**AlleleSeq: Analysis of Coordination of Allele-Specific Expression and Binding in a Network Framework**

Joel Rozowsky\*,†,1,2, Alexej Abyzov\*,1,2, Jing Wang2, Pedro Alves2, Debasish Raha3, Arif Harmanci1,2, Jing Leng2, Robert Bjornson4,5, Yong Kong5, Naoki Kitabayashi6, Nitin Bhardwaj1,2, Mark Rubin6, Michael Snyder7 and Mark Gerstein†,1,2,4.

1. Department of Molecular Biophysics and Biochemistry, Yale University, New Haven CT, USA.

2. Program in Computational Biology and Bioinformatics, 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.

7. Department of Genetics, Stanford University School of Medicine, Stanford CA, USA.

\* These authors contributed equally.

† Correspondence should be sent to pi@gersteinlab.org

**Abstract:**

To study allele-specific-expression (ASE) and binding (ASB), i.e. differences between the maternally and paternally derived alleles, we have developed a computational pipeline (AlleleSeq). Our pipeline initially constructs a diploid personal genome sequence using genomic sequence variants (SNPs, indels and structural variants), and then identifies allele-specific events with significant differences in the number of mapped reads between maternal and paternal alleles.  There are many technical challenges in the construction of and alignment of reads to a personal diploid genome sequence that we address, e.g. references biases. We have applied AlleleSeq to variation data for NA12878 from the 1000 Genomes Project as well as matched, deeply sequenced RNA-Seq and ChIP-Seq datasets generated for this purpose. In addition to observing fairly widespread allele-specific behavior within individual functional genomic datasets (including results consistent with X-chromosome inactivation), we can study the interaction between ASE and ASB. Furthermore we investigate the coordination between ASE and ASB from multiple transcription factors events using a regulatory-network framework. Using correlation analyses and network motifs we observe mostly coordinated allele-specific binding and expression.

**Introduction**

Due to rapidly increasing throughput and decreasing costs next-generation short read sequencing is fast replacing array based technology for performing functional genomic assays such as mapping locations of transcription factor binding or determining transcribed sequences in the genome. The initial analyses of high-throughput functional data using ChIP-Seq (Johnson *et al.* 2007 and Robertson *et al.* 2007) or RNA-Seq (Nagalakshmi *et al.* 2008 and Mortazavi *et al.* 2008) yields similar results that were obtained using tiling array based methodologies albeit with greater sensitivity and resolution, i.e. binding regions or regions of transcription. Also, with the developments in sequencing technologies there have been increasingly larger studies of the amount of sequence variation across the human population (The 1000 Genomes Consortium 2010). A natural area of recent focus has been looking at the degree of functional genomic differences across the human population (Kasowski *et al.* 2010, McDaniell *et al.* 2010, Montgomery *et al.* 2010, Pickrell et al. 2010, Gregg et al. 2010a, Gregg et al. 2010b). However, in order to understand population effects it is first useful to characterize the effects of functional variation within a single individual such as differences of expression and binding between alleles (i.e. allele-specific differences). When comparing functional data between individuals it is necessary to worry about normalization before any comparisons are performed; however, within a single individual there is a natural control of each allele against each other. By utilizing the actual sequence composition of the functional genomic sequence reads that overlap a heterozygous polymorphism it is possible to determine the sequences that originate from each allele separately (Degner *et al.* 2009, McDaniell *et al.* 2010, Montgomery *et al.* 2010, Pickrell *et al.* 2010, Lalonde *et al.* 2011). Thus it is possible to determine sites where transcription or transcription factor binding is originating predominately from one allele, i.e. allele-specific expression (ASE) or allele-specific binding (ASB); however, there are number of technical challenges which make this analysis challenging.

In each of the recently published studies that contained some level of allele-specific analysis, only one type of functional genomic assay was performed. A logical question is how these allele-specific events are coupled between assays. At first glance we expect significant coordination between binding of different transcription factors and expression of target genes. This has been previously been studied in a more limited fashion using array based technologies (Maynard et al. 2008). Here we address this question by analyzing a number of different functional genomic datasets using a pipeline that we have developed, AlleleSeq, for determining sites showing allele-specific behavior. For the first time we analyze allele-specific behavior for both transcription data using a very deeply sequenced RNA-Seq dataset (~160 million mapped reads) as well as matching deeply sequenced ChIP-Seq datasets (~30 - ~60 million mapped reads) for a number of different transcription factors (cFos, cMyc, JunD, Max, NfκB, CTCF) as well as polymerases (RNA Polymerase II and Polymerase III). These experiments were generated for the lymphoblastoid cell line GM12878, which has also been deeply sequenced together with both parents (as a trio) as part of the pilot II phase of the 1000 genomes project (The 1000 Genomes Consortium 2010). Thus, for these datasets we have a complete set of heterozygous variants (SNPs, indels, and structural variations (SVs)) for the individual NA12878, which can mostly be phased into maternal and paternal variants by comparing against the parents sequences. This is important for assessing the genome-wide amount of allele-specific behavior, which is severely limited by the number of identified heterozygous SNPs available (for instance Montgomery *et al.* (2010) and Pickrell *et al.* (2010) used HapMap III SNP calls which are approximately 10 fold fewer than those available from pilot II of the 1000 genomes project). Allele-specific behavior is presumably occurring also in regions devoid of heterozygous SNPs, where we cannot distinguish between the alleles. When assessing the number of comparisons of allele-specific behavior between transcription factor binding and expression, 10 fold fewer total number of heterozygous SNPs would only allow for approximately 100 fold fewer comparisons between ASB and ASE SNPs to be made.

There are numerous technical hurdles in determining allele-specific behaviour. Naively one would think that it would be possible to call heterozygous SNPs ab initio from the sequenced functional genomic datasets; however, as we will demonstrate it is necessary to have an independently determined set of SNPs, which are detected using direct sequencing of DNA such as was done by the 1000 Genomes Project. The same is likely true for other variants. Also, 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. We deal with this by constructing a diploid personal genome sequence by using the variation data (both for SNPs, indels and SVs) for NA12878 (The 1000 Genomes Consortium 2010, Mills *et al.* 2010). While the 1000 Genomes Project has created call-sets of sequence variants for each of the different genomes sequenced, they have not however assembled genome sequences (including NA12878) for each of the individuals sequenced. In the first part of our AlleleSeq pipeline, we generate a diploid genome sequence of maternal and paternal haplotypes by integrating the phased variation data (SNPs, indels and SVs) onto the reference genome sequence. In addition we filter out genomic sequences that are likely to correspond to copy number variants (CNVs) using read depth analysis (Abyzov *et al.* 2010). Construction of individual personal reference diploid sequences, as a first step for functional genomic analysis, will likely become standard in the near future.

In this paper we show that ASE of genes as well as novel transcribed regions, i.e. novel transcriptionally active regions (TARs) or transfrags (Kapranov *et al.* 2002, Rinn *et al.* 2003, Bertone *et al.* 2004), are coordinated with ASB of transcription factors and other DNA binding proteins located adjacent to the transcribed region. One can measure how well ASB and ASE are coordinated, by using a correlation plot of the two. However, representing the coordination between multiple allele-specific events is difficult. In order to facilitate this, we show how ASB for multiple transcription factors is coordinated with ASE of the target genes or novel TARs by visualizing this behavior using a simplifying regulatory network. We will see how certain allele-consistent regulatory motifs are enriched using network analysis. We will observe that ASB and ASE are not as coordinated as might have been naively expected and speculate on potentially complexities of allele-specific regulation.

**Results**

We start by assembling a set of sequence variants from the 1000 Genomes Project for the NA12878 individual. We then generated deeply sequenced ChIP-Seq datasets for cFos, cMyc, JunD, Max and RNA Polymerase II for the GM12878 cell line. We also created a matching deeply sequenced RNA-Seq dataset for the same cell line. We combined these datasets with previous published matching datasets for RNA Polymerase II, RNA Polymerase III, NfκB and CTCF (Raha *et al.* 2010, Kasowski *et al.* 2010, McDaniell *et al.* 2010). We summarize these datasets in Table 1 (see **Methods** for further details).

Determining allele-specific behavior from functional genomic data alone

Intuitively if one has performed a deeply sequenced functional genomic experiment such as RNA-Seq or ChIP-Seq from a single individual it should be possible to determine allele-specific behavior solely from the sequences obtained. The first step in this approach would be to determine the SNPs and other sequence variants directly from the sequence reads obtained. This might be true for certain regions sequenced at great depth; however, since functional genomic data (e.g. reads from a ChIP-Seq experiment) covers the genome with greatly varying sequence depth due to the nature of the functional assay. Thus, the accuracy of SNP (and other variant) calling from functional genomics data will necessarily vary across the genome. Conversely, the average sequencing depth across the genome for conventional genomic DNA sequencing is nearly uniform (with some differences to repeated regions and compositional biases).

We find that the accuracy of de novo SNP calling using reads from a functional genomic sequencing experiment such as RNA-Seq using a standard SNP caller package (e.g. SNVmix, Shah *et al.* 2009) is not as good as we would need for determining allele specific behavior. Any significant amount of miss calling of heterozygous SNPs will (obviously) lead to ill determined allele-specific behavior. There are a number of possible explanations for such mis-calls: first the very events we would like to find, SNPs within regions showing allele-specific expression, could potentially appear as homozygous using only the RNA-Seq sequence reads. If one experimentally only obtains sequences from a region that is expressed on one allele (due to ASE) then there is no way to know that any base within that region is polymorphic. Second, RNA editing could also lead to variations in RNA sequences that are not present at the DNA level. Lastly, sequencing of RNA involves additional experimental steps, like usage of reverse transcriptase that can increase chance of mis-sequencing.

Obviously, determining short indels from sequenced functional genomic data would be even harder and structural variants would be nearly impossible. Thus, while it might be possible to determine certain sequence variants from the functional genomic sequence reads, in order to generate a comprehensive set of polymorphic sites as well as other forms of sequence variation it is necessary to have an independently determined set from sequenced genomic DNA (such as from the 1000 Genomes Consortium).

Building an individual diploid reference genome for NA12878

It might not seem obvious but for a number of reasons reconstruction of a diploid personal genome sequence and using it instead of the reference genome is a critical step preceding allele-specific analysis. First, using reference genome introduces biases in read mapping – reads originated from non-reference allele are more susceptible to mismapping since, when aligned to the reference allele, they contain at least one mismatch (in case of SNPs) or gap (in case of indels) – the reference bias effect, i.e. both alleles are not treated equally by default. Second, expression or binding in regions of genome structural variation (SV) could be misinterpreted as allele-specific expression or binding. For example, duplication of an allele in the studied genome will double binding signal for the allele while signal for the allele on another haplotype will be unchanged. Last, but not least, SNP calling in the regions of SV are likely to be less precise and contain more false positives compared to non-SV regions (The 1000 Genomes Consortium 2010). We thus construct a personal diploid genome of NA12878 (see **Methods**), by utilizing genomic variations (see Table 2 for summary statistics) determined in the framework of the 1000 Genomes Project (The 1000 Genomes Consortium 2010) and, additionally, structural variants determined by sequencing of fosmid clones (Kidd et al., 2008).

To accomplish this we have developed a tool – vcf2diploid – for personal genome construction, which constitutes the first part of the AlelleSeq pipeline (see Figure 1a). The tool uses as input VCF files with all the SNPs, indels and SVs available for an individual of interest and outputs fasta sequences for each allele for each chromosome, along with equivalence map files (see Figure 1 and Supplementary Figure 1) that map nucleotide positions between paternal, maternal and reference haplotypes. It is important to be able to map annotation (i.e. genes) from the reference genome to the personal genome sequences. This is done using chain files, which facilitate the mapping of annotated regions between genomes using the liftOver tool (Rhead *et al.* 2010). This is particularly important for RNA-Seq where we also build maternal and paternal versions of the splice-junction library by mapping the GENCODE gene annotation (GENCODE 3c annotation is available from the UCSC Genome Browser, Harrow *et al.* 2006) onto the personal diploid genome.

The constructed diploid genome of NA12878 was different in 3,962,637 (~0.14%) bases from the reference for paternal and in 3,947,162 (~0.14%) for maternal alleles. The software package to perform personal genome sequence construction (the vcf2diploid tool and associated source code), the actual diploid sequence for NA12878, splice junction sequences for NA12878 and corresponding equivalence maps (between the maternal and paternal sequences as well as the reference genome, NCBI36/hg18) are available from alleleseq.gersteinlab.org. The diploid sequence for NA12878 is a valuable resource for anyone performing any sequence based analysis on this genome. The GM12878 cell lines is a primary tier one cell line under detailed investigation by the ENCODE Consortium. It should be also noted that a constructed personal genome is only as good and as complete as the variants used in construction. In light of this, the diploid genome of NA12878 that is presented here, is not perfect, but we believe it is the best possible sequence to date since it includes the most comprehensive set of variants. We intend to update this assembly as a resource as sequence variants are even more accurately determined.

In order to assess the effect of the differences between the maternal and paternal sequences compared to using the reference genome sequence on functional genomic data, we aligned the reads from the Pol II and CTCF ChIP-Seq data for GM12878 against each of the three sequences using BOWTIE (Langmead *et al.* 2009) (see supplementary Figure 2). In Table 3 we compare the POl II reads that align to each of the three genome sequences (reference, maternal and paternal haplotypes). We observe that by allowing up to two mismatches more reads (0.3% for paternal and 0.4% for maternal) align to the correct NA12878 as compared to the reference genome sequence (NCBI36). The major difference in numbers for paternal/maternal and reference haplotypes is due to reads that map to one haplotype but not the other. Namely, only about 0.1-0.2% of reads that map to the reference cannot be mapped to paternal or maternal haplotype, while a significantly higher fraction of reads (~0.5%) map to the paternal or maternal genome and cannot be mapped to the reference. Interestingly, for paternal and maternal haplotypes unmapped reads and reads with different mapping locations contribute roughly equally to the differences in overall mapping, presumably mostly due to short indels and SVs. We also see similar results for the same analysis done to the reads for CTCF ChIP-Seq (see Supplementary Table 1). This demonstrates that it is important to use a correctly assembled personal genome for aligning reads when performing an allele-specificity analysis.

Determining allele-specific expression and binding

The second part of the AlleleSeq pipeline determines ASE using RNA-Seq data and ASB using ChIP-Seq data. After the maternal and paternal derived haploid sequences are constructed, sequenced reads are aligned against the maternal and paternal sequences separately using BOWTIE (Langmead *et al.* 2009). Locations of mapping are determined by selecting the best alignment to both genome sequences. Read counts for the maternal and paternal alleles are then generated at each heterozygous SNP nucleotide positions, and ASE/ASB events are reported by applying a binomial test followed by correction for multiple hypothesis testing. SNP positions that by read-depth analysis (Abyzov *et al.* 2010) are determined to be potentially in a CNV (the read depth of genomic DNA reads in a 1kb window around the SNP is either less than 1 or greater then 3) are excluded (see **Methods**). We correct for multiple hypothesis testing by estimating the false discovery rate by explicit simulation of the number of false positives given an even null background (i.e. no allele-specific events) – see Figure 1b for a schematic of the second part of the pipeline (see **Methods** for further technical details). We also align reads to the maternal and paternal splice-junction libraries and determine splice junction ASE SNPs in a similar way.

Results for GM12878 RNA-Seq and ChIP-Seq data

We start our study of allele-specific phenomena by first focusing on analyses of individual events that occur within single experimental dataset. We then analyze the coordination between binding and expression in a pair wise fashion using direct correlation. Lastly, we investigate higher order coordination of allele-specific binding and expression using a regulatory network framework that will allow us to study the agreement between multiple regulatory interactions and target expression simultaneously.

We summarize the results of the AlleleSeq pipeline applied to the RNA-Seq data and ChIP-Seq data for GM12878 in Table 4. In the upper half of the table we present the results for all the autosomes and in the lower half for chromosome X (with the all the transcription factor combined). We observe that approximately 19.4% of autosomal detectable GENCODE genes exhibit allele-specific expression (this is the fraction of genes that were expressed at sufficient levels and contained at least one heterozygous SNP within an exon of an annotated transcript for that gene). We similarly find that 21.6% of heterozygous SNPs within splice-junctions of genes also show allele-specific expression. Similarly, we find that 9.3% of autosomally expressed novel TARs show allele-specific expression, we expect this number to be lower than genes as novel TARs correspond to exons of genes. We found that for genes that contained two or more heterozygous SNP showing allele-specific behavior, greater than 81% of the time all the SNPs would show consistent ASE from the same allele (significantly greater than expected by chance), some of the exceptions could be due to allele-specific alternate splicing. For the transcription factors the fraction of detectable autosomal binding sites that exhibit allele-specific behavior varies between 2% (for cMyc) and 11% (for Pol II). A possible explanation for the difference between binding and expression allele specificity is that even though these ChIP-Seq datasets have been sequence quite deeply (see Table 1), in order to have comparable power with the RNA-Seq data one would need to sequence an order or magnitude or two further. The number of overlapping sequence reads in binding site peaks for ChIP-Seq data depends on the efficiency of the antibody used and for most ChIP-Seq datasets we do not have sufficient read-depth within a binding site as compared to the read-depth within exons of highly expressed genes. As expected when restricted to the autosomes, both ASE for genes and novel TARs and ASB for all the transcription factors and polymerases are evenly divided between the maternal and paternal alleles.

In the lower half of Table 4 we present similar results for chromosome X. Since NA12878 is female there are two copies of chromosome X. We first observe that almost 80% of the expressed genes on chromosome X that contain a heterozygous SNP exhibit allele-specific behavior and 95% of these are expressed only on the maternal copy. This is consistent with our knowledge of X-chromosome inactivation (Lyon 1961, Goto and Monk 1998) where one copy of the two copies is shut-off. There are four genes on chromosome X that show allele-specific expression on the paternal copy; however, all of these are located in the pseudo-autosomal region of chromosome X which is known to escape X-chromosome inactivation (these include Xist, SLC25A6 and SFRS17A). We observe similar results on chromosome X for the allele-specific behavior of novel TARs as well as transcription factor binding where a greater fraction of sites exhibit allele-specific behavior compared to the autosomes and almost all are expressed on the maternal copy. It is interesting to note that most of the novel transcription and binding that shows paternal allele specificity are also in the pseudo-autosomal region similar to Xist and possibly have an associated functional role.

In order to assess whether our pipeline has some residual bias towards the reference allele versus the alternate we can plot the fraction of reads from the alternative allele for each heterozygous SNP location sequenced sufficiently deeply. If there were no bias we would expect that this distribution would be symmetric having as many reference allele-specific locations as for the alternate allele. In Figure 2 we plot the alternative allele fraction distribution for the RNA-Seq data, Pol II and the other transcription factors combined. We first observe that the overall distributions are fairly symmetric and that the allele-specific events (indicated in blue) are also symmetric – this indicates that there is no residual bias towards or against the reference allele. In Supplementary Figure 3 we observe similar distributions for the fractions of reads from the maternal allele for each heterozygous SNP location that could be phased and that was sequenced sufficiently deeply in the appropriate assay.

SNPs determined by the The 1000 Genomes Consortium 2010 each have a genotype likelihood score. In Supplementary Figure 5 we plot the distribution of all heterozygous SNPs and the subset of ASE SNPs versus this genotype likelihood score. We see a slighy enrichment of ASE SNPs will lower scores, however the majority of SNPs from both distributions have the highest possible score.

It has been reported by Degner *et al.* 2009 that heterozygous sites showing apparent allele specific behavior can be caused by regions in the genome that have been duplicated. Thus even though there might be an even amount of reads from each allele only one of the alleles will have uniquely mapping reads which would lead to seemingly allele-specific behavior. In order to assess the size of this effect on our results mapping to the diploid genome sequence, we used the mapped reads for the Pol II ChIP-Seq data. At each SNP location determined to show ASB we computed the fraction of reads mapped to one allele over the sum of reads mapped to each allele (allowing for the reads to map to each allele independently) (for a plot of the distribution of sites for this fraction see Supplemental Figure 4). We observe that only a minority of the sites (less than 15%) shows a significant skew in the mapping fraction towards one haplotype (a fraction greater than 0.6). As valid ASB sites that contain additional proximal sequence variants (such as additional polymorphisms or indels) would also exhibit a mapping skew towards one haplotype, we conclude that this is an upper bound on the size of this effect and do not consider it significant.

In order to assess the effect aligning the reads against the constructed diploid genome sequence for NA12878 versus using the reference genome sequence for determining the following analysis. For the RNA-Seq reads we also aligned the reads against the reference genome and determined allele-specific expression using an even binomial distribution (threshold applied in a similar manner). As an additional comparison we also the methodology of Montgomery *et al.* 2010 where a skewed binomial distribution is used with the reads aligned against the reference genome. Similar to Figure 2 we plotted the distribution of all expressed heterozygous SNPs (ASE SNPs in blue) for these two methods in Supplemental Figure 6. Using the naïve methodology with an even binomial we see the skew of the ASE SNPs towards the reference genome which is largely remove used the modified binomial test. When comparing our set of 5,862 ASE SNPs determined using the personal genome we find that only 69% are shared with those determined using the naïve 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.

We have also compared the ASB sites for the CTCF ChIP-Seq data from the AlleleSeq pipeline against those from McDaniell *et al.* 2010. Restricting the comparison to common heterozygous SNP between the two analysis (McDaniell *et al.* 2010 used an earlier set of SNPs calls for NA12878 from The 1000 Genomes Consortium 2010) we find that greater using a p-value threshold of 0.01 on their results greater than 85% of the ASB SNPs are in common.

The AlleleSeq pipeline determines allele-specific behavior for genomic elements (transcribed regions or binding sites) that contain heterozygous polymorphisms. It is also possible to determine allele-specific behavior for genomic elements that contain a heterozygous indel. In Supplementary Table 2 we show the results transcribed exons and novel TARs as well binding sites for Pol II and CTCF that can be determined to show allele-specific behavior using a heterozygous indel to distinguish the haplotypes. We will not include these in the following analyses.

We make available the complete list of SNPs that show ASB or ASE in VCF format as a resource from our website alleleseq.gersteinlab.org. We imagine that this file may be a useful resource for researchers interested in focusing in on allele-specific sites in less deeply sequenced functional genomic experimental datasets. This might be valuable even if the functional genomic experiment was not performed on the GM12878 cell-line as regions to investigate for allele-specific behavior.

Correlation of ASB in binding sites containing known motifs

In our analysis of allele-specific binding events, heterozygous SNPs are initially only used for distinguishing between the maternal and paternal alleles (presumably allele-specific behavior also occurs in genomic regions not containing heterozygous SNPs). However, in some locations the heterozygous SNP might be the causative reason for the difference in binding between the alleles, this would most likely occur in ASB binding sites where the heterozygous SNP is within a known DNA binding motif for a transcription factor. In order to see how ASB is correlated with perturbations to known DNA binding motifs, we compared the allele that is bound against the allele that matches better to the known motif. Thus, we first scanned ASB binding sites for known motifs, position weight matrices (PWMs) obtained from TRANSFAC (Matys *et al.* 2006) and JASPAR (Portales-Casamar *et al.* 2010) (see **Methods** for further details). We correlated the nucleotide of the allele, which is preferentially bound with the fitness score of the PWM. We observed a number of significant correlations between the PWM score for both alleles and the allele that is bound. The allele that is bound tends to correspond to the better match to the known PWM. In particular we see this for the known cMyc motif within both cMyc and Max binding sites (see Figure 3). This is interesting as we observe a correlation between the allele-specific behavious of cMyc motifs with Max binding sites indicating that these transcription factors tend to significantly regulate the same target genes. We also include in Supplementary Figure 7 additional examples of the correlation between motif fitness score and the allele being bound for CTCF binding sites containing CTCF motifs and cMyc motifs within Pol II binding sites. While all these examples do not contain many data points the trend is clear.

Correlation of ASE with protein structural fitness

Since some heterozygous SNPs that show allele-specific expression are within the protein coding sequence of genes it is natural to ask whether the allele that is expressed could track with allele-dependent structural changes (for SNPs in non-synonymous positions in the protein coding sequence). However, it is not clear that we expect to find a correlation between structural fitness and allele-specific expression, as many of these polymorphisms are not selected for in the human population in any case. In order to assess whether the allele that is expressed (for genes showing ASE) is correlated with the allele containing mutations deleterious to protein structure we performed the following analysis. We compared the occurrences of ASE SNPs within genes where the SNP corresponds to a non-synonymous substitution within the protein coding sequence of the gene. Using the tool SIFT (Ng et al. 2003) we can compare the preference of the allele that is expressed with the allele whose amino acid sequence shows better structural fitness. We find that of the 20 of the 37 genes that meet these criteria show expression on the allele that has the protein sequence that has better fitness. While we find slightly more genes where the allele with better structural fitness occurs on the same allele that is expressed, this result is not significant.

Correlating ASB with ASE

In Figure 4a we present an example of the gene SKA3 on chromosome 13 which has multiple heterozygous SNPs within exons which show consistent maternal allele-specific expression which agrees with the maternal allele-specific binding of another heterozygous SNP within a Pol II binding site proximal to the 5’ end of the gene. In this example we see coordinated maternal binding of Pol II with expression of the gene. In Figure 4b we present a similar example of a novel transcribed region on chromosome 4 where we see coordinated paternal binding of Pol II with the paternal expression of the novel TAR.

These two examples show how ASB is coordinated with ASE for a known gene and a novel TAR. To investigate this trend, we assess to what extent they are coordinated on a genome-wide scale. In Table 5 we tabulate the number of genes and novel TARs that have evidence for allele-specific expression and for proximal allele-specific binding. In table 5 we present the tabulated counts for Pol II separately from all the other transcription factors combined. We also compare against the number of genes and novel TARs that could potentially have been detected as having an ASB event proximal to an ASE transcribed region (requiring both heterozygous SNPs present as well as sufficient read depthes in the RNA-Seq and ChIP-Seq signals respectively). We find a number of genes (74 genes proximal to Pol II sites and 45 genes with proximal transcription factor sites) and novel TARs (55 and 15 for Pol II and other transcription factors respectively) in which we see both allele-specific binding proximal to allele-specific expression. While these numbers might seem relatively small, they reflect the low chance of having a heterozygous SNP in both a proximal binding site as well as the transcribed sequence of the gene. This underscores the need to sequence deeply and use a comprehensively determined set of SNPs or else we would have significantly fewer genes to be able to compare ASB and ASE.

In order to assess the degree of coordinated allele-specific behavior for the 74 genes that exhibit ASE that have a proximal ASB Pol II binding site; for each gene we can plot the maternal fraction read count for the most significant ASE SNP versus the most significant ASB SNP (if there are more that one significant heterozygous allele-specific SNP present). The maternal fraction read count is the fraction of reads overlapping an allele-specific SNP position that came from the maternal allele. In Figure 5 we see that there is a statistically significant correlation between allele-specific binding of Pol II to genes that exhibit allele-specific expression (Pearson correlation = 0.659, p-value = 3.0e-10). Similarly, as seen in Figure 5 there is also a statistically significant correlation between the 45 genes that exhibit ASE with ASB for all the combined transcription factors excluding Pol II (Pearson correlation = 0.577, p-value = 7.8e-5).

Further coordination of allele-specific behavior

As a further way to assess the coordination of allele-specific events within genes we combined all the ASE and ASB SNPs that occurred within a gene (from 2.5Kb upstream of the TSS to the TTS including introns). Using only genes that contained more that 10 SNPs showing ASE or ASB we could compute the fraction of SNPs that show maternal specificity. Ideally if all SNPs were perfectly coordinated then the fraction would be either zero or one. In the first panel of Figure 6 we see the actual distribution, most genes do show a high degree of coordination. Under a random null (where each ASE or ASB event could be maternal or paternal with equal chance) for the same genes each with the same number of SNPs we would expect a null distribution of maternal fraction computational simulated in panel 2 of Figure 6. Using a Kolmolgorov-Smirnov test we find significantly more coordination of ASB and ASE SNPs in genes than compared to the random null expectation (p-value = 6.43e-5) (see the third panel of Figure 6).

Representing ASE and ASB behavior on a regulatory network

In the previous analysis we showed that binding and expression were correlated in a pairwise fashion. Next we would like to investigate the coordination of allele-specific behavior between multiple factors and expression simultaneously. This is hard to represent in a two-dimensional correlation plot, thus we have developed a simplified representation of ASE and ASB in a regulatory network framework. Looking at the occurrences of network motifs (Milo *et al.* 2002) in this framework allows us to measure the coordination of allele-specific behavior between multiple transcription factors and target expression.

The network shown in Figure 7, represents the regulatory network of five transcription factors (cMyc, Max, cFos, JunD, NfκB and CTCF) and two polymerases (Pol II and Pol III). The network discretizes the allele-specific binding events into allele-specific regulation of target genes and novel TARs and their allele-specific expression. The edges in the network represent ASB of the TF or polymerase to the target gene or novel TAR (red edges indicate predominantly maternal regulation, blue edges indicate paternal regulation and grey edges indicate allele-specific regulation that could not be phased). ASE of the target genes is indicated by the color of the target gene or novel TAR (red – maternal, blue – paternal, grey – unphased allele-specific behavior). Thus the network contains all information on the allele specificity of the regulation and the expression of the targets. One can observe that there is a clear agreement between the allele specificity of the regulation and the expression of the target. When we tabulate the maternal/paternal regulation edges with maternal/paternal expressed genes or novel TARs (see first part of Table 6) we find that they are highly coordinated (p-value < 1e-3, Fisher exact test). Furthermore we can scan the network for coordinated regulation using multiple-input motifs (MIMs) and single-input motifs (SIMs) (Milo *et al.* 2002). MIMs are network motifs where at least two transcription factors are regulating the same target gene or novel TAR, while SIMs are motifs where a single transcription factor regulates a pair of targets. In the second part of Table 6 we tabulated the number of MIMs where the pairs of TFs (or polymerase) exhibit both maternal or both paternal regulation with the maternal or paternal expression of the target genes or novel TAR.

We find that for MIMs the regulatory allele-specific behavior is highly coordinated with the allele-specific expression of the target gene or novel TAR (p-value < 1e-3, Fisher exact test). In the third part of Table 6 we count the occurrences of single-input motifs where a TF (or polymerase) that exhibits maternal or paternal regulation for each of its targets, which can each be either maternally or paternally expressed. We similarly see a significant degree of coordination of allele-specific expression and regulation in SIMs as with MIMs.

**Discussion**

In this paper we have demonstrated that it in order to assess the effects of sequence variation on functional genomic data such as RNA-Seq or ChIP-Seq it is necessary to independently determine the sequence variants from sequenced genomic DNA such as by the 1000 Genomes Project (The 1000 Genomes Consortium 2010) . Determining sequence variation from sequenced functional genomic data directly is problematic especially if it is the same data that is being used to assess the effects of the variation. Studying allele-specific behavior is the simplest type of this analysis where it is possible to utilize the variation between the maternally and paternally derived alleles in order to detect sites of allele-specific binding and expression.

We have developed a pipeline for first building a personal genome sequence for an individual using the available sequence variants in order to construct a sequence containing both maternal and paternal haplotypes. Other groups (Adey et al. 2010, Fan et al. 2011 and Roach et al. 2011) have also been developing methods for constructing haplotypes from sequence variants with and without trios. In addition we have made available tools to enable a user to map annotation between alleles and the reference sequence from which it was derived. As more personal functional genomic data becomes available constructing a personal genome sequence will become the standard first step for analyzing the data. Also the method we have used to construct the personal genome, by overlaying sequenced variants onto a reference genome sequence, is more natural than de novo sequence assembly, given the short sequence reads generated from next generation sequence technology.

We observe that allele-specific expression (ASE) and allele-specific binding (ASB) are fairly widespread across the genome. Consistent with X-chromosome in inactivation we observe that on chromosome X the majority of the binding and expression occur on the maternally derived copy except for a couple exceptions in the pseudo-autosomal regions know to escape X-chromosome inactivation. Unlike earlier studies we were able to investigate the correlation between allele-specific binding and allele-specific expression. While not perfect we do see a significant degree of coordination between the two. It is worthwhile mentioning that not all allele-specific expression is necessary controlled by eQTLs, some might be due to imprinting or random mono-allelic expression (Gimelbrant *et al.* 2007).

Furthermore, by displaying on a regulatory network the allele-specific regulatory functions of the transcription factors and polymerases studied, together with the allele-specific expression of the target genes and novel TARs, we can investigate the coordination between multiple factors regulating shared target genes or novel TARs. We find that target genes or novel TARs that share multiple regulatory factors show highly coordinated allele-specific behavior.

In the future we imagine that the approaches presented here will be scaled up. The discovery of personal genomic sequence variants, such as being done by the 1000 Genomes Projects (The 1000 Genomes Consortium 2010) the types of experimental annotation being performed by the ENCODE Consortium (The ENCODE Project Consortium 2007) will merge into a hybrid “MyENCODE” endeavor focusing on explicit annotation of a personal genome.

**Acknowledgements**

We acknowledge Andrea Sboner and Lukas Habegger for valuable discussions. We acknowledge support from the NIH and from the AL Williams Professorship funds.

**Methods**

Experimental protocols for data generation

GM12878 cells were obtained from American Type Culture Collection (Expansion A for GM12878) and cultured by using standard conditions. RNA Pol II (8WG16) and Pol III antibodies were validated by both immunoprecipitation followed by Western blot (IP/Western) and mass spectrometry. Antibodies for cFos, cMyc, JunD, Max and NfκB were validated by IP/Western.

**ChIP-Seq.** ChIP DNA and matching input DNA control for each biological replicate was prepared from 5 × 107 formaldehyde cross-linked GM12878 cells, except after RNase and Proteinase-K digestion, ChIP DNA was purified by using QIAquick PCR Purification Kit (Qiagen). The adapters (Illumina) were ligated to ChIP DNA and sequenced. Peaks from the unique reads with two mismatches or less were scored using PeakSeq (Rozowsky *et al.* 2009) using default parameters.

**RNA-Seq.** The samples were prepared in accordance with the Illumina RNA sample preparation protocol (Part # 1004898 Rev. A September 2008). Briefly, mRNAs were fragmented at elevated temperature using divalent cations and transcribed into cDNA thereby generating a library of cDNA fragments. RNA-Seq adapters are then ligated to the blunt ends of the cDNA fragments. The library of cDNA fragments subsequently underwent a size-selection step in which cDNAs were first electrophoresed through a 2.5% agarose gel in TAE buffer. Then, the desired fragment size products (200 bp or 300 bp) were retrieved from the gel and subjected to PCR amplification using universal primer sites present at the end of the ligated adapters. The library was then subjected to quality control steps such as verification of fragment size and concentration measurements using the DNA 1000 Kit (Agilent Technologies) on an Agilent 2100 Bioanalyzer. All samples were sequenced using an Illumina Genome Analyzer II (GAII). Since the experiments were performed over several months as Illumina introduced advances to the GAII platform, the total number of reads and the read length vary (see Table 1). However, all samples were prepared following the same protocol.

All the RNA-Seq and ChIP-Seq data were generated as part of the ENCODE Consortium (The ENCODE Project Consortium 2007) and are available from the UCSC Genome Browser (Rhead *et al.* 2010). The CTCF dataset was published in McDaniell *et al.* 2010, the NfκB dataset was published in Kasowski *et al.* 2010 (GEO accession GSE19485), the Pol III and a subset of the Pol II reads were published in Raha *et al.* 2010 (GEO accession numbers GSE19549 and GSE19550). We sequenced the Pol II ChIP-Seq samples significantly deeper in order to perform our allele-specific analysis as well as the additional GM12878 ChIP-Seq and RNA-Seq datasets (in the process of being submitting to GEO). All the sequence data are also available from alleleseq.gersteinlab.org. A summary of the number total and mapped reads for these datasets is available in Table 1.

Construction of a diploid reference genome for NA12878

Construction of a personal diploid human genome can be performed, provided genomic sequence variants (SNPs, indels and SVs) are known with base pair resolution with respect to the reference genome. Information about personal genomic variants can be obtained from public databases (e.g. dbSNP or Database of Genomic Variants) or downloaded from projects aimed at the discovery and cataloging variant, e.g. HapMap or the 1000 Genomes Project. Construction of a diploid genome requires assigning each variant to one of the two (maternal/paternal) haplotypes or to both personal haplotypes, i.e. variant phasing. Variant phasing can be accomplished in few ways: i) by utilizing long reads spanning two or more variants; ii) by imputing from genotyped variants in the population; iii) by comparing variant genotypes in family trios (father, mother, and child). The latter one, while in principle simple, is also very accurate, for example, 89% of SNPs in NA12878 from the 1000 Genome Project (The 1000 Genomes Consortium 2010) could be unambiguously phased.

To construct the personal genome for NA12878 we used fosmid sequenced deletions (Kidd *et al.* 2008) and the genomic variants (SNPs, indels, and deletions) from the 1000 Genomes Project (The 1000 Genomes Consortium 2010), see Table 2. We have genotyped the fosmid sequenced deletions using a read-depth approach (Abyzov *et al.* 2010) (all other variants were genotyped). Subsequently, we have phased all variants (except SNPs that were already phased) using family trio genotypes.

The phased variants were incorporated into the reference genome using the vcf2diploid tool to yield the diploid genome for NA12878 (see Figure 1a). Random haplotypes was chosen for heterozygous variants that could not be phased. Due to the higher chance of SNP and indel miscalling and misgenotyping in SV regions we incorporated the SVs before the indels and the SNPs. For the same reason, we incorporated the indels before the SNPs. However, the vcf2diplod tool allows variants to be incorporated in any order if desired. During the construction, if a variant overlaps an already incorporated variant on the same haplotype (e.g., SNP within breakpoints of a deletion), then such a variant is not used (see last column in Table 2). The fraction of such variants was very small for NA12878.

Filtering SNPs in CNVs

We started from the human reference genome sequence, version hg18 (NCBI36). The mitochondrial chromosome, chromosome Y, alternative haplotypes and random genomic supercontigs were excluded from consideration. We considered SNPs for the remaining 23 chromosomes (chromosomes 1-22 and X) only. We additionally filtered out SNPs in genomic regions with abnormal read-depth; where the normalized mapped read-depth in symmetrical 2 kb window around each SNPs is less than 1 or more than 3 (the normalization factor of 2 indicates the diploid nature of the human genome). We filtered out SNPs that are more likely to be false positives or may represent duplicated or deleted regions which would complicate calling allele-specific behavior.

The allele-specific SNP processing pipeline

The pipeline has four main inputs: one or more collections of unmapped reads, a set of SNP positions, a personal genome, a set of known genes, listing transcription starts and stops, and exon coordinates.

The processing follows these steps. For each logical set of reads: 1) The reads are trimmed, if necessary, to remove ends that contain large numbers of errors and filtered to remove any reads containing N's. 2) SNP locations are converted to a standardized format that describes the alleles for all heterozygous SNPs in GM12878, including parental phasing, if possible. Phasing is possible for all heterozygous GM12878 SNPs except those in which both parents are also heterozygous. 3) The filtered reads are mapped, using bowtie, to the maternal and paternal genomes.  Bowtie was invoked with these flags:  --best --strata -v 2 -m 1, which returns only unique hits within a minimum number of mismatches, up to two. 4) The two sets of mapped reads are merged into a single set, with each read represented at most once, using the better mapping from the maternal or paternal haplotypes.  If the two mappings for the same read tie in quality, one is chosen at random. 5) Using Het-SNP file and the mapped reads, allele counts are generated for each Het-SNP location.  The resulting counts file contains the number of As,Cs,Gs, and Ts found in reads mapped over each SNP location.  Various other values are also generated for each Het-SNP location, including reference allele, maternal/paternal allele (if determinable), major and minor allele, and a binomial p-value assuming a 50/50 probability of sampling each of two alleles. 6) In order to calculate the false-discovery rate we perform an explicit computational simulation to correct for multiple hypothesis testing. We start with all the heterozygous SNP locations; for each SNP location we randomly assign each mapped read in the dataset to either allele. At a given p-value threshold (using the binomial test) we can determine the number of false positive allele-specific event calls (from the simulated data) and thus we can determine the false discovery rate as the number of false positive over the total number of observed positives. We require a false discovery rate (FDR) of less than 10% (which corresponds to a p-value of threshold between 0.004 for cMyc and 0.03 for Pol III). We intentionally apply a relaxed threshold in order to obtain a decent number of allele-specific events so as to perform genome-wide correlation analyses between ASB and ASE behavior. While we could apply a stricter FDR threshold, we found that the statistical significance of the Pearson correlations is dependant on the both the accuracy (greater accuracy using a stricter FDR threshold) as well as the statistical power determined by the number of observations made (more observations using a more relaxed FDR threshold). Thus, a strict threshold would increase the accuracy at the cost of fewer observed allele-specific events. There is a balance between the accuracy of the observations made as well as the number of observations made, in order to determine optimal correlation behavior. We found that the significance of the Pearson correlation between the observed ASB and ASE events was most significant when the FDR threshold set to approximately 10%.

In Table 4 we present the results for ASE and ASB calls for all the datasets. The first column of is the number of elements (genes, TARs or binding sites) that were sequenced sufficiently deeply in order for allele-specific activity to have been detected in that specific data set given the p-value used in order to obtain a 10% FDR threshold. For the RNA-Seq results 6x was sufficiently to obtain the maximum allowing p-value threshold. For the ChIP-Seq data sets the depth threshold required was 7x for Pol III, 8x for CTCF, JunD, cFos, NfkB and 9x for cMyc and Pol II.

The results of applying this threshold are outputed in the filtered counts file for each dataset.

Using the list of genes and all filtered counts files, information about all asymmetric Het-SNPs from any of the datasets are grouped together by gene.  The locations are annotated as being exonic or intronic.  The information about each SNP includes:  reference allele; maternal and paternal genotype; phasing if possible; A,C,G,T counts; biased-towards parent allele; q-value (false discovery rate) .

Comparison of ASB SNPs with know transcription factor motifs in binding sites

The motif consensus sequences were generated from the PWMs (position weight matrix) (source TRANSFAC (Matys *et al.* 2006) and JASPAR (Portales-Casamar *et al.* 2010)). The frequencies of the matrices were normalized if the original ones were not normalized.

The rules for creating IUPAC consensus sequences for TF motifs are as follows. A single nucleotide code is used if its frequency is greater than 50% and at least twice as high as the second most frequent nucleotide. A double-degenerate code is used if the combined frequencies of two nucleotides are more than 75% but each of them is present in less than 50%. A triple-degenerate code is used where one of the nucleotides does not show up at all in the sequence set and none of the aforementioned rules applied. The letter “N” represents all other frequency distributions. We scanned binding sites using TF PWMs. Genomic sequences defined by the binding sites are fetched (for both strands). The TF PWMs (and corresponding consensus motif) were used as queries to search the genome sequence, with 0 or 1 edit distance. Only those sites that include allele-specific heterozygous SNP locations that are phased are retained.

We compute the difference in binding strength of the motifs between the maternal and paternal alleles to compare against the fraction of maternally derived read counts. For the maternal and paternal alleles at position i:

delta(maternal - paternal) = log2[ P(maternal, i) / P(maternal) \* P(paternal) / P(paternal, i) ];

where

P(n, i): frequency for allele n at position i in the PWM (required to be greater than 0.01)

P(n): background frequency of allele n.

In Figure 3 we plot this difference in motif scores, delta (meternal – paternal) against the fraction of maternally derived reads overlapping the same heterozygous SNP in the ASB binding site. In small number of cases where there are multiple SNPs in the TF motif region, the best one is chosen where if the maternal read count fraction greater than half the best is equal to the biggest delta, while if fraction is less than half the smallest delta is chosen.

Building an allele dependent regulatory network

We decided to integrate the expression data for genes and TARs from the RNA-Seq experiment with the TF binding data from the (cFos, cMyc, JunD, Max, NfkB, CTCF, Pol II and III) ChIP-Seq experiments into a regulatory network. In order to construct a regulatory network determine the edge between a TF and a gene by assigning an ASB event to a target ASE gene if it lies within 2.5 Kb upstream of the annotated transcription start site (TSS) and the transcription termination site (TTS). For ASE novel TARs we do not know which strand is being expressed, thus we associate ASB binding events that occur within 2.5Kb of either end of the novel TAR. If it is allele-specific it could be further classified into: paternal, maternal, unphased. The “unphased” category represents the case where the experiments show allele specificity but it cannot be phased. After constructing the edges between the TFs and gene/novel TARs in the network we overlaid the gene/novel TAR allele-specific expression information onto the nodes. Each gene/novel TAR was categorized into three categories: paternal, maternal, or unphased allele-specific expression. After constructing the network we performed a network motif analysis on it, the results of which are shown in Table 6. We analyzed the occurrences of mutiple-input motifs (MIMs) where two TFs regulate the same gene/novel TAR and single-input motifs (SIMs) where a single TF regulates two different gene/novel TARs, taking into account the allele specificity of the regulation and the expression of the targets. Counting of occurrences of MIMs and SIMs were performed using Cytoscape (Cline *et al.* 2007).

**Figure Legends**

**Figure 1a.** Construction of personal genome by vcf2diploid tool is made by incorporating personal variants into the reference genome. Personal variants may require addition pre-processing, i.e., 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 though the array. Simultaneously, equivalence map (MAP-file format – see Supplementary Figure 1) between personal haplotypes and reference genome is being constructed. **1b.** 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 splice junction library.

**Figure 2.** For each heterozygous SNP location covered at a depth greater than six we can compute the fraction of reads derived from the alternative allele relative to 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 datasets 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.

**Figure 3.** We plot the difference of motifs scores (see **Methods**) between the maternal and paternal alleles against the fraction of maternally derived reads for ASB SNPs overlapping motifs within binding sites. Here we plot this for ASB SNPs in cMyc motifs that are located within Max binding sites. We see a strong correlation indicating that the motif with the stronger match tends to be on the allele that is preferentially bound.

**Figure 4.** Examples showing ASE and ASB for a gene (SKA3 on chromosome 13) and a novel TAR (on chromosome 4). Paternal SNPs exhibiting either ASE or ASB are indicated in blue and corresponding maternal SNPs are indicated in red. We also indicate the region of enriched Pol II binding in black. For these two examples we see coordinated maternal binding and expression for the known gene and coordinated paternal binding and expression for the novel TAR.

**Figure 5.** We plot the fraction of maternally derived reads for ASB SNPs proximal to genes (within 2.5Kb of the TSS to the TTS of the gene) against the maternally derived fraction for ASE SNPs for the same gene. For genes that contain multiple ASB or ASE SNPs we select the SNP with the greatest significance. We see a strong correlation between ASB and ASE for Pol II **(A)** as well as for all the other the transcription factors combined **(B)** (including Pol III but excluding Pol II). Genes that are in the upper right and bottom left quadrants of the plots show agreement between ASB and ASE,

**Figure 6.** We compare the degree of coordination in the maternal or paternal preference of ASB and ASE SNPs within a gene, to that of a random null distribution. Included are all genes that contain 10 or more such SNPs across all our GM12878 datasets. Using this set of genes and number of SNP per gene a null distribution is generated. The null hypothesis is that each SNP within a gene has an independent 50/50 chance of being maternal or paternally biased. The histograms show the distribution of maternal fraction across all genes, compared to that for the null distribution. The observed data shows a strong tendency toward either zero or one, indicating that, within a gene, the SNPs have a strong tendency to be either mostly maternal or paternal. The lower graph displays the results of a Kolmolgorov-Smirnov test to support the claim that the two distributions are significantly different, with a p-value of 6.432e-5 (maximal difference is indicate with a green line).

**Figure 7.** 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, blue for paternal and grey for unknown. As it can be observed there is significant agreement between allele-specific regulation and allele-specific binding.

**Supplementary Figure 1.** MAP file establishes equivalence of bases between reference haplotypes of personal genome and reference genome. Paternal haplotypes denoted as R, maternal as M, and reference as R. Nucleotides in each sequence are denoted as X. Ungapped block that is present in all three haplotypes is recorded by indices of the first bases in a block for each haplotype. Each position absent in either haplotype is recorded by base indices for haplotype(s) having a nucleotide in the position and by -1 for haplotype(s) with gap. Base numbering starts from 1.

**Supplementary Figure 2.** Procedure to compare read mapping to reference genome and haplotypes of NA12878. Pol II Chip-Seq reads were mapped to each haplotype using Bowtie. Then, for any pair of compared haplotypes read mappings were classified as i) equivalent, if a read maps at the equivalent positions; ii) different, if a read maps to non-equivalent positions; iii) not mapped, if a read does not map to either of the compared haplotypes.

**Supplementary Figure 3.** For each heterozygous SNP location covered at a depth greater than six we can compute the fraction of reads derived from the maternal allele relative to the paternal sequence. We then have plotted the distribution of maternal allele fraction for all heterozygous SNPs (significant allele-specific positions are indicated in blue) for the RNA-Seq, Pol II and remaining ChIP-Seq datasets combined. We observe that the distribution of all heterozygous SNPs as well as the allele-specific SNP positions are both quite symmetric and thus we do not see a significant bias towards either the maternal or paternal haplotype.

**Supplementary Figure 4.** 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. Sites with large skew (> 0.6) can be the result of read mapping bias suggested earlier (Degner *et al.* 2009), when reads on one haplotype couldn't be mapped due to non-unique mapping. However, the skew is still observed for a number of sites when the reads that map to only one haplotype are excluded from consideration (green line) to prevent mapping bias. To be conservative we considered an ASB site not affected by mapping bias if mapping skew does not change by more than 0.1 if all reads or only reads mapping to each haplotype are considered. This estimates that less than 15% of ASB sites can be affected by read mapping bias. Note, however, that alternative explanation is that those sites have other variants within the read length, and due to that read coming from one haplotype do not map to another.

**Supplementary Figure 5.** 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). 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). **References**

Abyzov, A., Urban, A.E., Snyder, M., and Gerstein, M. (2010). CNVnator: An Approach to Characterize and Genotype Atypical CNVs Using High-throughput Sequencing Coupled with Population and Family Structure. Genome Res (submitted).

Adey A, Morrison HG, Asan, Xun X, Kitzman JO, Turner EH, Stackhouse B, Mackenzie AP, Caruccio NC, Zhang X, Shendure J. (2010). Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biol. 2010 Dec 8;11(12):R119.

Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S, Gerstein M, Snyder M. (2004). Global identification of human transcribed sequences with genome tiling arrays. Science. 2004 Dec 24;306(5705):2242-6. Epub 2004 Nov 11.

Melissa S Cline, Michael Smoot, Ethan Cerami, Allan Kuchinsky, Nerius Landys, Chris Workman, Rowan Christmas, Iliana Avila-Campilo, Michael Creech, et al. (2007). Integration of biological networks and gene expression data using Cytoscape. Nature Protocols 2, 2366 - 2382 Published online: 27 September 2007 | doi:10.1038/nprot.2007.324.

Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, Pritchard JK. (2009). Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinformatics. 2009 Dec 15;25(24):3207-12. Epub 2009 Oct 6.

Fan HC, Wang J, Potanina A, Quake SR. (2011). Whole-genome molecular haplotyping of single cells. Nat Methods. 2011 Mar;8(3):242-5.

Gimelbrant A, Hutchinson JN, Thompson BR, Chess A. (2007) Widespread monoallelic expression on human autosomes. Science. Nov 16;318(5853):1136-40.

Goto T. and Monk M. (1998). Regulation of X-Chromosome Inactivation in Development in Mice and Humans. Microbiol Mol Biol Rev, June 1998, p. 362-378, Vol. 62, No. 2.

Gregg C, Zhang J, Weissbourd B, Luo S, Schroth GP, Haig D, Dulac C. (2010). High-resolution analysis of parent-of-origin allelic expression in the mouse brain. Science. 2010 Aug 6;329(5992):643-8. Epub 2010 Jul 8.

Gregg C, Zhang J, Butler JE, Haig D, Dulac C. (2010). Sex-specific parent-of-origin allelic expression in the mouse brain. Science. 2010 Aug 6;329(5992):682-5. Epub 2010 Jul 8.

Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JG, Storey R, Swarbreck D, Rossier C, Ucla C, Hubbard T, Antonarakis SE, Guigo R. (2006). GENCODE: producing a reference annotation for ENCODE. Genome Biol. 2006;7 Suppl 1:S4.1-9. Epub 2006 Aug 7.

Johnson DS, Mortazavi A, Myers RM, Wold B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science. 2007 Jun 8;316(5830):1497-502. Epub 2007 May 31.

Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP, Gingeras TR. (2002). Large-scale transcriptional activity in chromosomes 21 and 22. Science. 2002 May 3;296(5569):916-9.

Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, Habegger L, Rozowsky J, Shi M, Urban AE, Hong MY, Karczewski KJ, Huber W, Weissman SM, Gerstein MB, Korbel JO, Snyder M. (2010). Variation in transcription factor binding among humans. Science. 2010 Apr 9;328(5975):232-5. Epub 2010 Mar 18.

Jeffrey M. Kidd, Gregory M. Cooper, William F. Donahue, Hillary S. Hayden, Nick Sampas, Tina Graves, Nancy Hansen, Brian Teague et al. (2008). Mapping and sequencing of structural variation from eight human genomes. Nature 453, 56-64 (1 May 2008).

Lalonde E, Ha KC, Wang Z, Bemmo A, Kleinman CL, Kwan T, Pastinen T, Majewski J. (2011). [RNA sequencing reveals the role of splicing polymorphisms in regulating human gene expression.](http://www.ncbi.nlm.nih.gov/pubmed/21173033) Genome Res. 2011 Feb 1.

Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D108-10.

Ng P.C., Henikoff S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003 Jul 1;31(13):3812-4.

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

Lyon Mary. (1961). Gene Action in the X-chromosome of the Mouse (Mus musculus L.). Nature 190, 372 - 373 (22 April 1961).

Maynard ND, Chen J, Stuart RK, Fan JB, Ren B. Genome-wide mapping of allele-specific protein-DNA interactions in human cells. Nat Methods. 2008 Apr;5(4):307-9. Epub 2008 Mar 16.

McDaniell R, Lee BK, Song L, Liu Z, Boyle AP, Erdos MR, Scott LJ, Morken MA, Kucera KS, Battenhouse A, Keefe D, Collins FS, Willard HF, Lieb JD, Furey TS, Crawford GE, Iyer VR, Birney E. (2010). Heritable individual-specific and allele-specific chromatin signatures in humans. Science. Apr 9;328(5975):235-9. Epub 2010 Mar 18.

Mills, R.E. (2010). Mapping structural variation at fine-scale by population genome sequencing. Nature (in revision).

Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. (2002). Network motifs: simple building blocks of complex networks. Science. 2002 Oct 25;298(5594):824-7.

Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. Nature. 2010 Apr 1;464(7289):773-7. Epub 2010 Mar 10.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. (2008). Nat Methods. 2008 Jul;5(7):621-8. Epub 2008 May 30.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008 Jun 6;320(5881):1344-9. Epub 2008 May 1.

Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 464:768-72.

Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, Yusuf D, Lenhard B, Wasserman WW, Sandelin A. (2010). JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Res. 2010 Jan;38(Database issue):D105-10. Epub 2009 Nov 11.

Raha D, Wang Z, Moqtaderi Z, Wu L, Zhong G, Gerstein M, Struhl K, Snyder M. Close association of RNA polymerase II and many transcription factors with Pol III genes. Proc Natl Acad Sci U S A. 2010 Feb 23;107(8):3639-44. Epub 2010 Feb 5.

Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita P, Diekhans M, Smith KE, Rosenbloom KR, Raney BJ, Pohl A, Pheasant M, Meyer L, Hsu F, Hillman-Jackson J, Harte RA, Giardine B, Dreszer T, Clawson H, Barber GP, Haussler D, Kent WJ. (2010). The UCSC Genome Browser database: update 2010. Nucleic Acids Res.

Rinn JL, Euskirchen G, Bertone P, Martone R, Luscombe NM, Hartman S, Harrison PM, Nelson FK, Miller P, Gerstein M, Weissman S, Snyder M. (2003). The transcriptional activity of human Chromosome 22. Genes Dev. 2003 Feb 15;17(4):529-40.

Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, Shendure J, Drmanac R, Jorde LB, Hood L, Galas DJ. (2010). Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science. 2010 Apr 30;328(5978):636-9. Epub 2010 Mar 10.

Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, Thiessen N, Griffith OL, He A, Marra M, Snyder M, Jones S. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods. 2007 Aug;4(8):651-7. Epub 2007 Jun 11.

Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB. (2009). PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nature Biotechnology 27, 66 - 75.

Shah Sohrab P., Morin Ryan D., Khattra Jaswinder, Prentice Leah, Pugh Trevor, Burleigh Angela, Delaney Allen, Gelmon Karen, Giuliany Ryan, Senz Janine, et al. (2009). Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. Nature. vol461, 809-813. (2009).

The 1000 Genomes Consortium (2010). A map of human genome variation from population scale sequencing. Nature (in revision).

The ENCODE Project Consortium (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007 Jun 14;447(7146):799-816.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Data** | **Number of reads (millions)** | **Number of mapped reads (millions)** | **Read length; sequencing layout** | **Source** |
| RNA-Seq | 393.9 | 164.7 | 36 nt; single end  50 nt; single end  50 nt; paired end | this paper |
| Pol II ChIP-Seq | 128  (33) | 69.5  (13.2) | 36 nt; single end | this paper  + Raha *et al.* 2010  (shown in parentheses) |
| Pol III ChIP-Seq | 12 | 7.5 | 36 nt; single end | Raha *et al.* 2010 |
| cMyc ChIP-Seq | 125 | 65.5 | 36 nt; single end | this paper |
| Max ChIP-Seq | 79 | 46.1 | 36 nt; single end | this paper |
| JunD ChIP-Seq | 133 | 72.5 | 36 nt; single end | this paper |
| cFos ChIP-Seq | 84 | 30.4 | 36 nt; single end | this paper |
| NFκB ChIP-Seq | 62 | 35.5 | 36 nt; single end | Kasowski *et al.* 2010 |
| CTCF ChIP-Seq | 46 | 26.4 | 36 nt; single end | McDaniell *et al.* 2010 |

**Table 1. GM12878 RNA-Seq and ChIP-Seq Data Sets.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Source** | **Variant type** . **Counts** | **Total** | **Phased** | **Unphased** | **Inconsistent** | **Unutilized** |
| Fosmid sequencing | Deletions | 33 | 94% | 6% | 0% | 0 (0%) |
| 1000 Genomes Project | Deletions | 1,522 | 77% | 8% | 15% | 15 (1%) |
| Indels | 328,528 | 89% | 11% | 0% | 37 (0.1%) |
| SNPs | 2,766,607 | 89% | 11% | 0% | 1,794 (~0%) |

Table 2. Statistics on variants used to construct personal genome of NA12878. A variant can be phased (i.e., unambiguously assigned to a paternal or maternal haplotype), unphased (i.e., ambiguously assigned to either haplotype) or its genotyping can be inconsistent with genotyping in parents (e.g., heterozygous deletion in child but homozygous deletion in each parent), which precludes it from phasing. Due to overlap with other variants some variants are not used for genome construction of NA12878 (column “Unutilized”).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Haplotype** | **# of mapped reads** | **Equivalently mapped reads in** | | |
| **Reference** | **Paternal** | **Maternal** |
| 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** | **Paternal** | **Maternal** |
| 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** | **Paternal** | **Maternal** |
| 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%) |  |

**Table 3.** Comparison of read mappings to reference genome and paternal and maternal haplotypes of GM12878. Chip-Seq reads for Pol II were independently mapped to each haplotype (chromosomes 1-22 and X) and the best unambiguous mapping (no more than two mismatches) was selected for each read. More reads are mapped to either haplotypes of GM12878 than to the reference genome. The major difference in numbers for paternal/maternal and reference haplotypes is due to reads that map to one haplotype but not to other. Namely, only about 0.1-0.2% of reads that map to the reference cannot be mapped to paternal/maternal haplotype, while a significantly higher fraction ~0.5% of reads map to paternal/maternal genome and cannot be mapped to the reference. Interestingly, for paternal and maternal haplotypes unmapped reads and reads with different mappings contribute roughly equally to the discrepancy in overall mapping. See Supplementary Table 1 for the results for CTCF.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Genomic element** | | **Total number** | **Numberwith ASE or ASB** | **Allele-specific fraction** | **Maternal** | **Paternal** |
| Autosomes | Genes | | 4,829 | 935 | 0.19 | 491 | 424 |
| Splice Junctions | | 2,556 | 552 | 0.21 | 272 | 202 |
| Novel TARs | | 9,238 | 860 | 0.09 | 386 | 363 |
| Pol II | Binding Sites | 3,187 | 344 | 0.11 | 172 | 126 |
| Pol III | 46 | 2 | 0.04 | 0 | 2 |
| CTCF | 4,573 | 443 | 0.10 | 178 | 207 |
| NfκB | 1,300 | 56 | 0.04 | 22 | 27 |
| cFos | 378 | 36 | 0.10 | 12 | 12 |
| Max | 943 | 55 | 0.06 | 24 | 22 |
| cMyc | 1,542 | 36 | 0.02 | 15 | 15 |
| JunD | 313 | 25 | 0.08 | 15 | 6 |
|  | | | | | | | |
| Chromosome X | Genes | | 94 | 75 | 0.80 | 70 | 4 |
| Novel TARs | | 149 | 75 | 0.50 | 70 | 1 |
| Pol II Sites | | 110 | 48 | 0.44 | 47 | 1 |
| TFs Sites Combined | | 259 | 40 | 0.15 | 28 | 10 |

**Table 4**. List of ASE and ASB events for each dataset a) only autosomes b) only chr X. The first column indicates the number of elements (genes, novel TARs, splice junctions or binding sites) that are sequenced at a sufficient depth in order to detect allele-specific activity and containing a heterozygous SNP, see Methods for further details). The number of elements containing either ASE or ASB that can be phased are then split into maternal and paternal specific counts. We used the GENCODE 3c set of gene annotation and binding sites were determined using PeakSeq (Rozowsky *et al.* 2009) with default parameters.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Number of genes near binding sites (with het SNPs)** | **ASE genes near ASB sites** | **Number of novel TARs near binding sites**  **(with het SNPs)** | **ASE novel TARs near ASB sites** |
| Pol II | 3190 | 74 | 4845 | 55 |
| TFs | 7716 | 45 | 6758 | 15 |

**Table 5**. This table shows the association of transcriptional factor binding and expression of genes and novel TARs, broken down by autosomes and chrX. The association is defined by binding of TFs 2.5Kb upstream or downstream of a GENCODE gene or within 2.5Kb of a novel TAR.  Hets: with heterozygous SNPs. ASE: allele-specific expression. ASB: allele-specific binding.

|  |  |  |
| --- | --- | --- |
| **Single TF** | **Maternal Expression** | **Paternal Expression** |
| Maternal Regulation | 81 | 22 |
| Paternal Regulation | 31 | 64 |
|  |  |  |
| **Multiple TFs (MIM)** | **Maternal Expression** | **Paternal Expression** |
| Both Maternal Regulation | 40 | 0 |
| Both Paternal Regulation | 4 | 36 |
| Mixed Regulation | 3 | 2 |
|  |  |  |
| **Single TF (SIM)** | **Both Maternal Expression** | **Both Paternal Expression** |
| Both Maternal Regulation | 2,840 | 224 |
| Both Paternal Regulation | 254 | 1,232 |

**Table 6.** Here we tabulate first the number of transcription factors (or polymerases) that maternally or paternally regulate GENCODE genes or novel TARs that are maternally or paternally expressed. We see the maternal regulation is coordinated with maternal expression and similarly for paternal regulation with paternal expression. We also tabulate the breakdown of counts for two network motifs, multiple input motifs (MIMs) and single input motifs (SIMs), also see figure 7. A MIM is were two TFs regulate the same target gene or novel TAR and a SIM is where one TF regulates two different targets. We again observe coordinated regulation in these network motifs. For SIMs we also observe 1,910 cases of the form MP -> MP (opposite by coordinated regulation and expression) and 222 cases of MP -> PM (mixed regulation and expression).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Haplotype** | **# of mapped reads** | **Equivalently mapped reads in** | | |
| **Reference** | **Paternal** | **Maternal** |
| Reference | 26,322,823 |  | 26,287,466 (99.**87**%) | 26,311,193 (99.**96**%) |
| Paternal | (+0.24%) 26,386,899 | 26,287,466 (99.62%) |  | 26,334,565 (99.80%) |
| Maternal | (+0.33%) 26,411,779 | 26,311,193 (99.62%) | 26,334,565 (99.71%) |  |
|  | | **Differently mapped reads in** | | |
| **Reference** | **Paternal** | **Maternal** |
| Reference | 26,322,823 |  | 6,579 (0.02%) | 7,013 (0.03%) |
| Paternal | (+0.24%) 26,386,899 | 6,579 (0.02%) |  | 31,134 (0.**12**%) |
| Maternal | (+0.33%) 26,411,779 | 7,013 (0.03%) | 31,134 (0.**12**%) |  |
|  | | **Unmapped reads in** | | |
| **Reference** | **Paternal** | **Maternal** |
| Reference | 26,322,823 |  | 28,778 (0.11%) | 4,617 (0.02%) |
| Paternal | (+0.24%) 26,386,899 | 92,854 (0.**35**%) |  | 21,200 (0.08%) |
| Maternal | (+0.33%) 26,411,779 | 93,573 (0.**35**%) | 46,080 (0.17%) |  |

**Supplementary Table 1.** Comparison of read mappings to reference genome and paternal and maternal haplotypes of GM12878 (similar to Table 3). Chip-Seq reads for CTCF were independently mapped to each haplotype (chromosomes 1-22 and X) and the best unambiguous mapping (no more than two mismatches) was selected for each read.

|  |  |  |  |
| --- | --- | --- | --- |
| **ASE and ASB from Heterozygous Indels** | Total Count | Maternal | Paternal |
| Exons showing ASE | 128 | 75 | 53 |
| Novel TARs showing ASE | 233 | 126 | 107 |
| Pol II binding sites showing ASB | 123 | 53 | 70 |
| CTCF binding sites showing ASB | 52 | 22 | 30 |

**Supplementary Table 2.** In this table we present the number of additional sites exhibiting allele-specific behavior for RNA-Seq known exons and novel TARs determined by an exon or novel TAR overlapping a heterozygous indel. We also show the number of additional ASB binding sites for Pol II and CTCF for detemined for binding sites overlapping a heterozygous indel.