# Supporting Online Material for

## Heritable Individual-Specific and Allele-Specific Chromatin Signatures in Humans

Ryan McDaniell, Bum-Kyu Lee, Lingyun Song, Zheng Liu, Alan P. Boyle, Michael R. Erdos, Laura J. Scott, Mario A. Morken, Katerina S. Kucera, Anna Battenhouse, Damian Keefe, Francis S. Collins, Huntington F. Willard, Jason D. Lieb, Terrence S. Furey, Gregory E. Crawford,* Vishwanath R. Iyer,* Ewan Birney*

*To whom correspondence should be addressed. E-mail: greg.crawford@duke.edu (G.E.C.); vishy@mail.utexas.edu (V.R.I.); birney@ebi.ac.uk (E.B.)

Published 18 March 2010 on *Science* Express
DOI: 10.1126/science.1184655

**This PDF file includes:**

## Materials and Methods

**Cell Line Growth.** Lymphoblastoid cell lines used in this study came from the CEU (CEPH - Utah residents with ancestry from northern and western Europe) and YRI (Yoruba in Ibadan, Nigeria) sample populations that have been genotyped and resequenced by the HapMap and 1000 Genomes Project respectively. All cells were obtained from Coriell (Camden, NJ) and were cultured using standard growth procedures (http://ccr.coriell.org). Cells were grown in RPMI 1640 (2 mM L-glutamine) in 15% fetal bovine serum, and were monitored daily to maintain a cell density between 200k-500k viable cells per ml. Cells were split every other day into fresh media. Two biological replicates were grown on separate days for each cell line. At each harvest, cells were confirmed by trypan blue staining to be >99% viable.

**DNase I HS Assay.** DNase I hypersensitive (HS) sites were identified as previously described (S*1*). After gentle lysis by NP40, nuclei were digested with optimized concentrations of DNase I. DNase I-digested ends were blunted and ligated to biotinylated linkers containing an Mme I restriction site. Material was digested with MmeI, and linkers plus 20 bases of DNA flanking the DNase I-digested end were purified on a streptavidin column. MmeI-digested ends were ligated to a second set of linkers, material was amplified by PCR, and was sequenced by the Illumina (San Diego, CA) GA2 sequencer.

**CTCF Assay.** Chromatin immunoprecipitation (ChIP) for CTCF was carried out using previously described methods (S*2*). Sonicated chromatin containing an average of 500 bp DNA fragments was used to immunoprecipitate CTCF using an anti-CTCF antibody (07-729) from Millipore (Danvers, MA). After reversal of crosslinks, purified ChIP DNA was used to generate ChIP-seq libraries according to Illumina's recommended protocols. Purified ChIP-seq libraries were sequenced using the Illumina GA2 Sequencer at Duke University.

**Sequencing Statistics.** All the samples were sequenced using Illumina single-fragment sequencing protocols. Altogether we generated over 600 million sequences for this analysis. The number of sequences per cell line for DNase-seq and ChIP-seq are listed below in supplementary table S1. Individual useable sequences (total of 618,039,754 from DNase I HS and CTCF) have been filtered to remove linker-only sequences, and align less than 5 times to the human genome (hg18).

**Processing and Normalization.** This raw data was processed using a standard pipeline of alignment to the reference genome (May 2004, hg18/NCBI Build 36.1) using Maq (S*3*) followed by determination of enriched regions for each assay using the F-seq package (S*4*). Replicates showed between 0.88-0.96 Pearson correlation to each other for DNase I HS and CTCF respectively, and inter-replicate and inter-lymphoblastoid correlation values were far higher than correlations to other cell lines (K562, HepG2, see fig. S1 below). For female individuals, chromosome Y was omitted from the reference genome. For male individuals, the pseudoautosomal regions of chromosome Y were excluded as these were already represented on chromosome X. F-seq was used at a low threshold (4 standard deviations above the mean, -sd=4) to generate potential sites of enrichment across the genome from these mappings. For any particular analysis these were then clustered between replicates and individuals by single linkage clustering, in effect taking the union of positive positions on the genome. This process leads to large regions (> 5 Kb) in a small minority of regions, often with understandable and highly divergent biology, eg, the Hox cluster, and so these regions were discarded. When multiple F-seq signals were merged from one replicate due to this clustering the maximum peak was taken across linked sites.

The distribution of signal from both the chip-seq and DNase I HS is very complex both in the “background” low signal and the “foreground” high signal, and most definitely neither Gaussian nor a mixture of Gaussians. One expects the signal shape to scale approximately linearly with sequencing depth, but also by the efficiency of the precise assay and aspects such as sequencing accuracy, which will effect the amount of mappable reads. In the case of the CTCF, these distributions show different mid-signal behavior. For individual site definition, 3 replicates in the CTCF datasets were not used as their correlation to the overall mean suggested a different growth behavior and a small number of sites with very high variance between the cell lines were excluded (these were enriched in segmental duplication regions and CNVs). To normalize between cell lines between CTCF we used quantile-normalization as implemented in the limma BioConductor package which maps the ranked distribution of sites to a normal distribution. In case of DNase, the general shape of the distributions were similar, and so to normalize between replicates we set a standard quantile, in this case the 75% quantile, as 1 for each replicate and then scaled linearly the other peaks. This procedure therefore uses inherent distribution of the signal to provide this linear transformation, on the assumption that overall the underlying distribution of signals is comparable between replicates, and is less aggressive than quantile normalization. We then took the square root of this scaled number as to a first approximation the density of tags is a count-based statistic, and the square root of a Poisson style distribution is approximately normally distributed. This transformation provided better visualization of the data range (by compressing the higher signal and expanding the lower signal) but still remained non Gaussian. For both assays, a particular issue is the presence of a large number of “zero scored” regions in the low signal portion of both cases.

Because of the non-Gaussian nature of the data, we preferred non-parametric tests where ever possible (Wilcoxon-rank test for differences in levels; Spearman’s Rho for correlation). Only occasionally have we used parametric based tests, such as the use of Pearson’s correlation coefficient for the inter-sample correlation, and in this case we employed it as a metric for which we also estimated an empirical null by permutation. In our hands, parametric tests with a strong assumption of Gaussian behavior, such as the F-statistic did not behave well, even with the quantile normalized sets. Processing was done in Perl and R, and R data frames

of intermediate data are available on request. The number of constant, individual-specific and variable sites in the different cell line combinations are tabulated in table S2.

**Gene expression analysis.** RNA was obtained from lymphoblastoid cell lines at the same time they were harvested for DNase and ChIP experiments. Total RNA was isolated from these cells using Trizol extraction followed by cleanup on RNeasy columns (Qiagen, Valencia, CA) that included a DNase step. The RNA was checked for quality using a Nanodrop (Wilmington, DE) and an Agilent (Santa Clara, CA) Bioanalyzer. RNA (1µg) deemed to be of good quality was then processed according to the standard Affymetrix (Santa Clara, CA) Whole transcript Sense Target labeling protocol that included a riboreduction step. The fragmented biotin-labeled cDNA was hybridized over 16 h to Affymetrix Exon 1.0 ST arrays and scanned on an Affymetrix Scanner 3000 7G using AGCC software. Array .CEL files were normalized by RMA and gene level analysis was performed using Expression Console (Affymetrix). Data from replicates were averaged. Each binding site was assigned to its nearest gene. Correlation was calculated in R, and permuted datasets also generated. Binding sites were classified into four distance-based categories: "At TSS" = +/- 2.5 Kb; "Near TSS" = +/- 10 Kb of TSS; "Gene Locale" = +/- 100 Kb of TSS; "Distal to Gene" > 100 Kb of its closest TSS. Figure 3 shows only the split for TSS (< 2.5 Kb) or Near TSS. The full split by class is shown below in Fig. S11.

**Allele-specific sites.**
We assessed the allele-specific bias for each heterozygous SNP with more than 15 reads across all individuals and assayed 7,366 heterozygous DNase I HS sites and 9,192 heterozygous CTCF sites. Although there are more DNase I HS than CTCF sites in the genome, we obtained fewer reads within each DNase I HS site. A binomial P-value in conjunction with a false discovery rate (FDR) multiple testing correction threshold of 0.01 designated 7% of DNase I HS sites and 11% of CTCF sites as showing significant allele bias (Fig. 1C).
The procedure for identifying allele-specific sites in DNase I HS and CTCF sequencing data is described below under fig. S5. In addition, we screened out SNPs in repeat rich regions of the genome which are both potentially harder to call as heterozygous in the genomic sequence and in which complex interactions between repeat loci (in particular, loci not present on the reference, but present in these individuals) can occur. Using pooled data across all cell lines, with a null hypothesis of a 50:50 split and FDR correction (Benjamini & Hochberg, as implemented in the p.adjust function in R) on the raw binomial P values, 540 of the heterozygous sites in the DNaseI HS assay (7%) and 1,034 (11%) of the CTCF sites are allele specific. Each heterozygous site can be present or not in each cell line, and due the filtering for heterozygous positions seen in more than 1 sample to remove reference bias, each site is present in at least 2 lines. Table S3 shows the combinations of cell lines with these heterozygous sites, and how many of them were found to be significant at the 0.01 FDR level.

**Calculation of the difference in allele specificity across the CTCF PWM.** All SNPs showing a bias with $P <= 0.01$ were considered allele-specific (AS) for the purposes of Figure 4. In order to normalize the two classes of SNPs, the number of AS SNPs at each position was multiplied by the quotient of the total number of non-AS SNPs and the total number AS SNPs across all positions. The percentage of non-AS SNPs was then subtracted from the percentage of AS SNPs at each position.

**Verification of allele-specific binding by MALDI/TOF mass-spectrometry.** Experiments were carried out similar to what was previously described (*5*). SNPs were genotyped using iPlex Gold SBE (Sequenom, San Diego, CA). GM19240 CTCF ChIP library DNA was aliquoted at 0.16 ng/assay and GM19240 genomic control DNA was aliquoted at 6.6 ng/assay in 384-well format. For each SNP assay, 16 replicates of CTCF ChIP DNA and genomic control DNA were tested. Primer sequences for the SNP assays are shown below in supplementary table S4. MALDI-TOF analysis for each assay was performed sampling each matrix pad by rastering to nine independent positions on the pad accumulating ten laser shots per position. The genotypes were assigned using SpectroCaller software and the peak fitting and area under each allele peak was calculated by SpectroAcquire software (Sequenom, Inc.).
For each SNP we used the peak areas A and B of the lower and higher mass alleles, respectively, to estimate the proportion of the low mass allele =A/(A+B). As an alternative we adjusted the estimate of p to take into account the unequal proportions of the two alleles in heterozygous DNA samples. We calculated the sample mean of the ratios A/B from the heterozygous DNA replicates and estimated the adjusted proportion of the lower mass allele A as adj =A/(A+ *B). We compared the either the unadjusted or the adjusted proportion of the low mass allele between the CTCF and heterozygous DNA samples using a two-sample two-sided t-test with allowance for unequal variances (table S5 below). The t-test results were almost identical for the adjusted or unadjusted proportions of allele A. For ease of interpretation, all of the data presented in fig. S8B and table S5 below is based on the adjusted allele frequency.

**Supporting Online Figures**

**Supplemental Figure 1.**

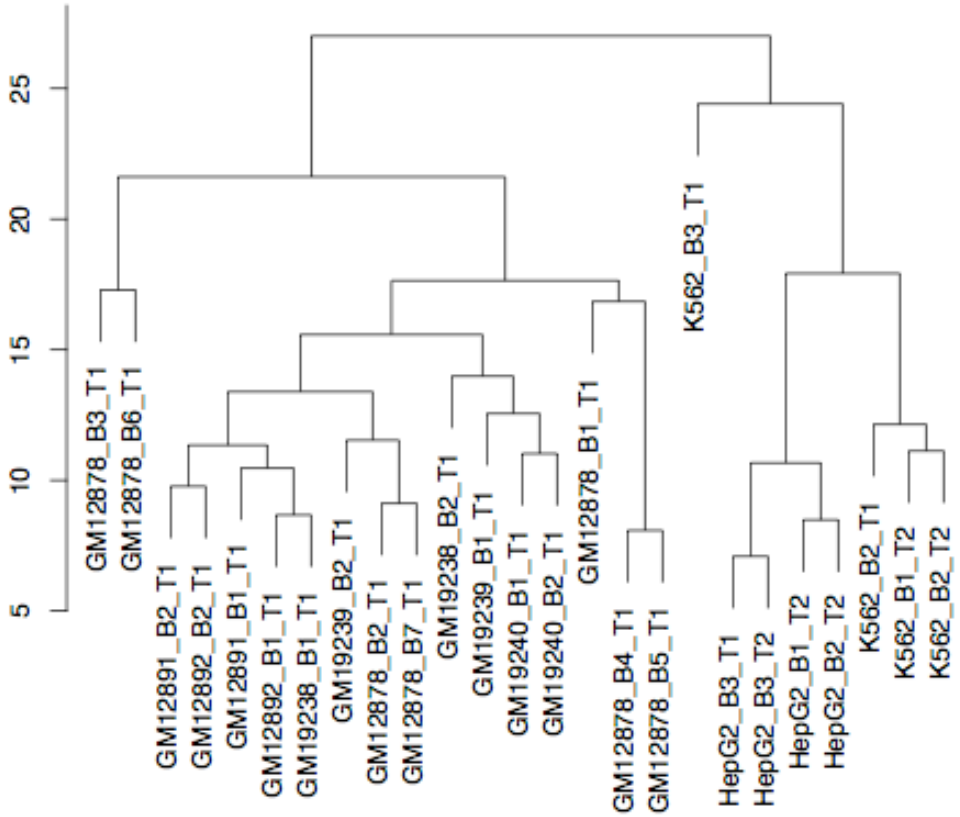


Figure S1. Comparison of lymphoblastoid replicates to other cell lines. The DNase I HS data was used with matched samples from HepG2 and K562. All 6 lymphoblastoid samples and these cell lines were jointly analyzed by performing single-linkage clustering of sites on the genome and discarding resulting regions > 5 Kb. Signals were summed across linked sites in one individual if needed. Distance based clustering of the replicates, either with or without normalization showed most of the lymphoblastoid lines falling in a cluster together, with HepG2 and K562 replicates often distinct. With some clustering parameters, K562 replicates partitioned between the HepG2 and lymphoblastoid samples, which is not surprising, given the common blood cell ancestry of these cell lines. No lymphoblastoid line clustered with HepG2 replicates.

**Supplemental Figure 2.**

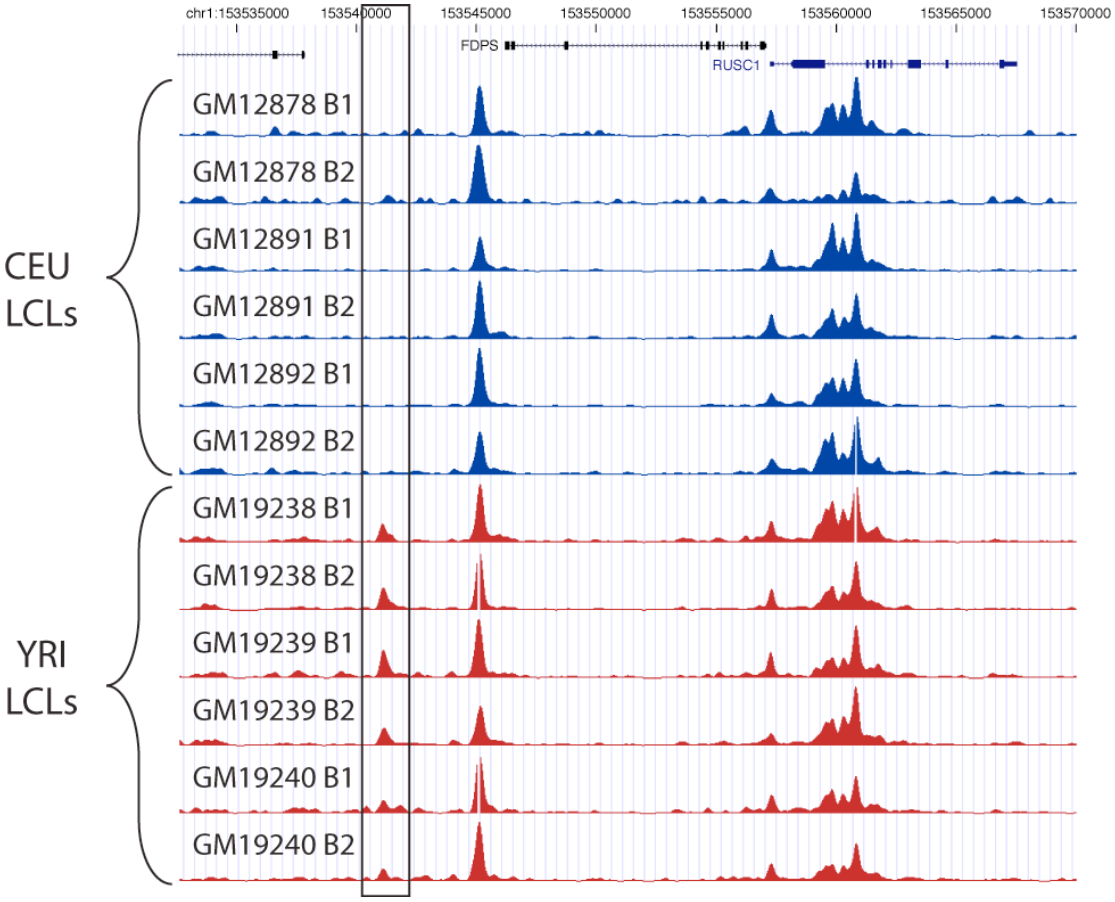


Figure S2. Example of individual-specific DNase I HS sites

**Supplemental Figure 3**.

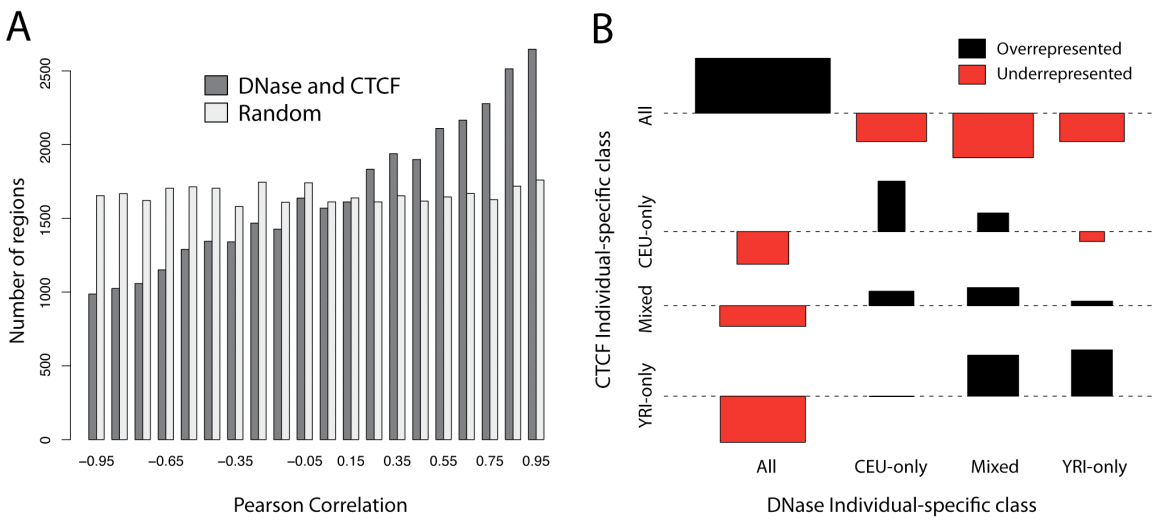


Figure S3. (A) Histogram of Pearson correlation values of DNase I HS site levels to CTCF site levels between individuals across sites. There was a significant over-representation of positive correlations and under-representation of negative correlations in the data compared to the flat distribution of randomly permuted data ($P < 10^{-15}$, Wilcoxon rank-sum test). (B) Association plot where the area of the rectangles is proportional to the number of sites and their height is proportional to the strength of the over/under representation. Individual-specific sites identified by a given assay in one population class were more likely to be overrepresented among sites identified by the other assay in the same population class (as indicated by higher black bars in the diagonal from top left to bottom right) rather than different populations (lower black and red bars). These differences were significant ($P < 10^{-15}$, chi-squared test).

**Supplemental Figure 4**.

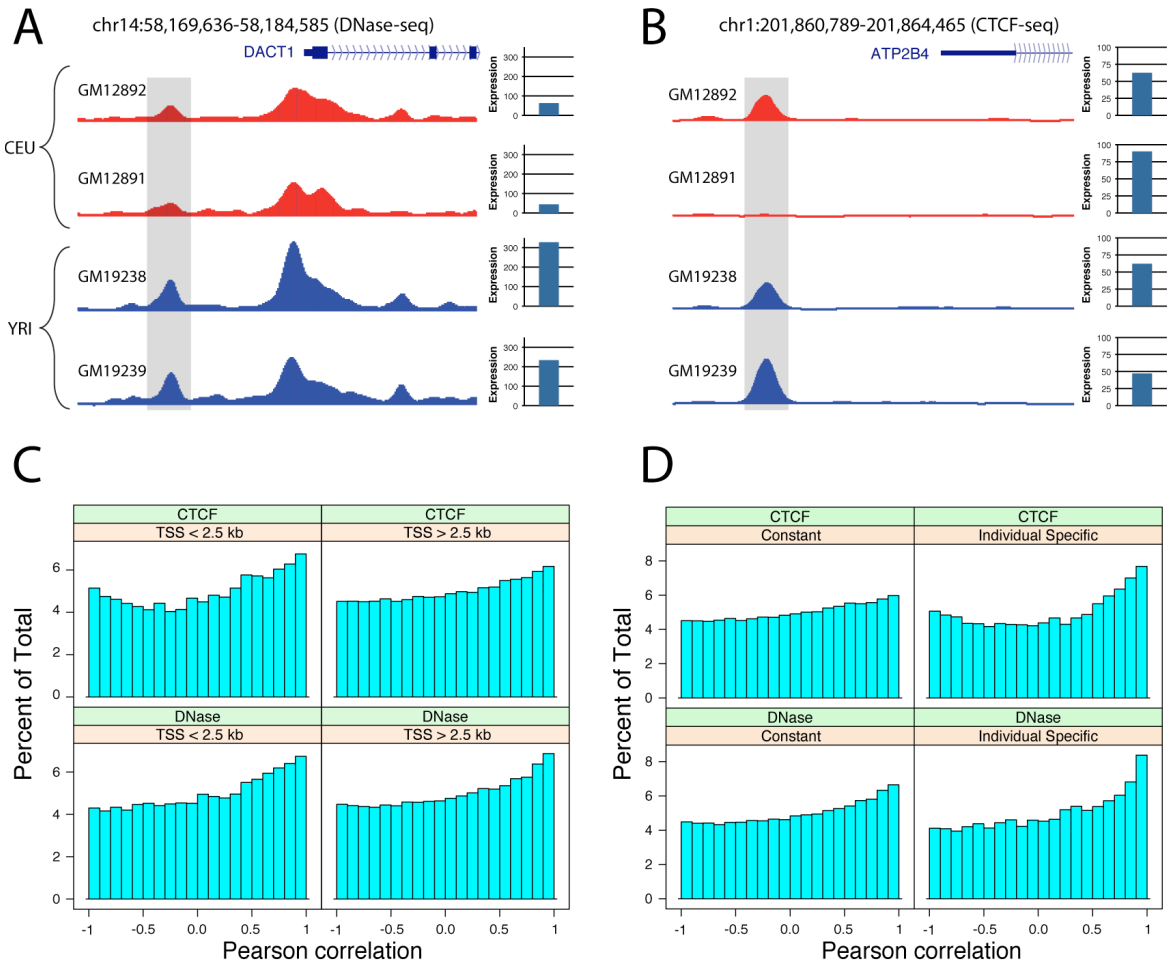


Figure S4. Chromatin sites correlate with gene expression. (A) Example of DNaseI HS site near the TSS (highlighted) whose strength was positively correlated with gene expression levels. Linear expression values for the DACT1 gene, averaged from 2 biological replicates and across all exons, are displayed on the graph on the right. (B) Example of CTCF site whose strength was anti-correlated with gene expression levels. (C) Distribution of correlation values between DNase or CTCF, and gene expression, across individuals. Sites were separated into those within 2.5 Kb of the nearest TSS, or more than 2.5 Kb away but less than 10 Kb away from the nearest TSS. (D) Distribution of correlation values between DNase or CTCF, and gene expression, across individuals. Sites were separated into constant or individual-specific as described in the text and Fig. 1.

**Supplemental Figure 5**.

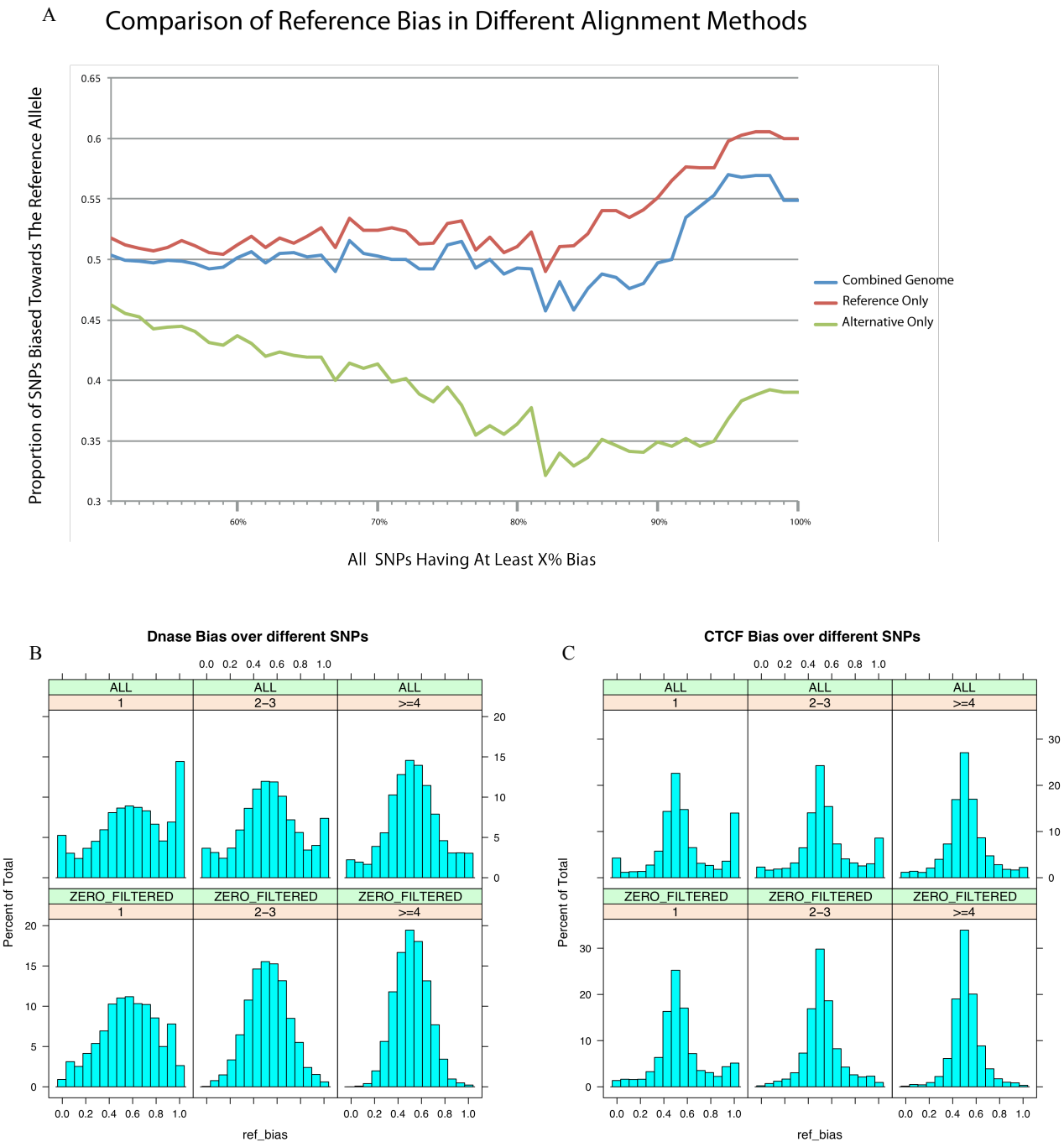


Figure S5. Bias towards reference allele and new mapping strategy. Alignment of reads to a single reference genome creates an artefactual bias towards the reference being used. While this paper was in review, a recent publication (J. F. Degner et al., Bioinformatics 25, 3207 (2009)) also noted a similar reference allele-bias in RNA seq data. We also found that masking the SNP base to "N" during alignment does not give satisfactory results for overcoming the reference allele bias. In order to get a more accurate number of reads originating from a heterozygous SNP position, we instead mapped all reads to two different genomes which were constructed from SNP calls obtained from the April 2009 1000 Genomes data release.
For each individual, two modified hg18 genomic reference files (named genome1 and genome2) were created. In the 1000 genomes dataset, two bases were given for each SNP position: either different (heterozygous), or the same (homozygous). For heterozygous SNPs, the two bases were randomly assigned to either genome1 or genome2 and replaced the base at the respective

position in hg18. In the cases of homozygous SNPs, both genomes reflected this change. All reads from each experiment were aligned to both of the individual genome1 and genome2 references using Maq. Typically, about 5% of reads changed position in comparison to the other genomic alignment. A combined alignment is generated whereby the reads that did not change position are counted only once, and the reads that changed location from one genome to the next are assigned to both positions.
(A) shows the proportion of SNPs biased towards the reference allele vs. the magnitude of the bias, using CTCF reads from GM19240 as an example. The naive alignments, where the reads are only aligned to the reference genome (red) or a genome where the reference base has been replaced by the other alternative allele (olive) diverge from an even 50/50 proportion. This divergence in the expected proportion is greatest for the most interesting SNPs, those showing a high degree of allele specificity. The combined genome1 and genome2 alignments, in (blue), show a greater degree of adherence to the expected proportion. Despite this improvement of mapping, there was still an appreciable bias towards reference alleles. As the reference is, in general, an arbitrary choice of the two potential alleles in any situation we looked for symmetrical behavior of reference bias to indicate that we had removed an overall reference effect. (B) and (C) show the distribution of reference bias for the different SNPs with different filters applied. One set of filters (Zero Filtered) was the basis of observing each allele at least once in a cell line in at least one assay. The other filter was to classify the SNP as occurring in >1 assay (the 2-3 or >4 panels above), in effect a filter on SNP heterozygous call quality. The combination of these two filters (the central lower panel in each case) shows symmetrical or near symmetrical reference bias for each assay, and this set of SNPs was the result set used for allele specific analysis.

**Supplemental Figure 6**.

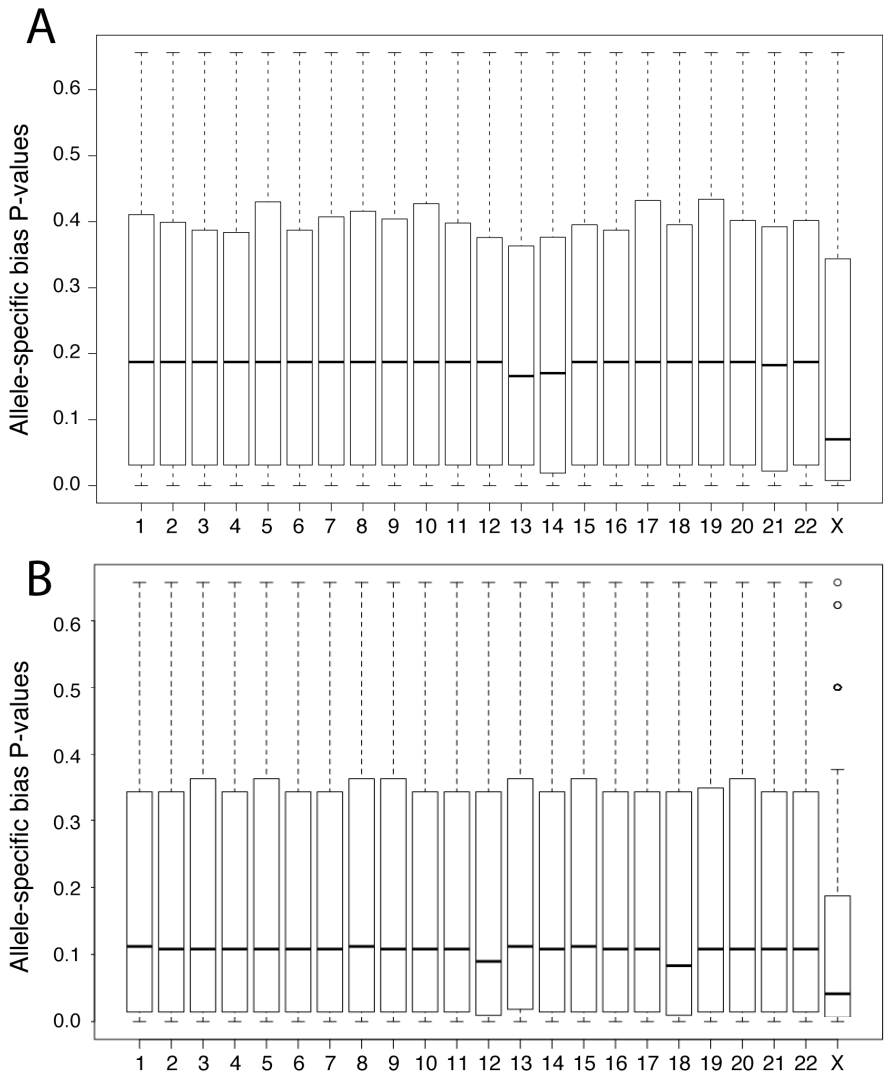


Figure S6. Increased allele-specific bias on chromosome X. Distribution of (A) CTCF and (B) DNaseI HS allele-specific bias P-values plotted by chromosome in 3 female lines GM12878, GM12892 and GM19240.

**Supplemental Figure 7**.

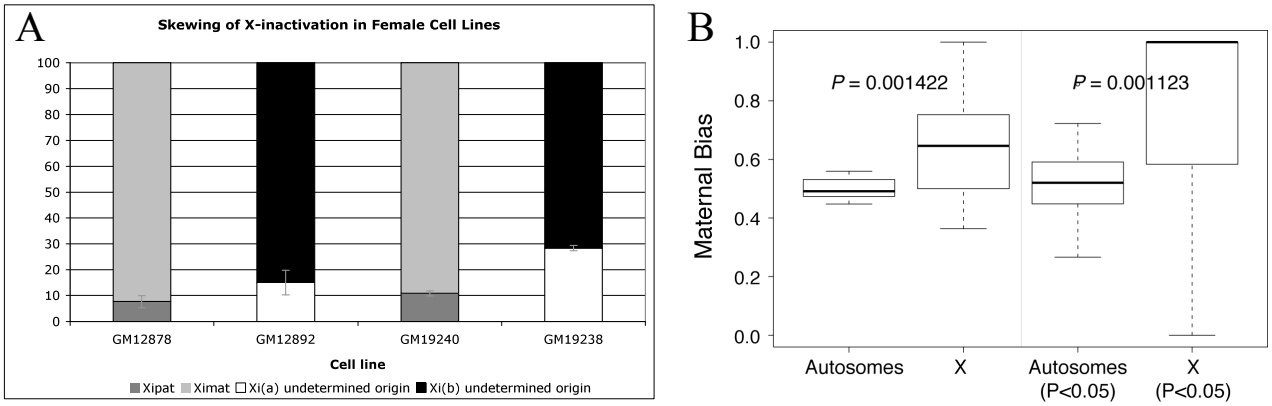


Figure S7. X-inactivation is skewed towards paternal X chromosome. X-chromosome inactivation in phenotypically unaffected females is generally accepted to be random, such that each X chromosome, Xpat and Xmat, is inactivated in approximately equal numbers of cells (*S5*). This is also expected to be reflected in lymphoblastoid cultures; however, skewing of inactivation in transformed cell lines may occur (*S6*). (A) We determined the extent and direction of skewing in the female cell lines by a PCR-based expression assay using heterozygous SNPs in two types of monoallelically expressed X-linked genes, genes that are subject to inactivation and the XIST gene. XIST is the only gene expressed solely from the inactive X chromosome (Xi) (*S7*), while genes subject to inactivation are expressed solely from the active X chromosome (Xa) (*S8*). The relative expression of each allele of monoallelically expressed genes is thus proportional to the fraction of each Ximat and Xipat in the culture. DNA and RNA from each lymphoblastoid line were isolated from fresh cells and stored at -20 and -80°C respectively. PCR was performed on DNA and cDNA where the amplicon for each tested gene contained a targeted heterozygous SNP. A quantitative Q-SNaPshot (*S8*), was employed at those SNPs. The Q-SNaPshot assay uses a primer extension reaction to incorporate a single fluorescent dideoxy nucleotide at the polymorphic site. The fluorescence output is then detected on an ABI 3100 instrument. The cDNA readout was normalized to the DNA signal with known 1:1 ratio of the two alleles to correct for biases in fluorescence output (*S8*). The direction of skewing toward Xipat or Ximat was determined for the second generation by tracing the parent of origin for each allele. Cell lines from both daughters are heavily skewed towards the paternally inherited inactive X. The X inactivation ratio (Xipat/Ximat) is 92/8% for GM12878 and 89/11% for GM19240. The extent of skewing is similar in the two mothers (GM12892 is 85/15%, and GM19238 is 72/28%). The origin of each Xi (pat/mat) cannot be determined in the first generation due to the lack of parental information. (B) Distribution of allele-specific CTCF binding bias toward the maternal allele in GM12878. Sites were binned along the genome (100 Mb for autosomes and 10 Mb for X), and the proportion of CTCF sites that were biased towards the maternal allele is plotted on the Y axis, for each of the groups shown on the X axis. The pair on the left hand side shows all assayable sites and the pair on the right hand side shows only allele-specific sites with significant bias at a binomial $P$ value < 0.05. There was a significantly increased proportion of maternally biased sites on the X chromosome (Wilcoxon rank-sum test).

**Supplemental Figure 8**.

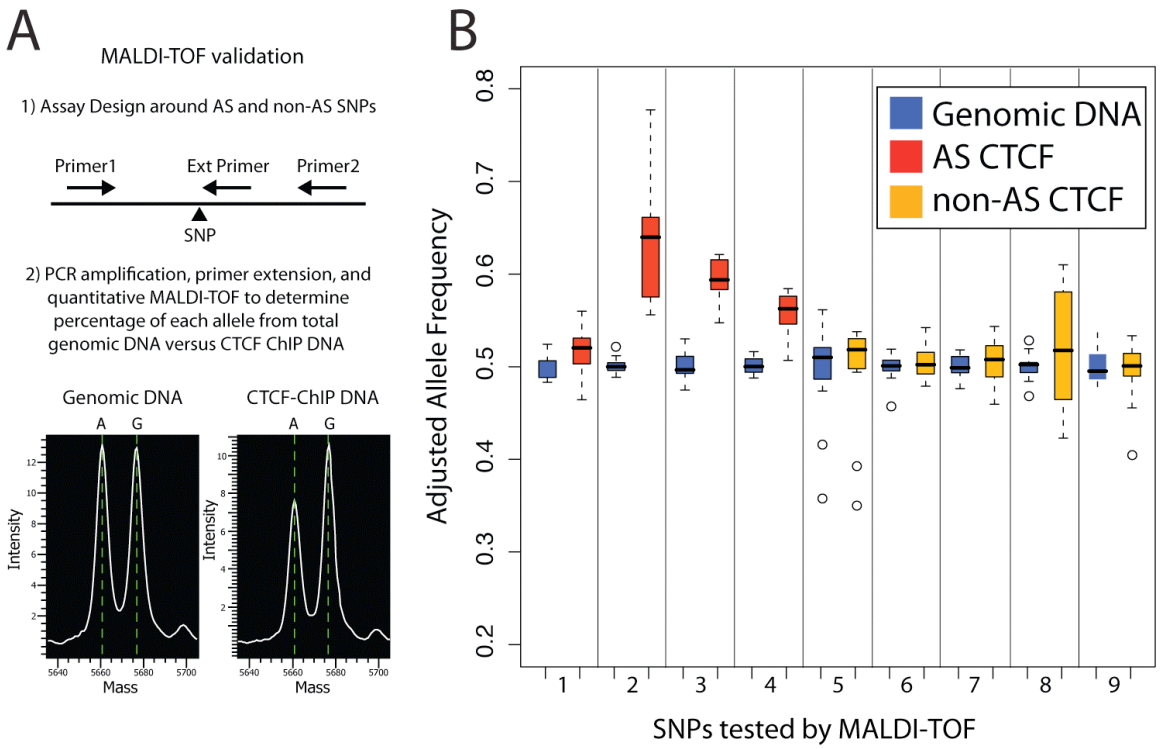


Figure S8. (A) Quantitative MALDI-TOF mass-spectrometry verification of CTCF binding. The scheme is shown on top. The bottom graphs show the relative peak areas for the control genomic and the CTCF ChIP sample after the MALDI-TOF assay. (B) The box-plot shows adjusted allele frequencies for nine sites which were identified as either allele-specific (AS, sites 1-4) or non-allele-specific (non-AS, sites 5-9) for CTCF binding by ChIP-seq. Each column shows the allele frequency of the allele that was enriched in the ChIP-seq data and represents the distribution of 16 replicate MALDI-TOF measurements. All four AS CTCF sites were significantly biased compared to the genomic DNA control, and towards the same allele indicated by ChIP-seq. None of the non-AS CTCF sites were significantly biased. SNP positions, primer sequences and P values are provided in Tables S4 and S5 below.

**Supplemental Figure 9**.

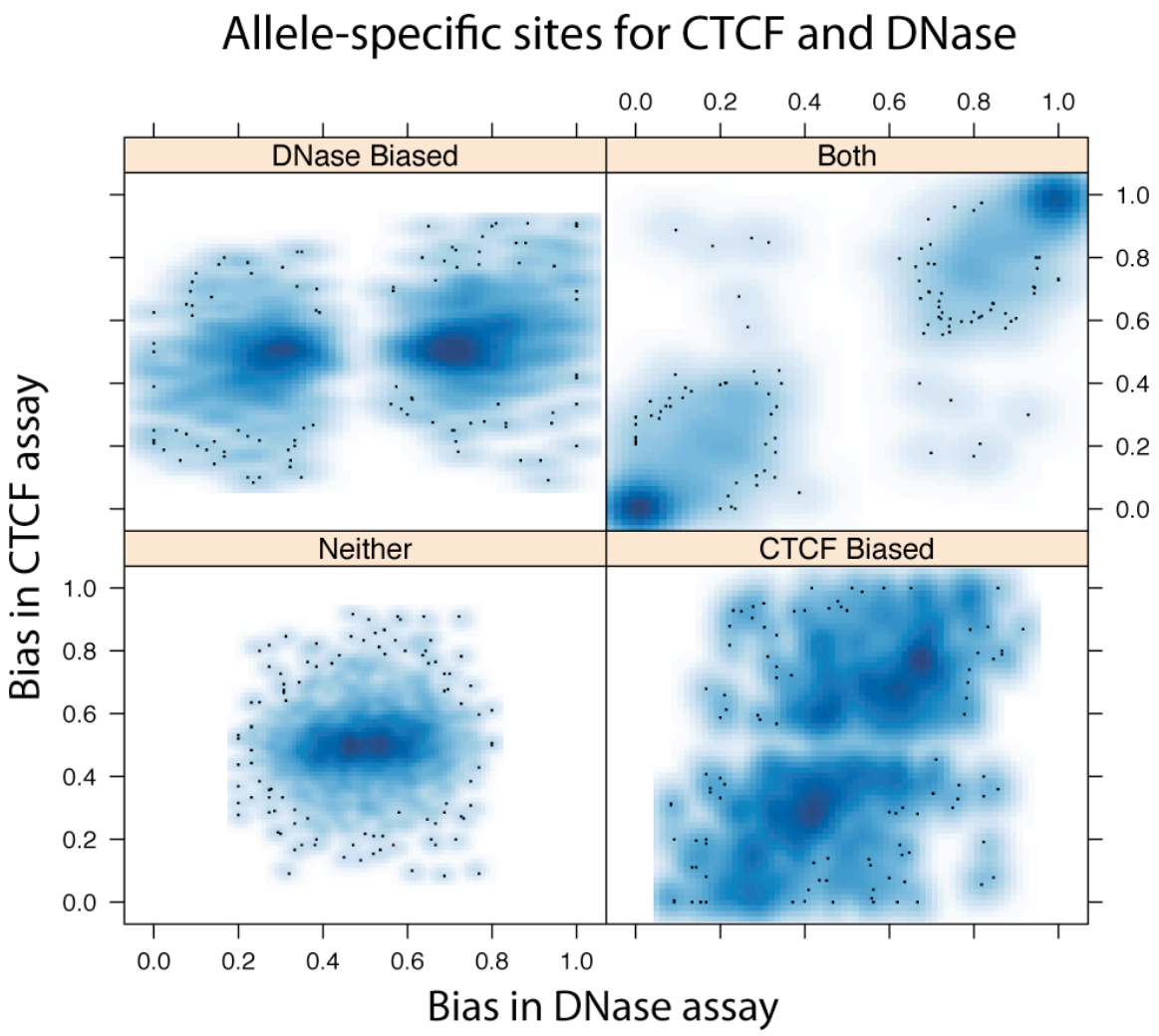


Figure S9. Allele-specific bias is concordant in both DNase I and CTCF assays. All DNase I HS and CTCF sites that contained an informative SNP in all individuals were classified as not being allele-specific in either method, or was allele-specific in DNase I HS data (DNase Biased), CTCF data (CTCF Biased), or allele-specific in both assays. X and Y-axes show the direction and relative degree of allele-specificity. The figure shows smoothed scatter-plots where the blue color and intensity represents the density of points. The black dots are individual points where the density is low. For the analysis shown in this figure, as well as in Fig. 3B, we assessed the proportion of sites showing discordant behavior (present in the top left and bottom right corners of each plot) by looking for a) different directional biases of a heterozygous SNP in two individuals and b) significant differences in that bias by a Fisher's Exact Test, again corrected for multiple testing by the FDR method, taking 0.01 FDR as the threshold. The majority of allele-specific regions detected in both assays are biased for the same allele (lower left and upper right quadrants), and only 2% were significantly biased for the opposite allele (upper left and lower right quadrants within a plot).

**Supplemental Figure 10**.

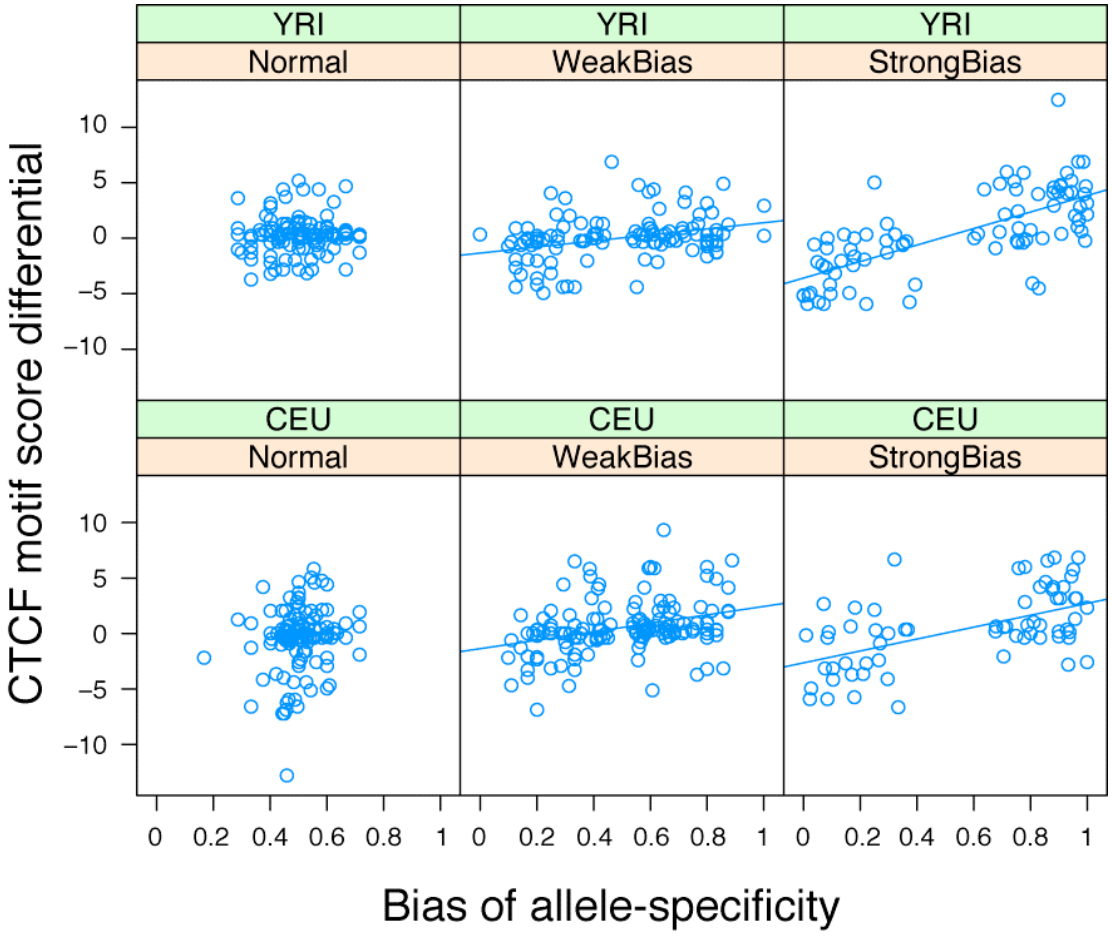


Figure S10. Motif Analysis of Allele Specific CTCF sites. To determine the effects of SNPs on the CTCF motif, the FIMO (Find Individual Motif Occurrence) (*S9*) software tool was used. For all SNPs in each individual that met our threshold of detection, we generated 2 sequences containing either SNP and the 39 bp window centered on the SNP. These sequences were then analyzed for the presence of the previously identified 20 bp CTCF binding motif (*11*). In cases where there were multiple overlapping motifs within the window, only the highest scoring motif was considered. We then cross-correlated the difference in the motif score between the two alleles with the degree of Allele specific bias for each SNP.

**Supplemental Figure 11**.

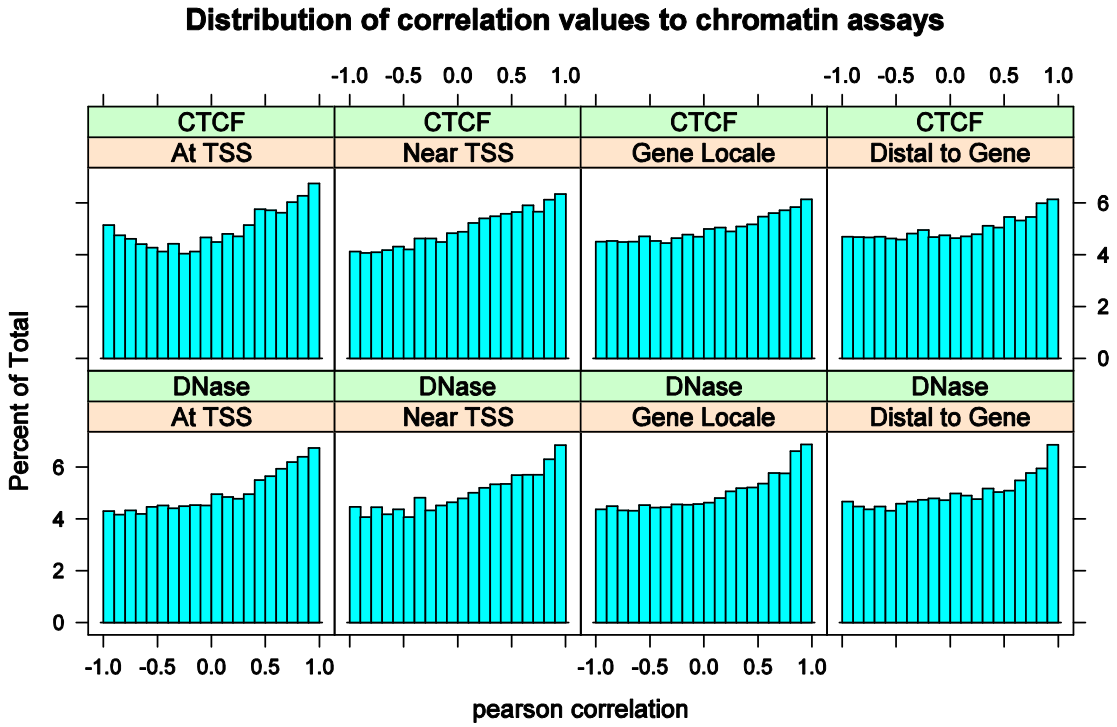


Figure S11. Distribution of Pearson correlation values between chromatin assays and gene expression. The four distance-based categories were: "At TSS" = +/- 2.5 Kb; "Near TSS" = +/- 10 Kb of TSS; "Gene Locale" = +/- 100 Kb of TSS; "Distal to Gene" > 100 Kb of its closest TSS. Figure 3 shows only the split for TSS (< 2.5 Kb) or Near TSS.

**Supplemental Table 1: Sequencing Statistics**

| | | DNase-seq | | | CTCF ChIP-seq | | |
|---|---|---|---|---|---|---|---|
| **Family Structure** | **Cell line** | **# of total sequences** | **# of useable sequences** | **% of total** | **# of total sequences** | **# of useable sequences** | **% of total** |
| CEU Father | GM12891 | 88,066,316 | 67,622,813 | 76.8% | 30,244,488 | 21,733,635 | 71.8% |
| CEU Mother | GM12892 | 91,200,528 | 69,682,443 | 76.4% | 44,885,150 | 34,494,412 | 77.0% |
| CEU Daughter | GM12878 | 169,842,640 | 120,719,864 | 71.1% | 32,547,270 | 25,846,561 | 79.4% |
| YRI Father | GM19239 | 90,090,626 | 68,375,492 | 76.0% | 26,628,402 | 20,232,825 | 76.1% |
| YRI Mother | GM19238 | 91,270,196 | 70,201,052 | 76.9% | 32,377,472 | 25,547,799 | 78.9% |
| YRI Daughter | GM19240 | 88,678,656 | 67,332,580 | 76.0% | 33,399,839 | 26,250,278 | 78.4% |
| | Total | 619,148,962 | 463,934,244 | | 200,082,621 | 154,105,510 | |

**Supplemental Table 2. The number of constant, individual-specific and variable sites in the different cell line combinations.**

| | Class in Fig. 1B | DNase | CTCF |
|---|---|---|---|
| **Total number of sites per cell line** | | | |
| GM12891 | | 66957 | 55154 |
| GM12892 | | 66906 | 52277 |
| GM19238 | | 62383 | 56348 |
| GM19239 | | 53836 | 50710 |
| **Individual Specific Sites** | | | |
| GM12891 | Singleton | 8326 | 1628 |
| GM12892 | Singleton | 10352 | 933 |
| GM19238 | Singleton | 7220 | 444 |
| GM19239 | Singleton | 3911 | 1112 |
| GM12891:GM12892 | CEU-only | 6802 | 823 |
| GM19238:GM19239 | YRI-only | 3239 | 809 |
| GM12891:GM19238 | Other combinations | 2237 | 2130 |
| GM12891:GM19239 | Other combinations | 1489 | 2582 |
| GM12892:GM19238 | Other combinations | 2276 | 1354 |
| GM12892:GM19239 | Other combinations | 732 | 1277 |
| **Constant Sites** | | | |
| Constant sites | Constant | 49702 | 58192 |

**Supplemental Table 3. Combinations of cell lines showing DNaseI HS and CTCF allele-specific sites. The Total column for each assay shows the number of sites at heterozygous SNPs that could be assayed for allele-specificity, and the Sig column shows how many of them were significant at the 0.01 FDR level.**

| | DNase | | CTCF | |
|---|---|---|---|---|
| **Cell line combinations where allele-specific sites occur** | **Total** | **Sig** | **Total** | **Sig** |
| GM12891:GM12892 | 1638 | 121 | 803 | 86 |
| GM19240:GM19238 | 707 | 61 | 1179 | 143 |
| GM19240:GM19239 | 598 | 38 | 673 | 65 |
| GM19238:GM19239 | 336 | 14 | 340 | 44 |
| GM19240:GM19238:GM19239 | 316 | 25 | 323 | 40 |
| GM12891:GM12878 | 250 | 8 | 728 | 74 |
| GM19238:GM12892 | 227 | 14 | 244 | 13 |
| GM12892:GM12878 | 214 | 32 | 778 | 73 |
| GM19239:GM12892 | 205 | 22 | 105 | 4 |
| GM19238:GM12891:GM12892 | 170 | 14 | 102 | 11 |
| GM19240:GM19238:GM12892 | 167 | 13 | 195 | 17 |
| GM19240:GM19239:GM12892 | 167 | 13 | 122 | 18 |
| GM19238:GM12891 | 159 | 13 | 183 | 18 |
| GM19238:GM19239:GM12892 | 146 | 17 | 96 | 11 |
| GM12891:GM12892:GM12878 | 139 | 12 | 352 | 69 |
| GM12891:GM19239:GM12892 | 138 | 10 | 62 | 6 |
| GM19240:GM19238:GM12891 | 137 | 6 | 117 | 17 |
| GM19240:GM12892 | 132 | 7 | 162 | 8 |
| GM19240:GM12891:GM19239 | 130 | 4 | 86 | 11 |
| GM12891:GM19239 | 117 | 3 | 87 | 5 |
| GM19240:GM12891:GM12892 | 108 | 7 | 60 | 3 |
| GM19238:GM12891:GM19239 | 101 | 4 | 68 | 8 |
| GM19240:GM19238:GM12891:GM12892 | 93 | 6 | 76 | 11 |
| GM19240:GM12891 | 89 | 6 | 131 | 8 |
| GM19240:GM19238:GM19239:GM12892 | 76 | 7 | 78 | 11 |
| GM19240:GM19238:GM12891:GM19239 | 73 | 8 | 57 | 9 |
| GM19240:GM12891:GM19239:GM12892 | 63 | 5 | 40 | 4 |
| GM19238:GM12891:GM19239:GM12892 | 61 | 3 | 40 | 6 |
| GM19239:GM12892:GM12878 | 40 | 5 | 108 | 6 |
| GM19240:GM19239:GM12892:GM12878 | 36 | 3 | 73 | 10 |
| GM19238:GM12891:GM12878 | 34 | 1 | 98 | 14 |
| GM19240:GM19238:GM12891:GM12878 | 34 | 2 | 89 | 17 |
| GM19240:GM19238:GM12891:GM19239:GM12892 | 34 | 3 | 31 | 3 |
| GM12891:GM19239:GM12878 | 30 | 0 | 86 | 11 |
| GM19240:GM12891:GM19239:GM12878 | 27 | 0 | 83 | 17 |
| GM19240:GM19238:GM12892:GM12878 | 27 | 2 | 98 | 18 |
| GM19238:GM12892:GM12878 | 26 | 3 | 121 | 8 |
| GM19240:GM12892:GM12878 | 24 | 3 | 78 | 11 |
| GM19240:GM19239:GM12878 | 24 | 1 | 63 | 5 |
| GM19238:GM12891:GM12892:GM12878 | 23 | 4 | 51 | 5 |
| GM19240:GM19238:GM12878 | 23 | 2 | 79 | 15 |
| GM19240:GM12891:GM12878 | 22 | 1 | 62 | 5 |
| GM19239:GM12878 | 21 | 1 | 88 | 4 |
| GM19240:GM12891:GM12892:GM12878 | 18 | 2 | 35 | 6 |
| GM19238:GM19239:GM12892:GM12878 | 17 | 1 | 55 | 5 |
| GM19240:GM19238:GM12891:GM12892:GM12878 | 17 | 2 | 52 | 6 |
| GM19240:GM19238:GM12891:GM19239:GM12878 | 17 | 0 | 52 | 8 |
| GM19238:GM19239:GM12878 | 16 | 1 | 45 | 1 |
| GM19240:GM12878 | 13 | 0 | 86 | 2 |
| GM19238:GM12878 | 12 | 1 | 118 | 10 |
| GM12891:GM19239:GM12892:GM12878 | 11 | 0 | 37 | 4 |
| GM19238:GM12891:GM19239:GM12878 | 11 | 1 | 47 | 14 |

| | | | | |
|---|---|---|---|---|
| GM19240:GM12891:GM19239:GM12892:GM12878 | 11 | 2 | 35 | 4 |
| GM19240:GM19238:GM19239:GM12892:GM12878 | 11 | 1 | 41 | 11 |
| GM19238:GM12891:GM19239:GM12892:GM12878 | 10 | 4 | 29 | 5 |
| GM19240:GM19238:GM12891:GM19239:GM12892: GM12878 | 10 | 0 | 28 | 7 |
| GM19240:GM19238:GM19239:GM12878 | 10 | 1 | 37 | 9 |
| Total | 7366 | 540 | 9192 | 1034 |

**Supplemental Table 4. SNPs and primer sequences used for MALDI-TOF verification of CTCF allele-specific binding, numbered as they appear in fig. S8B.**

| SNP | Position (genome_ chromosome_ coordinate) | PCR Primer 1 | PCR Primer 2 | Extension Primer |
|---|---|---|---|---|
| 1 | hg18_22_ 36155166 | ACGTTGGATGGGCAGG AGCGGATTAGGT | ACGTTGGATGACATGG TGTGTATCCCTGACAA | GGGAGCACTC TAATTTCT |
| 2 | hg18_X_ 129054023 | ACGTTGGATGGGGCGT CCCCTTGTGTAG | ACGTTGGATGTTCTCT GCCACTTCCTGTCC | AGGGAGTGCA GAGAGGG |
| 3 | hg18_X_ 150600094 | ACGTTGGATGGGGCAA GAGATGACAAGAAATG | ACGTTGGATGGCAGGC TCCAGCTCAGGTA | GAGCGCCACT GGCCTAC |
| 4 | hg18_X_ 67869737 | ACGTTGGATGGGAAGA CCCTGTGAAGGAAAAG | ACGTTGGATGGTTGCT CTGTGTGGGAAGATG | TGAGAAGACT TGGAGGC |
| 5 | hg18_11_ 64177304 | ACGTTGGATGCCTGAG AGGGGCTGGAATAC | ACGTTGGATGCCATTT GGCAGAAACTCACC | ACCCTCATCC TCTTTCC |
| 6 | hg18_15_ 62810775 | ACGTTGGATGATTTACA TCTCAGGCCCCTTGT | ACGTTGGATGCCCCAG GATCACACAGTCC | GGAGGCTGTG CCAGAGG |
| 7 | hg18_1_ 226016975 | ACGTTGGATGCAACGC AAGGACGAGTGT | ACGTTGGATGAACCAT GACGGATGTCTCA | TGTCAGGGAG GGACACC |
| 8 | hg18_3_ 13674549 | ACGTTGGATGTGTGAT GCTGGGTACAAAGAAG | ACGTTGGATGCTTACT CAAGCCCGTTCCAC | CAGTCCTCAC CCTCCGC |
| 9 | hg18_5_ 175933859 | ACGTTGGATGCAGGCA CCCTCTTTAGGTAAGC | ACGTTGGATGCTGACC GCCTTTGTGACC | CAATGTCCCC ACCCCAA |

**Supplemental Table 5. Two-sided t-test for adjusted allele frequency.**

| SNP | SNP position | *P* value |
|---|---|---|
| 1 | hg18_22_36155166 | 0.0214 |
| 2 | hg18_X_129054023 | 4.25E-07 |
| 3 | hg18_X_150600094 | 6.26E-14 |
| 4 | hg18_X_67869737 | 3.40E-09 |
| 5 | hg18_11_64177304 | 0.83 |
| 6 | hg18_15_62810775 | 0.34 |
| 7 | hg18_1_226016975 | 0.51 |
| 8 | hg18_3_13674549 | 0.24 |
| 9 | hg18_5_175933859 | 0.69 |

**References**


S1. A. P. Boyle *et al.*, *Cell* **132**, 311 (2008).
S2. J. Kim, J. H. Lee, V. R. Iyer, *PLoS ONE* **3**, e1798 (2008).
S3. H. Li, J. Ruan, R. Durbin, *Genome Res* **18**, 1851 (2008).
S4. A. P. Boyle, J. Guinney, G. E. Crawford, T. S. Furey, *Bioinformatics* **24**, 2537 (2008).
S5. J. M. Amos-Landgraf *et al.*, *Am J Hum Genet* **79**, 493 (2006).
S6. J. L. Rupert, C. J. Brown, H. F. Willard, *Eur J Hum Genet* **3**, 333 (1995).
S7. C. J. Brown *et al.*, *Nature* **349**, 82 (1991).
S8. L. Carrel, H. F. Willard, *Nature* **434**, 400 (2005).
S9. T. L. Bailey *et al.*, *Nucleic Acids Res* **37**, W202 (2009).