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

Jieming Chen[1,2], Joel Rozowsky[1,3], Jason Bedford[1], Arif Harmanci[1,3], Alexei Abyzov[1,3,6], Yong Kong[4,5], Robert Kitchen[1,3], Lynne Regan[1,2,3], Mark Gerstein[1,2,3,4]

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

Abstract

Large-scale sequencing of personal genomes has revealed multitudes of genomic variants, but for the majority, their functional impact is unknown. Here, we annotate the allele-specific behavior of many variants, including rare ones. Allele-specific behavior can be assessed by observing allelic imbalance in the readouts of ChIP-seq and RNA-seq experiments. To this end, we pool and uniformly reprocess many previous experiments, organizing the results into a database, AlleleDB. Overall, we detect 169,235 allele-specific binding SNVs and 144,083 allele-specific expression SNVs, representing 41% and 21% of SNVs accessible by the respective assays. Using the accessible non-allelic SNVs as controls, we identify genomic annotations (genes and categories of non-coding elements) significantly enriched or depleted in allele-specific behavior, such as the SNURF and FHIT genes and promoters with binding sites for RPB2 and PU.1 transcription factors. Finally, we find that allele-specific SNVs tend to be in regions under less purifying selection.

Deleted: provide functional annotation for

Deleted: , using allele

Deleted: , which

Deleted: re-process

Deleted: to

Deleted: a control

Introduction

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

Much of the characterization of variants so far has been focused on those found in the protein-coding regions, but the advent of large-scale functional genomic assays, such as chromatin immunoprecipitation sequencing (ChIP-seq) and RNA sequencing (RNA-seq), has facilitated the annotation of genome-wide variation. This can be accomplished by correlating functional readouts from the assays to genomic variants, particularly in identifying regulatory variants, such as mapping of expression quantitative trait loci (eQTLs)[7–9] and allele-specific (AS)[10,11] variants. eQTL mapping assesses the effects of variants on expression profiles across a large population of individuals and is usually used for detection of common regulatory variants. On the other hand, AS approaches assess phenotypic differences directly at heterozygous loci within a single genome. Using each allele in a diploid genome as a perfectly matched control for the other allele, AS variants can be detected regardless of their population allele frequencies. Therefore, AS approaches are very useful, in terms of functionally annotating personal genomes, for identifying cis-regulatory variants on a large scale.

Early high throughput implementations of AS approaches employed microarray technologies, and thus are restricted to a small subset of loci.[12–14] Later studies have used ChIP-seq and RNA-seq experiments for genome-wide measurements of AS variants but have been mostly limited to a single assay with a variety of individuals,[15] or a few individuals with deeply-sequenced and well-annotated genomes.[11,16] For instance, GM12878, a very well-characterized lymphoblastoid cell-line from a Caucasian female, has several RNA-seq datasets and a huge trove of ChIP-seq data for more than 50 transcription factors (TFs) distributed across multiple studies.[17–19] Merging these datasets is advantageous, be it increasing statistical power or simply having more features for more intra- and inter-individual comparisons (such as TFs and populations).

AS variant detection is extremely sensitive to the technical issues associated with variant calling RNA-seq and ChIP-seq experiments, such as heterozygous variant calling and read mapping.[20–23] For example, allele-specific SNVs detected in copy number variants have a higher rate of false positives, since copy number changes can easily masquerade as allelic imbalance. Moreover, studies with the appropriate datasets are typically designed with various goals.[24,25] These reasons portend that simply pooling results from multiple studies may not be optimal even for the same biological sample. The task of merging has to be carried out in a uniform and systematic 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 especially for allele-specific analyses. We detect more than 169K and 144K single nucleotide variants (SNVs) associated with allele-specific binding (ASB) and expression (ASE) events respectively. We are able to present a comprehensive survey of these detected AS SNVs in various categories of coding and non-coding genomic annotations. The variants and annotations are available in a resource, AlleleDB (http://alleledb.gersteinlab.org/). Finally, using our consolidated data, we investigate the extent

Deleted: all of

Deleted: mainly

Deleted: scans

Deleted: of

Deleted: in more than one study

Deleted: of

Deleted: and

Deleted: reads mapping to

Deleted: be erroneously regarded

Deleted: regions of

Deleted: , resulting in AS variants being falsely detected.

Deleted: for

Deleted: purposes, resulting in disparate sets of computational tools, strategies and threshold parameters used in the processing of data in each respective study

Deleted: meaningful

Deleted: specifically

Deleted: systematic

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

Results

AlleleDB Workflow

There are two layers of information with respect to an individual that needs to be integrated in order to more accurately detect AS SNVs: (1) the DNA sequence of the individual, and (2) reads from either the RNA-seq or ChIP-seq experiment to assess SNVs associated with ASB or ASE. Here, we implement a uniform pipeline to combine personal genomic, transcriptomic and binding data and to standardize our detection of potential AS SNVs (Figure 1). First, we construct a diploid personal genome for each of the 383 individuals, using variants from the 1000 Genomes Project. By mapping the reads to the respective personal genome instead of the human reference genome, we reduce reference bias that can potentially result in erroneous read mapping.[16] Next, we pool the reads from each individual's ChIP-seq or RNA-seq data and align them to each of the haploid genomes. In total, we reprocess 117 ChIP-seq and 475 RNA-seq datasets for 383 individuals. Lastly, the AS SNVs are detected based on allelic imbalance of reads between the two haplotypes at heterozygous loci. For ChIP-seq data, the SNVs are further pared down to those within peak regions. We also remove SNVs if they lie in regions predicted to be copy number variants (see Methods).

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

By comparing AS SNVs relative to the control SNVs in each genomic annotation (see Methods), we investigate the enrichment (or depletion) of AS SNVs in 953 categories of non-coding genomic elements, including DNaseI hypersensitivity sites and transcription factor binding motifs from the ENCODE rollout[26], and 19,257 autosomal protein-coding genes from GENCODE[27]. We also extend this analysis to gene elements, such as introns and promoter regions and six other gene categories, including housekeeping and imprinted genes (see Methods). Together, these provide a systematic survey of ASB and ASE with respect to various functional annotations in the human genome.

We build a database, AlleleDB (http://alleledb.gersteinlab.org/), to house the annotations, the predicted AS and accessible SNVs. AlleleDB can be downloaded as flat files or queried and visualized directly as a UCSC track in the UCSC Genome browser[28] as specific genes or genomic locations. This enables cross-referencing of AS variants with other track-based datasets and analyses, and makes it amenable to all functionalities of the UCSC Genome browser. Heterozygous SNVs found in the stipulated query genomic region are color-coded (AS SNVs are red, accessible SNVs are black) in the displayed track.

Deleted: and AlleleDB
Deleted: look for
Deleted: genome
Deleted: 142
Deleted: additionally
Deleted:
Deleted: statistical
Deleted: Integrative release
Deleted: 20,144
Deleted: Additionally,
Deleted: is also extended
Deleted: Figure 2,
Deleted: and
Deleted: candidate

ASB Inheritance analyses using CEU trio

The CEU trio is a well-studied family and particularly, many ChIP-seq studies were performed on different TFs. Previous studies have presented AS inheritance in a few TFs as a case-study.[11,18] Here, after uniformly processing datasets from multiple studies, we are able to analyze and compare the heritability of ASB across seven DNA-binding proteins in a consistent manner (Figure 2; see Methods). For the DNA-binding protein CTCF, we observe a high parent-child correlation (Figure 2), denoting great similarity in allelic directionality (Pearson's correlation, $r \geq 0.9$ in both parent-child plots). High inheritance of AS SNVs in the same allelic direction from parent to child also implies a sequence dependency in allele-specific behavior. Besides CTCF, PU.1, PAX5, POL2, SA1 also show AS inheritance. Comparatively, AS inheritance in MYC and RPB2 are not as apparent, indicating that inheritance of ASB behavior may not be a universal phenomenon in all TFs.

Allele-specific variants and enrichment analyses

Using the AlleleDB variants found in the personal genomes of the 2 parents of the trio and 380 unrelated individuals from Phase 1 of the 1000 Genomes Project, we focus on autosomal SNVs and detected 144,083 ASE and 169,235 ASB SNVs, representing 21% and 41% of the accessible SNVs respectively (Table 1). The higher number of ASB SNVs observed is in line with a previous study that showed more variability in binding than in expression among individuals.[19] Of great interest, is the annotation of these AS SNVs with respect to known genomic elements, both coding and non-coding. 56% of our candidate ASE SNVs and 6% of ASB SNVs are in the coding DNA sequences (CDS). From 953 non-coding categories, we observed statistical significance (Bonferoni-corrected $p < 0.05$) for 632 and 441 categories for ASB and ASE SNVs respectively. From 19,257 autosomal protein-coding genes, we observed statistical significance for 31 and 442 genes for ASB and ASE SNVs respectively (Supplementary files 1 and 2). Some genes are expected, while some are not evidently so. For example, SNURF is a maternally-imprinted gene, shown to be highly implicated in the Prader-Willi Syndrome, an imprinting disorder.[29] Thus, it is expected to be significantly enriched in allele-specific behavior in our analyses. On the other hand, FHIT is a tumor suppressor gene significantly depleted in allele-specific behavior. While it is known to be a sensitive locus implicated in a variety of cancers,[30,31] it is not obvious why allele-specific behavior is depleted in this gene.

Figure 3 shows the enrichment of AS SNVs to provide a survey of AS regulation in elements closely related to a gene model, namely enhancers, promoters, CDS, introns and untranslated regions (UTR). In general, both categories of AS SNVs are more likely found in the 5' and 3' UTRs, suggesting allele-specific regulatory roles in these regions. On the other hand, intronic regions seem to exhibit a dearth of allele-specific regulation. For SNVs associated with allele-specific expression (ASE), a greater enrichment in 3' UTR than 5' UTR regions might be, in part, a result of known RNA-seq bias.[32,33] For SNVs associated with allele-specific binding (ASB), we also observe an enrichment in promoters, suggesting functional roles for these variants found in TF binding motifs or peaks found near transcription start sites in the promoter regions to regulate gene expression. However, we see variable enrichments of ASB SNVs of particular TFs in promoter regions such as RPB2, while depletion in others, such as PU.1 (Figure 3, Supplementary file 3). These differences imply that some TFs are more likely to participate in allele-specific regulation than others.

Moved (insertion) [1]
Moved (insertion) [2]
Deleted: ~
Deleted: ~
Deleted: 716
Deleted: 467
Deleted: 20,144
Formatted: Font color: Auto
Formatted: Font color: Auto
Deleted: supp file).
Deleted: significantly enriched in allele-specific behavior in our analyses. It has
Deleted: ,
Deleted: and appears to be a sensitive locus with high occurrence of loss-of-heterozygosity and hypermethylation.
Field Code Changed
Deleted: 2
Deleted: the
Deleted: hinting at
Deleted: in
Deleted: 2, Supp

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)[34,35], namely imprinted genes,[36] olfactory receptor genes,[37] the major histocompatibility complex,[38] immunoglobulin genes and genes associated with T cell receptors.[39] We also include a list of genes found to experience random monoallelic expression (RME) in a study by Gimelbrant *et al* (2007).[40] As expected, most of the MAE gene sets have been found to be significantly enriched in both ASB and ASE SNVs, with the exception of the olfactory receptor, T cell receptor and RME genes, where enrichment is not observed in ASB. Interestingly, while a statistically significant enrichment of ASB SNVs is observed in the constitutively expressed housekeeping genes, there is no enrichment in ASE SNVs (Figure 3).

Rare variants and purifying selection in AS SNVs

To assess the occurrence of ASB and ASB SNVs in the human population, we consider the population minor allele frequencies (MAF). Table 1 shows the breakdown of the accessible and AS SNVs in seven ethnic populations and allele frequencies. Yoruba from Ibadan, Nigeria (YRI) contribute the most to both ASE and ASB variants at each allele frequency category. The number of rare AS SNVs (MAF ≤ 0.5%) is about two folds higher in the YRI (48% ASE SNVs and 34% ASB SNVs with MAF ≤ 5%) than the other European sub-populations of comparable (CEU, FIN) or larger (TSI) population sizes (see Methods for full explanation of population abbreviations). In general, rare variants do not form the majority of all the AS variants. Nonetheless, we observe a shift towards very low allele frequencies in AS SNVs, peaking at MAF ≤ 0.5% (Figure 4).

To examine selective constraints in AS SNVs, we consider the enrichment of rare variants with MAF ≤ 0.5%.[4,41] Our results show lower enrichment of rare variants in AS SNVs as compared to non-AS SNVs. This posits that, as a whole, AS SNVs are under lesser selective constraints than non-AS SNVs. Such weaker selection may be a result of accommodating varying levels of gene expression across individuals. In addition, ASB SNVs seem to be under less selective constraints than ASE SNVs, which aligns with more variability being observed in binding than expression.

Discussion

Much research on regulatory variants has been performed using eQTL mapping of common variants. AS analyses can provide a complementary approach for detecting regulatory variants. Firstly, we found a substantial number of very rare AS SNVs with MAF ≤ 0.5%. Rare SNVs are harder to assess by eQTL mapping and the number is expected to increase with more personal genomes. Secondly, in eQTL mapping, correlation is drawn between total expression measured between individuals in a population and their genotypes. This is allele-insensitive as the total expression across a single locus is measured. However, in an AS approach, even if the total expression is the same across genotypes, difference in allelic expression can still be detected. Such a within-individual control in an AS approach also alleviates normalization issues across multiple assays. Thirdly, eQTL mapping is contingent on population size for sufficient statistics, while the AS approach can detect AS effects *en masse* within a single 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.

Deleted:), namely imprinted genes,

Deleted: and three sets of genes known to undergo allelic exclusion: olfactory receptor genes,

Field Code Changed

Deleted: immunoglobulin,

Field Code Changed

Field Code Changed

Deleted: genes associated with T cell receptors and the major histocompatibility complex.

Field Code Changed

Deleted: Monoallelic exclusion is a process exhibiting monoallelic expression, whereby one allele is being expressed while the other is silenced or repressed.

Field Code Changed

Deleted: ·

Field Code Changed

Deleted: .

Deleted: 2

Deleted: contributes

Deleted: (~

Deleted: ~

Deleted: .

Deleted: 3

Deleted: degrees

Deleted: lesser

Deleted: ¶

Moved up [1]: ASB Inheritance analyses using CEU trio

Deleted: ¶

Moved up [2]: The CEU trio is a well-studied family and particularly, many ChIP-seq studies were performed on different TFs.

Deleted: Previous studies have presented AS inheritance in a few TFs as a case-study.[11,18] Here, we provide a more comprehensive and statistical investigation of the heritability of ASB (Figure 4 and Supp file). For the DNA-binding protein CTCF, we observe a high parent-child correlation (Figure 4), denoting great similarity in allelic directionality (slope of regression line, β=0.86 in both parent-child plots). High inheritance of AS SNVs with the sam ...

Formatted: Underline

Deleted: to detect

Deleted: This group of

Deleted: is

Deleted: access

Deleted: locus is measured. As such, effects fro ...

Deleted: eliminates

Our analyses focuses on relating allele-specific activity to known genomic annotations, such as CDS and various non-coding regions, and many diseases have been found to implicate ASE in particular genomic regions.[42–44] Therefore, our analyses can help to characterize genomic variants on two levels: firstly, at the single nucleotide level, where our detected AS SNVs can serve as an annotation to variant catalogs (e.g. 1000 Genomes Project) in terms of allele-specific cis-regulation; secondly, by associating AS SNVs with a genomic annotation, we might be able to define categories of genomic regions more attuned to allele-specific activity. This can help to prioritize downstream experimental characterization to determine if such allele-specific behavior do exist and if so, whether it leads to any phenotypic differences.[45] Additionally, high coordination between ASB in specific TFs and ASE in genes they regulate has been observed in previous studies.[16,46] By comparing the ASB and ASE enrichments within the same category of genomic region, we can provide some further insights into the coordination of ASB and ASE within a genomic annotation or category. For example, the high enrichment of AS SNVs in most loci associated with monoallelic expression can imply coordination of ASE events by ASB. The exceptions are the groups of RME, T cell receptor and olfactory receptor genes, where another mechanism (besides ASB) might be causing ASE in these genes.

Our current catalog of AS SNVs is detected from lymphoblastoid cell lines (LCLs), which is also the predominant cell-line type in the literature. However, it has already been known that there is considerable variability in regulation of gene expression in different tissues.[47] Data from projects, such as GTEx[47], which has more functional assays and sequencing in other tissues and cell lines can be incorporated to provide a more wholesome AS analysis. Furthermore, our search for datasets shows a dearth of personal genomes with corresponding ChIP-seq and RNA-seq data in non-European populations. It could be a strong reflection on the lack of large-scale functional genomics assays in specific ethnic groups – a concern echoed previously in population genetics and is recently being increasingly addressed.[48] 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.

In conclusion, there is great value and utility in integrating existing data. Even though an AS approach is able to detect many AS SNVs for a single personal genome, the increase in quantity and diversity of personal genomes will raise the number of rare AS SNVs detected. Additionally, more accurate datasets will be made available in the near future as allelic information becomes more precise with the advent of longer reads to help in haplotype reconstruction and phasing in next-generation sequencing.[49–51] As more diverse and accurate personal genomes and functional genomics data become available, a pipeline that processes them efficiently and in a uniform fashion is essential, and enables AlleleDB to be easily scaled to accommodate new individual genomes, tissue and cell types. Such should be especially valuable, not only for researchers interested in allele-specific regulation but also for the scientific community at large.

Materials and Methods

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

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

Field Code Changed

Deleted: [43–45…2–44] Therefore, our analyses can h

Moved down [3]: . For example, the high enrichment of AS SNVs in most loci associated with monoallelic expression can imply coordination of ASE events by ASB.

Deleted: The exceptions are the groups of RME and olfactory receptor genes, where another mechanism (besides ASB) might be causing ASE in these genes.…This can help to prioritize downstre

Field Code Changed

Deleted: [46…5]¶

Field Code Changed

Deleted: [47]

Moved (insertion) [3]

Deleted: Data from projects, such as ENCODE

Field Code Changed

Deleted: [36…7] and…ata from projects, such as

Deleted: we have shown that …here is great va

Deleted: ¶

Formatted: Font: Bold

Formatted: Font: Not Bold, Underline

Moved down [4]: **AlleleDB**

Deleted: ¶

Moved down [5]: The final data and results are organized into a resource, AlleleDB (http://alleledb.gersteinlab.org/), which conveniently interfaces with the UCSC genome browser for query and visualization. Since many in the scientific community are familiar with the genome browser

Deleted: The query results are also available fo

Moved down [6]: More in-depth analyses ca

Deleted: ¶

Moved down [7]: *al*. (2013)

Deleted: [54]), and data from distal regulatory

Moved down [8]: The lists can be found at

Deleted: [56] (http://enhancer.lbl.gov/).

Moved down [9]: ¶

Deleted: ¶

Finland (FIN), British in England and Scotland (GBR), Japanese from Tokyo, Japan (JPT), Toscani from Italy (TSI), and Yorubans from Ibadan, Nigeria (YRI) and 3 high-coverage genomes from the CEU trio family (average read depth of 30x from Broad Institute's, GATK Best Practices v3; variants are called by UnifiedGenotyper). Each diploid personal genome is constructed from the SNVs and short indels (both autosomal and sex chromosomes) of the corresponding individual found in the 1000 Genomes Project. This is constructed using the tool, *vcf2diploid*.[16] 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,[52] which calculates the average read depth within a defined window size, normalized to the genomic average for the region of the same length. For each low coverage genome, a window size of 1000 bp is used, while for the high coverage genomes, a window size of 100 bp is used. SNVs found within genomic regions with a normalized abnormal read depth <0.5 or >1.5 are filtered out, since these would mostly likely give rise to spurious AS detection.

RNA-seq and ChIP-seq datasets

RNA-seq datasets are obtained from the following sources: gEUVADIS[15], ENCODE[26], Lalonde *et al.* (2011)[53], Montgomery *et al.* (2010)[54], Pickrell *et al.* (2010)[7], Kilpinen *et al.* (2013)[18] and Kasowski *et al.* (2013)[19].

ChIP-seq datasets are obtained from the following sources: ENCODE[26], McVicker *et al.* (2013)[55], Kilpinen *et al.* (2013)[18] and Kasowski *et al.* (2013)[19].

Allele-specific SNV detection

AS SNV detection is performed by AlleleSeq.[16] 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,[56] No multi-mapping is allowed and only a maximum of 2 mismatches per alignment is permitted. Sets of mapped reads from various datasets are merged into a single set for allele counting at each heterozygous locus. Here, a binomial p-value is derived by assuming a null probability of 0.5 sampling each allele. To correct for multiple hypothesis testing, FDR is calculated. Since statistical inference of allele-specificity of a locus is dependent on the number of reads of the ChIP-seq or RNA-seq dataset, this is performed using an explicit computational simulation.[16] Briefly, for each iteration of the simulation, a mapped read is randomly assigned to either allele at each heterozygous SNV and performs a binomial test. At a given p-value threshold, the FDR can be computed as the ratio of the number of false positives (from the simulation) and the number of observed positives. An FDR cutoff of 10% is used for ChIP-seq data and 5% for RNA-seq data, since the latter is typically of deeper coverage. Furthermore, we allow only significant AS SNVs to have a minimum of 6 reads. For ChIP-seq data, AS SNVs

Field Code Changed
Deleted: [58]
Deleted: [59]
Deleted: [60]
Field Code Changed
Field Code Changed
Deleted: [61]
Field Code Changed
Field Code Changed
Deleted: [62]

have to be also within peaks. Peak regions are provided as per those called from each publication of origin, except for the dataset from McVicker *et al*. (2013), in which there are no peak calls. In the latter case, we determine the peaks by performing PeakSeq[57] using the unmapped control reads provided by McVicker *et al*. (2013) via personal communication with the author. We use PeakSeq version 1.2 with default parameters and mapability map for human genome (hg19) to call peaks. The peaks that pass q-value threshold of 0.05 are marked as significant and used in the analyses.

AlleleDB

The final data and results are organized into a resource, AlleleDB (http://alleledb.gersteinlab.org/), which conveniently interfaces with the UCSC genome browser for query and visualization. Since many in the scientific community are familiar with the genome browser, we hope that this would increase the accessibility and usability of AlleleDB. The query results are also available for download in the BED format, which is compatible with other tools, such as the Integrated Genome Viewer[58]. 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 it is identified as an AS SNV. AlleleDB also serves as an annotation of allele-specific regulation of the 1000 Genomes Project SNV catalog.

AS inheritance analyses

The conventional measure of 'heritability' allows the estimation of (additive) genetic contribution to a certain trait. The population genetics definition of 'heritability' in a parent-offspring setting is described by the slope, β, of a regression ($Y=\beta X + \alpha$), with the dependent variable being the child's trait value (Y) and the independent variable (X) being the average trait values of the father and the mother ('midparent').[59] This is a population-based measure typically performed on a large set of trios for a particular trait (e.g. height) and β is not necessarily bound between 0 and 1.

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

Given that Pearson's correlation coefficient, r, always gives a value between 0 and 1, we use r instead of β, as our measure of 'heritability'. We also compute and include β values in the Supplementary Table. The parent-parent comparison is provided as a source of comparison for two unrelated individuals with shared ancestry. For parent-parent β, the maternal allelic ratio is chosen arbitrarily to be the independent variable.

Deleted: [63]
Field Code Changed
Formatted: Font: Bold
Moved (insertion) [4]
Moved (insertion) [5]
Moved (insertion) [6]

Genomic annotations

Categories of gene elements from Figure 3, such as promoters, CDS regions and UTRs, and 19,257 autosomal protein-coding gene annotations (HGNC symbols) are obtained from GENCODE version 17.[27] Promoter regions are set as 2.5kbp upstream of all transcripts annotated by GENCODE.

Gene annotations also include 2.5kbp upstream of the start of gene. 953 categories of non-coding annotations are obtained from ENCODE Integrative release,[26] which includes broad categories such as TF binding sites and more specific annotations such as distal binding sites of particular TFs, e.g. ZNF274. Note that these TF binding sites are separate from those sites in promoter regions in Figure 2, which are based on the 59 TFs and peaks from the ChIP-seq experiments used in our pipeline.

Genes for random monoallelic expression are from Gimelbrant *et. al*. (2007)[40] The olfactory receptor gene list is from the HORDE database[37]; immunoglobulin, T cell receptor and MHC gene lists are from IMGT database[39]. We performed enrichment analyses on a number of enhancer lists, which are derived using the ChromHMM and Segway algorithms (Ernst and Kellis (2012)[60], Hoffman *et. al*. (2013)[61]), and data from distal regulatory modules from Yip *et al.* (2012)[62]. The result for the enhancers in Figure 3 is based on the union of these lists. The lists can be found at http://info.gersteinlab.org/Encode-enhancers. An additional enhancer list for experimentally validated enhancers is obtained from VISTA enhancer browser database[63] (http://enhancer.lbl.gov/). Housekeeping gene list is obtained from Eisenberg and Levanon (2013) (http://www.tau.ac.il/~elieis/HKG/)[64].

All enrichment analyses results with respect to these annotations are provided in the supplementary files.

Moved (insertion) [7]

Moved (insertion) [8]

Moved (insertion) [9]

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, such that accessible SNVs are all SNVs with number of reads, n = max(6,N). The control (non-AS) ASB or ASE SNVs are accessible SNVs excluding the respective ASB or ASE SNVs. Enrichment analyses are performed using the Fisher's exact test. P-values are Bonferroni-corrected and considered significant if < 0.05.

Deleted: **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 parent-child pair. Since AS events can only be detected at heterozygous sites, we consider all SNVs shared by parent and child from 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. We then calculate the slope using regression analysis, with the parent's allelic ratio as the independent variable (X) and the child's as the dependent variable (Y): $Y = \beta X + \alpha$. Pearson's correlation coefficients (r) are also provided in the supplementary table. For the parent-parent comparison, the maternal allelic ratio is chosen arbitrarily to be the independent variable; the correlation coefficient might be a better measure in this case.¶
¶
¶

Acknowledgements

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

Figure and table captions

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

Figure 2. Inheritance of allele-specific binding events is evident in some TFs but not so apparent in others. The left panel shows plots for the TFs CTCF (top row) and MYC (bottom row) being examined for inheritance in the CEU trio (Father: NA12891, blue; Mother: NA12892, red; Child: NA12878, green). Each point on the plot represents the allelic ratio of a common ASB SNV between the parent (x-axis) and the child (y-axis), by computing the proportion of reads mapping to the reference allele at that SNV. High Pearson's correlations, r, observed in both parent-child comparisons for CTCF ($r > 0.9$) signify strong heritability in allele-specific behavior. On the other hand, MYC has comparatively lower r values, suggesting that AS inheritance is not apparent. The table at the top right panel presents the r values for all seven TFs used in our analyses, in descending order of heritability.

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

Figure 4. Some genomic regions are more inclined to allele-specific regulation. We map variants associated with allele-specific binding (ASB; green) and expression (ASE; blue) to various categories of genomic annotations, such as coding DNA sequences (CDS), untranslated regions (UTRs), enhancer and promoter regions, to survey the human genome for regions more enriched in allelic behavior. Using the accessible non-AS SNVs as the expectation, we compute the log odds ratio for ASB and ASE SNVs separately, via Fisher's exact tests. The number of asterisks depicts the degree of significance (Bonferonni-corrected): *, $p<0.05$; **, $p<0.01$; ***,

Deleted: Rob

Deleted: ¶ ¶

Moved down [10]: References¶ ¶

Deleted: 1. Wheeler, D. A. *et al*. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452,** 872–6 (2008).¶ 2. Lupski, J. R. *et al*. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N. Engl. J. Med.* **362,** 1181–91 (2010).¶ 3. Levy, S. *et al*. The diploid genome sequence of an individual human. *PLoS Biol.* **5,** e254 (2007).' 4. Abecasis, G. R. *et al*. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).¶ 5. Muddyman, D., Smee, C., Griffin, H. & Kaye, J. Implementing a successful data-management framework: the UK10K managed access model. *Genome Med.* **5,** 100 (2013).¶ 6. Church, G. M. The personal genome project. *Mol. Syst. Biol.* **1,** 2005.0030 (2005).¶

Formatted

Formatted: Font: Times New Roman, 12 pt

Deleted: caption

Deleted: ¶ **Figure 2.**

Moved down [11]: **Some genomic regions are more inclined to allele-specific regulation.** We

Deleted: Using the accessible non-AS SNVs as the expectation, we compute the log odds ratio of

Moved down [12]: The number of asterisks depicts the degree of significance (Bonferonni-

Deleted: Genes known to be monoallelically expressed such as imprinted and MHC genes (CD

Moved down [13]: The minor allele frequency (MAF) spectra of ASB (green filled circle),

Moved down [14]: Comparing ASE+ to ASE- gives an odds ratio of 0.67 (hypergeometric $p < 2$

Deleted: **Figure 4.**

Deleted: Daughter

Deleted: ; blue

Deleted: ; green

Deleted: The correlation is evaluated by the slope of the regression, β.

Deleted: β values in CTCF

Deleted: shows low β

Deleted: β

Moved (insertion) [13]

Moved (insertion) [14]

Moved (insertion) [11]

Moved (insertion) [12]

p<0.001. For each transcription factor (TF) in AlleleDB, we also calculate the log odds ratio of ASB SNVs in promoters, providing a proxy of allele-specific regulatory role for each available TF. Genes known to be mono-allelically expressed such as imprinted and MHC genes (CDS regions) are highly enriched for both ASB and ASE SNVs. The actual log odds ratio of immunoglobulin genes for ASE SNVs and MHC genes for ASB and ASE SNVs are indicated on the bars.

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 population 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.

Supplementary Table

Supplementary Table 1
This table shows the slope and Pearman's correlation results for all seven DNA-binding proteins for parent-child and parent-parent comparisons. CTCF, PU.1, SA1, PAX5 and POL2 exhibit AS inheritance but MYC and RPB2 do not seem to have very apparent AS inheritance.

Supplementary Files

Supplementary File 1
This Excel file contains results from our AS analyses for 953 categories from ENCODE, including the Fisher's exact test odds ratios, p-values (original and Bonferroni-corrected), the number of AS SNVs and accessible non-AS SNVs found in each category. The results for five gene element categories from GENCODE and 16 enhancer categories are also included. 'NA' is marked in categories where odds ratio cannot be calculated due to insufficient numbers in non-AS SNVs. These are tabulated for ASB, ASE and AS SNVs, which is analysis based on the combined unique number of ASB and ASE SNVs.

Supplementary File 2
This Excel file contains results from our AS analyses for the 19,257 autosomal protein-coding genes (HGNC symbols) from GENCODE, including the Fisher's exact test odds ratios, p-values (original, Bonferroni-corrected), the number of AS SNVs and accessible non-AS SNVs found in the gene region. The results for housekeeping genes and 5 monoallelically-expressed gene categories are also included. 'NA' is marked in categories where odds ratio cannot be calculated due to insufficient numbers in non-AS SNVs. These are tabulated for ASB, ASE and AS SNVs, which is analysis based on the combined unique number of ASB and ASE SNVs.

Supplementary File 3
This Excel file contains the ASB enrichment in promoter regions for 59 TFs used in our database, including the Fisher's exact test odds ratios, p-values (original, Bonferroni-corrected), the

Formatted: Font: Bold, Underline
Deleted: ¶
Deleted: Figures
Deleted: Figures
Deleted: 5
Deleted: 20,144

number of ASB SNVs, accessible non-AS SNVs both found and not found in the gene region. ASB SNVs for each TF are contributed by different individuals. If either of the parents in the CEU trio is involved, ASB SNVs for NA12878 are not included. Those TFs with only ASB SNVs from NA12878 are annotated '1' under the column 'NA12878 only'. 'NA' is marked in categories where odds ratio cannot be calculated due to insufficient numbers in any of the last three columns.

References

Moved (insertion) [10]

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

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

3. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5,** e254 (2007).

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

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

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

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

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

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

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

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

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

13. Ge, B. *et al.* Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. *Nat. Genet.* **41,** 1216–22 (2009).

14. Lo, H. S. *et al.* Allelic variation in gene expression is common in the human genome. *Genome Res.* **13,** 1855–62 (2003).

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

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

17. Engström, P. G. *et al.* Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods* **10,** 1185–91 (2013).

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

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

20. Harismendy, O. *et al.* Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.* **10,** R32 (2009).

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

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

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

24. Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* **10,** 1177–84 (2013).

25. Engström, P. G. *et al.* Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods* **10,** 1185–91 (2013).

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

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

28. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).

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

30. Hallas, C. *et al.* Loss of FHIT expression in acute lymphoblastic leukemia. *Clin. Cancer Res.* **5,** 2409–14 (1999).

31. Zou, M., Shi, Y., Farid, N. R., Al-Sedairy, S. T. & Paterson, M. C. FHIT gene abnormalities in both benign and malignant thyroid tumours. *Eur. J. Cancer* **35,** 467–72 (1999).

32. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10,** 57–63 (2009).

33. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320,** 1344–9 (2008).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Formatted: Normal (Web), Indent: Left: 0", Hanging: 0.44"

Formatted: Font: Not Bold