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

Timothy E. Reddy[1,2], Jason Gertz[1], Florencia Pauli[1], Katerina S. Kucera[2], Katherine E. Varley[1], Kimberly M. Newberry[1], Georgi K. Marinov[3], Ali Mortazavi[3,4], 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

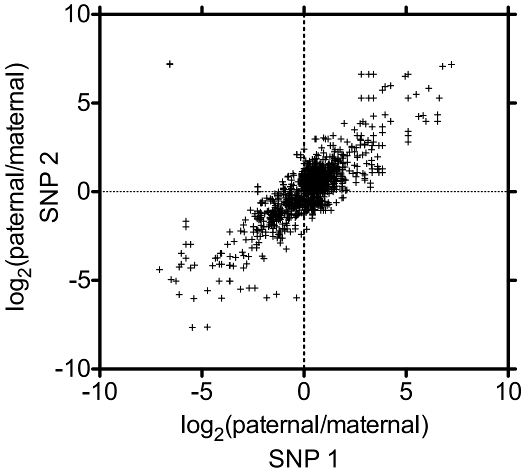
[2]Duke Institute for Genome Sciences & Policy, Duke University, Durham, NC, USA

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

[4]Developmental & Cell Biology, University of California, Irvine, CA, USA

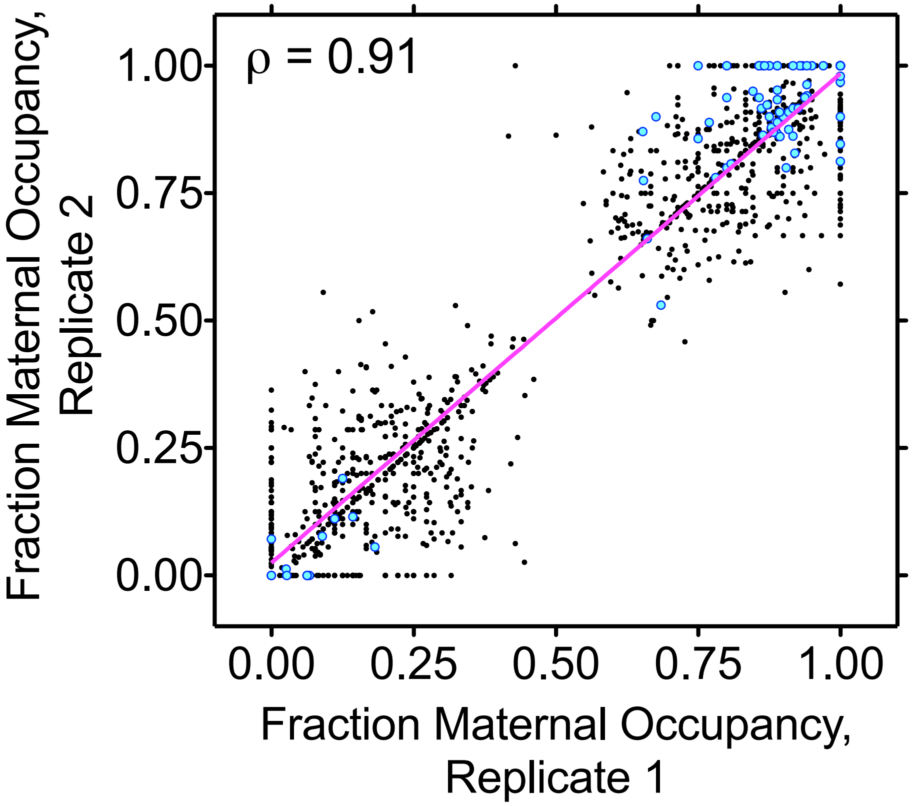
[*]To whom correspondence should be addressed (rmyers@hudsonalpha.org)

Supplementary Fig. 1:



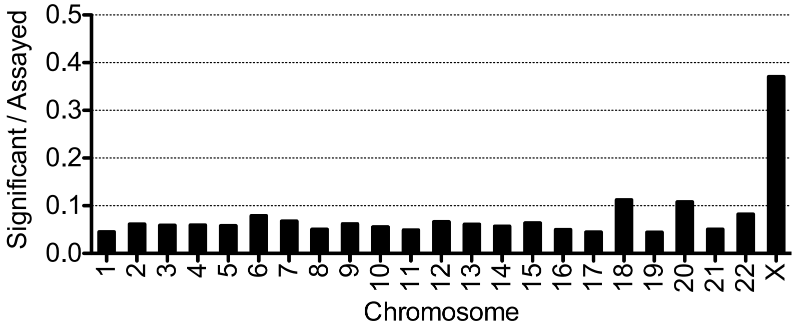
Intra-binding site correlation of differential allelic occupancy between SNPs. First, all binding sites that covered multiple SNPs with high coverage – at least 20 aligned reads – were identified. Then, the log of the ratio of paternal to maternal reads was calculated for each SNPs, and plotted for all pairs of SNPs in each binding site. Allelic biases between intra-peak SNPs were correlated with ρ = 0.65.

Supplementary Fig. 2:



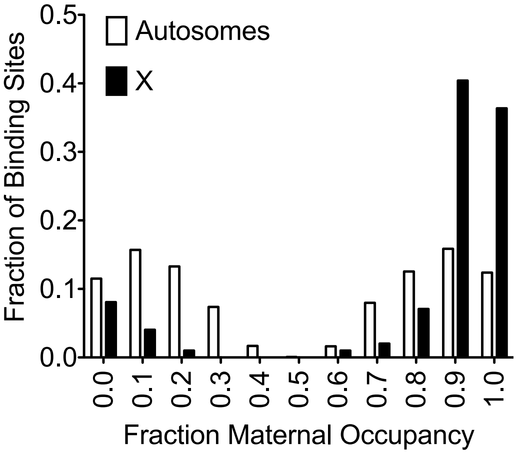
Reproducibility of allelic biases in occupancy. For each site of differential allelic occupancy, the fraction of reads aligning to the maternal allele is plotted for two biological replicates. Magenta line indicates linear regression between the replicates (ρ = 0.91). Black dots are autosomal sites (ρ = 0.91), and blue circles are X chromosomal sites (ρ = 0.70).

Supplementary Fig. 3:



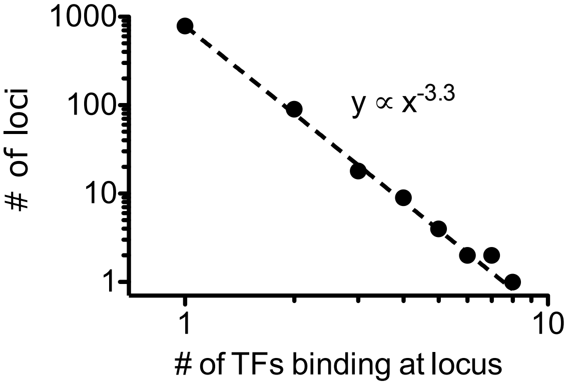
Per-chromosome distribution of allele-biased occupancy. For each chromsome (x-axis), the fraction of binding sites with significant allele-biased expression (y-axis) is plotted. Notably, binding sites on the X chromosome are far more likely to have allele-biased occupancy, as expected due skewed X inactivation in GM12878 cells.

Supplementary Fig. 4:



Histogram of maternal bias for all sites of differential allelic occupancy. White bars are autosomal sites, and black bars are X chromosomal sites.

Supplementary Fig. 5

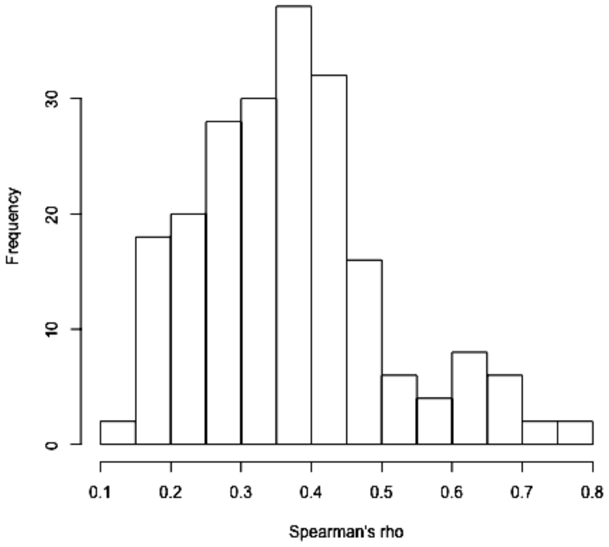


Overlap structure of differential allelic occupancy across the genome. Overlapping binding sites for different factors were clustered together on the genome. Each cluster of overlapping binding sites was referred to as a locus. Plotted is a histogram of the number of TFs binding in each locus. The dashed line indicates power-law distribution fit using maximum likelihood. The power-law fit was significant according to a Kolmogorov-Smirnov goodness-of-fit test with p = 0.40. Likelihood ratio tests to rule out fits to closely related distributions suggests that the distribution is more likely to be a power-law than an exponential distribution (p = 0.03) or a Poisson distribution (p = 0.009), and did not have sufficient power to distinguish from a Weibull distribution (p = 0.4) or from a log-normal distribution (p = 0.43). Fitting the power-law distribution, goodness-of-fit tests, and likelihood ratio tests were all performed as described in (Clauset et al. 2009) using code provided by the authors.

Supplementary Fig. 6: (Included as external image due to size)

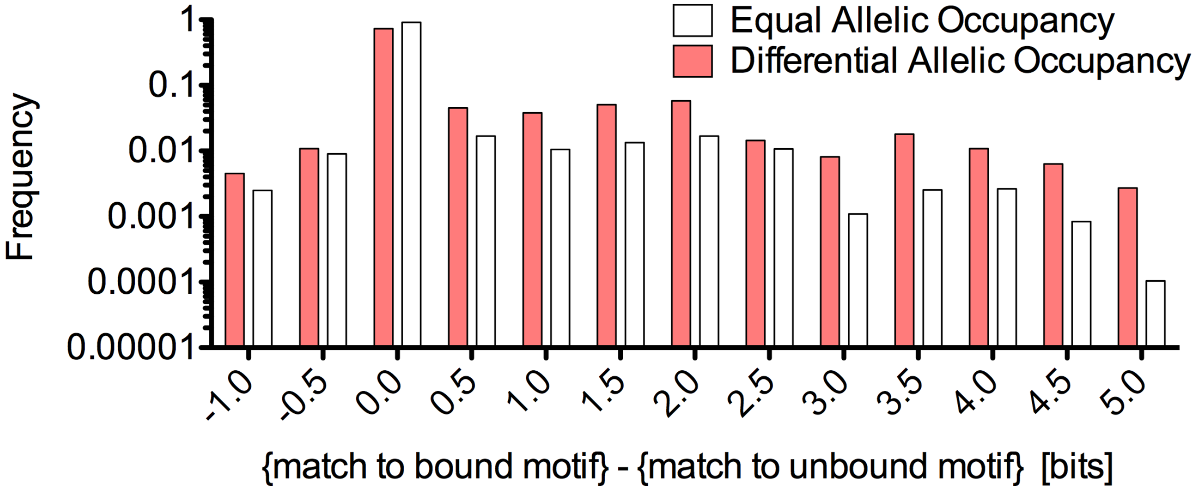
Scatter-plot of allele-biased occupancy at co-bound SNPs for all pairs of transcription factors. Color indicates the amount of correlation, is significant, using the same color bar as in Fig. 1b. White indicates no significant correlation (p > 0.05). Matrix is organized as for Fig. 1b.

Supplementary Fig. 7:



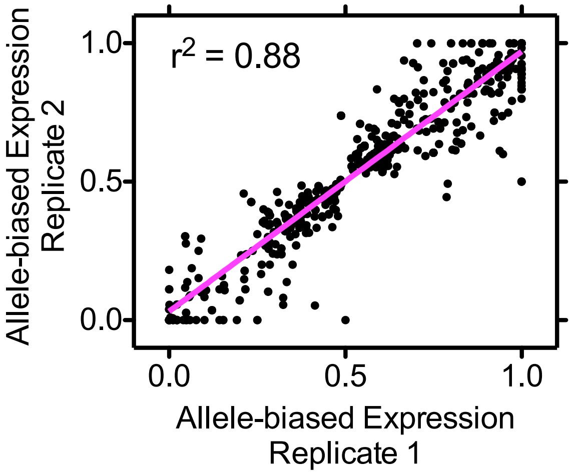
Distribution of coordinated differential allelic TF occupancy. For all pairs of factors with significant (p < 0.05) correlation in differential allelic occupancy, the number of such pairs (y-axis) is plotted as a function of the Spearman correlation coefficient (x-axis).

Supplementary Fig. 8:



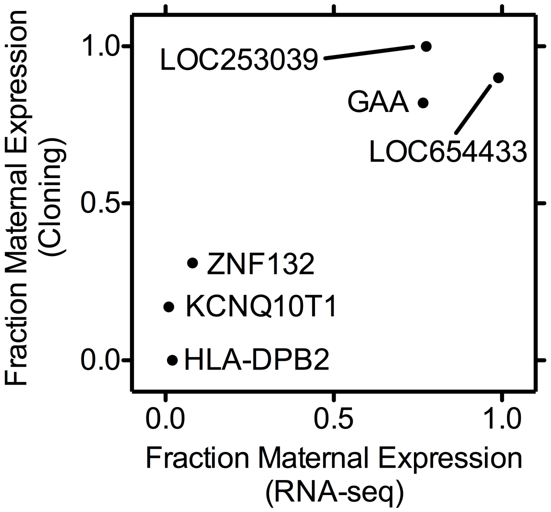
Distribution of differences in TF binding motif similarity between bound and unbound alleles (plotted as log scale on the x-axis) for differentially bound (red) and equally bound (white) sites. Overall there is more similarity on the bound allele, even when the difference in allelic occupancy between the two alleles was not significant. However, when allelic differences in binding were significant, the difference was overall larger. Note that the x-axis is on a log scale and that subtle differences between red and white bars are indeed substantial.

Supplementary Fig. 9:



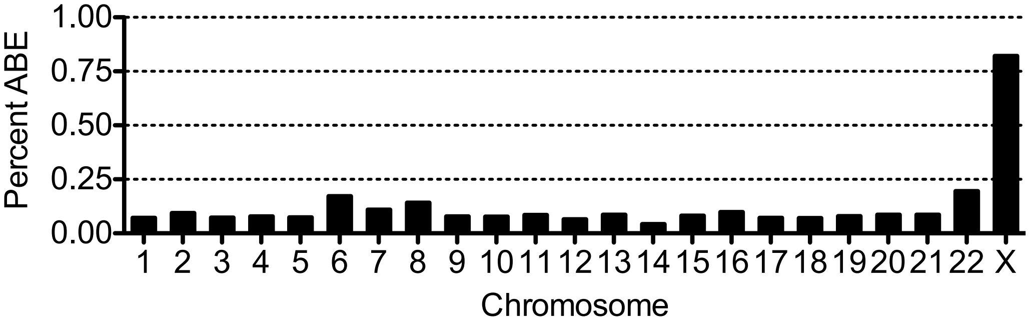
Reproducibility of measurement of differential allelic expression. For each gene with differential allelic expression, the fraction of maternal expression was plotted for two biological replicates (x- and y- axis). Magenta line indicates linear regression between the replicates.

Supplementary Fig. 10:



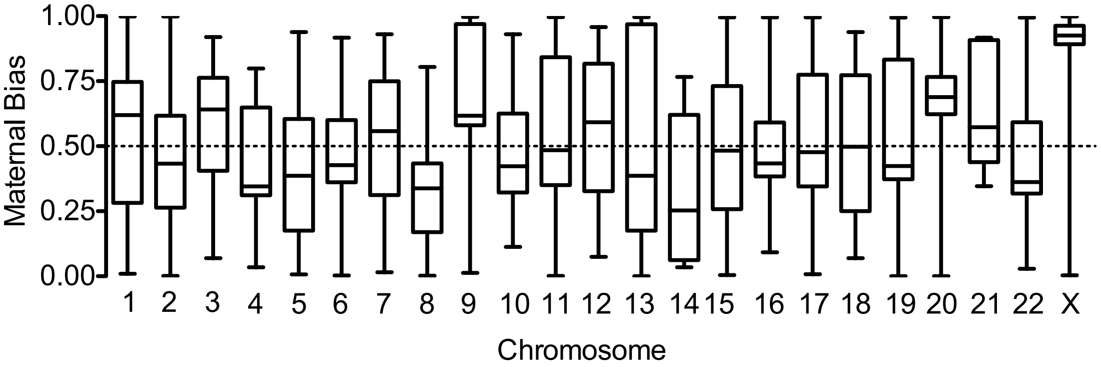
Validation of differential allelic expression. To validate differential allelic expression from RNA-seq, we used PCR to amplify fragments of genomic DNA and RT-PCR to amplify the same fragments of expressed mature RNAs. We then cloned fragments into a sequencing plasmid, transformed the plasmid into *E. coli*, and grew the transformed *E. coli* on selective media. We picked colonies, isolated the plasmids, and sequenced the cloned inserts. Plotted is the fraction of maternal expression determined by RNA-seq (x-axis) against the fraction of maternal expression determined by cloning (y-axis). Differential allelic expression of all six genes tested matched differential allelic expression by RNA-seq. For five of the six genes, the differential allelic bias was significant compared to the genomic DNA sequencing ($p < 0.5$, two-sided binomial test). For the sixth gene (*ZNF132*), 11 of 16 colonies matched the paternal allele and additional sequencing may provide additional statistical power to show significance.

Supplementary Fig. 11:



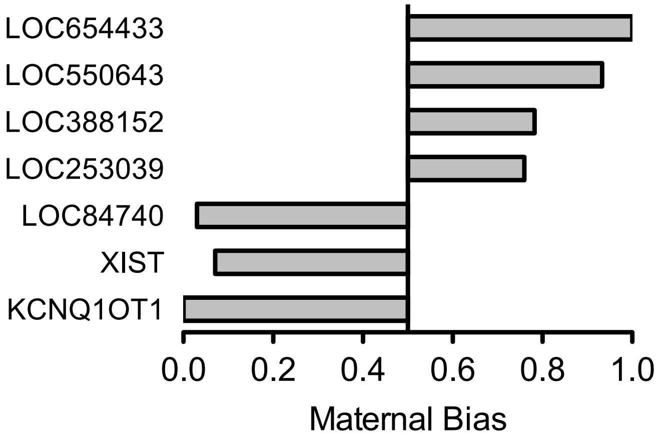
Per-chromosome distribution of allele-biased expression. For each chromosome (x-axis), the fraction of heterozygous genes with allele-biased expression is indicated on the y-axis. As expected, allele-biased expression is far more prevalent on the X chromosome.

Supplementary Fig. 12:



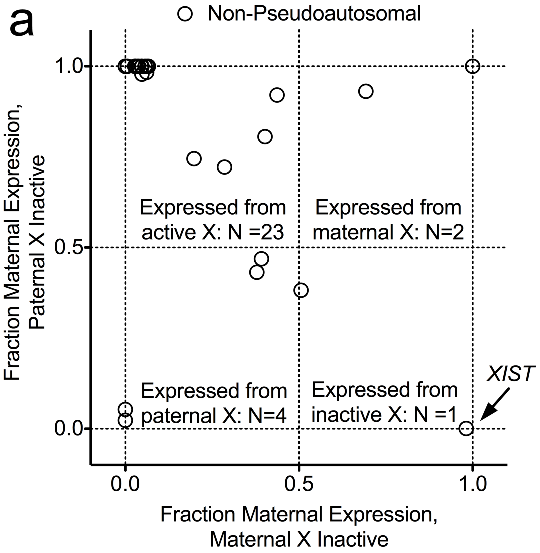
Per-chromosome distribution of maternal bias (i.e. the fraction of expression arising from the maternal allele) for all genes with allele-biased expression. For each chromosome (x-axis), the median maternal bias (y-axis) is close to 0.5, indicating as much biased expression arises from the maternal allele as for the paternal allele. For the X chromosome, however, the majority of expression arises from the maternal (predominantly active) allele. Error bars indicate the full range of the data.

Supplementary Fig. 13:



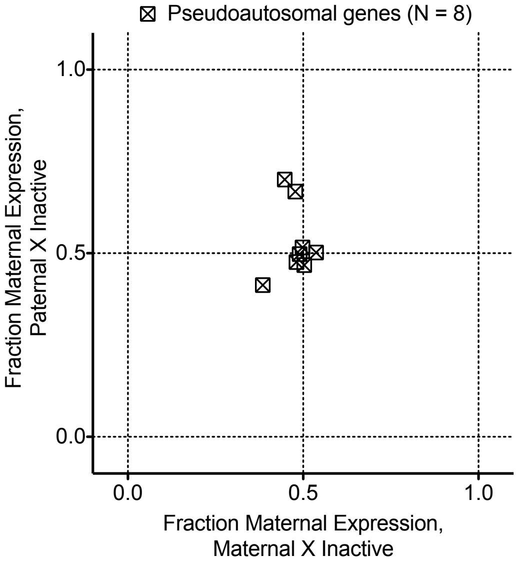
Fraction of maternal expression (x-axis) for all long non-coding RNAs with allele-biased expression (y-axis). *XIST* and *KCNQ1OT1* are well known to have allele-biased expression. The others are novel.

Supplementary Fig. 14



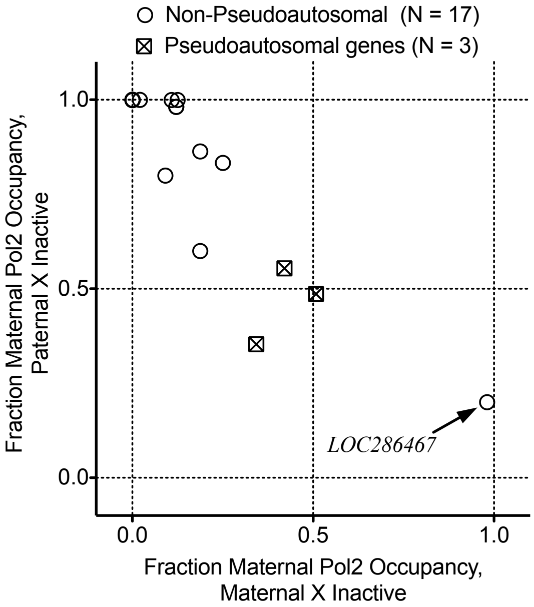
Comparison of differential allelic expression of X chromosomal, non-pseudoautosomal (i.e. subject to inactivation) genes between clonal isolates of GM12878 with the maternal X inactivated (x-axis) and isolates with the paternal X inactivated (y-axis). Genes in the top-left corner are those always expressed from the active X, whereas *XIST*, in the bottom-right corner is always expressed from the inactive X (as expected). Genes in bottom-left and top-right quadrants are always expressed paternally or maternally, respectively, and genes in the center appear to escape inactivation.

Supplementary Fig. 15:



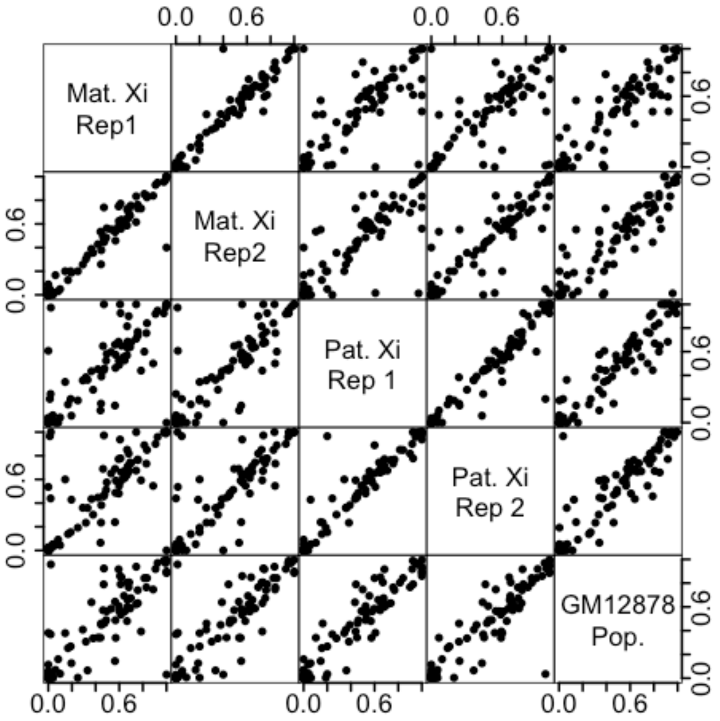
Similar to above, but for genes located in the pseudoautosomal region of the X that is not subject to inactivation.

Supplementary Fig. 16:



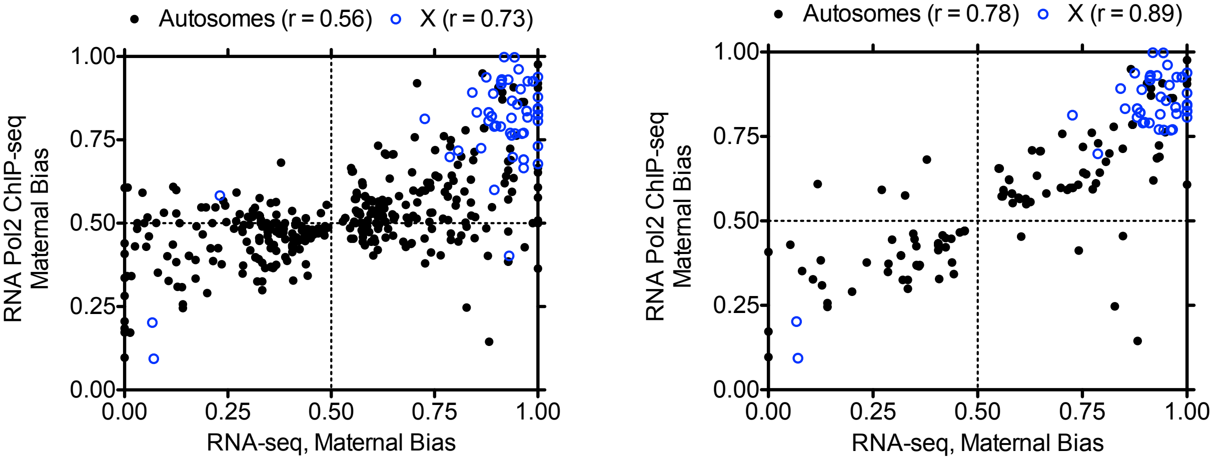
Pol2-predicted biases for X chromosomal genes in inactivated region (circles) and pseudoautosomal regions (squares), plotted as in above plots for RNA-seq. Substantially fewer and shorter high-throughput sequencing reads were produced for these experiments, explaining the fact that only 20 genes were covered at the same coverage threshold.

Supplementary Fig. 17:



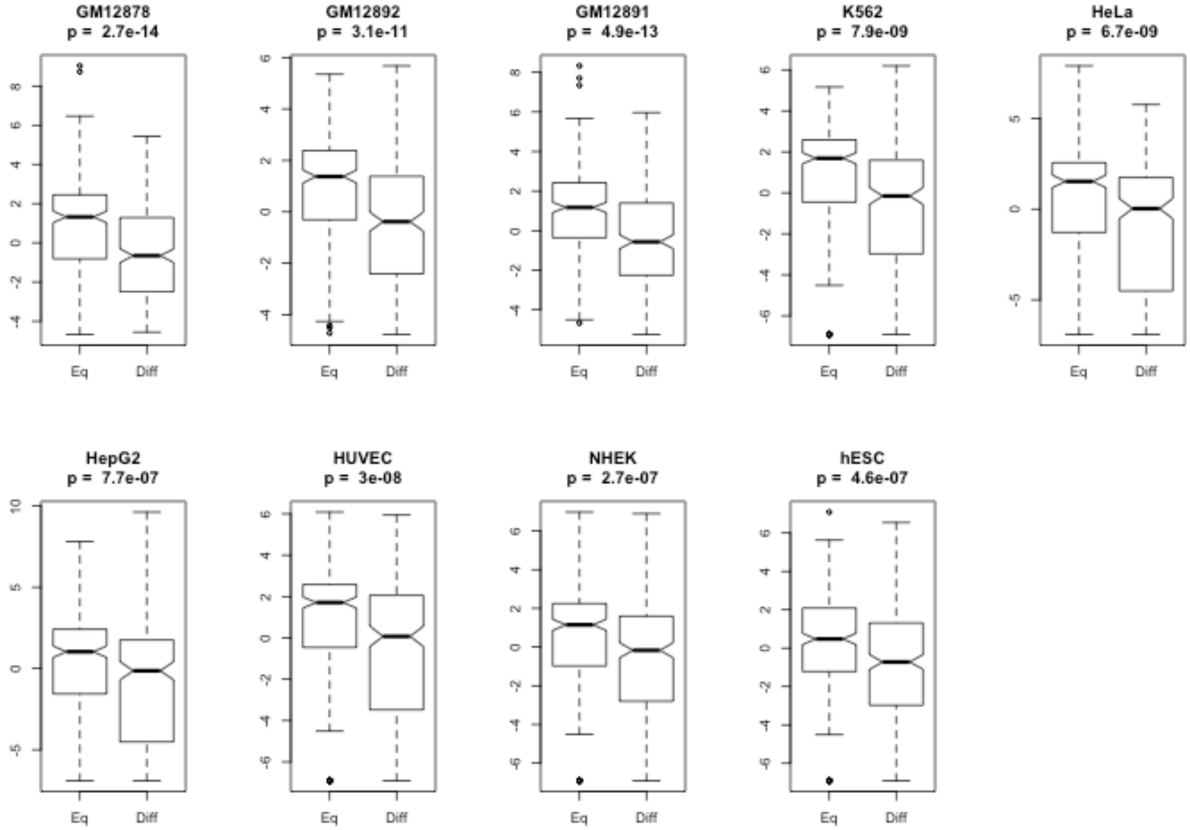
The fraction of maternal expression for autosomal genes in clonal isolates of GM12878 and compared to the original GM12878 population. Labels refer to the X inactivation state of the isolated clones and the replicate of the clonal isolation and propagation. Spearman correlation coefficients range from 0.77 to 0.95.

Supplementary Fig. 18:



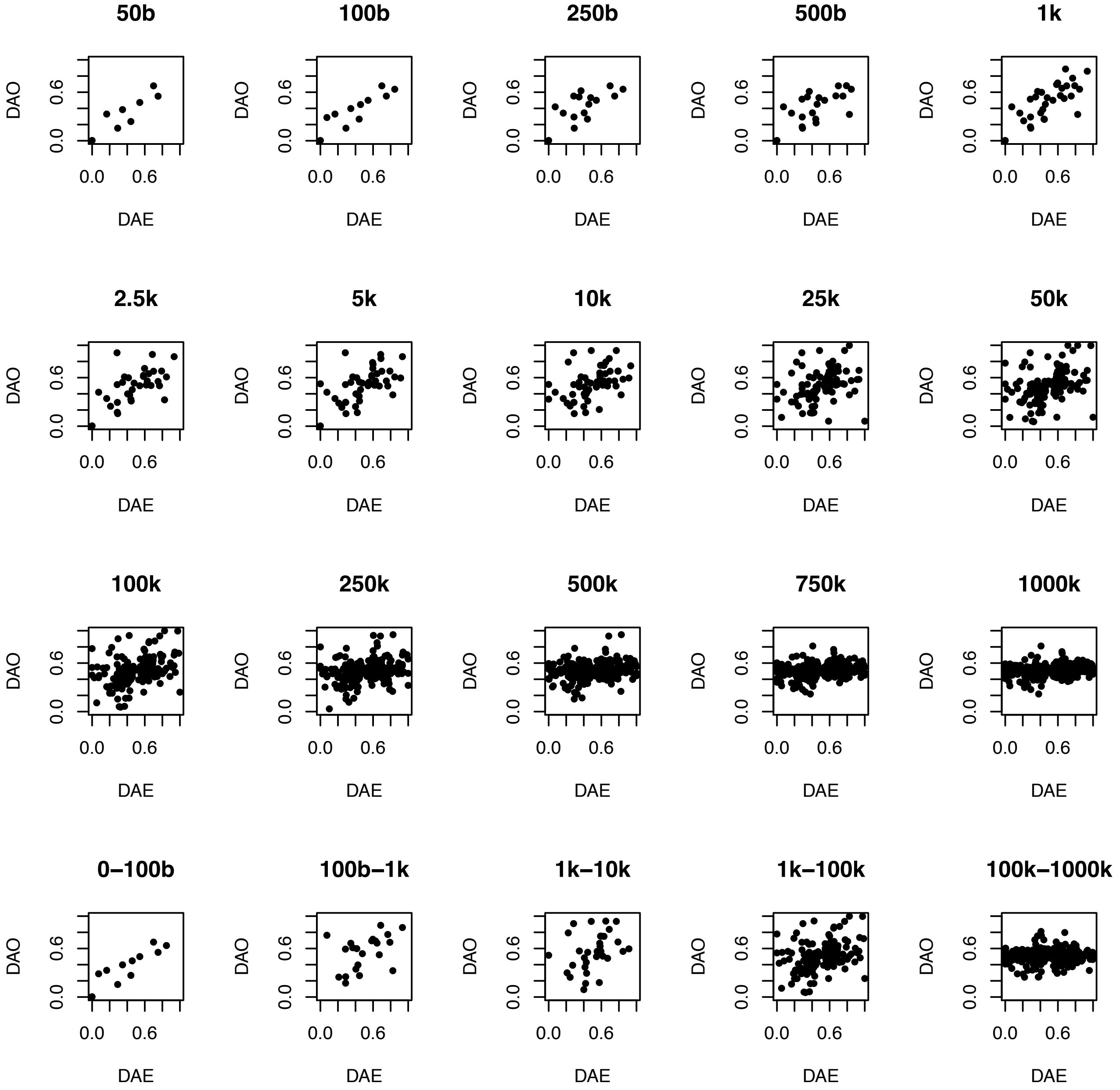
Fraction of maternal Pol2 occupancy (y-axis) plotted against the fraction of maternal expression (x-axis). Left panel shows all genes with significant (FDR < 0.05) differential allelic expression according to RNA-seq. Right panel shows all genes that have significant differential allelic expression in both RNA-seq and RNA Pol2 ChIP-seq. Blue circles indicate X chromosomal genes.

Supplementary Fig. 19:



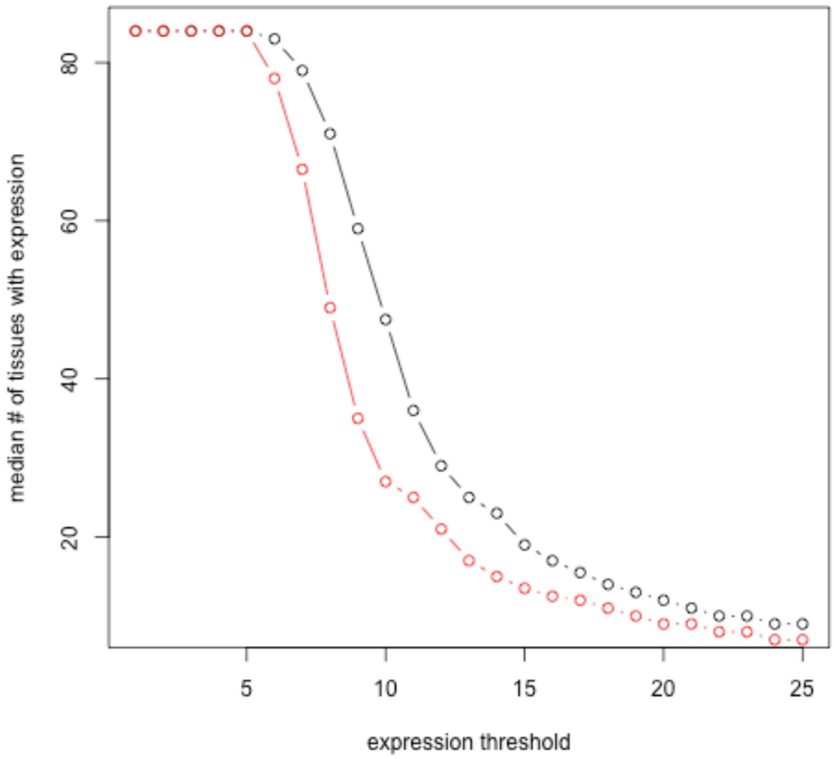
Comparison across nine cell lines of the expression of genes with differential allelic expression to that of genes with equal allelic expression in GM12878. Genes were identified as differentially expressed if the percent of maternal occupancy for RNA Pol2 was less then 25% or greater than 75%, and equally expressed otherwise. Unbiased genes were far more numerous than biased genes, and therefore we randomly sampled such that the final sets had the same number of genes. We performed the analysis in GM12878, K562, HeLa, HepG2, HUVEC, NHEK, and hESC cell lines. P-values were calculated using the median p-value reported from the Wilcoxon test applied to 10,000 random samplings of the unbiased genes.

Supplementary Fig. 20:



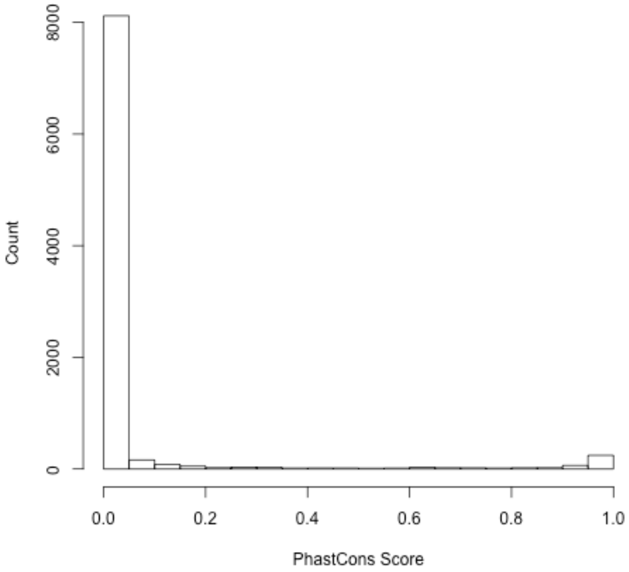
Scatter plots of correlation between differential allelic expression ("DAE", x-axes) and differential allelic occupancy of sequence-specific TFs ("DAO", y-axes). Each point represents a gene with significant (FDR < 5%) differential allelic expression that has significant differential allelic occupancy within the window indicated in each plot title. For the top three rows, plots include aggregation of all allelic binding signal within the distance from transcription start sites indicated in the title. In the bottom row, data is shown for distance windows as indicated in the title. For instance, "1k-10k" indicates all allelic occupancy more than 1 kb but less than 10 kb (inclusive) from transcription start sites.

Supplementary Fig. 21:



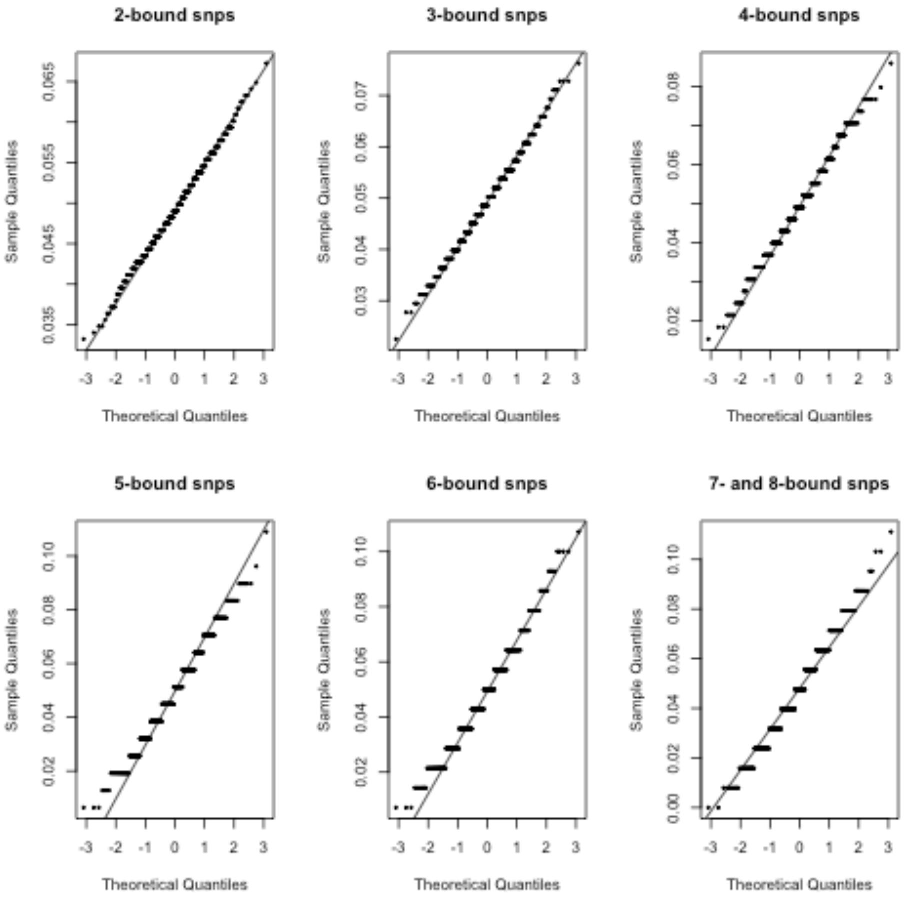
Genes with differential allelic expression (red) are expressed in fewer tissues than genes without evidence of differential allelic expression (black). We obtained gene expression measurements from a broad selection of human tissues from (Su et al. 2004). To avoid artifacts arising from selecting an arbitrary expression threshold at which to classify a gene as expressed or not expressed in a given tissue, we instead selected a range of gene expression thresholds (x-axis), and calculated the number of tissues in which each gene is expressed above that threshold. We report, on the y-axis, the median of that distribution for both sets of genes. The genes with differential allelic expression are always expressed in fewer tissues, independent of the chosen expression threshold. It may be that genes with differential allelic expression are found in fewer tissues because that have overall lower expression. Therefore, we also reasoned that genes with a greater degree of tissue-specific expression would have more variable expression across the entire panel of tissues. Indeed, the coefficient of variation (CV) for genes with differential allelic expression (median CV = 10.7) was significantly greater than the CV for genes without (median CV = 4.2) with $p = 1.3 \times 10^{-6}$ according to a one-sided Wilcoxon rank sum test.

Supplementary Fig. 22:



For all TF-bound variants in our study, the number of variants at each phastCons score. Scores close to 0 indicate low evolutionary conservation, and scores close to 1 indicate high evolutionary conservation.

Supplementary Fig. 23:



Quantile-quantile plots of the distribution of the mean percent of variants under conservation in sampled sub-sets of the uniquely-bound variants. For each plot, the number of variants in the test set (e.g. all variants bound by 2 factors) were sampled 500 times from the uniquely-bound variants, and the average number of variants under conservation in each sample set was reported. Plotted in each panel are the associated quantile-quantile plots.

Supplementary Table 1: High-throughput sequencing depth and antibody used for ChIP-seq experiments.

Factor	Antibody	Replicate 1 Aligned Reads (M)	Replicate 2 Aligned Reads (M)	Total Aligned Reads (M)
ATF3	sc-188[1]	18	23	41
BATF	sc-100974[1]	19	20	39
BCL11A	ab19489[2]	19	21	40
BCL3	sc-185[1]	18	28	46
BCLAF1	sc-101388[1]	32	27	59
EBF1	sc-137065[1]	32	18	50
EGR1	sc-110[1]	18	18	36
ELF1	sc-631[1]	20	19	39
EP300	sc-585[1]	31	18	49
ETS1	sc-350[1]	20	22	42
GABPA	sc-28312[1]	32	21	53
IRF4	sc-6059x[1]	18	21	39
LEF1	sc-8592[1]	30	21	51
NRSF	Custom	34	14	48
PAX5	sc-1974[1]	39	15	54
PBX3	sc-891[1]	20	21	41
POU2F2	sc-233[1]	46	51	97
SIX5	sc-55706[1]	23	27	50
SP1	sc-7824[1]	32	20	52
SPI1	sc-22805[1]	35	31	66
SRF	sc-335	18	35	53
TCF12	sc-357[1]	19	18	37
USF1	sc-229[1]	22	31	53
YY1	sc-281[1]	21	23	44
ZBTB33	sc-23871[1]	19	16	35
Total		**635**	**579**	**1,214**

[1] Santa Cruz Biotechnology
[2] Abcam

Supplementary Table 2: The amount of genomic binding and allele-biased occupancy observed for all sequence-specific transcription factors and P300 in the study. Average binding site size varied between factors tested, and explained the variation in the percent of binding sites with ≥ 7x coverage at heterozygous variants.

Factor	Binding Sites	Average Binding Site Size (bp)	Binding Sites with ≥ 7 Het. Reads	% Het.	ABO Binding Sites	% Allele-biased Occupancy
ATF3	2,192	575	166	7.57%	2	1.20%
BATF	17,639	607	2,147	12.17%	205	9.55%
BCL11A	6,662	711	829	12.44%	23	2.77%
BCL3	1,962	971	394	20.08%	8	2.03%
BCLAF1	2,122	1,654	467	22.01%	11	2.36%
EBF1	16,331	761	2,475	15.16%	235	9.49%
EGR1	3,498	718	287	8.20%	10	3.48%
ELF1	12,118	1,103	1,871	15.44%	36	1.92%
EP300	983	762	148	15.06%	6	4.05%
ETS1	3,494	858	425	12.16%	7	1.65%
GABPA	3,688	868	494	13.39%	35	7.09%
IRF4	5,051	713	716	14.18%	5	0.70%
LEF1	1,122	523	117	10.43%	1	0.85%
NRSF	3,346	829	509	15.21%	28	5.50%
PAX5	7,827	779	1,101	14.07%	63	5.72%
PBX3	4,720	629	430	9.11%	17	3.95%
POU2F2	3,705	1,189	770	20.78%	28	3.64%
SIX5	3,085	627	331	10.73%	7	2.11%
SPI1	19,977	510	2,438	12.21%	191	7.91%
SP1	6,227	825	975	15.66%	13	1.33%
SRF	2,547	572	227	8.91%	12	5.29%
TCF12	9,575	782	1,140	11.91%	66	5.79%
USF1	4,582	584	478	10.45%	39	8.35%
YY1	14,209	1,105	1,981	13.95%	44	2.27%
ZBTB33	924	715	97	10.50%	2	2.06%
Total	**157,586**	**799**	**21,013**	**13.34%**	**1,094**	**5.22%**

Supplementary Table 3: Reference biases in ChIP-seq alignment

		Percent of unique reads aligning to the reference allele		
Factor	# of het. SNPs with >= 7x coverage	Mean	p[1]	adjusted p[2]
ATF3	368	49.3%	0.35	1.00
BATF	4,936	49.7%	0.57	1.00
BCL11A	2,207	50.6%	0.23	1.00
BCL3	3,054	50.8%	0.02	0.46
BCLAF1	8,466	49.9%	0.56	1.00
EBF1	6,832	50.2%	0.33	1.00
EGR1	1,203	50.7%	0.20	1.00
ELF1	4,358	50.6%	0.09	1.00
EP300	1,062	49.5%	0.55	1.00
ETS1	1,302	50.5%	0.24	1.00
GABPA	1,821	51.0%	0.10	1.00
IRF4	2,902	49.6%	0.34	1.00
LEF1	313	51.3%	0.36	1.00
NRSF	1,523	50.7%	0.19	1.00
POU2F2	12,857	50.0%	0.96	1.00
PAX5	5,734	50.2%	0.51	1.00
PBX3	1,795	50.5%	0.28	1.00
Pol2	19,583	50.2%	0.25	1.00
SIX5	700	50.3%	0.78	1.00
SP1	4,182	50.0%	0.94	1.00
SPI1	6,423	49.9%	0.60	1.00
SRF	961	50.5%	0.52	1.00
TCF12	3,109	50.9%	0.01	0.33
USF1	2,770	50.7%	0.14	1.00
YY1	6,040	50.0%	0.93	1.00
ZBTB33	366	50.4%	0.54	1.00
Total	**104,867**	**50.1**		

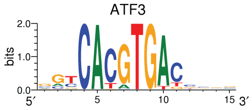
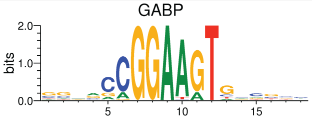
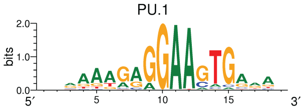
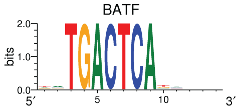
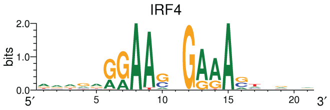
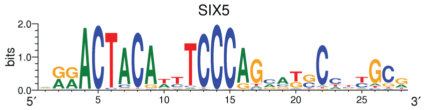
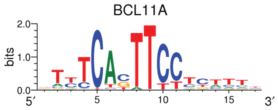
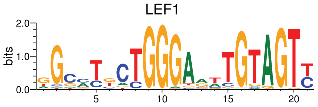
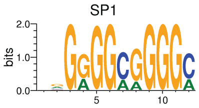
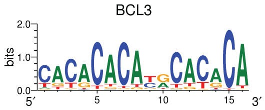
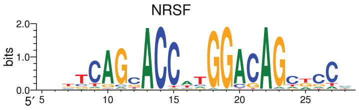
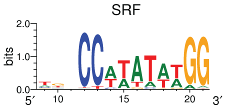
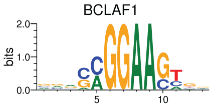
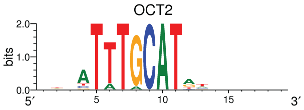
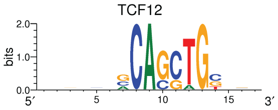
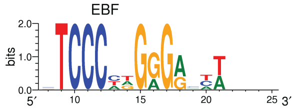
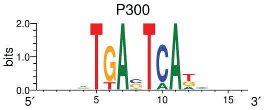
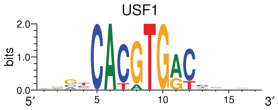
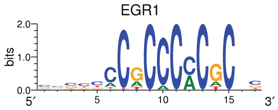
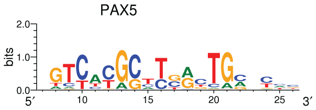
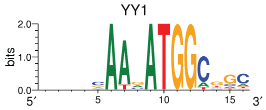
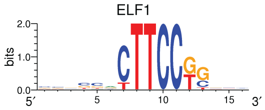
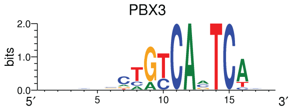
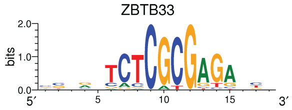
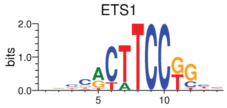
[1] One sample Wilcoxon test against null hypothesis that median = 50%. Only reads mapping to heterozygous positions are considered.

[2] Adjusted for multiple hypotheses by the method of Holm (1979)

Supplementary Table 4: Distribution of TF and co-factor co-occupancy at heterozygous variants in GM12878 cells.

# of TFs with differential allelic occupancy binding at the same locus	# of loci	# of TF:DNA interactions	Fraction of all TF:DNA interactions
-1	774	774	70%
2	91	182	17%
3	17	51	5%
4	9	36	3%
5	5	25	2%
6	2	12	1%
7	1	7	<1%
11	1	11	<1%
Total	**900**	**1,098**	**100%**

Supplementary Table 5: DNA binding motifs identified for all sequence-specific transcription factors in the study.

TF	Motif	TF	Motif	TF	Motif
ATF3		GABP		PU1	
BATF		IRF4		SIX5	
BCL11A		LEF1		SP1	
BCL3		NRSF		SRF	
BCLAF1		OCT2		TCF12	
EBF		P300		USF1	
EGR1		PAX5		YY1	
ELF1		PBX3		ZBTB33	
ETS1					

Supplementary Table 6: Sequencing statistics for RNA-seq and RNA Pol2 ChIP-seq experiments

RNA-seq	Total Paired-end 75bp Reads (M)	Reads Aligned to Refseq (M)	Percent Aligned to Refseq
Replicate 1	44.0	14.7	33%
Replicate 2	24.7	10.4	42%

RNA Pol2 ChIP-seq	Total Single-end 36bp Reads (M)	Total Aligned Reads (M)	Fraction Aligned Reads
Replicate 1	43.8	32.7	75%
Replicate 2	45.2	29.0	64%

RNA Pol2 ChIP-seq	Total Paired-end 100bp Reads (M)	Total Aligned Reads (M)	Fraction Aligned Reads
Replicate 2, Forward Read	79.6	64.0	80%
Replicate 2, Reverse Read	79.6	63.8	80%

Supplementary Table 7: List of genes with discordant allelic expression between clonal isolates of the GM12878 cell line.

id	gene	xi_mat_rep1	xi_mat_rep2	xi_pat_rep1	xi_pat_rep2
NM_000104	CYP1B1	0.77	0.82	0.20	0.25
NM_000575	IL1A	0.47	0.74	1.00	0.97
NM_001025197	CHI3L2	0.56	0.55	0.14	0.24
NM_001122898	CD99	0.46	0.49	0.65	0.69
NM_001134418	LEPREL1	0.00	0.02	0.79	0.97
NM_001145088	WDR67	0.81	0.88	0.17	0.42
NM_001159280	ADAL	0.09	0.20	0.32	0.54
NM_001979	EPHX2	0.00	0.00	0.52	0.57
NM_002145	HOXB2	0.00	0.04	0.73	0.68
NM_002460	IRF4	0.78	0.73	0.44	0.59
NM_003070	SMARCA2	0.02	0.01	0.50	0.59
NM_004642	CDK2AP1	0.11	0.21	0.39	0.51
NM_004973	JARID2	0.18	0.17	0.58	0.45
NM_005832	KCNMB2	0.00	0.00	0.20	0.66
NM_005860	FSTL3	0.14	0.20	0.34	0.60
NM_014971	EFR3B	0.44	0.53	0.10	0.06
NM_017444	CHRAC1	0.76	0.79	0.48	0.59
NM_017852	NLRP2	1.00	0.40	0.00	0.00
NM_018026	PACS1	0.66	0.69	0.49	0.40
NM_022488	ATG3	0.00	0.02	0.61	0.54
NM_030915	LBH	1.00	1.00	0.65	0.45
NM_144594	GTSF1	0.02	0.01	0.97	1.00
NR_026892	LOC84740	0.01	0.05	0.20	0.97

Supplementary Table 8: Evidence that lower expression of genes with differential allelic expression is robust to thresholds and significant after controlling for differences in overall intensity of RNA Pol2 ChIP-seq signal.

Sampling criteria			Rep1			Rep2			
				Median expression (RPKM)			Median expression (RPKM)		
r	N	w	p	EAO	DAO	p	EAO	DAO	p_{cov}
0.15	25	6.5	0.05	2.00	0.65	0.08	2.19	0.62	0.10
0.15	100	67	0.05	3.55	1.01	0.03	3.91	1.23	0.23
0.2	30	7.5	0.02	2.34	0.96	0.01	2.35	0.62	0.90
0.2	50	14	0.13	2.53	0.98	0.07	2.78	0.86	0.39
0.2	80	40	0.01	3.36	0.42	0.01	3.66	0.44	0.87
0.2	120	60	0.01	3.99	0.76	0.01	4.30	0.64	0.27
0.25	30	6	0.05	2.30	1.39	0.02	2.38	1.29	0.28
0.25	50	14	0.01	2.43	0.54	0.01	2.71	0.84	0.17
0.25	80	40	0.02	3.30	0.98	0.04	3.61	0.88	0.72
0.25	120	60	0.07	3.97	1.33	0.05	4.29	1.28	0.06

Sampling criteria:

r: Threshold on differential allelic RNA Pol2 occupancy. E.g. r = 0.15 defines differential allelic occupancy as one allele having less than 15% of total Pol2 occupancy, and equal allelic Pol2 occupancy as no allele having less than (50 – 15 = 35%) Pol2 occupancy.

N, w: Read depth threshold. All genes must have allelic Pol2 coverage of N +/- w reads.

p: probability that genes with differential allelic expression have equal overall expression as genes with equal allelic expression.

P_{cov}: probability that overall RNA Pol2 occupancy at heterozygous positions is equal between differential and equally occupied genes in the chosen sets.

EAO: genes with equal allelic RNA Pol2 occupancy

DAO: genes with differential allelic RNA Pol2 occupancy

Supplementary Table 9: Overview statistics of permutation tests for TF occupancy in GWAS regions.

	Total Number of Binding Sites	Unique Tagged GWAS Variants	Unique Linked Variants	Unique Binding Sites with Linked Variant	Number of Instances of TF Binding at a GWAS-Linked Variant
GWAS Catalog:					
Not Significant Allele Biased Occupancy	19,839	123	155	274 (1.4%)	438
Significant Allele Biased Occupancy	1,115	10	14	12 (1.1%)	21
1,000 Minor allele frequency (MAF)-matched variant sets[1]:					
Not Significant Allele Biased Occupancy			66.8 ± 8.2	149 ± 25	
Significant Allele Biased Occupancy			7.7 ± 2.7	9.7 ± 4.3	
1,000 variant sets with matched distance to nearest TSS[2]:					
Not Significant Allele Biased Occupancy			86.8 ± 9.1	195 ± 28	
Significant Allele Biased Occupancy			10 ± 3.1	13.2 ± 5.2	
150 TSS and MAF-matched variants sets[3]:					
Not Significant Allele Biased Occupancy			88.8 ± 9.0	200 ± 29	
Significant Allele Biased Occupancy			9.8 ± 3.8	12.7 ± 5.2	
Notes:					
1: Minor allele frequency matching required a less than 5% absolute difference in minor allele frequence between GWAS snps and permutations					
2: Nearest TSS matching required a less than 1 kb absolute difference in the distance to the nearest TSS between GWAS snps and permutations					
3: Combined matchining required a less than 2 kb difference in distance to the nearest TSS, and a less than 10% difference in MAF.					

Supplementary Table 10: Disease-associated variants bound by TFs with differential allelic occupancy.

GWAS Variant	Linked Variant	Factor	Maternal Occupancy	Disease	PMID	Etiology
rs9271100	rs9271170	YY1	60%	Systemic lupus erythematosus	19838193	Autoimmune
rs10484561	rs17533167	SP1	62%	Follicular lymphoma	20639881	Various
rs9272346	rs1063355	EBF	67%	Type 1 diabetes	18978792	Autoimmune
	rs1063355	TCF12	31%		17554300	
rs6806528	rs6776027	BATF	25%	Celiac disease	20190752	Autoimmune
	rs6784841	BATF	25%			
rs9273349	rs1063355	EBF	67%	Asthma	20860503	Various, incl. autoimmune
	rs1063355	TCF12	31%			
rs12928822	rs12162021	PAX5	68%	Celiac disease	20190752	Autoimmune
	rs12162021	PAX5	74%			
	rs12162021	TCF12	66%			
	rs12918017	EBF	68%			
rs9976767	rs9976479	EBF	64%	Type 1 diabetes	18840781	Autoimmune
rs1557351	rs1557351	BATF	7%	Multiple sclerosis (age of onset)	19010793	
	rs1557351	PU.1	37%			
	rs12457489	BATF	7%			
	rs12457489	PU.1	37%			
	rs1557352	PU.1	37%			
rs7993214	rs9603612	EBF	79%	Psoriasis	18369459	Autoimmune
rs674313	rs2097432	SP1	62%	Chronic lymphocytic leukemia	21131588	Various
	rs3129763	SP1	62%			

Supplementary Table 11:

TF or protein	SNP ID	Chrom.	Position	Pat. Reads	Mat. Reads	Fraction of Pat. Occupcy	p-val	FDR	CNV Status
BCLAF1	**NA12878.350874**	**chr17**	**41,625,958**	**14**	**10**	**0.58**	**0.54**	**1.00**	**amp**
POU2F2	**NA12878.350874**	**chr17**	**41,625,958**	**16**	**8**	**0.67**	**0.15**	**0.88**	**amp**
Pol2	**NA12878.350874**	**chr17**	**41,625,958**	**42**	**6**	**0.88**	**0.00**	**0.00**	**amp**
USF1	**NA12878.350874**	**chr17**	**41,625,958**	**60**	**24**	**0.71**	**0.00**	**0.00**	**amp**
YY1	**NA12878.350874**	**chr17**	**41,625,958**	**73**	**27**	**0.73**	**0.00**	**0.00**	**amp**
PAX5	**rs2240759**	**chr17**	**41,603,192**	**17**	**7**	**0.71**	**0.06**	**0.34**	**amp**
POU2F2	NA12878.321425	chr14	105,397,056	16	5	0.76	0.03	0.47	het.del
Pol2	NA12878.321425	chr14	105,397,056	88	30	0.75	0.00	0.00	het.del
SPI1	**NA12878.391263**	**chr22**	**21,357,646**	**23**	**1**	**0.96**	**0.00**	**0.00**	**het.del**
Pol2	rs10136437	chr14	105,373,980	10	10	0.50	1.00	1.00	het.del
Pol2	rs10139433	chr14	105,374,744	16	8	0.67	0.15	0.52	het.del
SPI1	**rs11090173**	**chr22**	**21,393,645**	**68**	**1**	**0.99**	**0.00**	**0.00**	**het.del**
Pol2	rs12184945	chr14	105,378,795	12	10	0.55	0.83	0.98	het.del
Pol2	rs12885461	chr14	105,373,354	12	6	0.67	0.24	0.59	het.del
Pol2	**rs1467858**	**chr22**	**20,841,719**	**2**	**12**	**0.14**	**0.01**	**0.11**	**het.del**
SPI1	rs2073453	chr22	20,846,998	10	14	0.42	0.54	0.92	het.del
Pol2	rs2075590	chr15	72,496,619	6	10	0.38	0.45	0.83	het.del
Pol2	rs2256346	chr14	105,391,137	104	102	0.50	0.94	1.00	het.del
Pol2	rs2753488	chr14	105,376,305	6	10	0.38	0.45	0.83	het.del
POU2F2	**rs5757106**	**chr22**	**20,841,467**	**7**	**13**	**0.35**	**0.26**	**0.97**	**het.del**
POU2F2	rs5757107	chr22	20,841,496	13	17	0.43	0.58	1.00	het.del
Pol2	rs6003229	chr22	21,370,389	64	0	1.00	0.00	0.00	het.del
Pol2	rs7153502	chr14	105,374,068	14	12	0.54	0.85	0.98	het.del
Pol2	rs7153935	chr14	105,374,072	12	12	0.50	1.00	1.00	het.del
EBF1	rs765267	chr12	107,550,030	9	13	0.41	0.52	0.88	het.del
Pol2	rs765267	chr12	107,550,030	10	10	0.50	1.00	1.00	het.del
Pol2	rs7925131	chr11	810,268	12	14	0.46	0.85	0.98	het.del

Table of variants that overlap regions of copy number variation as determined by Illumina Human1M-Duo DNA Analysis BeadChips. Results from the arrays and the methods used are available from the UCSC Genome Browser for the hg18 version of the human genome, and listed under “Common Cell CNV”. Variants overlapping CNVs are reported in the table. Entries in bold indicate variants where differential allelic occupancy may arise from copy number variation. For most heterozygous deletions, both alleles were observed indicating that the endpoints of the deletion did not include the variant in question.

Supplementary Table 12: Experiment identifiers of ChIP-seq data.

All data can either be downloaded from the UCSC Genome Browser

http://genome-test.cse.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeHaibTfbs

Some data is currently in submission, and therefore all data used in the study is also available at http://mendel.hudsonalpha.org/Tim/Effects_of_seqvar_on_TF_and_exp/

Identifiers in blue indicate the 2 PCR version of the ChIP-seq protocol was used.

Identifiers for Sequence Specific Factors and P300:

GM12878	**Rep1**	**Rep2**	**Rep3**
ATF3	SL1269	SL1508	
BATF	SL839	SL985	
BCL11A	SL650	SL976	
BCL3	SL652	SL1018	
BCLAF1	SL1509	SL2128	
EBF1	SL745	SL988	
EGR1	SL482	SL3579	
ELF1	SL2254	SL3352	
EP300	SL551	SL564	
ETS1	SL1507	SL1655	
GABPA	SL203	SL205	
IRF4	SL838	SL951	
LEF1	SL1597	SL1791	
NRSF	SL202	SL204	SL852
PAX5	SL675	SL735	
PBX3	SL615	SL647	
POU2F2	SL851	SL614	SL648
SIX5	SL1061	SL1200	
SP1	SL746	SL846	
SPI1	SL612	SL963	SL649
SRF	SL292	SL3578	
TCF12	SL673	SL1019	
USF1	SL448	SL483	
YY1	SL1475	SL2129	
ZBTB33	SL814	SL923	

	Rep1	Rep2
Pol2	SL748	SL847

GM12891	**Rep1**	**Rep2**
GABP	SL750	
OCT2	SL918	SL802
PAX5	SL2131	SL1662
PU.1	SL977	SL948
YY1	SL2130	SL2388
Control	SL1782	SL812

GM12892	**Rep1**	**Rep2**
GABP	SL751	
OCT2	SL919	
PAX5	SL2133	SL1664
PU.1	SL947	SL837
YY1	SL2132	SL3584
Control	SL1783	SL818

Supplementary Table 13: Location of raw data for RNA-seq experiments

Note: all files listed below can be found at:
http://hgdownload-test.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCaltechRnaSeq/

	Rep1
GM12891	wgEncodeCaltechRnaSeqGm12891R2x75Il200FastqRd1Rep1.fastq.gz
	wgEncodeCaltechRnaSeqGm12891R2x75Il200FastqRd2Rep1.fastq.gz
	wgEncodeCaltechRnaSeqGm12891R2x75Il200FastqRd1Rep2.fastq.gz
	wgEncodeCaltechRnaSeqGm12891R2x75Il200FastqRd2Rep2.fastq.gz
GM12892	wgEncodeCaltechRnaSeqGm12892R2x75Il200FastqRd1Rep1.fastq.gz
	wgEncodeCaltechRnaSeqGm12892R2x75Il200FastqRd2Rep1.fastq.gz
	wgEncodeCaltechRnaSeqGm12892R2x75Il200FastqRd1Rep2.fastq.gz
	wgEncodeCaltechRnaSeqGm12892R2x75Il200FastqRd2Rep2.fastq.gz
K562	wgEncodeCaltechRnaSeqK562R2x75Il200FastqRd1Rep1.fastq.gz
	wgEncodeCaltechRnaSeqK562R2x75Il200FastqRd2Rep1.fastq.gz
	wgEncodeCaltechRnaSeqK562R2x75Il200FastqRd1Rep2.fastq.gz
	wgEncodeCaltechRnaSeqK562R2x75Il200FastqRd2Rep2.fastq.gz
HeLa	wgEncodeCaltechRnaSeqHelas3R2x75Il200FastqRd1Rep1.fastq.gz
	wgEncodeCaltechRnaSeqHelas3R2x75Il200FastqRd2Rep1.fastq.gz
	wgEncodeCaltechRnaSeqHelas3R2x75Il200FastqRd1Rep2.fastq.gz
	wgEncodeCaltechRnaSeqHelas3R2x75Il200FastqRd2Rep2.fastq.gz
HepG2	wgEncodeCaltechRnaSeqHepg2R2x75Il200FastqRd1Rep1.fastq.gz
	wgEncodeCaltechRnaSeqHepg2R2x75Il200FastqRd2Rep1.fastq.gz
	wgEncodeCaltechRnaSeqHepg2R2x75Il200FastqRd1Rep2.fastq.gz
	wgEncodeCaltechRnaSeqHepg2R2x75Il200FastqRd2Rep2.fastq.gz
HUVEC	wgEncodeCaltechRnaSeqHuvecR2x75Il200FastqRd1Rep1.fastq.gz
	wgEncodeCaltechRnaSeqHuvecR2x75Il200FastqRd2Rep1.fastq.gz
	wgEncodeCaltechRnaSeqHuvecR2x75Il200FastqRd1Rep2.fastq.gz
	wgEncodeCaltechRnaSeqHuvecR2x75Il200FastqRd2Rep2.fastq.gz
NHEK	wgEncodeCaltechRnaSeqNhekR2x75Il200FastqRd1Rep1.fastq.gz
	wgEncodeCaltechRnaSeqNhekR2x75Il200FastqRd2Rep1.fastq.gz
	wgEncodeCaltechRnaSeqNhekR2x75Il200FastqRd1Rep2.fastq.gz
	wgEncodeCaltechRnaSeqNhekR2x75Il200FastqRd2Rep2.fastq.gz
hESC	wgEncodeCaltechRnaSeqH1hescR2x75Il200FastqRd1Rep1.fastq.gz
	wgEncodeCaltechRnaSeqH1hescR2x75Il200FastqRd2Rep1.fastq.gz
	wgEncodeCaltechRnaSeqH1hescR2x75Il200FastqRd1Rep2.fastq.gz
	wgEncodeCaltechRnaSeqH1hescR2x75Il200FastqRd2Rep2.fastq.gz

References

Clauset A, Shalizi CR, Newman MEJ. 2009. Power-law distributions in empirical data. *arXiv:07061062v2*.

Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* 101: 6062-6067.