previously developed, AlleleSeq,[cite] to

Finally, using our consolidated data, we are able to present a systematic and unbiased 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

AlleleDB, a resource for allele-specific behavior genome annotation

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 443,316 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 intentionally choose a set of control SNVs, which we termed 'accessible' SNVs. These SNVs are heterozygous and possess at least the minimum number of reads (for each dataset) that is detectable statistically but are not identified to be allele-specific (Table 1). In other words, these controls matched by both heterozygosity and statistical accessibility to the allele-specific variants. Altogether, we identified 665,860 and 409,708 accessible SNVs for ASE and ASB SNVs respectively.

We build a database, AlleleDB (http://alleledb.gersteinlab.org/), to house the candidate allele-specific and accessible SNVs. AlleleDB can be downloaded as flat files or queried and visualized directly, in terms of gene or genomic locations, as a UCSC track in the UCSC Genome browser (Figure 1). [cite] 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. All heterozygous SNVs found in the stipulated query genomic region, including accessible SNVs, are color-coded (AS SNVs are red, others 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. Only ~56% of our candidate ASE SNVs and ~6% of ASB SNVs are found in coding DNA sequences (CDS). Using the AlleleDB variants found in 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 with respect to the control set of accessible SNVs within the regions tested. 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.

Figure 2 shows the enrichment of elements more closely related to a gene structure, namely enhancers, promoters, CDS, introns and untranslated regions (UTR). In general, both categories of AS SNVs are more likely found in the 3' and 5' UTRs, suggesting 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

Deleted: **Candidate ASE and ASB variants in 383 individuals¶**

Deleted: ¶ ¶

Deleted: 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

Moved down [2]: . 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 ≤ 5%) than the other European sub-populations of comparable (CEU, FIN) or larger (TSI) population sizes.

Moved down [3]: Common SNVs (MAF > 5%) constitute the majority of the ASE and ASB variants in all populations.

Deleted: 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¶**

Deleted: these putative

Deleted: and it directly processes and displays the results

Deleted: 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.

Moved down [4]: Since many in the scientific community are familiar with the genome browser, we hope that this would increase the accessibility and usage of AlleleDB. The query results are also ...

Deleted: in the human genome

Deleted: .

Deleted: against a 'matched' set of control SNVs, which we termed 'accessible' SNVs (see Methods). Accessible SNVs are heterozygous and possess at ...

Deleted: For categories

Deleted: coding DNA sequences (

Deleted:),

Deleted:) (Figure 2

Deleted: strongly indicating

bias.[cite] 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 observe variable enrichments of ASB SNVs in particular categories of binding sites for TF families in promoter regions such as xx, xx and xx (Figure 2, Supp fig). 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. Several studies have found that many TFs bind in the protein-coding regions, for instance to regulate codon usage. [cite, Supp table] More ASB SNVs found in these regions might suggest an allele-specific mechanism to such regulatory roles of the TFs.

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], and three sets of 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]. (5) A list of genes found to experience random monoallelic expression found in a study by Gimelbrant *et al* is also included. [cite]. Expectantly, 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).

Allele frequency analyses

To examine the occurrence of ASE and ASB SNVs in the human population, we consider the population minor allele frequencies (MAF) from Phase 1 of the 1000 Genomes Project. Table 1 shows the breakdown of the AS SNVs in seven sub-populations and some MAF categories. YRI contributes the most to both ASE and ASB variants at each allele frequency category. 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 ≤ 5%) than the other European sub-populations of comparable (CEU, FIN) or larger (TSI) population sizes.

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.

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 (Supp Table); for ASB, it is highly dependent on the chosen TFs.

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. Unifying these studies and pooling the data 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 variants derived from high-coverage genomes of the CEU family trio, we investigate the inheritance of allele-specific binding events in eight DNA-binding proteins (Figure 4 and Supp fig). 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 boxes in Figure 4) compared to the A and D quadrants (grey boxes in Figure 4), denoting great similarity in allelic directionality (binomial p=5.7e-48 and p=2.0e-54). The inheritance of AS SNVs in the same allelic direction from parent to child implies a sequence dependency in allele-

Deleted: . This is expected, since many

Deleted: can be

Deleted: transcription factor

Formatted: Font color: Auto

Deleted: suggest

Deleted: Surprisingly, an enrichment

Deleted: is

Deleted: transcription factors (

Deleted:)

Moved down [5]: 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]

Deleted:] and

Deleted: a

Deleted: .

Deleted: most of

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

Deleted: ¶
¶
Minor

Deleted: **Frequency (MAF)**

Deleted: plot the distributions of allele frequency of ASB and ASE SNVs, based on

Moved (insertion) [2]

Moved (insertion) [3]

Deleted: experiments

Deleted: transcription factors (

Deleted:). This

Deleted: from

Deleted: Each plot compares two individuals in

Deleted: 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

Deleted: between 2 individuals

Deleted: , denotes

Deleted: . 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 ...

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 genetic similarity of the same population and the sequence heritability of AS behavior. Besides CTCF, PU.1 (p=xx) and POL2 (p=xx) also show AS inheritance. On the contrary, MYC (binomial p=8.2e-5 and p=1.1e-7), PAX5 (p=xx), RPB2 (p=xx) and SA1 (p=xx) exhibit enrichment of points in quadrants B and C with very much lower statistical significance, indicating that AS inheritance is not as apparent in some TFs – inheritance of AS behavior may not be a universal phenomenon.

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. Research on regulatory variants so far has also focused on eQTL mapping and consequently common variants.

There are several immediate advantages of AS approaches compared 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 two aspects: the vast number of rare variants found to be present in the human population by the 1000 Genomes Project and a substantial proportion of AS SNVs are very rare variants, as observed in our study and others.[cite] These make allele-specific variant detection a valuable asset in annotating cis-regulatory variants in personal genomes. Second, 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. Third, 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.

Despite its ability detect AS variants in just a single diploid genome, the enrichment of rare variants and considerable inter-individual variability in gene regulation provide impetus for larger sample size, especially in the budding field of personal genomics. [cite] This is because as more personal genomes are being sequenced, more rare and private SNVs need to be annotated. Larger sample sizes allow inter-sample comparisons and more genomes to be annotated properly. 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. Here, we have devised a processing pipeline to capitalize on 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.

Limited by the availability of personal genomes with ChIP-seq and RNA-seq data, a part of our strategy is the selection of individuals that are found in the 1000 Genomes Project. When we distinguish samples by their ancestry, we found that there is only 1 individual each for CHB and JPT. It could be a strong reflection on the lack of large-scale functional genomics assays in specific ethnic groups – concerns echoed by many other studies.[cite Bustamente review on research diversity] Since many AS variants have been found to be rare, it is of great interest and importance that more samples of diverse ancestries be represented in a dataset.

Deleted: 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

Formatted: Font color: Auto

Deleted: 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),

Deleted:), implying

Deleted: However

Deleted: hypergeometric

Deleted: 0.61

Deleted: 0.4

Deleted: ,

Deleted: statistically insignificant

Deleted: . This shows that allele-specific behavior might

Deleted: necessarily be

Deleted: . Also, the heritability of AS SNVs in the same allelic direction implies a sequence dependency in allele-specific behavior.

Deleted: ¶

Deleted: As such, a systematic characterization of

Moved (insertion) [1]

Deleted: both ASB and ASE on a large

Deleted: size is valuable in the understanding of

Deleted: variability caused by allele-specific events

Deleted: ASE when exploring allele-specific behavior,

Deleted: and

Deleted: TFs

Deleted: are able to re-use

Formatted: Font: Bold, Underline

Deleted: A

Deleted: We

Deleted: and

Deleted: are

Deleted: that we did not scour the literature hard enough for ChIP-seq and RNA-seq data

Deleted: non-Caucasian and non-African samples, or it could reflect

Deleted: research, in general,

Deleted: .

Our enrichment analyses 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, more susceptible to allele-specific activity. For instance, we found enrichments of ASB SNVs in many regions such as promoter and UTR regions, which might suggest an allele-specific mechanism to regulatory roles of the TFs. We also observe a depletion in ASB SNVs 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 regulation of these genes [can we find any articles?]. Interestingly, a statistically significant enrichment is also observed in the housekeeping genes. 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].

In our analyses, we also assess the enrichment of 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, AS SNVs 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, there is a substantial number of rare SNVs with MAF ≤ 0.5% (~17% in ASE and ~5% in ASB, Figure 3) among the AS SNVs (ASB and ASE); these will be inaccessible to eQTL mapping.

The final data and results are centralized in AlleleDB, which plugs into 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 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.

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 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 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 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.[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,

Deleted: 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

Moved (insertion) [5]

Deleted: with more allele-specific activity.

Deleted: tests are also performed on

Deleted: . [

Deleted: they

Formatted: Font color: Auto

Deleted: (

Formatted: Font color: Auto

Deleted: %)

Formatted: Font color: Auto

Formatted: Font color: Auto

Deleted:) denotes that

Formatted: Font color: Auto

Deleted: 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

Formatted: Font color: Auto

Deleted: appropriate

Formatted: Font color: Auto

Deleted: better investigate allele-specific behavior

Formatted: Font color: Auto

Moved (insertion) [4]

Deleted: We

Deleted: extant

Deleted: ,

Deleted: current

Shendure, Church, Illumina]. In light of these datasets and new technologies, AlleleDB is intended as a scalable resource. As technology evolves, more personal genomes are sequenced with more functional genomics data becoming available, this study provides an infrastructure that can accommodate new samples (of potentially varied ancestries), tissue and cell types, which can be similarly processed. 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

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]

Construction of diploid personal genomes

There are a total of 383 genomes used in this study: 382 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 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*, [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.

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

Deleted: public

Deleted: , so that as

Deleted: and

Deleted: experimental

Deleted: becomes

Deleted: and the detected allele-specific SNVs can be included in this central repository

Deleted: ¶
¶

Formatted: Underline

Formatted: Font: Bold

Moved (insertion) [6]

Formatted: Font: Bold, Font color: Auto

Formatted: Font: Bold

Deleted: 380

Deleted: unrelated genomes

Deleted: CEPH

Deleted: , which is part of the AlleleSeq suite

Formatted: Font color: Red

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 Figure 2).

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

Acknowledgements

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

Deleted: the figures

Moved up [6]: 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]¶
¶

Formatted: Font: Bold, Font color: Auto

Formatted: Font: Bold

Deleted: 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