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

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.

Deleted: specific binding

Deleted: expression variation in

Deleted: individuals

Deleted: 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¶

Deleted: Allele-

Deleted: there is

Deleted: Genomic variants associated with allele-specific events constitute

Deleted: 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

Deleted: identified by large-scale

Deleted: , such as ChIP-seq and RNA-seq

Deleted: studies on

Deleted: are

Deleted: aggregate

Deleted: across these multiple studies. In doing so, we gain not only statistical power, but also a wider range of analyses that can be performed.

Deleted: because of the integrative nature of

Deleted: variant

Deleted: , it

Deleted: Therefore

Deleted: desirable

Deleted: Henceforth

Deleted: of

Deleted: these separate data sources

Deleted: 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

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 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 more inter-individual comparisons (such as TFs and populations).

~~Unfortunately~~, allele-specific variant detection is extremely sensitive to the technical issues from SNV calling and RNA-seq and ChIP-seq experiments, 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. Finally, using our consolidated data, we are able to present a systematic and unbiased survey of these

Deleted: moderately rare (~ 58% are of population allele frequency < 0.5%).[

Deleted: huge

Deleted: thus is constrained

Deleted: their ability to detect very rare

Deleted: Henceforth

Deleted: samples

Deleted: sample

Deleted: at least 10 separate

Deleted:] Aggregation

Deleted: TFs

Deleted: sample

Deleted: .

Deleted: because several layers of data are integrated in AS

Deleted: , it

Deleted: details of

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 two 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. 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 SNVs are a set of SNVs not identified to be allele-specific but is derived from the accessible SNVs. In other words, these controls are matched by both heterozygosity and statistical accessibility to the allele-specific variants. Altogether, we identified a total of 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 personal genomes of 380 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 ~20,000 CDS regions of protein-coding genes from GENCODE and 953 categories of non-coding genomic elements, including annotations from GENCODE, DNaseI hypersensitivity sites [cite], enhancers [cite] and transcription binding motifs from ENCODE.[cite, Methods] The comparisons are performed using Fisher's exact test with respect to the control set of accessible SNVs within the regions tested. Overall, for ASB SNVs, we observe statistical significance ($p \leq 0.05$) for 787 non-coding categories and 15 genes and for ASE SNVs, 598 non-coding categories and 831 genes.

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 5' and 3' 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

Deleted: several

Deleted: intentionally choose

Deleted: control SNVs, which we termed

Deleted: . These SNVs

Deleted: (

Deleted:) that is detectable statistically but are

Deleted: (Table 1).

Deleted: low-coverage

Deleted: 382

Deleted: 954

Deleted: gene

Deleted: Subsequently

Deleted: use the Fisher's exact test to estimate the odds ratios and

Deleted: values of finding AS variants in these regions, relative to the expected odds provided by the control SNVs.

Deleted: 3'

Deleted: 5'

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, 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 of particular TFs in promoter regions such as RPB2 and SA1, while depletion in others, such as PU.1 and POL2 (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 (MAE), 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 in a study by Gimelbrant *et al* is also included. [cite]. As expected, all of these MAE gene sets have been found to be significantly enriched in ASE SNVs, especially when compared to the constitutively expressed housekeeping genes, which show no enrichment at all (Figure 2). We also see that the overall enrichment of CDS regions in all genes (Figure 2) is largely a consequence of highly enriched MAE gene sets (Supp fig). However, in terms of ASB SNVs, almost all the MAE gene categories are enriched except for the olfactory receptor genes. This might imply that allele-specific expression is coordinated by allele-specific binding events at most MAE loci, while at the olfactory receptor genes, another mechanism might predominate for allele-specific expression. It could also be due to missing TFs from our dataset that might be coordinating the allele-specific expression. Also interestingly, a statistically significant enrichment of ASB SNVs is observed in the housekeeping genes, without resulting in any allele-specific expression.

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 accessible and AS SNVs in seven ethnic populations and allele frequencies. YRI (Yoruba from Nigeria) contributes the most to both ASE and ASB variants at each allele frequency category. While very rare AS SNVs with MAF ≤ 0.005 comprise a considerable proportion in each population, 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 in AS SNVs, peaking at MAF ≤ 0.5%, compared to other MAF categories (Figure 3). 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 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]. In the high-coverage personal genome of NA12878, 2% of heterozygous sites (~2,000 SNVs) potentially tag for ASE (Supp Table) and 1%.

ASB Inheritance analyses using CEU trio

Deleted: in
Deleted: categories of binding sites for TF families
Deleted: xx, xx
Deleted: xx
Formatted: Font color: Auto
Deleted: ,
Deleted: found
Deleted: Expectantly,
Deleted: (except for olfactory receptors),
Deleted: sub-
Deleted: some MAF categories. YRI
Deleted: Interestingly, while
Deleted: substantial
Deleted: all populations
Deleted: % in AS SNVs,
Deleted: of MAF
Deleted: On average, in each person, ~0.1
Formatted: Font color: Red
Deleted: tags
Deleted:); for ASB, it is highly dependent on the chosen TFs.

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-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 (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. AS approaches provide several advantages that can be complementary 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. 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 sizeable proportion of AS SNVs that are rare variants, as observed in our study and others.[cite] Second, in eQTL mapping, correlation is drawn between total expression measured between individuals and their genotypes, that is, allele-insensitive. As such, effects from 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, difference in expression between alleles can still be detected. Such a within-individual control in an AS approach also eliminates normalization issues across multiple assays, since factors that phenotypic differences between individuals 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 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.

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.

Deleted: genetic similarity of the same population and

Formatted: Tab stops: 2.25", Left

Deleted: ¶
¶
There are

Deleted: immediate

Deleted: of AS approaches compared

Deleted: , so very rare variants can be possibly detected.

Deleted: substantial

Deleted: very

Deleted: cite] These make allele-specific variant detection a valuable asset in annotating cis-regulatory variants in personal genomes.

Deleted: samples

Deleted: trans-acting

Deleted: samples

Deleted: a heterozygous site can be directly associated with a differential readout by comparing

Deleted: the two different

Deleted: (within one individual).

Deleted: sample

Deleted: samples

Deleted: sample.

Deleted: Despite its ability detect

Deleted: variants in just

Deleted: single diploid

Deleted: provide impetus for larger sample size, especially in the budding field of personal genomics. [cite] This

Deleted: because as

Deleted: are being sequenced,

Deleted: rare and

Deleted: need to be annotated. Larger sample sizes

Deleted: sample

Deleted: and more genomes to be annotated properly

Deleted: Pol2. Here, we have devised

Deleted: processing

Deleted: large

Deleted: (

Deleted: purposes) solely

A 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.[cite Bustamente review on research diversity] 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 elements, such as CDS and various non-coding regions. Together, these aid in the enduring effort of characterizing genomic variants on two levels: firstly, at the single nucleotide level, where 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 gene models such as promoters and enhancers, we might be able to define categories of genomic regions more susceptible to allele-specific activity. A comparison between enrichment of ASB and ASE SNVs in the same category of genomic region can also help shed light on coordination of regulation by allele-specific TF binding and allele-specific expression in genes. 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 actually 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). There is a substantial number of very rare AS SNVs with MAF ≤ 0.5% (~17% in ASE and ~5% in ASB, Figure 3). This group of SNVs is inaccessible to eQTL mapping and the number is expected to increase with more personal genomes. Also, the enrichment of rare variants has been shown to be a considerable indicator for negative selection, where conserved, and probably functional, regions are shown to possess an enrichment of rare variants.[cite] Even though rare variants constitutes a considerable proportion in each AS and non-AS allele frequency spectrum, 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. This observation was also noted in previous studies using only a high-coverage single individual [cite encodenets, alleleseq]. A weaker selection may also account for more toleration to varying gene expression profiles across individuals.

The final data and results are centralized in AlleleDB, which conveniently 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 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 [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, our current catalog of AS SNVs are detected in 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.[cite] More extensive projects are already underway to involve functional assays and sequencing in other tissues and cell lines, such as GTex [cite]

Deleted: Limited by the availability

Deleted: , a part of our strategy is the selection of individuals that are found

Deleted: the 1000 Genomes Project. When we distinguish samples by their ancestry, we found that there is only 1 individual each for CHB

Deleted: JPT

Deleted: concerns

Deleted: by many other studies.[

Deleted: Since

Deleted: samples

Deleted: in a dataset

Deleted: enrichment

Deleted: emphasize

Deleted: in characterization of

Deleted: elements

Deleted: ,

Deleted: For instance, we found enrichments

Deleted: many regions such as promoter and UTR regions, which might suggest an

Deleted: mechanism to regulatory roles of the TFs. We also observe a depletion in ASB SNVs in the olfactory genes but enrichment in imprinted

Deleted: MHC genes, possibly indicating differences in allelic exclusion mechanisms and the role of transcription factors in

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

Formatted: Font color: Auto

Deleted: This

Deleted: have a very high fraction

Deleted: Our

Deleted: than

Deleted: suggests

Deleted: 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.

Deleted: usage

Deleted: 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 certai...

Deleted: collection is tissue specificity. Here, our

Deleted: all

Deleted:). Most

and ENCODE [cite]. 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, [cite phased HeLa, Shendure, Church, Illumina, PacBio]. In light of the evolution of technology, and the availability of more personal genomes and functional genomics data, AlleleDB is intended as a scalable resource. This study provides an infrastructure that can accommodate new individual genomes (of potentially diverse 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

Categories of genomic regions are obtained from (1) GENCODE and (2) ENCODE, (3) Genes for random monoallelic expression are from Gimelbrant *et. al*. [cite] (4) The olfactory receptor gene list is from [cite]; (5) immunoglobulin, T cell receptor and MHC gene lists are from [cite]. Enhancer lists are 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: 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 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 the family of CEU trio, for each child's genome, these haploid genomes become the maternal and paternal genomes, since the parental genotypes are known. Subsequently, at a heterozygous locus in the child's genome, if at least one of the parents has a homozygous genotype, the parental allele can be known. However, for each of the genomes of the 380 unrelated individuals, the alleles, though phased, are of unknown parental origin.

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.

Allele-specific SNV detection

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.

Deleted: information
Deleted: these datasets and new technologies, AlleleDB is intended as a scalable resource. As
Deleted: evolves,
Deleted: are sequenced with more
Deleted: becoming available, this
Deleted: samples
Deleted: varied
Deleted: SNV/Gene lists data provenance:
Formatted: Font color: Auto
Deleted: ;
Formatted: Font color: Auto
Deleted: ;
Formatted: Font color: Auto
Deleted: genes
Formatted: Font color: Auto
Formatted: Font color: Auto
Formatted: Font color: Auto
Formatted: Font color: Auto
Deleted: list
Formatted: Font color: Auto
Deleted:]; (6) loss-of-function variants from 1000 Genomes Project as annotated by Variation Annotation Tool [cite].
Formatted: Font color: Auto
Deleted: also
Formatted: Font color: Auto
Formatted: Font color: Auto
Deleted: 382
Deleted: a
Deleted: (CEU),
Deleted: These personal genomes are provided in AlleleDB.
Deleted: .[
Deleted: This is achieved by calculating
Deleted: -
Deleted: bin
Deleted: detection. Read depths for each heterozygous SNV in each genome are also provided in AlleleDB
Formatted: Font color: Auto

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?].

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. 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 (Supp Figure), 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 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 then ranked to determine a 'degree' of allele-specific inheritance.

Acknowledgements

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

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 chromosome

Deleted: highly
Deleted: size
Deleted: are
Deleted: ;
Deleted: then
Deleted: control and
Deleted: For
Deleted: n
Deleted: We can obtain
Deleted: a
Deleted: computed for each n and each number of successes
Deleted:).
Deleted: (denoted by asterisks in Figure 2).

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

Figure 2. Some genomic regions are more susceptible to allele-specific regulation. We can map variants associated with allele-specific binding (ASB; green) and expression (ASE; blue) to various genomic categories, such as coding DNA sequences (CDS), untranslated regions (UTRs), enhancer and promoter regions, to survey the human genome for susceptibility to allelic behavior. Using the accessible non-AS SNVs as the expectation, we can 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 olfactory receptor genes (CDS regions) are highly enriched for both ASB and ASE SNVs (>1.5); 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 inset zoomed in on the histogram at MAF < 3%.

Figure 4. Inheritance of allele-specific binding events is evident in some TFs but not so apparent in others. Here, the TFs CTCF (top row) and MYC (bottom row) are being examined for inheritance. For each TF, three plots compares two individuals in the CEU trio (Father: NA12891, Mother: NA12892, Daughter: NA12878), with the identity of the individual on the x-axis denoted by blue and that on the y-axis by red. 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 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.