Supplementary Methods for "The effects of genome sequence on differential allelic transcription factor occupancy and gene expression"

Timothy E. Reddy[1#], Jason Gertz[1], Florencia Pauli[1], Katerina S. Kucera[2], Katherine E. Varley[1], Kimberly M. Newberry[1], Georgi K. Marinov[3], Ali Mortazavi[3], Brian A. Williams[3], Lingyun Song[2], Gregory E. Crawford[2], Barbara Wold[3], Huntington F. Willard[2], Richard M. Myers[1*]

[1]HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA

[2]Duke Institute for Genome Sciences & Policy, Duke University, Durham, NC, USA

[3]Department of Biology, California Institute of Technology, Pasadena, CA, USA

[#]Current Address: Department of Biostatistics & Bioinformatics, Duke University, Durham, NC, USA

[*]To whom correspondence should be addressed (rmyers@hudsonalpha.org)

Supplementary Methods

Filtering of Biased or Discordant Variants

Results from our experiments are susceptible to sequencing errors in the reference genome. To limit the effect of such errors, we removed suspicious variants identified by three orthogonal methods. First, we removed from analysis all variants with a >7-fold bias towards a single allele in control sequencing of DNA that we formaldehyde fixed but did not immunoprecipitate (N = 82 variants). The threshold was chosen to ensure that variants at our minimum 7x coverage threshold were candidates for exclusion.

Second, we removed from analysis all variants that were discordant with an independent re-sequencing of the GM12878 genome by Complete Genomics (N = 14,748). Combining the two lists, we removed a total of 14,828 variants from all analysis. These variants are included as supplemental datasets.

Third, we removed from analysis a set of 10 variants that were found to be in regions of copy number variation according to microarray experiments. These variants are listed in Supplementary Table 11.

Determination that TF binding sites with differential allelic occupancy overlap more often than sites of equal allelic occupancy

To test for differences in the overlap structure between sites of differential allelic occupancy and sites of equal allelic occupancy, we used a permutation-based testing strategy. The amount of overlap within a set of sites strongly depends on the number of binding sites in that set. Therefore, to control for such effects, we compare overlaps in the significant differential allelic binding sites with an equal number (N = 1,115) of randomly chosen binding sites that lack a significant allelic bias. As a basic strategy, we compared overlaps using three complementary tests: a comparison of the fraction of binding sites that overlapped at least one other site in the same set, a comparison of the log-sum of all the cluster sizes in the set, and a Wilcoxon sign-rank test. Finally, to account for differences that may arise as a result of systematic biases between the sites with and without differential allelic occupancy, we also performed the same tests against sets matched (within 10% difference) for the amount of ChIP-seq signal and sets matched against the reported fold enrichment over background signal from a formaldehyde treated but not-immunoprecipitated sample. For each such test, we performed 500 randomized samplings.

Among the sites of differential allelic occupancy, 30% overlapped another such site. Among the 500 randomly chosen allele-balanced sites, between 10% and 20% of sites -- with of average 15.5% of sites -- overlapped another site in the same set. Sets matched on the number of aligned ChIP-seq reads,

and the fold-enrichment over background signal, had nearly identical distributions. In the randomly chosen, the ChIP-signal matched, and the fold-enrichment matched sets, the percent of overlapping sites followed a normal distribution ($p = 0.20$, $p = 0.22$, $p = 0.13$ respectively, Shapiro-Wilk test), and all had standard deviation of ~0.015, and the 30% overlap observed among the sites of differential allelic occupancy was about 2,000 standard deviations from the permutations.

Comparing sets based on the sum of the log of all cluster sizes (i.e. a score that places more weight on larger clusters, and no weight on non-overlapping sites), we reached very similar conclusions: the randomly chosen and matched sets of binding sites with equal allelic occupancy had scores of 60 ± 6 SD, whereas the sites with differential allelic occupancy had a score of 111, a value 8.5 standard deviations away from the background.

As a third comparison, we performed Wilcoxon sign-rank tests to compare the distribution of cluster sizes in the sites of differential allelic occupancy to the random and matched control sets. Overall, the sites of differential allelic occupancy differed from the control sets with median p-value of 2×10^{-5}, and a range of p-values from 0.02 to 1×10^{-11}.

Finally, the increased overlap of sites with differential allelic occupancy may be caused sites that are overall larger (i.e. cover more nucleotides in the genome). However, this is not the case as the allelically imbalanced sites (with median length of 490 bp) are overall shorter than the sites without a significant allelic difference in occupancy (with median length of 679 bp).

Identification of TF binding motifs

To identify binding motifs for each factor, we first extracted genome sequence for the 50 bp flanking the summit of ChIP-seq signal. Then, to identify an initial seed motif, we applied BioProspector (Liu et al. 2001) to the 200 peaks with the strongest ChIP-seq signal using numerous motif widths. We then selected the most likely seed motif to be the one most similar to a known motif (Matys et al. 2003; Bryne et al. 2008) or, if no known motif was available, according to the maximal BioProspector score. Finally, the motif was refined by using BioOptimizer (Jensen and Liu 2004) on an expanded number of sequences.

Scoring sequences with TF binding motifs

For every TF, the determined binding motif was converted into a position weight matrix (PWM) with each nucleotide *i* at each position j defined as:

$$I_{i,w} = p_{i,w} \log_2\left(p_{i,w}/q_i\right)$$

where $p_{i,w}$ is the probability of nucleotide *i* at position *w* in the motif, and q_i is the background probability of nucleotide *i* occurring. Then we defined the maximum score $S_{\max}$ of the motif as the sum of the relative entropy of the motif:

$$S_{\max} = \sum_{i=A}^{T}\sum_{j=1}^{W} I_{i,j}$$

Finally, we scored each position *R(l)* in the binding site *R* as the sum of scores for each nucleotide at each corresponding position in the PWM:

$$S_l = \sum_{i=l}^{l+W} I_{i,\{\text{nucleotide at } R(l)\}}$$

Calculating dM/dI

To calculate the rate of heterozygosity in motif (dM) versus non-motif (dI) intergenic positions, we first located instances of the TF binding motif in each ChIP region based on similarity at constrained positions. For every TF binding motif, we defined a constrained position as one at which any nucleotide in the PWM has score >0.60. Then, we labeled every position in a binding region that had PWM score >0.75 x S_{max} as an instance of the binding motif. We labeled all constrained positions as motif (*M*), and all positions either not in a binding sequence or in a non-constrained position within a binding sequence as non-motif (*I*). Then, to calculate dM/dI for a set of sequences, we divided the fraction of motif positions with a heterozygous SNP (dM) by the fraction of non-motif positions with a heterozygous SNP (dI). To compare between sites with and without differential allelic occupancy, we defined differential allelic occupancy using our standard 5% FDR, and defined equal allelic occupancy as sites with >25% FDR.

Reporting and Statistical Comparisons of dM/dI

Mean dM/dI values reported in main text are a weighted mean that was calculated by combining all counts across all individual dM/dI calculations and then calculating an overall dM/dI. For clarity, however, distributions in Fig. 2b are not weighted and simply represent distribution over factors in our study. To calculate the statistical significance of the enrichment for motif-disrupting variants, it was

necessary to account for sample size differences between the motif and non-motif sequences, as well as for the occasionally small number of motif-disrupting variants. To do so, we used a simulation approach. Specifically, for each factor and for the combination of all factors, we first empirically calculate dM and dI for differentially- and equally-bound sequences. Then, we simulate the distribution of dM – dI for the equally-bound sequences under the sample size of differentially-bound sequences by sampling from binomial distributions based on empirical estimates. We sample 1,000 times, and the resultant data was normally distributed ($p > 0.05$, Shapiro-Wilk test), and we fit a normal distribution to the sampled values. Finally, we calculate how likely it would be for a same or greater dM/dI (of the differentially-bound regions) to occur in the equally-bound sequences based on the sampled distribution.

Predicting Differential Allelic Expression from RNA Pol2 ChIP-seq

To predict differential allelic expression with RNA Pol2 ChIP-seq, we aligned RNA Pol2 reads to the personalized GM12878 reference genome, as described in the main text. Because RNA Pol2 ChIP-seq signal was more noisy, we require coverage at three different SNPs for all genes for which we make a determination.

As shown in the main text (Fig. 3c) all genes on the X chromosome with significant differential allelic RNA Pol2 occupancy by our measures agree with known details of X-inactivation. However, many genes to not reach our FDR threshold: they may be false negatives (perhaps due to sequencing errors or low sequence coverage), or genes that escape inactivation. To understand how many may escape inactivation, we compared to a recent study evaluating genome-wide escape from X inactivation from many individuals (Carrel and Willard 2005). Of the 61 high-coverage genes (>20X reads at heterozygous positions) that do not meet our FDR threshold, 24 were assayed for escape from X-inactivation in (Carrel and Willard 2005). Of those, 5 (*XG*, *NLGN4X*, *KAL1*, *GPM6B*, and *ARSD*) were previously shown to escape inactivation in all or all but one individual of the individuals previously tested. The overlap was only weakly suggestive ($p = 0.17$) of potential enrichment for inactivation escaping genes in the negatives. Descriptions of these analyses are included in the main and supplemental text.

Clonal isolation of GM12878

Clonal isolated of GM12878 with homogeneous paternal or maternal X inactivation state were obtained by serial dilution. Colonies were expanded, and X inactivation state and homogeneity was tested using a quantitative PCR-based single nucleotide extension assay (Carrel and Willard 2005) to detect relative allelic expression levels at *XIST* rs1620574 (Kucera et al. 2011).

Evaluation of Potential for Random Monoallelic Expression

To evaluate the potential for random monoallelic expression in the GM12878 population (Gimelbrant et al. 2007), we identified autosomal genes with discordant allelic expression among the cell lines clonally derived form GM12878. To identify discordant allelic expression were genes we required that (i) allelic expression was biased towards different alleles in any pair of clonal lines, (ii) the allelic was significant (FDR < 0.05) in at least one of the two lines, (iii) there was an absolute difference of at least 20% in allelic imbalance between the pair of clonal lines (i.e. 45% maternal in one line and 55% maternal in another line would not satisfy our criteria, but a 40%-60% difference would), and (iv) at least seven reads aligned to a heterozygous variant or variants in each of the two different lines. We reasoned that, if indeed differential allelic expression is predominantly attributable to random monoallelic expression in the GM12878 cell line, than we would expect there to be many differences in differential allelic expression between the remaining clones. Of the 170 autosomal genes with significant differential allelic expression in any line, 23 (13.5%) showed evidence of random monoallelic expression. The discordant genes are listed in Supplementary Table 7.

Measurement of Reference Bias

To measure reference bias, we counted the number of reads aligning the reference allele and to the alternate allele at each variant. We then calculated the mean reference bias for each factor by summing over all variants. To calculate how likely a equal or greater bias could occur by random, we compared the distribution of maternal and paternal coverage over all variants using a wilcoxon sign-rank test. Datasets for which we observed a significant bias ($p < 0.05$ after correction for multiple hypotheses) were excluded from downstream analysis.

Physical Interactions between TFs with Positively Correlated Allelic Occupancy

To evaluate if pairs of TFs that we found to have positively correlated allelic occupancy may co-bind, we searched the homoMINT database (Persico et al. 2005) for pairs of interacting proteins. Interactions

reported (ELF1-SRF-SIX5) and (EP300-EGR1-SP1-GABPA) did not correspond to significant positive correlations in allelic occupancy.

Measuring Evolutionary Conservation at Multiply-bound Variants

To measure evolutionary conservation at variants bound by multiple transcription factors, we classified variants by the number of TFs bound. Then for each class, we obtained phastCons conservation scores using the SeattleSeq annotation server. PhastCons scores range from 0 to 1, and variants are predominantly non-conserved (<<0.5) or conserved (>>0.5) (Supplementary Fig. 22). For the purposes of our study, we consider a nucleotide to be conserved when the phastCons score was greater than 0.5.

For each class of variants, we first calculate the percentage of variants that are conserved. Next, to determine if there were significant differences in conservation between the different classes, we used a sampling approach to estimate the variance in the set of uniquely-bound variants normalized to the sample size of the test set. Specifically, for each set of multiply-bound variants, we randomly sampled an equal number of variants from the singly-bound set, and calculate the percentage of conserved variants in that sample. We then repeat the process 500 times. The resulting values follow a normal distribution (Supplementary Fig. 23). Finally, for each set of multiply-bound variants, we calculate the percentage under constraint. Then, according to the mean and standard deviation of the sample, we then determine as our p-value the probability that a greater fraction of singly-bound variants in a sample of the same size are also under constraint.

Inheritance of Differential Allelic Expression

To determine the extent to which allelic expression may be inherited, we performed RNA-seq in the GM12891 and GM12892 LCLs that were derived from the parents of GM12878. RNA-seq was performed as described earlier, except with 50 bp paired end reads from an Illumina HiSeq instrument. To calculate expression, we aligned to RefSeq transcripts, and normalized for the length of the sequence and the number of reads aligned (i.e. RPKM). We then filtered genes with low expression (RPKM < 1.0 in all of GM12891, GM12892, and GM12878), and genes that were known to be imprinted. Finally, for all genes with significant differential allelic expression, we used Spearman correlations to compare the $\log_2$ ratio of expression in the maternal to the paternal cell lines [i.e. $\log_2$(GM12892/GM12891)] to the $\log_2$ ratio of maternally to paternally aligned reads from the GM12878 RNA-seq or RNA Pol2 ChIP-seq experiments. Correlation and significance were calculated in R.

Determination of Autosomal Dosage Compensation

To assay for evidence of autosomal dosage compensation, we evaluated if evidence of differential allelic expression of RNA Pol2 was associated with increased or decreased expression as measured by RNA-seq. The motivations for basing the determination of differential allelic expression on RNA Pol2 were many-fold. First, RNA Pol2 ChIP-seq allows us to predict differential allelic expression in a greater number of genes owing to increased heterozygosity in introns. Second, estimating differential allelic expression based on RNA Pol2 occupancy allows us to control for read depth independent of expression measurements. Specifically, when comparing expression between genes with differential and equal allelic RNA Pol2 occupancy, we required the overall number of RNA Pol2 reads at heterozygous positions to be statistically similar between the two sets of genes. Doing so has the effect of biasing us towards genes with similar levels of expression, thus giving us a conservative estimate of differences in gene expression.

First, we select a threshold $\mathbf{r} < 0.25$ indicating the extent of differential allelic occupancy, and classify genes as differentially expressed if the fraction of maternal reads is less than or greater than **r**, and unbiased if the fraction of maternal reads are between $0.5 - \mathbf{r}$ and $0.5 + \mathbf{r}$. Then, to control for read depth, we select a minimum and maximum coverage at heterozygous positions $\mathbf{n} - \mathbf{w}$ and $\mathbf{n} + \mathbf{w}$, respectively, and only consider genes with RNA Pol2 coverage within $\mathbf{n} \pm \mathbf{w}$, inclusive. As an additional control, we compare the distribution of coverage between the two classes, and only consider our test valid if there is no evidence that the two distributions are statistically different ($p > 0.5$ according to a two-sided Wilcoxon test). Finally, we compare the median expression according to our RNA-seq experiments between the two sets of genes, and test for a statistically significant difference between the sets using a Wilcoxon sign-rank test.

We performed this analysis for many choices of **r**, **n**, to show the results are insensitive to the specific parameters chosen. Generally, **w** was chosen to ensure that at least 5 genes were in each case to ensuring enough statistical power to make a comparison while maintaining the requirement of similar levels of coverage. All calculations were performed using the R statistical package.

Identification of TF and co-factor occupancy at variants associated with transcriptional regulation

To determine if TF occupancy was enriched at variants in the genome previously shown to correlate with regulation of gene expression (i.e. expression quantitative trait loci, or eQTLs), we retrieved the list

of all genetic linkage measurements from (Montgomery et al. 2010). There were 102 unique TF-bound variants that were occupied by a TF or cofactor in our study. At five of those variants, we also observed differential allelic occupancy. To determine how many overlaps would be expected by random, we randomly permuted significance values in the eQTL data 500 times, and repeated the analysis. For both the overlap with all TF binding sites ($p = 7.4 \times 10^{-21}$ based on a normal approximation to the null distribution) and the overlap with sites of differential allelic occupancy ($p = 0.06$ based on the empirical distribution) the overlap was less than would be expected by random. That the overlap with differential allelic occupancy was very small is likely due a combination of the use of tagged variants and haplotype structure in the original study (i.e. not identifying the causative variants); that many of the identified eQTLs were not heterozygous in GM12878 cells; and that there are very many active regulatory factors in GM12878 that we did not assay in our study. Nonetheless, the statistically significant overlap is encouraging that we are indeed recovering functional relationships between TF occupancy.

Detailed Illumina Sequencing Library Construction Protocol

DNA fragments recovered from ChIPs or reverse cross-linked chromatin were repaired, ligated to adapters, size selected and PCR-amplified to make the library for sequencing. Illumina DNA Library Construction Kit reagents were substituted in this protocol with reagents from NEB and Finnzymes, except for the adapter oligo mix and the PCR primers, which can be ordered from Illumina. We used paired-end adapters for library construction, even though the ChIP libraries were sequenced with a single-end sequencing run. The one exception was one lane of RNA Pol2 ChIP-seq, which was sequenced as paired-end 100bp reads from an Illumina HiSeq 2000.

For end repair, the following were mixed in a PCR tube on ice, spun down briefly in a microfuge to mix and then incubated at 20°C in a thermal cycler for 30 minutes: 10 µl 10X T4 DNA ligase buffer (supplied with T4 DNA Ligase, NEB M0202), 4 µl 10 mM dNTP mix (NEB N0447), 75 µl recovered DNA fragments from ChIP, 5 µl T4 DNA polymerase (NEB M0203), 5 µl T4 Polynucleotide Kinase (NEB M0201) and 1 µl Klenow DNA polymerase (NEB M0210). The reaction was purified on one QIAquick PCR cleanup column (Qiagen 28106) and eluted with 32 µl EB warmed to 55°C. The EB was allowed soak the filter in the column for 1 minute before spinning for 1 minute to collect the DNA. For dA addition, the following were mixed in a PCR tube on ice, spun down briefly in a microfuge to mix and then incubated at 37°C in a thermal cycler for 30 minutes: 32 µl end-repaired DNA fragments, 10 µl 1mM dATP (NEB N0440S), 5 µl 10X NEBuffer2, and 3 µl Klenow Fragment (3’ to 5’ exo-; NEB

M0212). The reaction was purified on one QIAquick PCR cleanup column and eluted with 42 µl EB warmed to 55°C. The EB was allowed soak the filter in the column for 1 minute before spinning for 1 minute to collect the DNA. For adapter ligation, the following were mixed in a PCR tube on ice, spun down briefly in a microfuge to mix and then incubated at 20°C in a thermal cycler for 15 minutes: 42 µl DNA recovered from dA addition, 5 µl T4 DNA Ligase buffer (supplied with T4 DNA Ligase, NEB M0202), 1 µl of a 1:10 dilution of Adapter oligo mix (Illumina) and 2 µl T4 DNA ligase (NEB M0202). The reaction was purified on one QIAquick PCR cleanup column and elute with 50 µl EB warmed to 55°C. The EB was allowed soak the filter in the column for 1 minute before spinning for 1 minute to collect the DNA. Alternatively, this purification step was skipped when proceeding directly to gel size selection.

Gel purification and size selection was carried out to remove the extra sequencing adapters that were not ligated to ChIP DNA and to isolate 150-300 bp fragments allowing for higher density of productive clusters on the sequencing flowcells. A 2% low-melting agarose gel (SeaPlaque) in 1X TAE with EtBr (final concentration in gel is 0.4 µg/ml) was poured and run in a 4°C cold room. PCR Marker DNA Ladder (NEB N3234L) and the DNA products were loaded on the gel and the gel was run at ~115 V until the loading dye migrated 6 cm (~2 hours). The gel region from 150 bp to 300 bp was excised from the gel for each sample. Due to the low concentration of library DNA fragments at this step, they were not visible on the gel. The adapters, however, were visible and were carefully excluded from the extracted gel fragment. Image the gel before and after excision of the library, if desired. The DNA fragments were extracted from the gel with QIAquick Gel Extraction Kit (Qiagen 28704) columns. The Qiagen protocol was followed, with the exception of warming the gel and Buffer QG to 55°C to melt the gel. Instead, this was done at room temperature by vortexing every 2-3 minutes until the gel dissolved. We included the optional step of washing the column with 0.5 ml of Buffer QG before adding Buffer PE and eluted with 25 µl 50°C EB. The EB was allowed soak the filter in the column for 1 minute before spinning for 1 minute to collect the DNA.

For sequencing library amplification, the following were mixed in a PCR tube on ice: 24 µl DNA fragments (from gel size selection), 25 µl Phusion DNA Polymerase Mix (Finnzymes F531), 0.5 µl PCR primer 1.1 (Illumina) and 0.5 µl PCR primer 2.1 (Illumina). The reaction was spun down briefly in a centrifuge to mix and amplified in a thermal cycler with the following protocol: 98°C for 30 sec, 15 cycles of 98°C for 10 sec, 65°C for 30 sec and 72°C for 30 sec, 72°C for 5 min, 4°C hold. Some samples (denoted in blue in Supplementary Table 12) were amplified with 25 cycles of PCR before gel

extraction and 15 cycles after gel extraction. The reaction was purified on one QIAquick PCR cleanup column and eluted with 32 µl EB warmed to 55°C. The EB was allowed soak the filter in the column for 1 minute before spinning for 1 minute to collect the DNA. The final product was quantified with a Qubit fluorometer with a high-sensitivity (HS) kit before sequencing on the Illumina GAIIX platform.

References

Bryne JC, Valen E, Tang MH, Marstrand T, Winther O, da Piedade I, Krogh A, Lenhard B, Sandelin A. 2008. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res* **36**(Database issue): D102-106.

Carrel L, Willard HF. 2005. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* **434**(7031): 400-404.

Gimelbrant A, Hutchinson JN, Thompson BR, Chess A. 2007. Widespread monoallelic expression on human autosomes. *Science* **318**(5853): 1136-1140.

Jensen ST, Liu JS. 2004. BioOptimizer: a Bayesian scoring function approach to motif discovery. *Bioinformatics* **20**(10): 1557-1564.

Kucera KS, Reddy TE, Pauli F, Gertz J, Logan JE, Myers RM, Willard HF. 2011. Allele-specific distribution of RNA polymerase II on female X chromosomes. *Hum Mol Genet*.

Liu X, Brutlag DL, Liu JS. 2001. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput*: 127-138.

Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV et al. 2003. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* **31**(1): 374-378.

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**(7289): 773-777.

Persico M, Ceol A, Gavrila C, Hoffmann R, Florio A, Cesareni G. 2005. HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. *BMC Bioinformatics* **6 Suppl 4**: S21.