Allele-specific binding and expression: a uniform survey over many individuals and assays

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 functionally annotate many variants, including rare ones, using allele-specific behavior. This 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, and organize the results into a database, AlleleDB. Overall, for binding and expression, we detect 7,462 and 85,742 allelic SNVs, representing 6% and 16% of SNVs accessible by the respective assays. Using the accessible SNVs as controls, we identify genomic annotations (genes and groups 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 POL2 and PU.1 transcription factors (TF). We also identify xxx SNVs that seem to break or gain TF motifs, thus having a potential to change TF occupancy. Finally, we find that allele-specific SNVs tend to be in genomic regions under less purifying selection.

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 even at low 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 analysis is extremely sensitive to the technical issues associated with variant calling and processing, RNA-seq and ChIP-seq experiments, such as thresholding 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.[17,24] 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 7K and 85K 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

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

Results

AlleleDB Workflow

In general, the AlleleDB workflow uniformly processes two pieces of information from each individual: the DNA sequence, and reads from either the ChIP-seq or RNA-seq experiment to assess SNVs associated with ASB or ASE respectively (Figure 1). Briefly, it starts by **(1)** constructing a diploid personal genome for each of the 382 individuals, using DNA variants from the 1000 Genomes Project. **(2)** It then aligns the ChIP-seq or RNA-seq dataset to each of the haploid genomes instead of the human reference genome. This reduces reference bias that can potentially result in erroneous read mapping.[16] Because each individual can have multiple ChIP-seq or RNA-seq datasets, the alignment is performed sequentially in two ways. **(2a)** First, the alignment is performed for each of 287 ChIP-seq and 993 RNA-seq datasets to calculate a measure of overdispersion, ρ (see Discussion and Methods). We observe that if there is a greater overdispersion in the allelic ratio distribution of a dataset, there is a higher tendency for a larger number of sites to possess allelic imbalance. This will likely result in the detection of more false positives (Figure 2). There are varying degrees of overdispersion in our datasets, with RNA-seq datasets generally less overdispersed than ChIP-seq datasets. By performing alignment to the personal haplotypes for each dataset and calculating allelic ratios for each heterozygous SNV, we were able to remove datasets that are deemed to be highly dispersed in allelic ratio distributions, leaving 184 ChIP-seq and 955 RNA-seq datasets for AS detection (Supp Table 1). **(2b)** The second alignment is performed by pooling ChIP-seq and RNA-seq datasets that has not been filtered in Step 2a. This is performed for each individual and each transcription factor (for ChIP-seq). **(3)** Finally, the pooled alignment is used in the detection of potential allele-specific SNVs based on a betabinomial test. These SNVs are heterozygous loci with imbalance in the read counts between the two haplotypes. 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).

This serves as a dataset-filtering step to remove datasets with low percentage of aligned reads to either or both haplotypes (<50%) or with unusually high number of SNVs with allelic imbalance.

We build a database, AlleleDB (http://alleledb.gersteinlab.org/), to house the annotations, the AS and accessible SNVs. AlleleDB can be downloaded as flat files or queried and visualized directly as a UCSC track in the UCSC Genome browser[25] as specific genes or genomic locations. This enables cross-referencing of AS variants with other track-based datasets and analyses, and makes it amenable to all functionalities of the UCSC Genome browser. Heterozygous SNVs found in the stipulated query genomic region are color-coded (ASB SNVs are red, ASE SNVs are black) in the displayed track.

ASB and ASE Inheritance analyses using CEU trio

The CEU trio is a well-studied family and with multiple ChIP-seq studies performed on different TFs. Previous studies have also presented AS inheritance.[11,18] Here, after uniformly processing

datasets from multiple studies, we are able to analyze and compare the heritability of ASE and ASB across two DNA-binding proteins in a consistent manner (Figure 3; see Methods). For the DNA-binding protein CTCF and PU.1, we observe a high parent-child correlation (Figure 3, Supp Table 2), denoting great similarity in allelic directionality (Pearson's correlation, $r \geq 0.77$ in both parent-child plots). We also observe considerable heritability in ASE, but to a lesser degree. In general, the high inheritance of AS SNVs observed in the same allelic direction from parent to child also implies a sequence dependency in allele-specific behavior.

Allele-specific variants and enrichment analyses

Using the AlleleDB variants found in the personal genomes of the 2 parents of the trio and 379 unrelated individuals from Phase 1 of the 1000 Genomes Project, we focus on autosomal SNVs and detected 85,742 ASE and 7,462 ASB SNVs, representing 16% and 6% of the accessible SNVs respectively (Table 1). 15% of our candidate ASE SNVs and 3% of ASB SNVs are in the coding DNA sequences (CDS).

Of great interest, is the annotation of these AS SNVs with respect to known genomic elements, both coding and non-coding. We calculate the enrichment of ASB and ASE SNVs in various genomic categories. To do so, we further define sets of 'control' SNVs. This is especially pertinent to our enrichment analyses, since the Fisher's exact test is dependent on the choice of the null expectation (i.e. 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 needed to be statistically detectable for allelic imbalance; in other words, the control SNVs are well-matched in power to the detected allele-specific SNVs. The accessible SNVs are determined for each ChIP-seq (grouped by individual and TF, not by study) or RNA-seq dataset (Table 1).

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 679 unique categories of non-coding genomic elements, including DNaseI hypersensitivity sites and transcription factor binding motifs from the ENCODE project[26], and 19,257 autosomal protein-coding genes from GENCODE[27]. Together, these provide a systematic survey of AS regulation with respect to various functional annotations in the human genome. From 679 unique non-coding categories, we observed statistical significance (Bonferroni-corrected $p \leq 0.05$) for 632 and 441 categories for ASB and ASE SNVs respectively (Supp file 1). From 19,257 autosomal protein-coding genes, we observed statistical significance for 71 and 352 genes for ASB and ASE SNVs respectively (Supp file 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.[28] 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,[29,30] it is not obvious why allele-specific behavior is depleted in this gene.

Additionally, we extend this analysis to gene elements, such as introns and promoter regions and seven other gene categories, including housekeeping and imprinted genes. Figure 4 shows the enrichment of AS SNVs in elements closely related to a gene model, namely enhancers, promoters, CDS, introns and untranslated regions (UTR). In general, ASE SNVs are more likely

found in the 5' and 3' UTRs, suggesting allele-specific regulatory roles in expression 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 slightly greater enrichment in 3' UTR than 5' UTR regions might be, in part, a result of known RNA-seq bias.[31,32] For SNVs associated with allele-specific binding (ASB), we also observe an enrichment in the 5' UTRs. This is in line with an enrichment of ASB SNVs in promoters, suggesting functional roles for these variants found in TF binding motifs or peaks found near transcription start sites to regulate gene expression. However, we see variable enrichments of ASB SNVs in the peaks of particular TFs such as POL2, SA1 and CTCF in promoter regions, while depletion in others, such as PU.1 (Figure 4, Supp file 3). These differences might imply that some TFs are more likely to participate in allele-specific regulation than others. Overall in CDS regions, there is a general depletion of ASE SNVs but enrichment of ASB SNVs.

We also specifically investigated gene categories known to be involved in monoallelic expression (MAE)[33,34], namely imprinted genes,[35] olfactory receptor genes,[36] the major histocompatibility complex (MHC),[37] immunoglobulin genes and genes associated with T cell receptors.[38] As expected, most of the MAE gene sets have been found to be significantly enriched in both ASB and ASE SNVs (except for ASB SNVs in MHC). We additionally include a list of genes found to experience random monoallelic expression (RME) in a study by Gimelbrant *et al* (2007)[39], and we show that the category is only enriched in ASE SNVs. Interestingly, there is a depletion in ASE SNVs for the constitutively expressed housekeeping genes (Figure 4).

Variants affecting TF occupancy in TF binding motifs
<insert text here>

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 six ethnic populations (we combined the results for CHB and JPT) 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 $\leq$ 5%) is about two folds higher in the YRI than the other European sub-populations of comparable (CEU, FIN) or larger (TSI) population sizes (see Methods for full explanation of population abbreviations). However, the percentage of AS SNVs (in accessible SNVs) remain fairly consistent. In general, rare variants do not form the majority of all the AS variants. Nonetheless, we observe a shift of the allele frequency spectrum towards very low allele frequencies in AS SNVs (compared to accessible, non-AS SNVs), peaking at MAF $\leq$ 0.5% (Figure 5).

To examine selective constraints in AS SNVs, we consider the enrichment of rare variants with MAF $\leq$ 0.5%.[4,40] We limit our analyses for ASE SNVs to only those found in CDS regions and ASB SNVs to only those found within known TF motifs (among the 679 non-coding categories in Supp File 1). Our results in Figure 5 show a statistically significant lower enrichment of rare variants in ASE SNVs as compared to non-ASE SNVs (Fisher's exact test odds ratio=0.2, p<2.2e-16) but statistically insignificant higher enrichment of rare variants in non-ASB SNVs than ASB SNVs (Fisher's exact test odds ratio=1.4, p=0.04). This posits that ASE SNVs are

under lesser selective constraints than non-ASE 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 well with the results in a previous study where more variability is being observed in binding than expression [19].

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 $\leq 0.5\%$. Rare SNVs are harder to assess by eQTL mapping. However, 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.

To obtain a conservative set of AS SNVs in AlleleDB, we introduce the use of the overdispersion parameter, ρ, in the betabinomial probability density function (pdf), for two purposes: (1) to account for the overdispersion in the statistical inference of AS SNVs, and (2) as a means to remove datasets that are highly overdispersed. The binomial test is typically used to provide statistical significance for the identification of AS SNVs [cite ~~alleleseq~~ ~~geuvadis~~ etc]. However, previous studies have observed a deviation from the binomial distribution in read count distributions in ChIP-seq and RNA-seq datasets, which in turn results in broader allelic ratio distributions, i.e. overdispersed. [cite] While datasets with low overdispersion give very similar results between binomial and betabinomial tests (Figure 2A), datasets with higher overdispersion tend to give a higher number of detected SNVs, which can be accounted for by ρ in the betabinomial test (Figure 2B). This appears to be a consequence of a bias in datasets with greater overdispersion in allelic ratio distributions, where there is an increased number of SNVs with allelic imbalance (towards the two ends of the distribution) (Figure 2B). Hence, we adopt a serial two-step approach of first filtering individual datasets with high overdispersion ($\rho>0.34$; rationalize this as arbitrary or provide another sup figure to show flip in ends?), and then pooling the resultant datasets (by individual and TF) for AS detection using the betabinomial test. In addition, we also provide a more confident set of AS SNVs, which are found to be in the same allelic direction in more than 1 individual in AlleleDB. We are also able to identify population-specific sets of these AS SNVs which are of higher confidence.

Our downstream 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.[41–43] 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 and assigning a proxy measure of allele-specific behavior, we are able to define categories of genomic regions more attuned to allele-specific activity. The quantification allows the use of weighting schemes based on allele-specific activity in downstream computational pipelines. This also enables prioritization of future experimental characterization to determine if such allele-specific behavior do exist and if so, whether it leads to any phenotypic differences.[44] Additionally, high coordination between ASB in specific TFs and ASE in genes they regulate has been observed in previous studies.[16,45] 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 and MHC, where another mechanism (besides ASB) might be the main cause of 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.[46] Data from projects, such as GTEx[46], 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.[47] 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.[48–50] 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. AlleleDB is 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 381 genomes used in this study: 379 unrelated genomes, of low-coverage (average depth of 2.2 to 24.8) from Utah residents in the United States with Northern and Western European ancestry (CEU), Han Chinese from Beijing, China (CHB), Finnish from Finland (FIN), British in England and Scotland (GBR), Japanese from Tokyo, Japan (JPT), Toscani from Italy (TSI), and Yorubans from Ibadan, Nigeria (YRI) and 3 high-coverage genomes from the CEU trio family (average read depth of 30x from Broad Institute's, GATK

Best Practices v3; variants are called by UnifiedGenotyper). Each diploid personal genome is constructed from the SNVs and short indels (both autosomal and sex chromosomes) of the corresponding individual found in the 1000 Genomes Project. This is constructed using the tool, *vcf2diploid*.[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 379 unrelated individuals, the alleles, though phased, are of unknown parental origin.

CNV genotyping is also performed for each genome by CNVnator,[51] 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)[52], Montgomery *et al.* (2010)[53], 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)[54], Kilpinen *et al.* (2013)[18] and Kasowski *et al.* (2013)[19]. In total, we reprocess 287 ChIP-seq and 993 RNA-seq datasets for 381 individuals.

Read alignment and estimation of ρ in individual and pooled datasets

Reads are aligned against each of the derived haploid genome (maternal/paternal genome for trio) using Bowtie 1.[55] No multi-mapping is allowed and only a maximum of 2 mismatches per alignment is permitted. This enables the calculation of the proportion reads that align to the reference allele, or the allelic ratio, at each heterozygous SNV.

To estimate ρ, we adopt a three-step approach. We first obtain the empirical histogram for the allelic ratios of all heterozygous SNVs with read counts ≥ 6. Next, we calculate the expected null distribution (where there is no allelic imbalance) using the probability density function (pdf) of the betabinomial distribution using the R package, VGAM [cite]:

$$P_{betabin}(X = k|n, a, b) = \binom{n}{k} \frac{B(k + a, n - k + b)}{B(a, b)}$$

where n represents the total number of reads at a particular locus, B(x,y) represents the beta function with variables x and y, a and b represent the shape parameters of the beta distribution. For computational efficiency, if $n \geq 1000$, we set it to a maximum of 1000, but retain the allelic

ratio at the SNV. The VGAM betabinomial routines require the input of the overdispersion parameter, ρ, and probability of success (also the mean of the beta distribution), which we fix at p=0.5 since the null hypothesis assumes no allelic imbalance. We then obtain the expected betabinomial distributions for ρ=0 to ρ=1 with increment of 0.1, and choose ρ that minimizes the least sum of squared errors (LSSE) between the empirical and the expected distributions. Lastly, to further refine our estimate, we iterate a bisection method to arrive at a LSSE with the following R pseudo-code:

```
while (previous_LSSE ≠ current_LSSE within 3 significant figures)
{
        previous_LSSE = current_LSSE
        start_ρ = prev_ρ – (prev_increment / 2)
        end_ρ = prev_ρ + (prev_increment / 2)
        current_increment = prev_increment / 4

        range = seq(start_ρ, end_ρ, by=current_increment)

        for (values in range)
        {
                obtain_betabinomial_distribution
                calculate_LSSE_between_betabinomial_and_empirical_distributions

                if(current_LSSE_within_for_loop > previous_LSSE_within_for_loop)
                {
                        current_LSSE = previous_LSSE_within_for_loop
                        break_out_of_for_loop
                }
        }

}
```

We calculate ρ for each 287 ChIP-seq and 993 RNA-seq individual datasets. In addition to 13 ChIP-seq and 6 RNA-seq datasets that have insufficient read alignments, we removed 55 ChIP-seq and 32 RNA-seq datasets with an arbitrary threshold of ρ>0.34.

Using the resultant 219 ChIP-seq and 955 RNA-seq datasets, we pool datasets by TF and individual for ChIP-seq and by individual for RNA-seq and re-calculate ρ for each pooled dataset. This final ρ is used in the betabinomial test for allele-specific SNV detection.

Allele-specific SNV detection

AS SNV detection is performed on the pooled datasets, as mentioned above. Here, a betabinomial p-value is derived based on the VGAM R package as described in the previous section. Similarly for computational efficiency, if $n \geq 1000$, we set it to a maximum of 1000, but retain the allelic ratio at the SNV. 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 betabinomial test using the estimated ρ. 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 empirical 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 determined by first performing PeakSeq[56] for each of the personal haploid genome. Subsequently, the coordinates are re-mapped to the reference genome and then finally being merged between the haploid genomes. 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 [AH].

AS detection for all TFs and gene expression of 381 individuals took about 600 days in CPU time (1.6 years), but the pipeline is highly parallelizable, thereby streamlining the process.

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 BED format, which is compatible with other tools, such as the Integrated Genome Viewer[57]. 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').[58] 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 Supplementary Table 2. 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.

Genomic annotations

Categories of gene elements from Figure 4, 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. 679 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 4, which are based on the 44 TFs and peaks from the ChIP-seq experiments used in our pipeline.

Genes for random monoallelic expression are from Gimelbrant *et. al*. (2007)[39] The olfactory receptor gene list is from the HORDE database[36]; immunoglobulin, T cell receptor and MHC gene lists are from IMGT database[38]. Imprinted genes are from the Catalog of Parent-of-origin Effects (http://igc.otago.ac.nz/home.html).[59] 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 4 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, which are provided for download on the AlleleDB website (http://alleledb.gersteinlab.org/download/).

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 betabinomial 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 betabinomial 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.

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.

References

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

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

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

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

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

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

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

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

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

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

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

30. 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).

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

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

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

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

35. 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).

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

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

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

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

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

41. 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).

42. 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).

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

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

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

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

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

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

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

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

51. 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).

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

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

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

55. 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).

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

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

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

59. Morison, I. M., Ramsay, J. P. & Spencer, H. G. A census of mammalian imprinting. *Trends Genet.* **21,** 457–65 (2005).

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).

Figure and table legend

Figure 1. Workflow for uniform processing of data from 382 individuals and construction of AlleleDB. For each of the 381 individuals, (1) a diploid personal genome is first constructed using the variants from the 1000 Genomes Project. Next, reads from individual (2a) and pooled (2b) ChIP-seq or RNA-seq data are mapped onto each of the haploid genome of the diploid genome. In (2a), overdispersion (OD) is measured for each dataset and used to filter out highly dispersed datasets. (2b) The resultant datasets are pooled and the overdispersion parameter is estimated based on the pooled datasets. To determine if a SNV is allele-specific (AS), a statistical significance is computed (after multiple hypothesis test correction) using the betabinomial test at each heterozygous SNV, by comparing the number of reads that map to either allele (allelic ratio). 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. Comparing the effects of the binomial and betabinomial tests in datasets with low and intermediate level of overdispersion. The grey bars represent the empirical allelic ratio distribution, while the red and blue lines represent the expected allelic ratio distribution using the

binomial and betabinomial tests respectively. Figure 2A shows the empirical and expected distributions for one of the individual RNA-seq datasets for the individual HG00096. It has a low overdispersion parameter, ρ=0.0205. The empirical distribution does not have heavy tails and the binomial and betabinomial tests give very similar results. This differs from Figure 2B, which shows the empirical and expected distributions for one of the individual RNA-seq datasets for the individual NA11894. Overdispersion increases to ρ=0.1234, and the betabinomial null distribution provides a clearly better fit to the empirical allelic ratio distribution than the binomial distribution. The empirical distribution (grey bars) also show heavier tails, signifying more SNVs with allelic imbalance.

Figure 3. Inheritance of allele-specific binding events is evident in some TFs but not so apparent in others. The left panel shows plots for the TF CTCF (top row) and ASE (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 ($r \geq 0.77$) signify strong heritability in allele-specific behavior. ASE also shows considerably strong evidence of heritability but has comparatively lower r values. The table at the top right panel presents the r values for ASB in two TFs and ASE in our analyses.

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 (Bonferroni-corrected): *, $p<0.05$; **, $p<0.01$; ***, $p<0.001$. 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 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 ASB SNVs in imprinted genes, both ASB and ASE SNVs in immunoglobulin genes and ASE SNVs for MHC genes are indicated on the bars.

Figure 5. 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 0.2 (hypergeometric $p < 2.2e-16$), while comparing ASB+ to ASB-, gives an odds ratio of 1.4 ($p=0.04$), signifying statistically significant depletion of AS variants relative to non-AS variants in ASE SNVs but the opposite in ASB SNVs. Statistically significant depletion in ASE suggests that ASE SNVs are under less purifying selection.

Table 1. Breakdown of SNVs in each ethnic population: heterozygous (HET), accessible (ACC) and ASE SNVs in Table 1A and ASB SNVs in Table 1B for 380 unrelated individuals. Table 1C shows the same HET, ACC and both ASE and ASB SNVs detected in a single individual of NA12878, who is also part of trio family. 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 number of individual datasets being filtered out due to insufficient reads and overdispersion parameter ρ>0.34.

Supplementary Table 2
This table shows the slope and Pearson's correlation results for two DNA-binding proteins and ASE for parent-child and parent-parent comparisons.

Supplementary Files

Supplementary File 1
This Excel file contains results from our AS analyses for 679 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; the latter is the results for 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; the latter is the combined unique number of ASB and ASE SNVs.

Supplementary File 3
This Excel file contains the ASB enrichment in promoter regions for 44 TFs used in our database, including the Fisher's exact test odds ratios, p-values (original, Bonferroni-corrected), the 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.