98 PubMedCentral citations
Mol Sys Bio: 5-year IF (2015) = 12.3

Molecular Systems Biology 7; Article number 522; doi:10.1038/msb.2011.54
**Citation: *Molecular Systems Biology* 7:522**

www.molecularsystemsbiology.com

molecular systems biology

# AlleleSeq: analysis of allele-specific expression and binding in a network framework

**Joel Rozowsky[1,2,8,*], Alexej Abyzov[1,2,8], Jing Wang[2], Pedro Alves[2], Debasish Raha[3], Arif Harmanci[1,2], Jing Leng[2], Robert Bjornson[4,5], Yong Kong[5], Naoki Kitabayashi[6], Nitin Bhardwaj[1,2], Mark Rubin[6], Michael Snyder[7] and Mark Gerstein[1,2,4,*]**

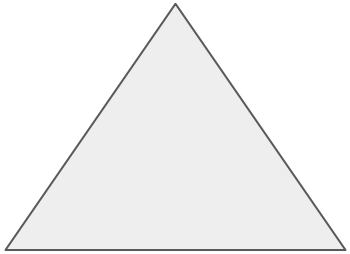
[1] Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA, [2] Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA, [3] Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, CT, USA, [4] Department of Computer Science, Yale University, New Haven, CT, USA, [5] Keck Biotechnology Resource Laboratory, Yale University, New Haven, CT, USA, [6] Department of Pathology and Laboratory Medicine, Weill Cornell Medical Center, New York, NY, USA and [7] Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA
[8] These authors contributed equally to this work
* Corresponding authors. J Rozowsky or M Gerstein, Program in Computational Biology and Bioinformatics, Yale University, 266 Whitney Ave Bass 432, New Haven, CT 06520, USA. Tel.: +1 203 432 5405; Fax: +1 203 432 5175; E-mail: joel.rozowsky@yale.edu or Tel.: +1 203 432 6105; Fax: +1 203 432 5175; E-mail: mark.gerstein@yale.edu

Received 21.12.10; accepted 7.7.11

Presenter: Charlotte Darby

# Big Picture

“The interaction between allele-specific binding and allele-specific expression are investigated, revealing clear coordination.”

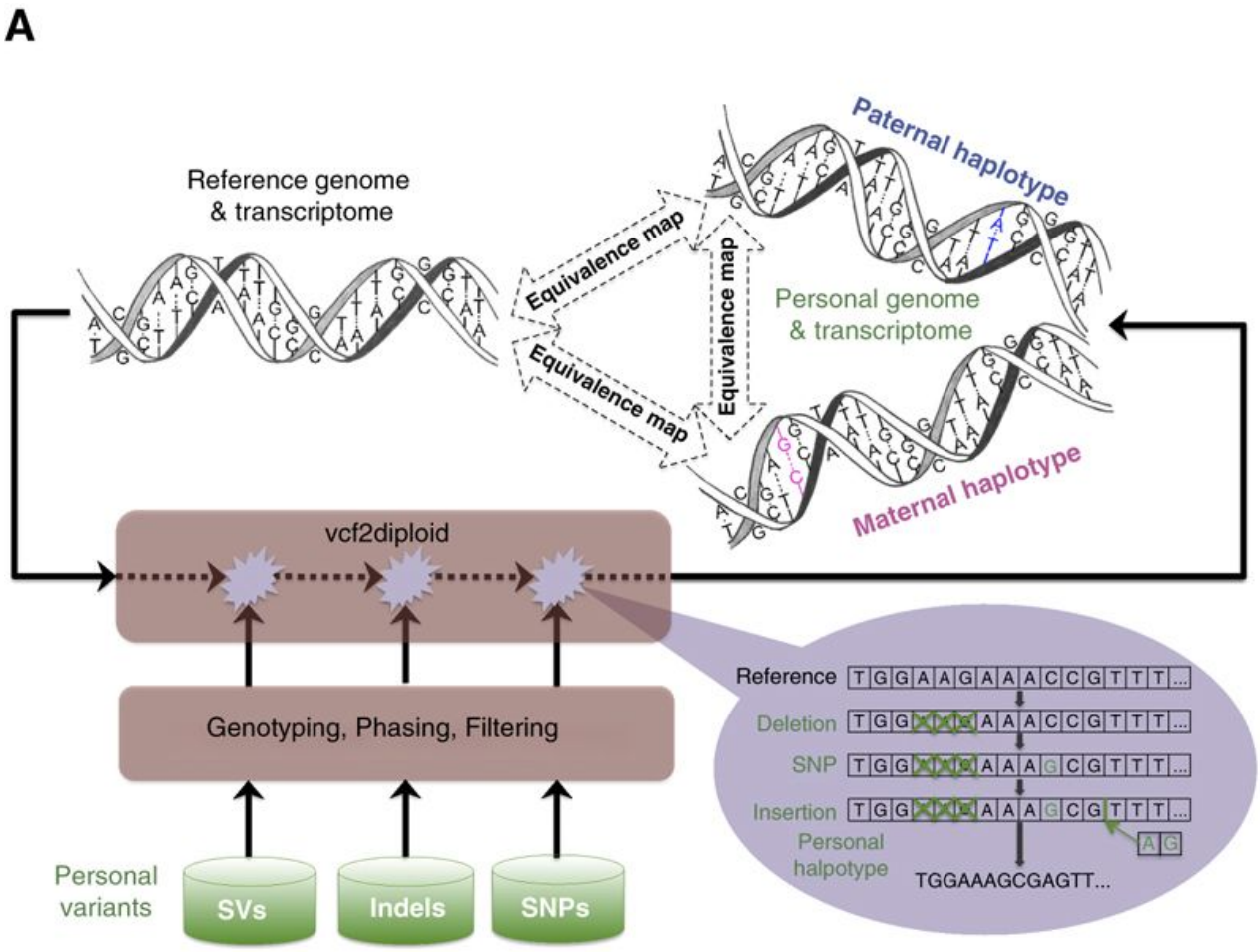


## Technical contributions

“Software was developed for building a personal diploid genome sequence, and determining sites of allele-specific binding and expression (AlleleSeq).”

## Biological findings

“This computational pipeline was used to analyze variation data, and deeply sequenced RNA-Seq and ChIP-Seq datasets, for individual NA12878 from the 1000 Genomes Project.”




(**A**) Construction of a personal genome by vcf2diploid tool is made by incorporating personal variants into the reference genome. Personal variants may require additional pre-processing, that is, filtering, genotyping, and/or phasing. The output is the two (paternal and maternal) haplotypes of personal genome. During the construction step, the reference genome is represented as an array of nucleotides with each cell representing a single base. Iteratively, the nucleotides in the array are being modified to reflect personal variations. Once all the variations have been applied, a personal haplotype is constructed by reading through the array. Simultaneously, equivalence map (MAP-file format—see Supplementary Figure 1) between personal haplotypes and reference genome is being constructed. This can similarly be done for a personal transcriptome.

# Using the vcf2diploid pre-processing utility

## Why?

“One might think that it is possible to simply map the sequenced reads against the reference genome in order to determine allele differences; however, **this introduces reference biases.** Most analyses of human genomic data use the reference genome sequence for comparison; nevertheless, when genome scale analysis of allele-specific behavior is performed we show that it is necessary to **align reads against a diploid sequence** for that individual.”

## Pros and cons

+ Plug-and-play with a VCF and a reference genome
- Chain and Map file formats are cryptic and poorly documented, but there isn’t really an alternative
- Doesn’t incorporate complex SVs, actually only insertions/deletions

# A detour into reference bias...

Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data, Degner et al. *Bioinformatics*, 2009

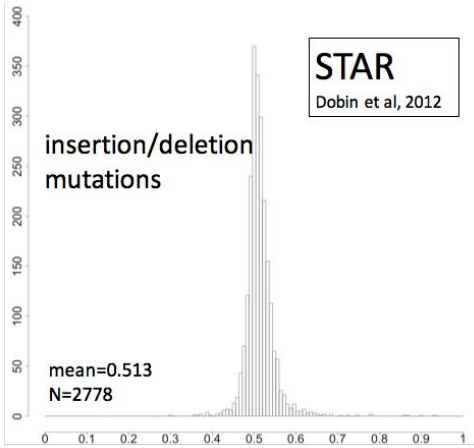
"We generated 16 million 35 bp reads from mRNA of each of two HapMap Yoruba individuals. When we mapped these reads to the human genome we found that, at heterozygous SNPs, there was a significant bias toward higher mapping rates of the allele in the reference sequence, compared with the alternative allele [using MAQ, Bowtie, BWA]. Masking known SNP positions in the genome sequence eliminated the reference bias but, surprisingly, did not lead to more reliable results overall. We find that even after masking, **~5–10% of SNPs still have an inherent bias toward more effective mapping of one allele. Filtering out inherently biased SNPs removes 40% of the top signals of ASE.**"

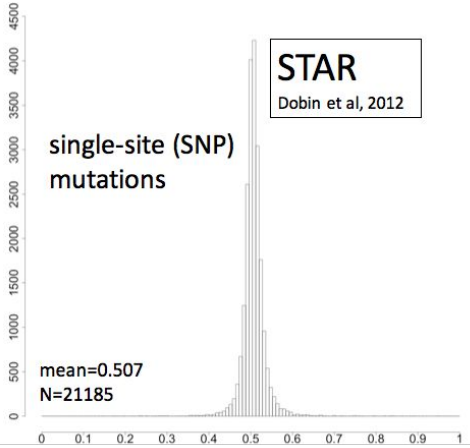
# A detour into reference bias...

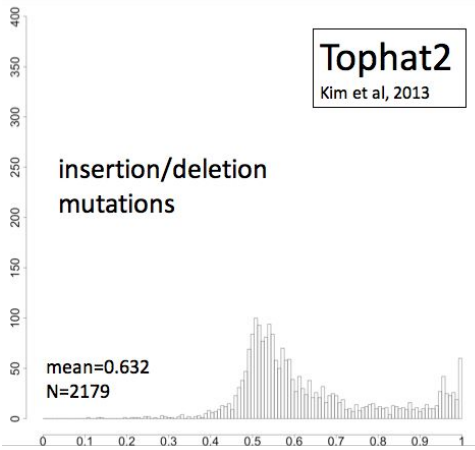
Allelic mapping bias in RNA-sequencing is not a major confounder in eQTL studies, Panousis et al. *Genome Biology*, 2014

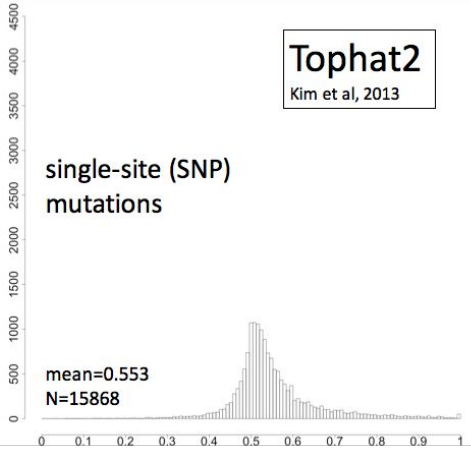
“After alignment of these reads to the reference genome with BWA, we found that as much as **13.86% of the unique start sites showed unequal mapping of reads carrying different alleles**."

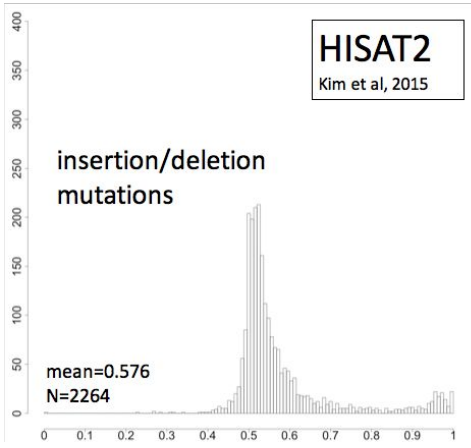
“However, removing potentially biased RNA-seq reads from an eQTL dataset of 185 individuals has a **very small effect on gene and exon quantifications and eQTL discovery**. We detect only a handful of likely false positive eQTLs, and overall eQTL SNPs show no significant enrichment for high mapping bias.”

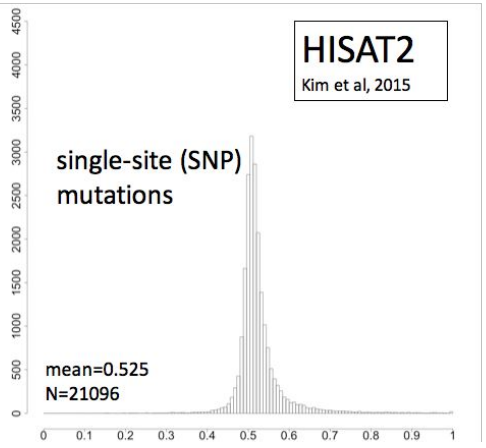



| Experiment | % reads map to reference | % alignments correct* | % reads map to origin genome | % alignments correct* |
|---|---|---|---|---|
| 1% mutation, Tophat | 77.4 | 86.6 | 99.4 | 98.6 |
| 1% mutation, STAR | 99.9 | 85.8 | 100. | 98.1 |
| 1% mutation, HISAT2 | 93.7 | 87.8 | 99.9 | 99.0 |
| 0.1% mutation, Tophat | 98.4 | 95.1 | 99.4 | 96.4 |
| 0.1% mutation, STAR | 99.9 | 95.2 | 99.9 | 96.4 |
| 0.1% mutation, HISAT2 | 99.8 | 96.3 | 99.9 | 97.4 |

*correct = leftmost positions of both mates align +/-4 base pairs from the simulated locations

# Ameliorating reference bias

Filter sites with likely bias

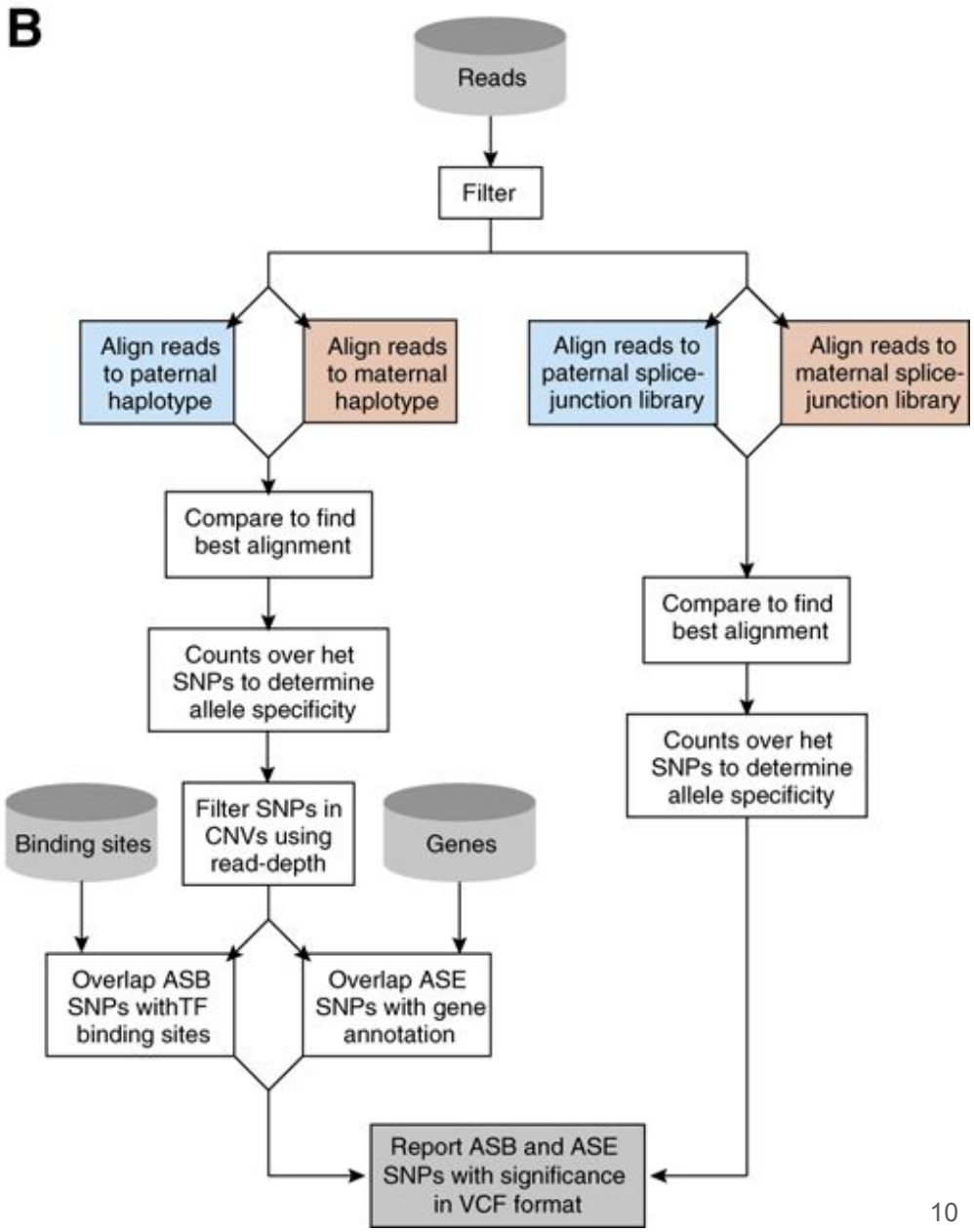
Filter reads with likely bias (WASP, van de Geijn et al. bioRxiv 2014)

Assume the “null distribution” is not exactly 50-50 read mapping (Castel et al., Genome Biology 2015)

Variant or haplotype-aware alignment (HISAT2, GSNAP)

# That was all without a diploid reference … but assuming we have the resources to build one, the authors say it has to be high-quality

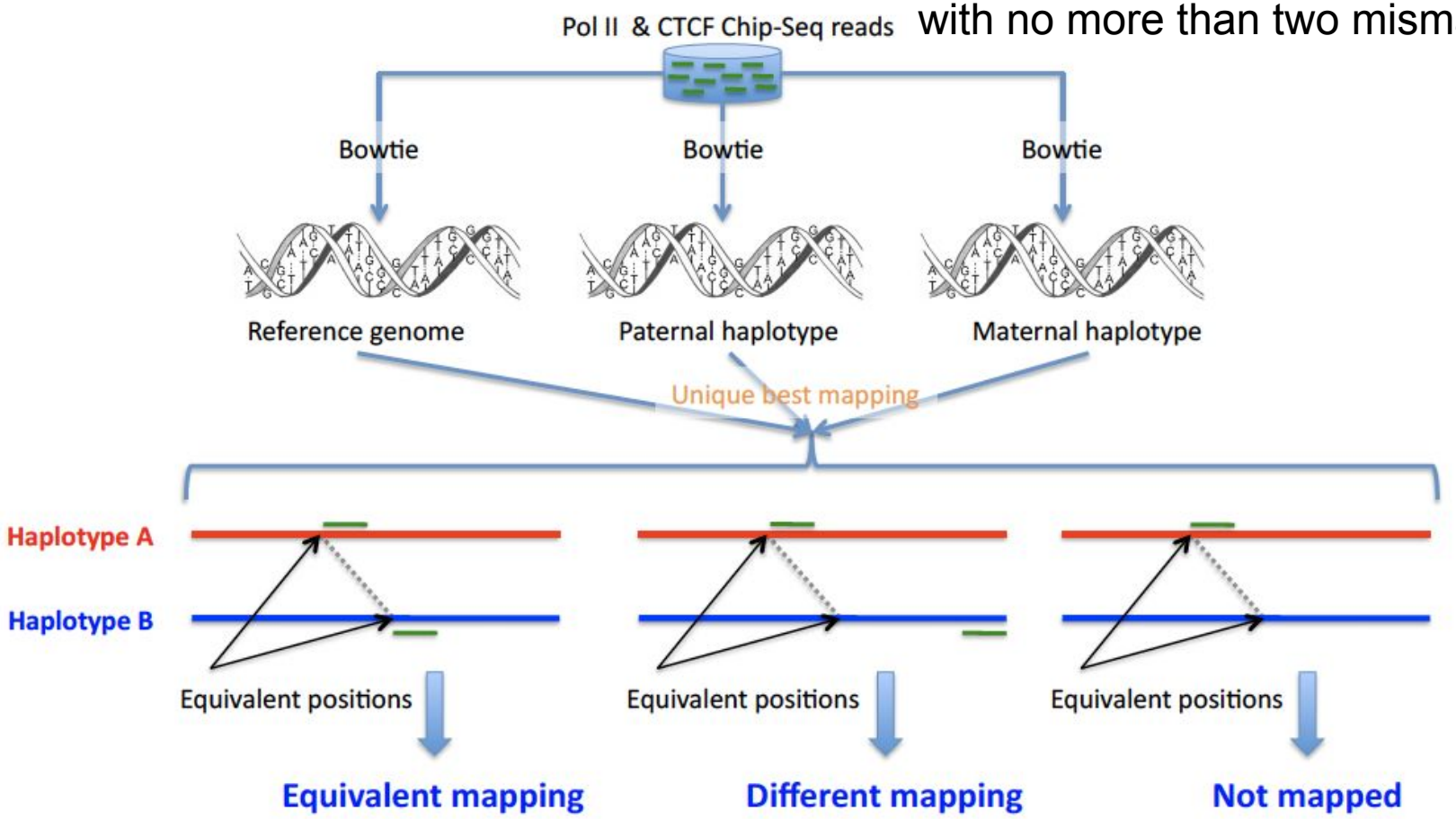
- Can’t use the functional assay itself to call variants (Supplementary Table 1)
- Genotyping arrays don’t [didn’t?] have enough variants
- Variants need to be phased, as well as called
- NA12878 was probably [one of] the only confidently phased genomes available with trio WGS at the time this work was done




(**B**) AlleleSeq pipeline for determining allele-specific binding (ASB) and allele-specific expression (ASE) aligning reads against the personal diploid genome sequence as well as a diploid-aware gene annotation file (including splice-junction library).

# How they “compare to find best alignment”

Best unambiguous Bowtie mapping with no more than two mismatches



# How they “compare to find best alignment”

**Table III** Comparison of read mappings to reference genome and paternal and maternal haplotypes of GM12878

| Haplotype | No. of mapped reads | Reference | Paternal | Maternal |
|---|---|---|---|---|
| | | Equivalently mapped reads in | | |
| Reference | 69 086 591 | | 68 942 501 (99.79%) | 69 034 357 (99.92%) |
| Paternal | (+0.3%) 69 296 783 | 68 942 501 (99.49%) | | 69 099 705 (99.72%) |
| Maternal | (+0.4%) 69 394 995 | 69 034 357 (99.48%) | 69 099 705 (99.58%) | |
| | | Differently mapped reads in | | |
| Reference | 69 086 591 | | 18 248 (0.03%) | 18 291 (0.03%) |
| Paternal | (+0.3%) 69 296 783 | 18 248 (0.03%) | | 113 796 (0.16%) |
| Maternal | (+0.4%) 69 394 995 | 18 291 (0.03%) | 113 796 (0.16%) | |
| | | Unmapped reads in | | |
| Reference | 69 086 591 | | 125 842 (0.18%) | 33 943 (0.05%) |
| Paternal | (+0.3%) 69 296 783 | 336 034 (0.48%) | | 83 282 (0.12%) |
| Maternal | (+0.4%) 69 394 995 | 342 347 (0.49%) | 181 494 (0.26%) | |

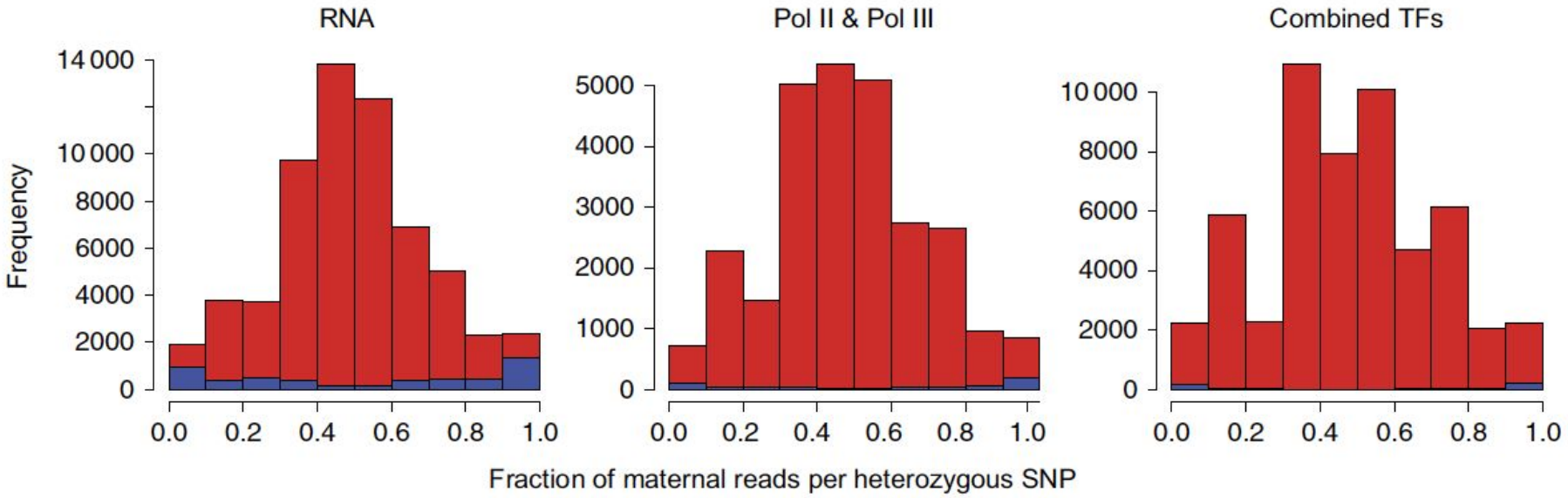
Does a < 1% better mapping rate matter? These reads are probably enriched for overlapping variant sites.

Why are there reads that map only to reference?

# Determining ASE/ASB from maternal:paternal ratio

Binomial test over read counts at heterozygous SNPs

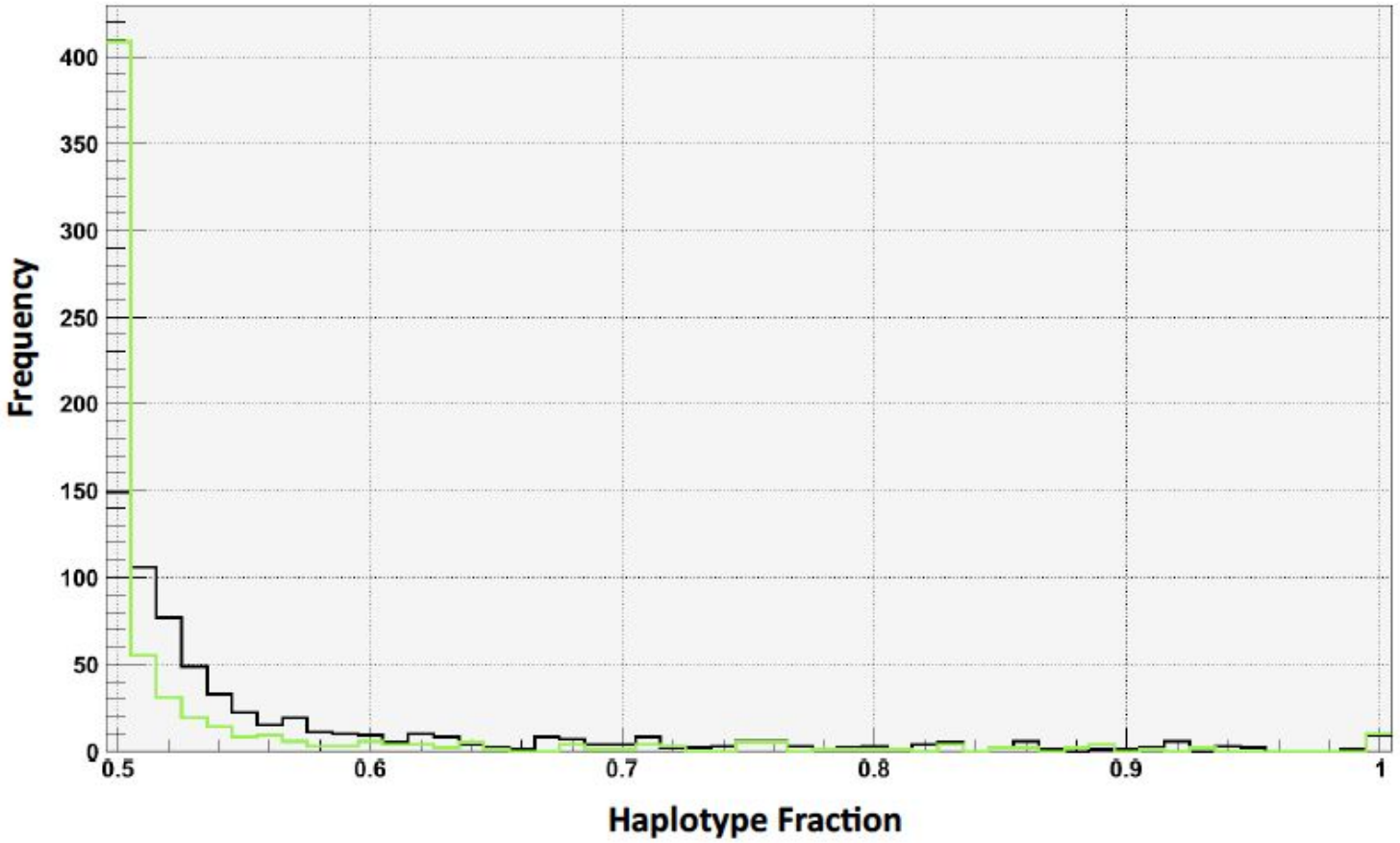
exclude CNV; correct for FDR with random read assignment simulation

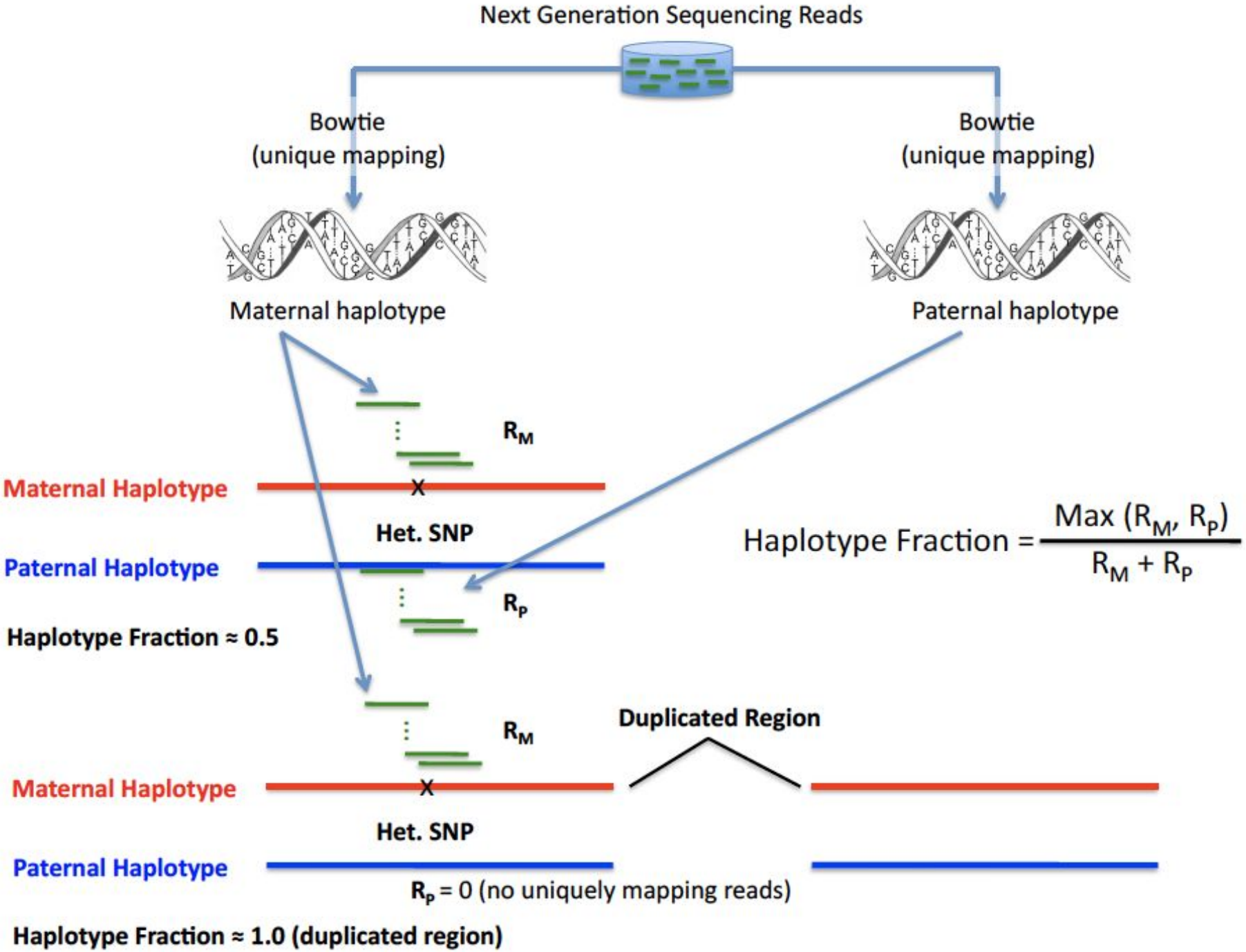



For each heterozygous SNP location covered at a depth greater than six, we can compute the fraction of reads the reference sequence. We then plotted the distribution of alternative allele fraction for all heterozygous SNPs (significant allele-specific positions are indicated in blue) for the RNA-Seq, Pol II, and remaining ChIP-Seq data sets combined. We observe that the distribution of all heterozygous SNPs as well as the allele-specific SNP positions is quite symmetric; and thus, we do not see a significant reference bias.

# CNV is not too big of a concern anyway

Supplementary Figure 6. Read mapping skew towards either haplotype at ASB site for Pol II. For each ASB site the skew was calculated as a fraction of reads mapped to one haplotype over the sum of reads mapped to each haplotype. At each site the maximum value of skew is used. Black line shows the distribution of skew when reads mapping uniquely to each either haplotype are considered. Green line shows the distribution of skew when read mapping to each haplotypes are used.




Next Generation Sequencing Reads
Bowtie
(unique mapping)
Bowtie
(unique mapping)
Maternal haplotype
Paternal haplotype
$R_M$
Maternal Haplotype
Het. SNP
Paternal Haplotype
$R_P$
Haplotype Fraction ≈ 0.5
$R_M$
Duplicated Region
Maternal Haplotype
Het. SNP
Paternal Haplotype
$R_P$ = 0 (no uniquely mapping reads)
Haplotype Fraction ≈ 1.0 (duplicated region)
Haplotype Fraction = $\frac{\text{Max}(R_M, R_P)}{R_M + R_P}$

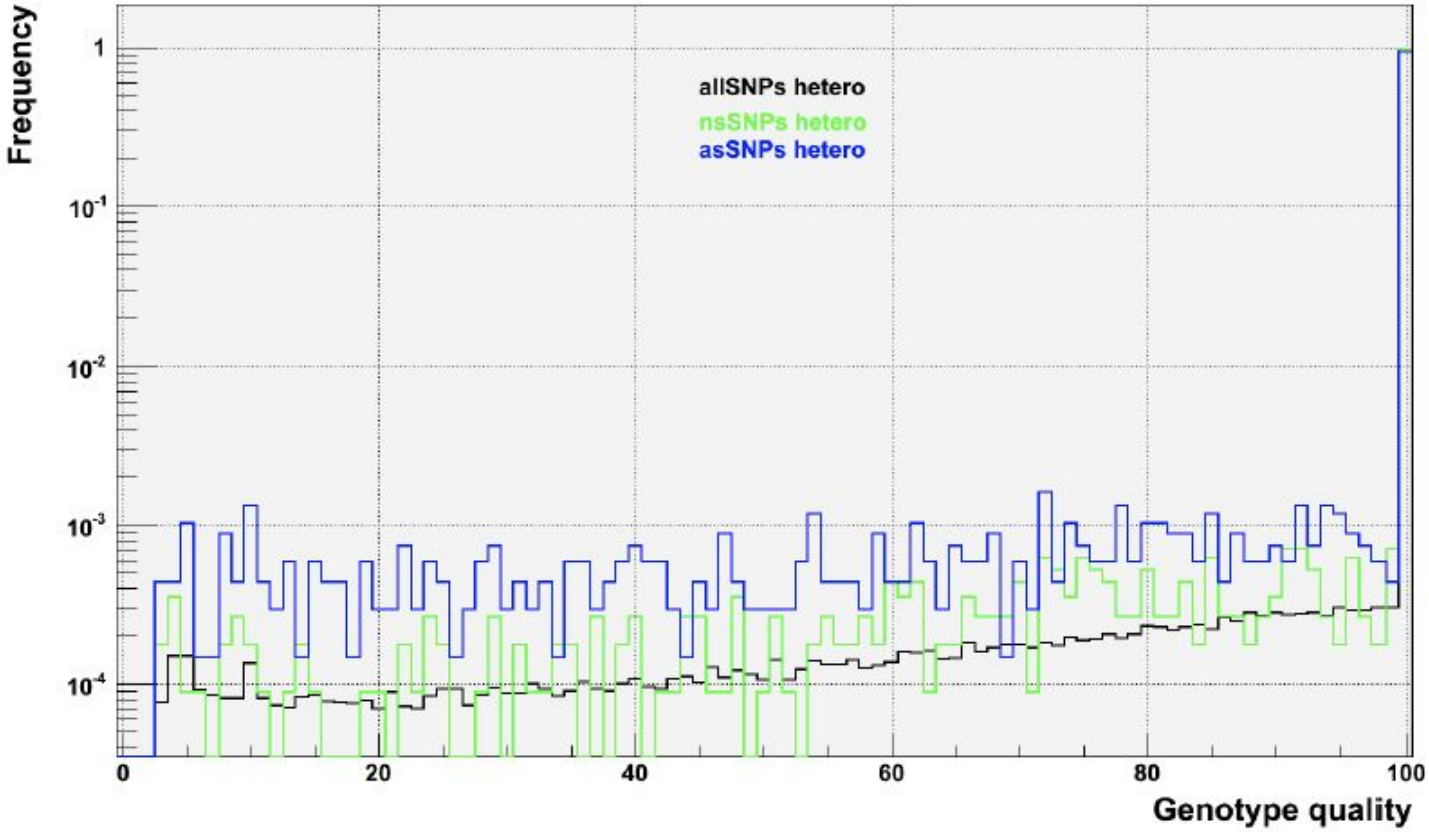
# Findings in NA12878 using the AlleleSeq pipeline

“We observe that 19.4% of all autosomal GENCODE genes that are accessible for allele-specific behavior exhibit ASE.We similarly find that 21.6% of accessible heterozygous SNPs within splice junctions of genes also show ASE. Similarly, we find that 9.3% of autosomally expressed accessible novel TARs [transcriptionally active regions] show ASE.”

# Known results are corroborated

Autosomes show equal allele-specific activity from each parent
Observe dynamics of X-chromosome inactivation
85% of CTCF sites in (McDaniell et al., Science 2010) are also found

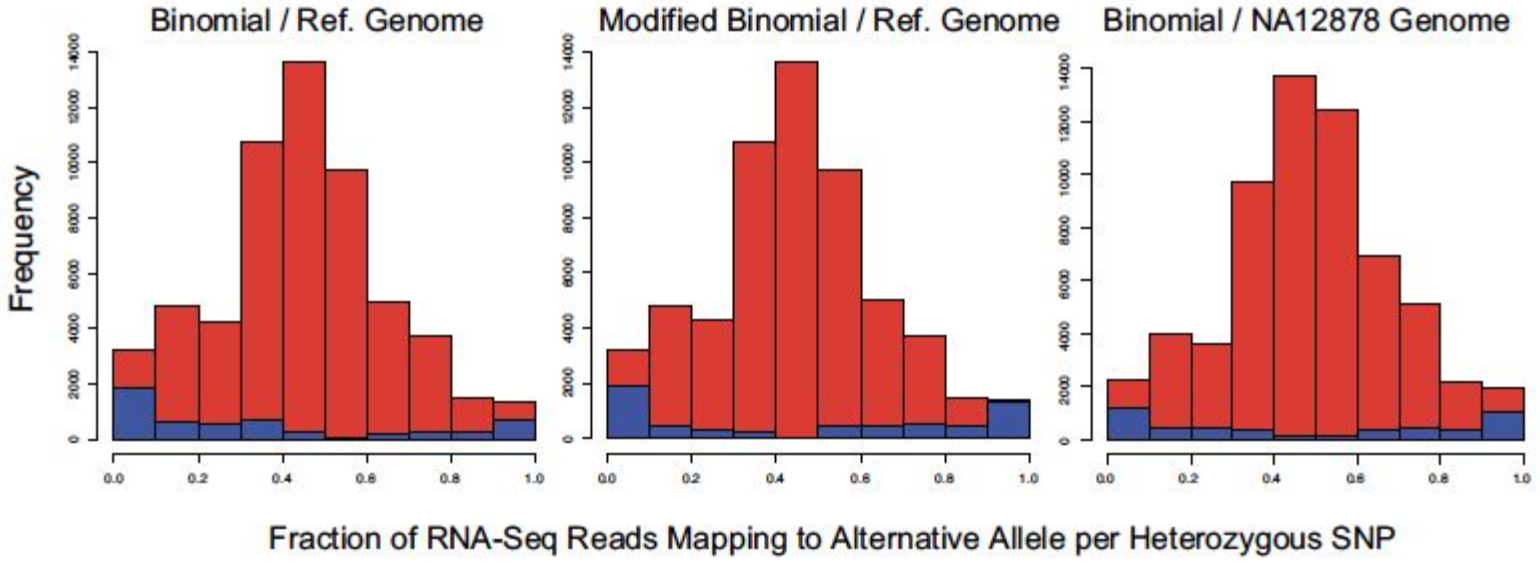
# ASE SNPs are lower quality, but not too much




Supplementary Figure 4. Density distribution of heterozygous SNPs in NA12878 by genotyping qualities (all SNPs -- black line, SNPs that are identified as allele specific in this study -- blue line, non-synonymous SNPs – green line). AS SNPs are on average of slightly lower quality than all SNPs, however the vast majority of them (~99%) are confidently genotyped (quality score > 30). Non-synonymous SNPs also exhibit a very similar distribution.

# Using a binomial test with modified null hypothesis isn't good enough

"Using the naive methodology with an even binomial we see the skew of the ASE SNPs toward the reference genome which is largely removed using the modified binomial test. When comparing our set of 5862 ASE SNPs determined using the personal genome we find that only 69% are shared with those determined using the naive approach. Using the modified binomial methodology from Montgomery et al (2010), we see an improvement (75% in common); however, we still are detecting a significant number of ASE sites that were missed aligning to the reference genome and only modifying the binomial test versus using the correct diploid genome to align against."



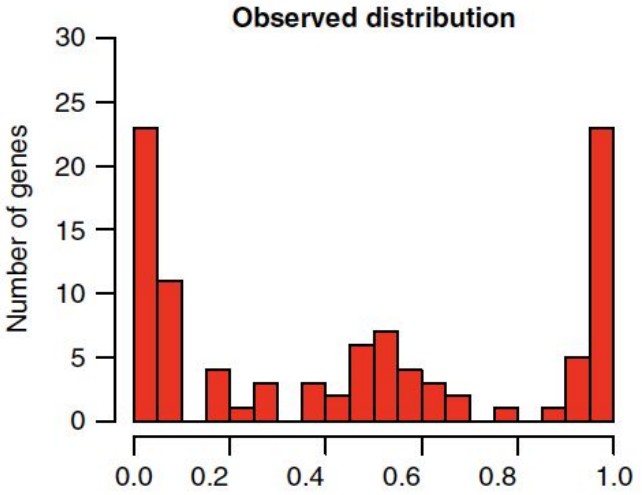
# Findings are generally intuitive

For motifs (containing a SNP) the allele that is closer to the “known PWM” (better fitness) generally has more binding

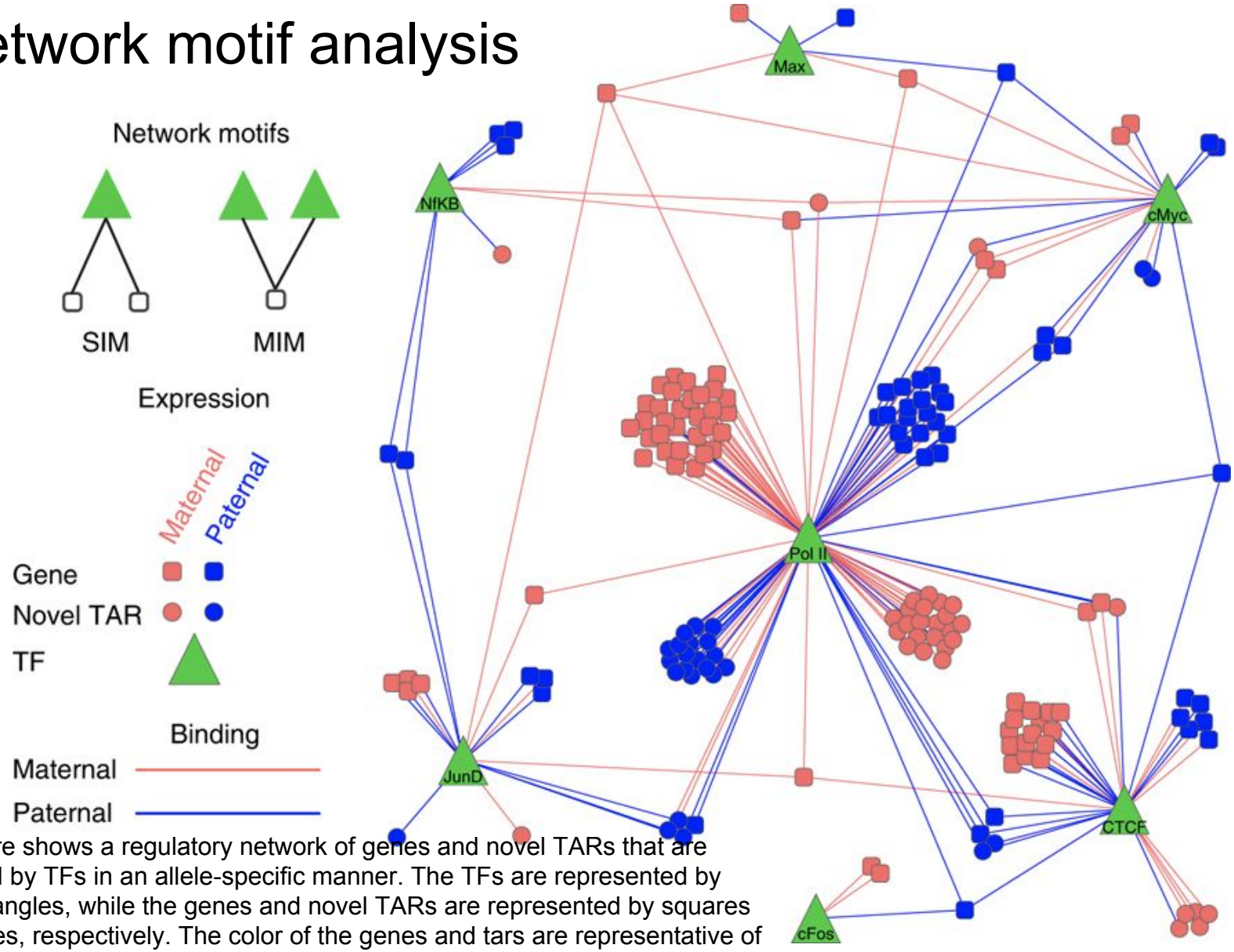
Proteins with a non-synonymous mutation don’t correlate with ASE

ASE and ASB are correlated, as one might expect

Allele-specific SNPs tend to be from one haplotype or the other



# Network motif analysis

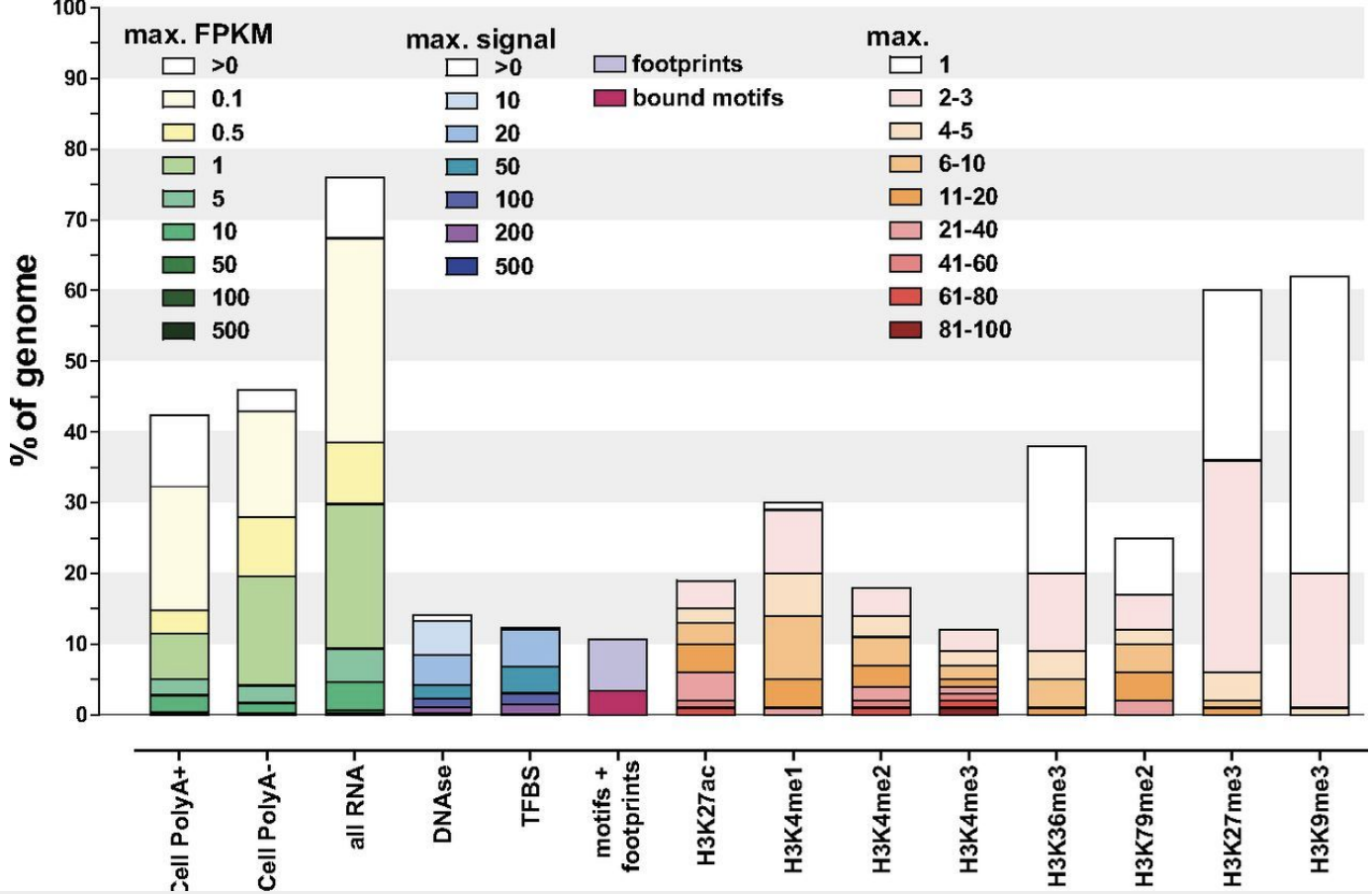



This figure shows a regulatory network of genes and novel TARs that are regulated by TFs in an allele-specific manner. The TFs are represented by green triangles, while the genes and novel TARs are represented by squares and circles, respectively. The color of the genes and tars are representative of their allele-specific expression and the edges from TFs, which represent regulation by TFs, to them likewise; the colors used are pink for maternal, and blue for paternal. As it can be observed, there is significant agreement between allele-specific regulation and allele-specific binding.

# Bringing the analysis up-to-date

More phased diploid genomes and functional assays (Entex)

Improved read-mapping and read-generating technologies: is it as bad now?




Summary of the coverage of the human genome by ENCODE data. The fraction of the human genome covered by ENCODE-detected elements in at least one cell line or tissue for each assay is shown as a bar graph. All percentages are calculated against the whole genome, including the portion that is not uniquely mappable with short reads and thus is invisible to the analysis presented here (see Fig. S1). A more detailed summary can be found in Fig. S2. For transcripts, coverage was calculated from RNA-seq–derived contigs (104) using the count of read fragments per kilobase of exon per million reads (FPKM) and separated into abundance classes by FPKM values. Note that FPKMs are not directly comparable among different subcellular fractions, as they reflect relative abundances within a fraction rather than average absolute transcript copy numbers per cell. Depending on the total amount of RNA in a cell, one transcript copy per cell corresponds to between 0.5 and 5 FPKM in PolyA+ whole-cell samples according to current estimates (with the upper end of that range corresponding to small cells with little RNA and vice versa). “All RNA” refers to all RNA-seq experiments, including all subcellular fractions (Fig. S2). DNAse hypersensitivity and transcription-factor (TFBS) and histone-mark ChIP-seq coverage was calculated similarly but divided according to signal strength. “Motifs+footprints” refers to the union of occupied sequence recognition motifs for transcription factors as determined by ChIP-seq and as measured by digital genomic footprinting, with the fuscia portion of the bar representing the genomic space covered by bound motifs in ChIP-seq. Signal strength for ChIP-seq data for histone marks was determined based on the *P* value of each enriched region (the –log10 of the *P* value is shown), using peak-calling procedures tailored to the broadness of occupancy of each modification (*SI Methods*).



Defining functional DNA elements in the human genome, Kellis et al. (PNAS 2014)