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

To study allele-specific expression (ASE) and binding (ASB), that is, differences between the maternally and paternally derived alleles, we have developed a computational pipeline (AlleleSeq). Our pipeline initially constructs a diploid personal genome sequence (and corresponding personalized gene annotation) 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 and alignment of reads to a personal diploid genome sequence that we address, for example, bias of reads mapping to the reference allele. 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 data sets generated for this purpose. In addition to observing fairly widespread allele-specific behavior within individual functional genomic data sets (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. Correlation analyses and network motifs show mostly coordinated ASB and ASE.
Molecular Systems Biology **7**: 522; published online 2 August 2011; doi:10.1038/msb.2011.54
Subject Categories: Functional genomics; Computational methods
Keywords: allele-specific; ChIP-Seq; networks; RNA-Seq

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; Robertson *et al*, 2007) or RNA-Seq (Mortazavi *et al*, 2008; Nagalakshmi *et al*, 2008) yield similar results that were obtained using tiling array-based methodologies albeit with greater sensitivity and resolution, that is, 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 Project Consortium, 2010). A natural area of recent focus has been looking at the degree of functional genomic differences across the human population (Gregg *et al*, 2010a, b; Kasowski *et al*, 2010; McDaniell *et al*, 2010; Montgomery *et al*, 2010; Pickrell *et al*, 2010). 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 SNP, 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, that is, allele-specific expression (ASE) or allele-specific binding (ASB); however, there are number of technical issues 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 manner using array-based technologies (Maynard *et al*, 2008). Here, we address this question by analyzing a number of different functional genomic data sets 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 data set (~160 million mapped reads) as well as matching deeply sequenced ChIP-Seq data sets (~30 to ~60 million mapped reads) for a number of different transcription factors (cFos, cMyc, JunD, Max, NfκB, and 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 Consortium (2010). Thus, for these data sets 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 ~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 ~100-fold fewer comparisons between ASB and ASE SNPs to be made.

There are numerous technical hurdles in determining allele-specific behavior. 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 (Mills *et al*, 2011; The 1000 Genomes Project Consortium, 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*, 2011). 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, that is, 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 data sets for cFos, cMyc, JunD, Max, and RNA Polymerase II for the GM12878 cell line. We also created a matching deeply sequenced RNA-Seq data set for the same cell line. We combined these data sets with previously published matching data sets for RNA Polymerase II, RNA Polymerase III, NfκB, and CTCF (Kasowski *et al*, 2010; McDaniell *et al*, 2010; Raha *et al*, 2010). We summarize these data sets in Table I (see Materials and methods for further details).

Table I GM12878 RNA-Seq and ChIP-Seq data sets

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)

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) cover 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 (see Supplementary Table 1 for the results of *de novo* SNP calling of heterozygous SNPs). Any significant amount of miscalling of heterozygous SNPs will (obviously) lead to ill determined allele-specific behavior. There are a number of possible explanations for such miscalls; the very events we would like to find, SNPs within regions showing ASE 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. Finally, 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 SVs 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, that is, both alleles are not treated equally by default. Second, expression or binding in regions of genome SV could be misinterpreted as ASE or ASB. 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 is likely to be less precise and contain more false positives compared with non-SV regions (The 1000 Genomes Project Consortium, 2010). Thus, we construct a personal diploid genome of NA12878 (see Materials and methods), by utilizing genomic variations (see Table II for summary statistics) determined in the framework of The 1000 Genomes Project Consortium (2010) and, additionally, SVs 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 gene annotation (including, most importantly, splice-junction library) by mapping the GENCODE 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

Table II Statistics on variants used to construct personal genome of NA12878

Source	Variant type	Counts	Total	Phased (%)	Unphased (%)	Inconsistent (%)	Unutilized
Fosmid sequencing	Deletions		33	94	6	0	0 (0%)
1000 Genomes Project	Deletions		1522	77	8	15	15 (1%)
	Indels		328 528	89	11	0	37 (0.1%)
	SNPs		2 766 607	89	11	0	1794 (~0%)

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').

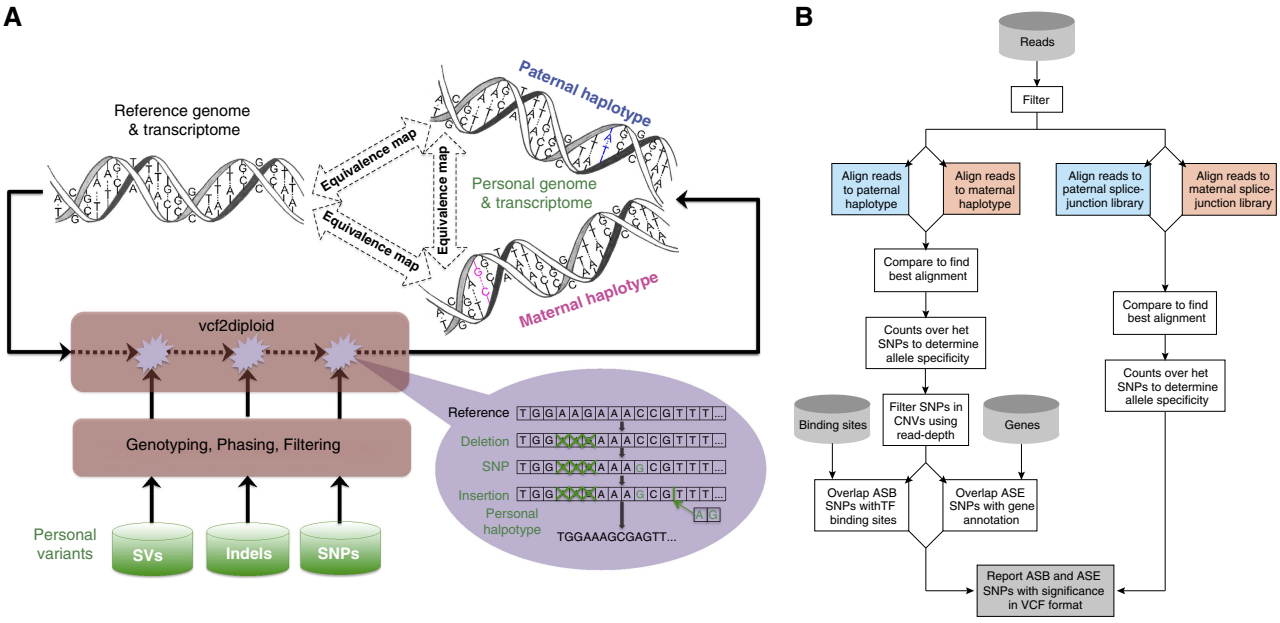

Figure 1 (**A**) Construction of a personal genome by vcf2diploid tool is made by incorporating personal variants into the reference genome. Personal variants may require additional pre-processing, that is, filtering, genotyping, and/or phasing. The output is the two (paternal and maternal) haplotypes 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. (**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).

package to perform personal genome sequence construction (the vcf2diploid tool and associated source code), the actual diploid sequence for NA12878, splice-junction sequences and personalized gene annotation for NA12878 and corresponding equivalence maps (between the maternal and paternal sequences as well as the reference genome, NCBI36/hg18) are available from http://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 are 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 with 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 III, 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 with 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. 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 2). This demonstrates that it is important to use a correctly assembled personal genome for aligning reads when performing an allele specificity analysis.

Similarly, transcription factor binding sites also overlapped more when aligned to the maternal and paternal genomes of NA12878, rather than the reference sequence. For this comparison, we used the set of independently mapped reads for all three genome sequences to determine binding sites using PeakSeq (Rozowsky *et al*, 2009), and performed a pair-wise nucleotide overlap of the binding sites between the three genome sequences (Supplementary Table 3). In addition, we observe that the differences in binding sites, among the three genomes, are greater than the underlying differences in read mapping.

Determining ASE and ASB

The second part of the AlleleSeq pipeline determines ASE using RNA-Seq data and ASB using ChIP-Seq data. After the

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%)	

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.

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*, 2011) are determined to be potentially in a CNV (the read depth of genomic DNA reads in a 1-kb window around the SNP is either <1 or >3) are excluded (see Materials and methods). We correct for multiple hypothesis testing by estimating the false-discovery rate (FDR) 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 Materials and 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 data set. We then analyze the coordination between binding and expression in a pair-wise manner using direct correlation. Finally, we investigate higher order coordination of ASB and ASE 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 IV. In the upper half of the table, we present the results for all the autosomes and in the lower half for chromosome X (with all the transcription factor combined). In the second column of Table IV, we list the number of genomic elements (genes, novel TARs, and binding sites) that are accessible for allele-specific behavior—that is, those that we could have detected allele-specific activity in. This is the set of genomic elements that contain at least one heterozygous SNP and are sequenced at sufficient depth in order to detect allele-specific activity, see Materials and methods for further details. The third column shows that number of genomic elements that we determine to show allele-specific behavior. The fourth column shows the fraction of genomic elements that are accessible for allele-specific behavior that do show either ASE or ASB. Finally for allele-specific events that can be phased we report those that are maternal or paternal.

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 show ASE, 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, >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 accessible 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 data sets have been sequence quite deeply (see Table I), 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 data sets we do not have sufficient read depth within

Table IV List of ASE and ASB events for each data set (a) only autosomes (b) only chr X

Genomic element	Number of elements accessible for allele-behavior	Number with ASE or ASB	Fraction with allele-specific behavior	Maternal	Paternal
Autosomes					
Genes	4829	935	0.19	491	424
Splice junctions	2556	552	0.21	272	202
Novel TARs	9238	860	0.09	386	363
Binding sites					
Pol II	3187	344	0.11	172	126
Pol III	46	2	0.04	0	2
CTCF	4573	443	0.10	178	207
NfκB	1300	56	0.04	22	27
cFos	378	36	0.10	12	12
Max	943	55	0.06	24	22
cMyc	1542	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

The first column indicates the number of elements (genes, novel TARs, splice junctions, or binding sites) that are sequenced at a sufficient depth and containing a heterozygous SNP in order to be accessible to detect allele-specific behavior, see Materials and 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.

a binding site as compared with 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 IV, 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 accessible genes on chromosome X exhibit allele-specific behavior and 95% of these are expressed 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 ASE 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 with 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.

We make available the complete list of SNPs that show ASB or ASE in VCF format as a resource from our website http://alleleseq.gersteinlab.org. We imagine that this file may be a useful resource for researchers interested in focusing on allele-specific sites in less deeply sequenced functional genomic experimental data sets. 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.

There are a number of technical issues associated with determining allele-specific behavior including various biases that can be introduced as part of the analysis. We investigate some of these potential effects in detail below:

1. *Reference bias*: In order to assess whether our pipeline has some residual bias toward 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 toward 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.
2. *Correlation with SNP quality (genotype likelihood scores)*: SNPs determined by The 1000 Genomes Project Consortium (2010) each have a genotype likelihood score. In Supplementary Figure 4, we plot the distribution of all heterozygous SNPs and the subset of ASE SNPs versus this genotype likelihood score. We see a slight enrichment of ASE SNPs will lower scores; however, the majority of SNPs from both distributions have the highest possible score. For comparison, non-synonymous SNPs also show a similar distribution.
3. *Relation to genome duplications* (*effect of* Degner *et al*, 2009): It has been reported by Degner *et al* (2009)

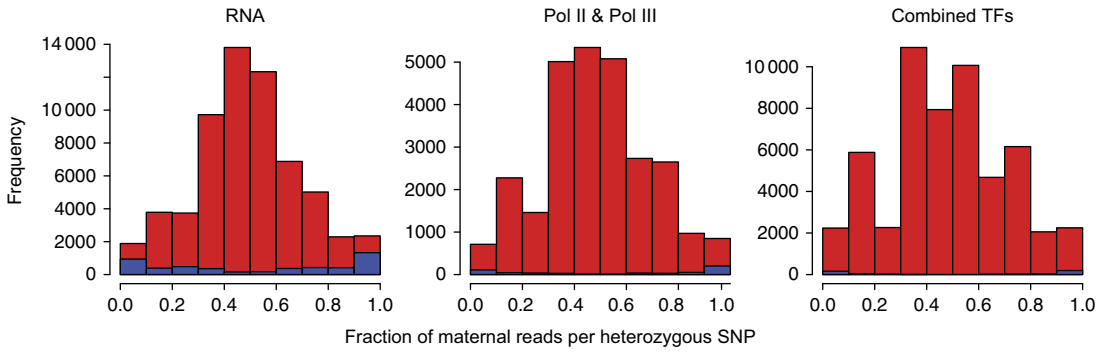

Figure 2 For each heterozygous SNP location covered at a depth greater than six, we can compute the fraction of reads derived from the alternative allele relative to the reference sequence. We then plotted the distribution of alternative allele fraction for all heterozygous SNPs (significant allele-specific positions are indicated in blue) for the RNA-Seq, Pol II, and remaining ChIP-Seq 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.

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 a similar number of reads coming from each allele, only one of the alleles would have uniquely mapping reads which would lead to seemingly allele-specific behavior (see Supplementary Figure 5 for a schematic comparing region without a duplication to regions that have been duplicated). In order to assess the size of this effect on our results, we used the Pol II ChIP-Seq reads mapped uniquely and independently to each of the maternal and paternal genomes. This is as opposed to the normal mapping procedure in the AlleleSeq pipeline, where we map independently to both haplotypes and then select the allele with the better mapping location. At each heterozygous SNP location determined to show ASB we computed the haplotype fraction, the fraction of reads mapped to one allele over the sum of reads mapped to both alleles (we choose the haplotype fraction corresponding to the allele that has the greater fraction, see Supplementary Figure 5). For sites that have not been duplicated the haplotype fraction should have a value close to 0.5, while for duplicated regions exhibiting the Degner effect the fraction would be close to 1 (where all the uniquely mapped reads would come from one allele). In Supplementary Figure 6, we plot the distribution of haplotype fractions for all Pol II ASB sites. We observe that only a minority of the sites ($<15\%$) shows a significant skew in the haplotype fraction toward one haplotype (a fraction >0.6). As valid ASB sites that contain additional proximal sequence variants (such as additional SNPs or indels) would also exhibit a fraction biased toward one haplotype, we conclude that this is an upper bound on the size of this effect and do not consider it significant.

4. *Modified binomial test*: In order to assess the effect of aligning reads against the constructed diploid genome sequence for NA12878 versus using the reference genome sequence we perform the following analysis. For the RNA-Seq reads, we also aligned the reads against the reference genome and determined ASE using an even binomial distribution (threshold applied in a similar manner). As an additional comparison, we also applied the methodology of Montgomery *et al* (2010) where a skewed binomial distribution is used with the reads aligned against the reference genome (they composite for the reference bias induced by mapping to the reference genome by modifying the binomial distribution). Similar to Figure 2, we plotted the distribution of all expressed heterozygous SNPs (ASE SNPs in blue) for these two methods in Supplementary Figure 7. 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.
5. *Comparison with* McDaniell *et al*, (2010): 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 with common heterozygous SNP between the two analysis (McDaniell *et al*, 2010 used an earlier set of SNP calls for NA12878 from The 1000 Genomes Project Consortium, 2010) we find that greater using a *P*-value threshold of 0.01 on their results $>85\%$ of the ASB SNPs are in common.
6. *Allele-specific elements using heterozygous indels*: The AlleleSeq pipeline determines allele-specific behavior for genomic elements (transcribed regions or binding sites) that contain heterozygous SNPs. It is also possible to determine allele-specific behavior for genomic elements that contain a heterozygous indel. In Supplementary Table 4, we show the results for 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.

Correlation of ASB in binding sites containing known motifs

In our analysis of ASB 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 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 with the known motif. Thus, we first scanned ASB sites for known motifs, position weight matrices (PWMs) obtained from TRANSFAC (Matys *et al*, 2006) and JASPAR (Portales-Casamar *et al*, 2010) (see Materials and 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 behavior 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 8 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.

Correlation of ASE with protein structural fitness

Some heterozygous SNPs within genes can result in one allele losing its ability to function as a transcript (i.e., heterozygous loss of function). Additionally, non-synonymous SNPs within the protein-coding sequence can cause a difference in the structure fitness of the protein coded from each allele. We studied the coordination between these effects and the genes that show ASE.

We first investigated the overlap between genes that exhibit ASE with genes that show loss of function on one allele due to a premature stop codon, a frame-shift or a disruption of an intron–exon splice site (all caused by heterozygous SNPs). We find four cases of genes that show ASE as well as heterozygous loss of function and in all four cases the allele that is expressed is opposite to the allele that has lost its ability to code for a protein. We speculate that in some of these cases the transcript from the allele suffering from a disruption might be degraded due to non-sense-mediated decay, which, in turn, might cause the ASE from the other allele.

Since some heterozygous SNPs that show ASE 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 ASE, as many of these SNPs 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 and Henikoff, 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 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.

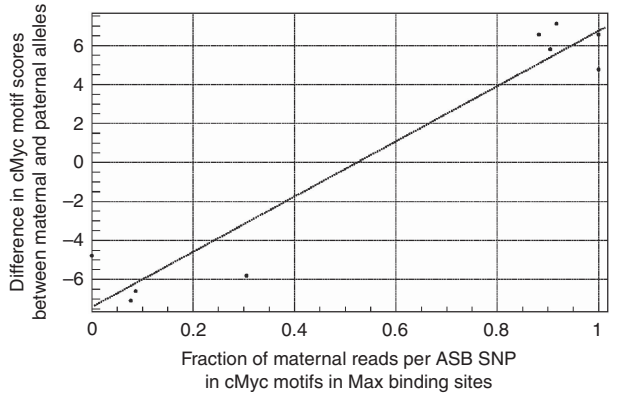

Figure 3 We plot the difference of motif scores (see Materials and methods) between the maternal and paternal alleles against the fraction of maternally derived reads for ASB SNPs overlapping motifs within binding sites. Here, we plot this for ASB SNPs in cMyc motifs that are located within Max binding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

Correlating ASB with ASE

In the upper panel of Figure 4, we present an example of the gene *SKA3* on chromosome 13 which has multiple heterozygous SNPs within exons which show consistent maternal ASE which agrees with the maternal ASB 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 the lower panel of Figure 4, 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 allele-specific behaviors detected using heterozygous SNPs are coordinated on a genome-wide scale. In Table V, we tabulate the number of genes and novel TARs that have evidence for ASE and for proximal ASB. We also tabulate the total counts of genes that have a proximal binding site where both the gene and binding are jointly accessible for detecting allele-specific behavior. We perform a similar calculation for novel TARs and their proximal sites.

In Table V, we present the tabulated counts for Pol II & Pol III and CTCF separately from all the other transcription factors combined. We find a number of genes (74 genes proximal to Pol II & Pol III sites, 29 genes proximal to CTCF binding sites, and 44 genes with proximal transcription factor sites other than CTCF) and novel TARs (55, 8, and 15 for Pol II & Pol III,

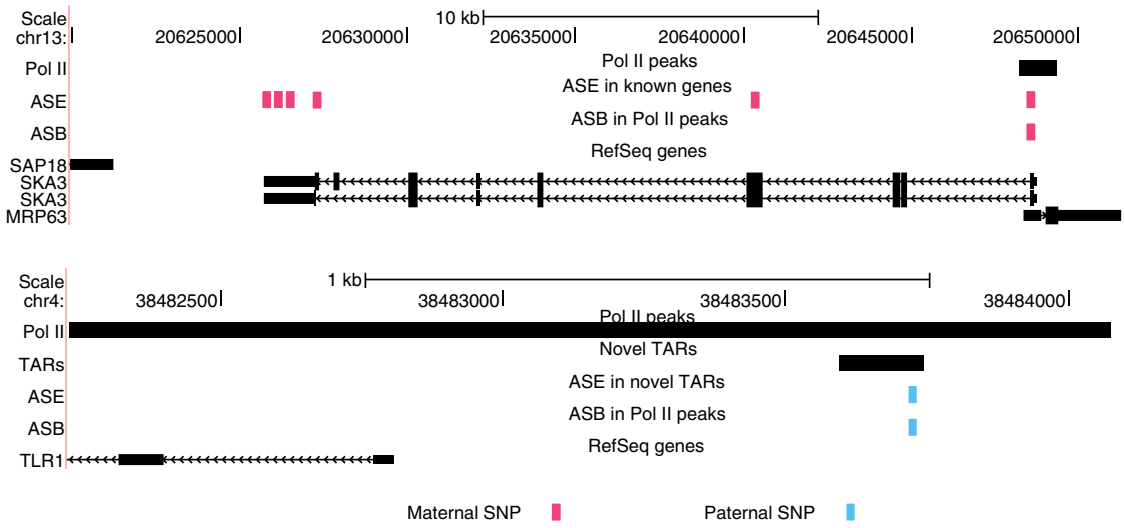

Figure 4 Examples showing ASE and ASB for a gene (*SKA3* on chromosome 13) and a novel TAR (on chromosome 4). Paternal SNPs exhibiting either ASE or ASB are indicated in blue and corresponding maternal SNPs are indicated in red. We also indicate the region of enriched Pol II binding in black. For these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

Table V Association of transcriptional factor binding (for Pol II & Pol II, CTCF, and the other TFs combined) and expression of genes and novel TARs

	Number of genes near binding sites jointly accessible for allele-specific behavior	ASE genes near ASB sites	Number of novel TARs near binding sites jointly accessible for allele-specific behavior	ASE novel TARs near ASB sites
Pol II & Pol III	3190	74	4845	55
CTCF	1739	29	99	8
Other TFs	7716	44	6758	15

The association is defined by binding of TFs 2.5 kb upstream or downstream of a GENCODE gene or within 2.5 kb of a novel TAR. Het SNPs: with heterozygous SNPs.

CTCF, and remaining transcription factors, respectively) in which we see both ASB proximal to ASE. We separate CTCF from the remaining transcription factors because of its function as an insulator. While these numbers might seem relatively small, they reflect the low chance of having both a proximal binding site as well as an expressed gene with both of them jointly accessible for the detection of allele-specific activity. 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 we performed the following analysis. For each gene, we tabulated the allele-specific behavior of the most significant ASE SNP versus the most significant ASB SNP (if there are more than one significant heterozygous allele-specific SNP present). In Table VI , we tabulate maternal and paternal ASB of binding sites against maternal and paternal ASE of proximal genes (we do this separately for Pol II & Pol III, CTCF, and the remaining transcription factors combined). We see that there is a statistically significant coordination between ASB of Pol II & Pol III to genes that exhibit ASE (Fisher's exact test, *P*-value=1.4e−3). Similarly, as seen in Table VI there is also a statistically significant correlation between the 45 genes that exhibit ASE with ASB for all the combined transcription factors excluding CTCF (Fisher's exact test, *P*-value=1.8e−5). We do not however, see a significant correlation of ASB with ASE for CTCF which is probably due to its role as an insulator.

Table VI We tabulate the ASB SNPs proximal to genes (within 2.5 kb of the TSS to the TTS of the gene) containing ASE SNPs

	Maternal binding	Paternal binding
ASB for Pol II and Pol III versus ASE		
Maternal expression	35	5
Paternal expression	7	19
ASB for transcription factors combined (excluding CTCF) versus ASE		
Maternal expression	14	6
Paternal expression	2	19
ASB for CTCF versus ASE		
Maternal expression	8	10
Paternal expression	4	6

For genes that contain multiple ASB or ASE SNPs, we select the SNP with the greatest significance for each. We separately tabulate binding sites that exhibit either maternal or paternal ASB against genes that have maternal or paternal ASE. We do this separately for Pol II and Pol III, CTCF, and the remaining transcription factors combined. Using a Fisher's exact test, we see a significant coordination between Pol II & Pol II ASB versus ASE (*P*-value=1.4e−3) and between the other TFs showing ASB versus ASE (*P*-value=9.3e−5). We do not observe significant coordination between CTCF ASB and proximal ASE, which is expected given the role of CTCF as an insulator.

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 2.5 kb upstream of the transcription start site (TSS) to the transcription termination

site (TTS) including introns). Using only genes that contained >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 5, 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 two of Figure 5. Using a Kolmolgorov–Smirnov test, we find significantly more coordination of ASB and ASE SNPs in genes than compared with the random null expectation (*P*-value=8.45e−8; see the third panel of Figure 5).

Representing ASE and ASB behavior on a regulatory network

In the previous analysis, we showed that binding and expression were correlated in a pair-wise manner. 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 6 represents the regulatory network of six transcription factors (cMyc, Max, cFos, JunD, NfκB, and CTCF) and two polymerases (Pol II and Pol III). The network discretizes the ASB events into allele-specific regulation of target genes and novel TARs and their ASE. 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 gray 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, and gray—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 mater-

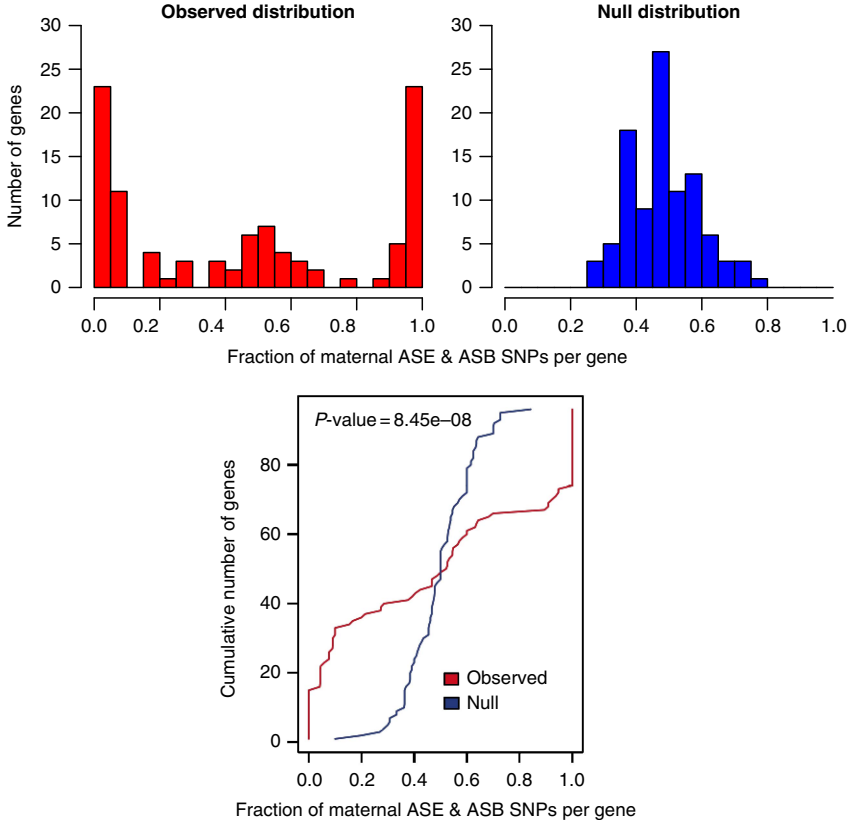

Figure 5 We compare the degree of coordination in the maternal or paternal preference of ASB and ASE SNPs within a gene, to that of a random null distribution. All genes that contain 10 or more such SNPs across all our GM12878 data sets are included. Using this set of genes and number of SNP per gene, 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 with that for the null distribution. The observed data show 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 8.45e−8 (maximal difference is indicate with a green line).

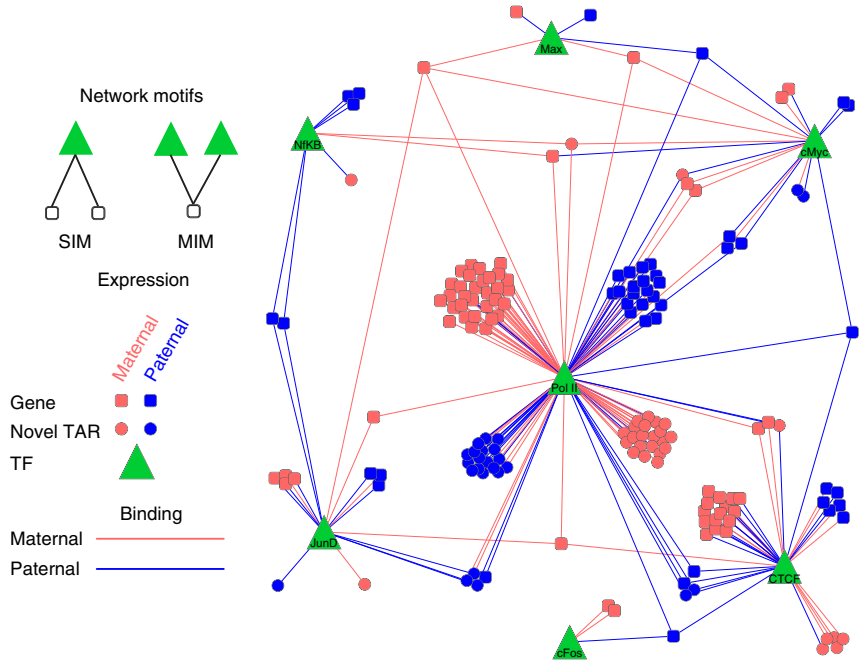

Figure 6 This figure shows a regulatory network of genes and novel TARs that are regulated by TFs in an allele-specific manner. The TFs are represented by green triangles, while the genes and novel TARs are represented by squares and circles, respectively. The color of the genes and tars are representative 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.

Table VII Number of transcription factors (or polymerases) that maternally or paternally regulate GENCODE genes or novel TARs that are maternally or paternally expressed

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	2840	224
Both paternal regulation	254	1232

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 6. An MIM is where two TFs regulate the same target gene or novel TAR and an SIM is where one TF regulates two different targets. We again observe coordinated regulation in these network motifs. For SIMs, we also observe 1910 cases of the form MP→MP (opposite by coordinated regulation and expression) and 222 cases of MP→PM (mixed regulation and expression).

nal/paternal expressed genes or novel TARs (see first part of Table VII) we find that they are highly coordinated (P-value $<1e{-}3$, Fisher's 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 VII, 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 ASE of the target gene or novel TAR (P-value $<1e{-}3$, Fisher's exact test). As we can see in Figure 6, most MIMs correspond to the coordinated regulation of Polymerase II and a transcription factor of a target gene or novel TAR. In the third part of Table VII, we count the occurrences of SIMs 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 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 ASB and ASE.

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; Roach *et al*, 2010; Fan *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 ASE and ASB are reasonable common in the regions we are able to assess allele-specific behavior. 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 ASB and ASE. We do see a significant degree of coordination between the two. It is worthwhile mentioning that not all ASE is necessary correlated with 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 ASE 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 Project Consortium (2010) the types of experimental annotation being performed by The ENCODE Project Consortium (2007) will merge into a hybrid 'MyENCODE' endeavor focusing on explicit annotation of a personal genome.

Materials and 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 were prepared from 5×10^7 formaldehyde crosslinked 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 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 I). However, all samples were prepared following the same protocol.

All the RNA-Seq and ChIP-Seq data were generated as part of The ENCODE Project Consortium (2007) and are available from the UCSC Genome Browser (Rhead *et al*, 2010). The CTCF data set was published in McDaniell *et al* (2010) (GEO accession GSE19622), the NfκB data set 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 data sets (GEO accession GSE30401). All the sequence data are also available from http://alleleseq.gersteinlab.org. A summary of the number total and mapped reads for these data sets is available in Table I.

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, for example, 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, that is, 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 Genomes Project 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 Consortium (2010), see Table II. We have genotyped the fosmid

sequenced deletions using a read-depth approach (Abyzov *et al*, 2011) (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 were 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 II). 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 SNP is <1 or >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 FDR, 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 data set 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 FDR as the number of false positive over the total number of observed positives. We require a FDR of $<10\%$ (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's correlations is dependent on both the accuracy (greater accuracy using a stricter FDR threshold) and 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's correlation between the observed ASB and ASE events was most significant when the FDR threshold set to $\sim 10\%$.

In Table IV, we present the results for ASE and ASB calls for all the data sets. The second column is the number of elements (genes, TARs, or binding sites) that are accessible for the detection of allele-specific behavior, that is, they contain a heterozygous SNP as well as are 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 $6\times$ was sufficiently to obtain the maximum allowing *P*-value threshold. For the ChIP-Seq data sets, the depth threshold required was $7\times$ for Pol III, $8\times$ for CTCF, JunD, cFos, and NfkB, and $9\times$ for cMyc and Pol II. The results of applying these thresholds are outputed in the filtered counts file for each data set.

Using the list of genes and all filtered counts files, information about all asymmetric Het-SNPs from any of the data sets 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 toward parent allele; *q*-value (FDR). Supplementary data includes the list of all allele-specific SNPs (Supplementary Dataset 1) as well as the list of all genes that show ASE (Supplementary Dataset 2), these are also available from http://alleleseq.gersteinlab.org/.

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

The motif consensus sequences were generated from the PWMs (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 $>50\%$ 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 $>75\%$ but each of them is present in $<50\%$. 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 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:

$$\text{delta}\,(\text{maternal}-\text{paternal}) = \log2(P(\text{maternal}, i) / P(\text{maternal}) \times P(\text{paternal}) / P(\text{paternal}, i))$$

where $P(n, i)$ is the frequency for allele n at position i in the PWM (required to be >0.01) and $P(n)$ is the background frequency of allele n.

In Figure 3, we plot this difference in motif scores, delta (maternal–paternal) against the fraction of maternally derived reads overlapping the same heterozygous SNP in the ASB 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 to 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 TSS and the TTS. For ASE novel TARs we do not know which strand is being expressed, thus we associate ASB events that occur within 2.5 kb of either end of the novel TAR. If it is allele specific then it could be further classified into paternal, maternal, and 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 ASE information onto the nodes. Each gene/novel TAR was categorized into three categories: paternal, maternal, or unphased ASE. After constructing the network, we performed a network motif analysis on it, the results of which are shown in Table VII. We analyzed the occurrences of MIMs where two TFs regulate the same gene/novel TAR and 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 was performed using Cytoscape (Cline *et al*, 2007).

Supplementary information

Supplementary information is available at the *Molecular Systems Biology* website (www.nature.com/msb).

Acknowledgements

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

Author contributions: JR and MG conceived of the study. DR and NK performed the experiments. JR, AA, JW, PA, AH, JL, RB, YK, and NB performed the analysis and developed the code. JR, AA, and MG drafted the manuscript. MR, MS, and MG supervised the study.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Abyzov A, Urban AE, Snyder M, Gerstein M (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* **21:** 974–984

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* **11:** 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* **306:** 2242–2246

Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B, Hanspers K, Isserlin R, Kelley R, Killcoyne S, Lotia S, Maere S, Morris J, Ono K, Pavlovic V, Pico AR *et al* (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc* **2:** 2366–2382

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* **25:** 3207–3212

Fan HC, Wang J, Potanina A, Quake SR (2011) Whole-genome molecular haplotyping of single cells. *Nat Methods* **8:** 242–245

Gimelbrant A, Hutchinson JN, Thompson BR, Chess A (2007) Widespread monoallelic expression on human autosomes. *Science* **318:** 1136–1140

Goto T, Monk M (1998) Regulation of X-chromosome inactivation in development in mice and humans. *Microbiol Mol Biol Rev* **62:** 362–378

Gregg C, Zhang J, Butler JE, Haig D, Dulac C (2010a) Sex-specific parent-of-origin allelic expression in the mouse brain. *Science* **329:** 682–685

Gregg C, Zhang J, Weissbourd B, Luo S, Schroth GP, Haig D, Dulac C (2010b) High-resolution analysis of parent-of-origin allelic expression in the mouse brain. *Science* **329:** 643–648

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* **7** (Suppl 1)**:** S4.1–S4.9

Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of *in vivo* protein-DNA interactions. *Science* **316:** 1497–1502

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* **296:** 916–919

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* **328:** 232–235

Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, Haugen E, Zerr T, Yamada NA, Tsang P, Newman TL, Tüzün E, Cheng Z, Ebling HM, Tusneem N, David R *et al* (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature* **453:** 56–64

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. *Genome Res* **21:** 545–554

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

Lyon MF (1961) Gene action in the X-chromosome of the mouse (Mus musculus L.). *Nature* **190:** 372–373

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* **34** (Database issue)**:** D108–D110

Maynard ND, Chen J, Stuart RK, Fan JB, Ren B (2008) Genome-wide mapping of allele-specific protein-DNA interactions in human cells. *Nat Methods* **5:** 307–309

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* **328:** 235–239

Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, Chinwalla A, Conrad DF, Fu Y, Grubert F, Hajirasouliha I, Hormozdiari F, Iakoucheva LM, Iqbal Z, Kang S, Kidd JM *et al* (2011) Mapping copy number variation by population-scale genome sequencing. *Nature* **470:** 59–65

Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network motifs: simple building blocks of complex networks. *Science* **298:** 824–827

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* **464:** 773–777

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5:** 621–628

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* **320:** 1344–1349

Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31:** 3812–3814

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:** 768–772

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* **38** (Database issue)**:** D105–D110

Raha D, Wang Z, Moqtaderi Z, Wu L, Zhong G, Gerstein M, Struhl K, Snyder M (2010) Close association of RNA polymerase II and many transcription factors with Pol III genes. *Proc Natl Acad Sci U S A* **107:** 3639–3644

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 *et al* (2010) The UCSC genome browser database: update 2010. *Nucleic Acids Res* **38** (Database issue)**:** D613–D619

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* **17:** 529–540

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* **328:** 636–639

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* **4:** 651–657

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. *Nat Biotechnol* **27:** 66–75

Shah SP, Morin RD, Khattra J, Prentice L, Pugh T, Burleigh A, Delaney A, Gelmon K, Guliany R, Senz J, Steidl C, Holt RA, Jones S, Sun M, Leung G, Moore R, Severson T, Taylor GA, Teschendorff AE, Tse K *et al* (2009) Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461:** 809–813

The 1000 Genomes Project Consortium (2010) A map of human genome variation from population scale sequencing. *Nature* **467:** 1061–1073

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