Chapter 2

The genetics of CCCTC binding factor

Collaboration note. *This chapter contains work in collaboration with Yunyun Ni, Sander W. Timmer and others in the research groups of Gregory E. Crawford, Jason D. Lieb, Vishwanath R. Iyer and Ewan Birney. This work is published in PLoS Genetics (Ding et al., 2014). I am the lead author alongside the other two joint first authors Yunyun Ni and Sander W. Timmer. My contribution in this work includes data production and quality control, quantifying CTCF binding regions, genotype production, association mapping, and jointly with Yunyun Ni allele specific analysis. The manuscript also contains a novel discovery of three distinct CTCF binding modes on X chromosome, which was primarily conducted by Sander W. Timmer, and is not presented here.*

2.1 Overview

In the past decade a large number of variants have been discovered associated with traits or disease. Although they provide important hints, it is not at all straightforward to understand the underlying biological mechanisms. The majority of the loci that have been found are in non-protein coding DNA sequences, suggesting regulatory roles often responsible for the phenotypic effect (The 1000 Genomes Consortium, 2010). Sequence conservation based approaches can identify the regulatory regions that are under selection pressure, possibly due to binding of a protein factor or other regulators (Lindblad-Toh et al., 2011). Recent sequencing based technologies can give a more direct measure for regulatory events, such as the binding of CCCTC factor (Kunarso et al., 2010; Schmidt et al., 2010) and a number of other factors, e.g. Noonan and McCallion, 2010 and McVicker et al., 2013, revealing the landscape of the regulatory elements in human genome.

Studying the effect of genetic variants on gene regulation has become an important approach to find intermediates between genotype and whole organism phenotype. Using DNase I hypersensitivity and binding assays for the CTCF transcription factor on two family trios with known genome sequences, McDaniell et al. (2010) showed that allele-specific binding patterns consistent with strong genetic effects could be readily measured at heterozygous sites. Other studies have shown allele specific binding of RNA polymerase and NF-κB binding measured across a small number of individuals (Kasowski et al., 2010), or of a wider range of transcription factors in a single cell line (Reddy et al., 2012). Similarly, differences between mouse strains in binding of PU-1 and CEBPa at enhancer regions correlate with sequence differences and adjacent gene expression (Heinz et al., 2013). Intriguingly, some sites with prominent SNPs in the binding motifs of CTCF did not show a genetic effect in a study of its binding across an extended family (Maurano et al., 2012). Reciprocally, differences in transcriptor factor binding were seen between closely related species even where there was no sequence difference in the binding region (Stefflova et al., 2013).

In order to examine these phenomena further, and infer potential causative connections to disease GWAS results, we need to identify specific cases where a genetic variant affects binding. To do this we can use genetic association mapping. When applied to transcript expression levels as the measurements on 60 or more samples, this approach has identified thousands of expression quantitative trait loci (eQTLs) (Spielman et al., 2007; Stranger et al., 2007; Pickrell et al., 2010). A QTL study of human open chromatin (Degner et al., 2012) found 8,902 DNase I hypersensitivity sites that were correlated with genetic variants. However, there are currently no systematic association studies of how genetic variation in human populations affects the binding pattern of a specific transcription factor. Here we carry out such a study.

To identify transcription factor binding QTLs, we measured the binding of CTCF across a panel of lymphoblastoid cell lines (LCL). Previous studies have shown that there is resemblance between LCLs and the parent lymphocytes at a variety of molecular levels including transcription factor binding according to accumulated observations (See review Sie et al., 2009). Despite of some inherit limitations, such as aneuploidy, gene mutations and reprogramming, often associated with telomerase activity, which can be controlled experimentally to a certain level, LCLs has still been instrumental in general as a resource for functional screening that offers acceptable fidelity and is scalable compared to clinical trails or *in vivo* systems.

CTCF is a highly conserved multifunctional protein that serves both as a transcription factor as well an insulator binding protein, preventing interactions between enhancers and promoters and demarcating chromatin domains. Working with cohesin, CTCF can also mediate chromosomal looping interactions, and is involved in imprinting as well as X-inactivation (see Lee et al., 2012; Merkenschlager and Odom, 2013 for reviews). There have been extensive locus specific studies (Bell et al., 1999; Bell and Felsenfeld, 2000; Yusufzai et al., 2004; Splinter et al., 2006; Stedman et al., 2008; van de Nobelen et al., 2010; Sopher et al., 2011) and specific genome wide screens (Cuddapah et al., 2009; Phillips and Corces, 2009).

Schmidt et al. (2010) showed in breast cancer cell lines and hepatocellular carcinoma cell lines CTCF appears to work independently to cohesin. Schmidt et al. (2010) compared CTCF binding patterns across five species and showed that its binding variation correlates with the evolution distances between species. Previous studies have shown the extent of genetic effects on CTCF binding in families (McDaniell et al., 2010; Maurano et al., 2012), although specific loci underlying these effects have not been identified.

We used ChIP-seq to measure CTCF binding in 51 lymphoblastoid cell lines (LCLs) from the HapMap CEU population, each of which had already been sequenced as part of the 1000 Genomes Project (The 1000 Genomes Consortium, 2010) and had been subjected to RNA-seq analysis (Montgomery et al., 2010). Our data and analysis identified thousands of CTCF binding QTLs across the human genome. These data, together with the available full genome sequence of the cell lines, allowed us to explore parameters of genetic effects on protein-DNA binding. For example, we defined the relationship of the QTL location to the TF binding motif, estimated the relative impact of substitutions and insertions/deletions (INDELs), and measured whether allele-specific differences are indicative of population-wide variation.

2.2 Measuring CTCF binding in HapMap cell lines

ChIP-seq Chromatin immunoprecipitation was done at the University of Texas Austin and sequenced at the Wellcome Trust Sanger Institute. Cells were cross-linked with 1% formaldehyde for 7 min at room temperature. Formaldehyde was deactivated by adding glycine. Chromatin from harvested cells was sonicated with a Bioruptor to an average size of 500 bp DNA. Immunoprecipitation was performed using sonicated chromatin by adding anti-CTCF antibody (Millipore 07-729). For a subset of eight samples, including day replicates GM12891 and GM12892, the same procedure was applied but without using the anti-CTCF antibody, which gives information for estimating the input background. ChIP DNA was used to generate a ChIP-seq library according to the standard Illumina protocol. The library was then sequenced using the Illumina HiSeq platform in 50bp paired end reads. On average ~85.5M reads were produced per sample. Data have been submitted to the European Nucleotide Archive, available with accession number ERP002168. They are also deposited in ArrayExpress with accession number E-ERAD-141. Sequence lanes were assessed for multiple quality metrics including total yield, read quality, mapping quality, GC content distribution and duplication rate. All sequencing reads were aligned to the human reference sequence (GRCh37) using BWA v0.5.9-r16 (Li and Durbin, 2009) using default parameter settings. Duplicate reads were marked by the "MarkDuplicates" function of the software Picard (v1.47 http://picard.sourceforge.net/) and removed. We reason that as the binding interface is much smaller than the fragment size and we used paired-end sequencing, duplicates are more likely to be technical than biological. We applied a stringent filter by removing all the reads with mapping quality score below

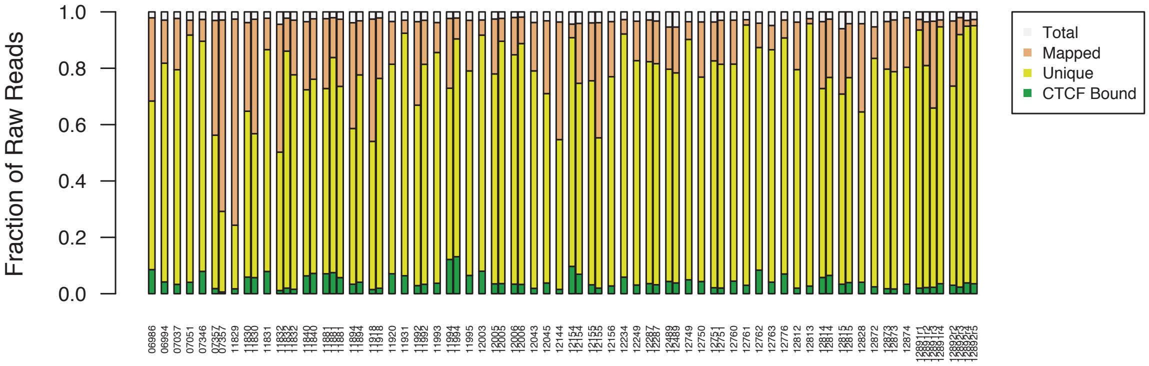

Figure 2.1: ChIP-seq production. The proportions of the mapped fragments, unique fragments, and CTCF bound fragments are plotted for each samples.

30, improperly paired (with 0x2 flag set in the BAM format), or with mate pairs more than 1kb apart (Figure 2.1). For allele specific analysis, we further performed local realignment using a variant-aware aligner glia (https://github.com/ekg/glia), which aligns reads against paths in a variant graph built by combining the reference sequence and known variants.

Binding region calling We performed binding region identification using a Parzen kernel density window algorithm that we applied in previous studies and achieved good performance (Shivaswamy et al., 2008; Lee et al., 2012). This procedure was applied to both experimental and input datasets after combining lanes and replicates into cell-line sample sets. Local maxima of these Parzen scores were used to define binding peak positions, and the interquartile range of the kernel density profile was used to determine the corresponding binding site of highest read density. The resulting set of candidate CTCF binding sites was then subjected to input correction, filtering for copy number artifacts, and determination of statistical significance. A input profile was built using data from a subset of eight samples that went through the exact same production except that the antibody was not used. First, in order to normalize for background represented by the input control, each binding site was paired with the corresponding input site with the highest read count within 200 bp. A binomial P-value was computed for each binding site under the null hypothesis that ChIP and input reads were equally likely. The ratio of total ChIP to input reads for each sample was used to normalize for differences in sequencing depth before calculating the binomial P-value, with the library having higher sequencing depth always scaled downward. Binding sites falling in previously defined genomic regions with aberrantly high signal due to copy number differences were discarded (Boyle, Davis et al. 2008, http://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=334775099&c=chrX&g=wgEncodeMapability). Binding sites dominated by input were also discarded, retaining only sites where the ChIP read count scaled by sequencing depth exceeded input.

The resulting set of filtered peak P-values was subjected to multiple hypothesis testing using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995). Next, binding regions for the cell lines at various significance levels were merged using bedtools v2.17.0 (Quinlan and Hall 2010) in such a way as to preserve the set of calling cell lines (bedtools merge -nms -scores collapse -n). We employed several metrics in order to determine an appropriate significance cutoff, including the relationship between binding region count and P-value (Figure 2.2) and the number of calling cell lines for each binding region (Figure 2.3). Raw P-values were used to define significant sites once the P-value threshold was determined. Binding regions with BH-adjusted P-value $\leq$ 1E-5 were initially retained as significant (n=127,351), as that value appeared to be the inflection point in the binding region versus P-value curve and had the largest reduction in binding regions called in just one sample.

Finally, in order to assess the quality of binding regions called in only one cell line, we used bedtools (bedtools intersect –c) to identify binding regions containing the extended CTCF motif (Figure 2.4). Binding regions called in only one cell line showed a significantly lower occurrence of the CTCF motif as compared to binding regions called by two or more cell lines. Therefore, we discarded binding regions called in only one cell line and retained the 63,753 merged binding regions at adjusted P-value

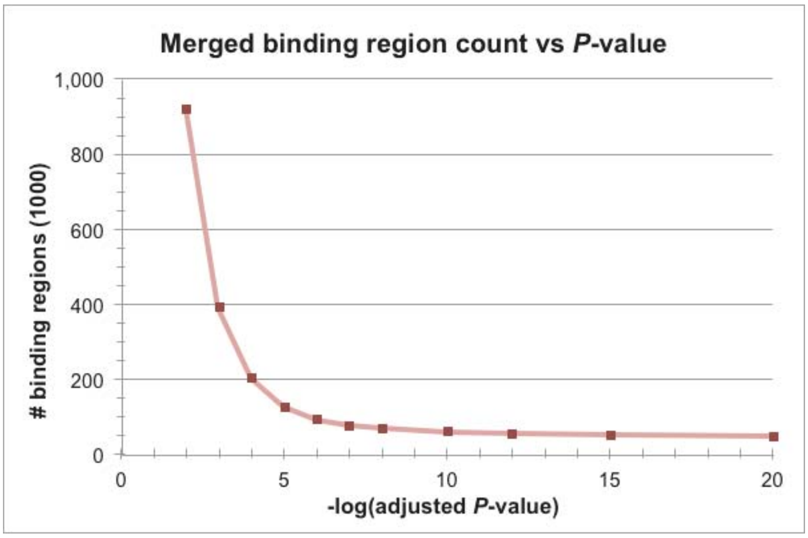

Figure 2.2: Number of merged binding regions plotted as a function of –log(BH-adjusted binomial P-value).

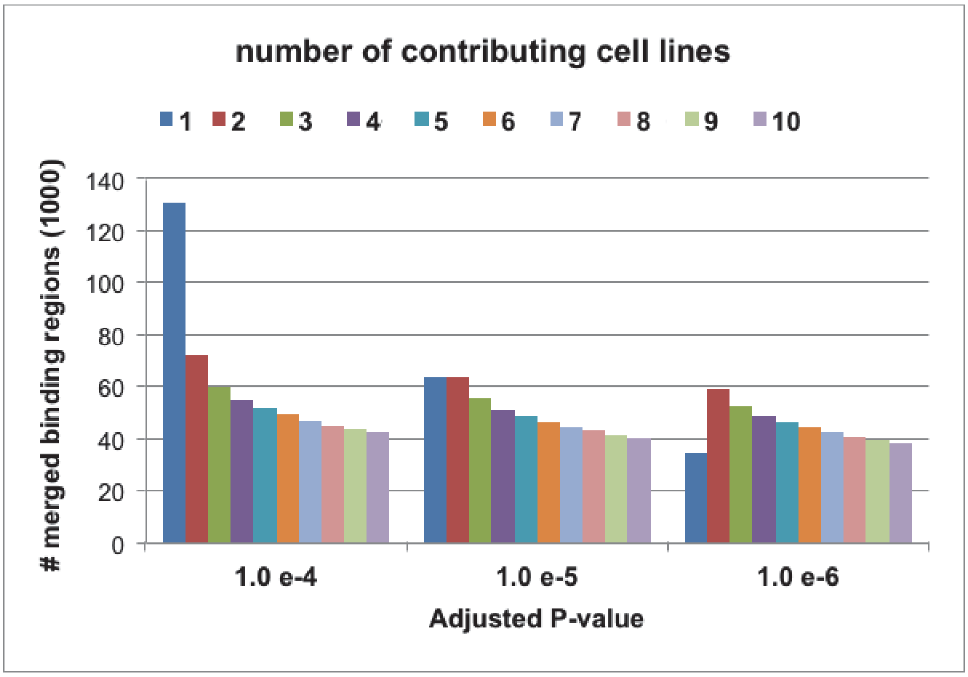

Figure 2.3: Number of merged binding regions as a function of number of calling cell lines, at three adjusted P-values.

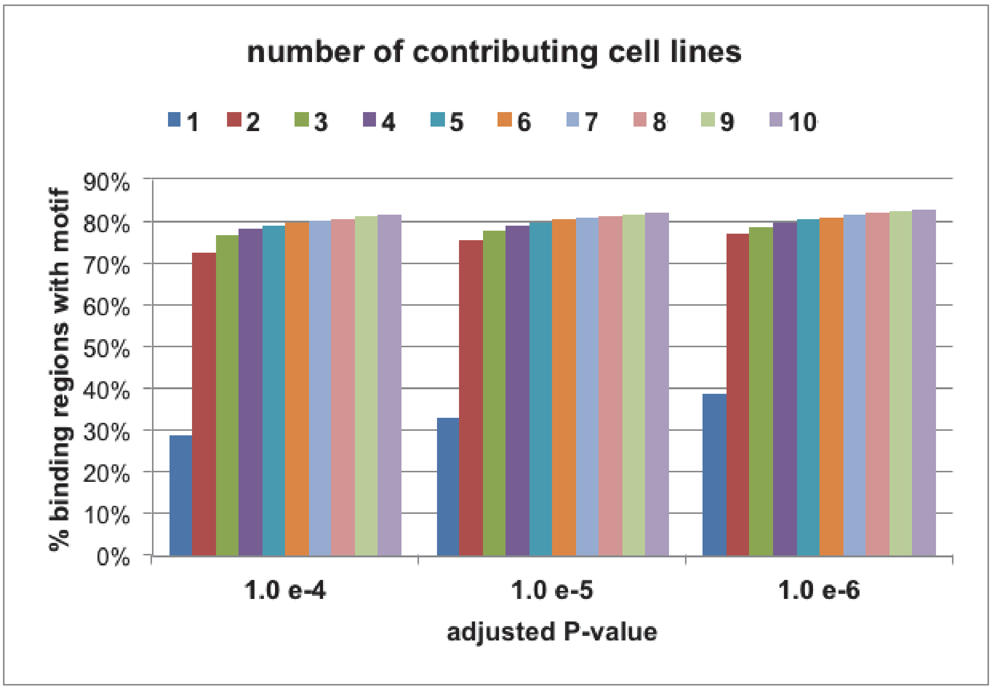

Figure 2.4: Proportion of merged binding regions as a function of number of calling cell lines, at three adjusted P-values.

1E-5 with two or more cell lines.

Blacklisting regions Out of 63,753 binding regions identified, we removed 2,898 binding regions falling in repeat sequences or in the Immunoglobulin heavy chain locus or major histocompatibility complex (MHC). In detail, 2,578 binding regions lie completely within repeat sequences marked by a merged set consisting of “Repeat Masker”, “Segmental Dup” or “Simple Repeat” from the table browser of the UCSC Genome Browser, 35 binding regions lie within the Immunoglobulin heavy chain locus (chr14:106053226-106330470) and 285 fall in the MHC region (chr6:28477797-33448354).

Motif word identification We searched for instances of CTCF motif in the discovered binding regions using the CTCF canonical 19bp position weight matrix downloaded from the JASPAR database (Sandelin et al., 2004, http://jaspar.binf.ku.dk/). We extracted DNA sequences at the identified binding regions from human genome reference GRCh37 to construct a sequence database. The search was then performed using the software FIMO(Grant et al., 2011) of the MEME tool suite (Bailey et al., 2009) using parameter "–threshold 1E-4". This process identified at least one motif instance in 45,867 of our 57,428 binding regions. For the ones with multiple motif instances, we selected the motif with highest matching score as the nominal binding motif for the region for some analysis.

2.3 Quantification of CTCF binding

With the peak profile identified above, we quantified the signal for each binding region by counting the number of sequencing fragments (read pairs) when alignment overlapping the region. We applied stringent criteria by only counting the properly aligned read pairs with quality score at least 30 and excluding all the duplicated reads (samtools view -f 0x42 -F0x604 –q 30). We used Bedtools (v2.16.2) (Quinlan and Hall, 2010) to count the intersection between fragments and identified binding regions. This produced an $N \times M$ matrix, where N is the number of samples and M is the number of binding regions. To evaluate the variation in the ChIP experiments, for two samples we collected replicated data on four consecutive days. Using binding sites defined previously, we compared the correlation between replicates grown on consecutive days and the correlation between all other samples. We found a mean pairwise correlation coefficient of 0.83 and 0.82 for the replicate sets for NA12891 and NA12892, respectively, while the mean pairwise correlation coefficient between samples was 0.17. This suggests a good signal to noise ratio in the experiment. This could be considered as covariates in linear model. However, in our data, we do not see much deviation from uniform in the test results from our random control (results shown in 2.11), for simplicity, we do not add additional variables to our tests.

For the subsequent genetic analysis, we are interested in the binding regions that have good signal and also vary between individuals. The mean and variance of binding intensities are correlated by the nature of the Poisson process for the sequencing. We found a group of 4,516 binding regions (7% of the total binding regions identified) with little signal or variation - defined as binding regions mapped with fewer than 6 fragments on average per sample and SD < 5.14. The cut-off was chosen as it delineates clear groups of background intensity and signal intensity with distinct strengths (Figure 2.5). These binding regions were excluded from further analysis.

Normalization Previous studies (Montgomery et al., 2010; Degner et al., 2012) have shown that appropriate normalization can substantially enhance genetic association signals by removing confounding non-genetic sources of variation. Potential sources of confounding variation include experimental batch effects, GC bias in sequencing library construction and latent unknown technical or biological factors that have systematic effects across large numbers of binding regions. To address these issues, we normalized the raw binding intensity using the following five step approach to generate a normalised adjusted binding intensity (NABI).

1. Rescale by sequence depth.

$$X_{i,j} = \frac{R_{i,j} Mean(S_j)}{S_j},\ i = 1...M, j = 1...P$$

where $R_{i,j}$ is the raw intensity of the ith binding region of the jth lane, and S_j is the sum of intensity across all binding regions for the jth lane. $R_{i,j}$ is scaled by a factor of the proportion of mean of S across all P lanes over S_j.

2. Remove variance introduced by GC composition. We adjusted for GC bias in sequencing library construction by forming percentile bins for GC composition of all binding regions and normalising the binding intensities within each bin. Since the fragment length is much larger than the motif length, this bias is not strongly influenced by the motif sequence.

$$X_{i,j} = \frac{X_{i,j}}{Median(X_{k,j}; k \text{ same GC bin as } i)}$$

where i, j, k are the indices for binding region, lane, and GC bin respectively.

3. Merge lanes of a same individual by taking the mean. A subset of our samples

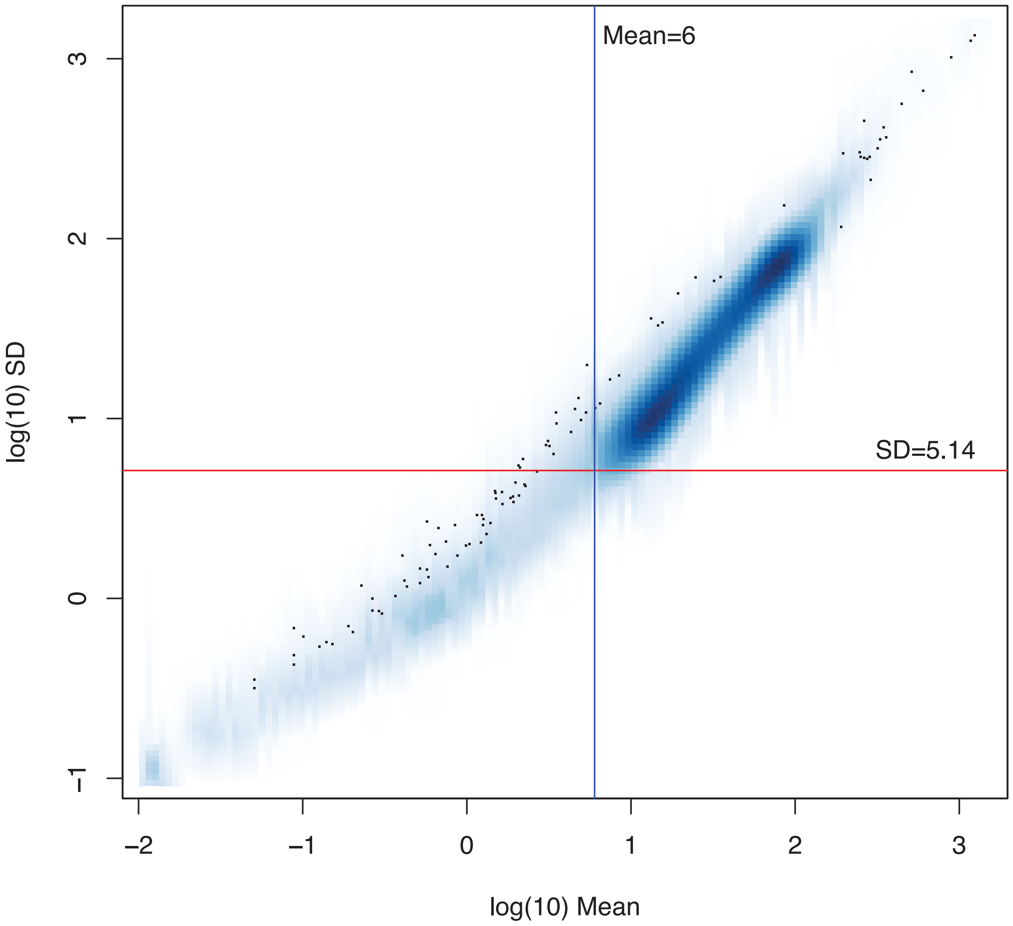

Figure 2.5: Quality control by raw signal intensity and inter cell line variability. For each binding region we counted the overlapping sequencing fragments (identified by a properly paired read pair) and used it as a measure for the raw binding intensity. We plot the log of the variance of the binding intensities across 51 individuals versus the log of the mean of the binding intensities using the R function *smoothScatter*. The degree of blue is proportional to the density of data points. As a Poisson process the mean and variance correlate to each other. There exists a natural cutoff between the lower left tail and the majority at mean 6 and standard deviation 5.14. These lower left tail binding regions are the sites with very low intensity and also low variability. We removed these sites, 4,516 binding regions in total, before further analysis.

were sequenced on multiple lanes and in these cases we took the mean value across lanes as the measurement of the individual.

$$D_{i,l} = Mean(X_{i,j}; j \text{ lanes of } l),\ l = 1...N$$

where $X_{i,j}$ is the measure from the previous step, i, j, l are indices for the binding region, lane and samples, respectively. N is the total number of samples.

4. Centre-scale binding intensity for each binding region. We then scaled the binding intensity for each binding region by subtracting the mean and then dividing by the standard deviation. This transforms the measures of each binding region into zero mean and unit variance, which is needed for the quantile normalization to be less affected by the different variances of different binding regions

$$Z_{i,l} = \frac{D_{i,l} - Mean(D_i)}{StDev(D_i)}$$

where i, l are indices for binding region and sample.

5. Quantile normalize each sample data to a normal distribution. The distribution of binding intensities for each individual is complex. Previous studies have shown that quantile normalization, initially developed for normalising the microarray signals of gene expression, can assist statistical analysis by converting the distributions of each sample to a reference distribution. The linear regression model used to identify QTL in our study assumes a Gaussian distribution of binding measures within each genotype class. We therefore mapped the measures across all binding regions of each sample to the corresponding normal quantiles. This produces a matrix that is essentially a perturbation permutation of the normal quantiles

$$\tilde{Z}_{i,l} = \Phi^{-1}(\frac{\sum_{m=1}^{M} I\{Z_{m,l} < Z_{i,l}\}}{M+1})$$

where Φ is the cumulative normal density function and M is the total number of binding regions. I is an indicator function that returns 1 if the condition is met and 0 otherwise.

6. Remove confounding variation by principal component analysis (PCA). The

Variables	PC1	PC2
Sequencing mapping rate	0.049	0.50
Duplication rate	0.065	0.031
Sequencing depth	0.043	0.012
ChIP batch	0.31	0.097
ChIP batch with sequencing batch regressed out	0.47	0.075
Epstein–Barr virus load	0.12	0.022

Table 2.1: Correlations between PC1, PC2 and the experimental variables. In association tests PC1 was removed.

measures of binding for each individual can be confounded by a number of hidden factors due to either biological or technical factors, or both. We performed PCA and saw that the first factor explained 24.1% of the variance in the data, substantially more than later components (Figure 2.6). Further investigation of this component showed that it was correlated with ChIP batch date, and it was therefore removed (Table 2.1).

2.4 Imputing missing genotypes

Our 51 samples consist of 35 individuals present in the 1000 Genomes Phase 1 release (v3 20101123) (The 1000 Genomes Consortium, 2012), 11 individuals in the 1000 Genomes Pilot, 2 individuals in 1000 Genomes high coverage Trio (NA12891 and NA12892) and 3 individuals in the HapMap III (Stranger et al., 2012). The eleven 1000 Genomes Pilot samples have low coverage. We calculated the genotype likelihood for each of the Phase 1 sites using samtools (Li et al., 2009) and then performed imputation using BEAGLE (Browning and Yu, 2009) and IMPUTE2 (Howie et al., 2009) with the 1000 Genomes Phase 1 data as a reference panel. Using Illumina Omni 2.5M SNP array genotypes (available ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120131_omni_genotypes_and_intensities/) as a validation set, we obtained good accuracy from this procedure with a mean non-reference discordance rate of 2.33% and an average genotype dosage R^2 of 0.956. We also imputed the three HapMap III samples, using their genotype data on the Omni 2.5M array as

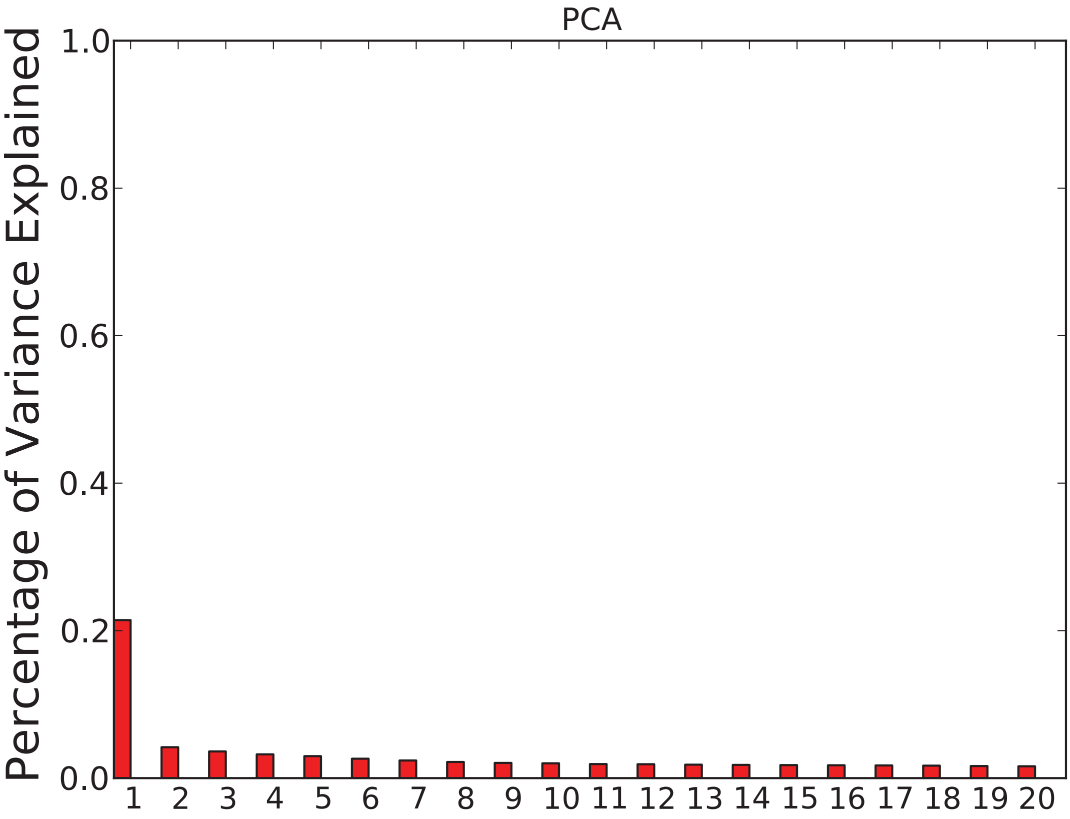

Figure 2.6: Proportion of phenotypic variance explained by each principal component (PC). We performed principal component analysis (PCA) on the normalized data to discover latent factors that explain large proportion of phenotypic variation. We saw that the first principal component explains substantially more variance than the others. When we looked at the correlation between the first principal component and technical and experimental variables, we found that it correlates with ChIP batch at ρ=0.47. The first principal component is removed from the data before further analysis.

the imputation panel and the 1000 Genome Phase 1 as the reference panel. We then integrated data from each source and obtained a consolidated genotype set for all 51 individuals. For association mapping, we filtered variants by requiring >5% minor allele frequency, P value for Hardy-Weinberg Equilibrium (HWE) >1E-4 and position within 50kb to either side of a binding region being mapped. The window size was chosen to be 100kb as we are primarily interested in *cis* regulation but also allowing possibility that there may exist multiple binding sites with variable affinity strengths in the window. Finally, 4,687,317 variants entered analysis, with 4,250,881 SNPs and 436,436 INDELs.

The 1000 Genome Phase 1 release gives a comprehensive ascertainment of the genetic variants. However, it is still possible that some variants private to this study cohort are yet to be found. To address this concern, we performed variant calling for the CTCF binding regions using ChIP-seq data. The calling was done by using samtools mpileup with parameters "-DV -C50 -q 30 -Q 30 -d 10000 –u -l $qtl_regions -b $bam_list -f $reference", followed by BCF tools with parameter "-t $qtl_regions -mv". This is independent from the previous variant calling and gives information private to the ChIP-seq data. We filtered on the quality of the calling by keeping only variants with QUAL score greater than 20. We also kept only the variants that are private to the new call set and are absent in the 1000 Genomes Phase 1 data. In the end, we obtained 4,756 variants are within binding regions with 2,282 SNPs and 2,474 INDELs. It is a small additional quantity compared to the variant set of the 1000 Genome Phase1 release, but is enriched for INDELs (52%). When we conducted the same association scan using only these additional variants, we discovered 55 QTL binding regions associated with 60 variants, out of which only 8 QTL binding regions are new and no variants were found within motif. Thus the effect of this additional variant set is minimum in our QTL scan.

2.5 Association testing

We applied linear regression for association testing. For each binding region, we tested the association between the binding intensities and the genotypes of the variants that are within 50kb of the binding region by linear regression: $y_{il} = \beta_k x_{lk} + \epsilon_{ilk}$, $i \in \{1, ..., 57428\}$ where y_{il} is the normalized binding intensity for the lth individual, x_{lk} is the genetic dosage, represented as the minor allele count, for variant k and individual l, and ϵ is the non-genetic noise term assumed to follow distribution $N(0, \sigma^2)$. The parameters $(\hat{\beta}, \sigma^2)$ can be fitted using maximum likelihood methods. For each loci k, we tested the null hypothesis $\beta = 0$ using test statistics $t = \frac{\hat{\beta}}{\sqrt{var(\hat{\beta})}}$.

We estimated the FDR by a q value method (Storey and Tibshirani, 2003), which establishes P<7.1E-5 as an FDR of 1%. We further filtered the associated SNPs by requiring the P value to be within one order of magnitude to that of the P value of the lead SNP. We report these cluster variants as associated to the target binding region. We also reported results when a more stringent Bonferroni threshold was applied. The threshold was calculated at a significant level of α=0.05 corrected for 13,293,727 tests, which gives 3.8E-9 for the actual threshold.

2.6 Allele specific analysis

Read counts at each allele were counted for the 5.6M SNPs within 50kb of a binding region. Heterozygous SNPs with significant allele-specific CTCF binding were identified. In detail, for each individual at each site, we calculated a binomial P value at all heterozygous SNPs with the null hypothesis that the two allele counts are equal. We then performed multiple testing adjustment for all heterozygous SNPs that have at least 2 reads at each allele and at least 2 reads difference between the two alleles using the Benjamini&Hochberg (Benjamini and Hochberg, 1995) method. Significant allele-specific binding was determined with an FDR 5%.

2.7 Results

Analysis of CTCF binding in 51 genotyped individuals reveals thousands of binding QTLs We performed ChIP-seq on extracted chromatin from genotyped LCLs as previously described (Lee et al., 2012) except that we sequenced the DNA fragments from both ends (Figure 2.7). We quantified binding to binding regions similarly to previous work (Lee et al., 2012) but pooled all the samples and identified a composite set of binding regions with detectable CTCF binding at low threshold. We then counted the sequence fragments that overlap each binding region in each individual, and normalized the signal to correct for systematic biases as in Degner et al., 2012. We discarded binding regions that showed very little inter-individual variance or had only one or two individuals with significant binding scores. Overall, our normalized data showed effectively enhanced signal noise ratio and motivated QTL analysis (Figure 2.7B, 2.9, and 2.11).

To measure the variance due to growth differences between the cells, we grew two individual cell lines as four independent cultures started on four consecutive days. There was higher correlation between these biological replicates from the same individual than between samples from different individuals, although all data sets were modestly correlated as expected for CTCF ChIP-seq (Figure 2.8). We next examined the data to see whether there were any systematic biases between samples. A principal component analysis identified some systematic variance, with a particularly strong first component (explained 24.1% of the variance, Figure 2.6) that on investigation was correlated to known experimental batches. We therefore removed the first principal component, significantly improving the recovery of QTLs (Figure 2.9). This is in general a good practice from previous studies, e.g. Degner et al., 2012, as methods such as PCA could not enhance random noise. We used the resulting normalized adjusted binding intensity (NABI) for subsequent analyses.

To discover QTLs, we correlated SNPs and small biallelic insertion or deletion (INDEL) variants within 50 kb of the binding region with the NABI metric, using a linear model (Table 2.2, example in Figure 2.10). As expected, the majority of variants do not have a significant association with variation in CTCF binding, with the linear model P-value distribution following the expected distribution (>95% of tests, fraction of the overlap between the black line and red line, Figure 2.11). When samples are permuted, the distribution of the test statistic falls on the expected line. Using a non-parametric statistic we saw similar P values (Figure 2.11). Using a Bonferroni adjusted threshold of $P < 3.8E-9$ we find 509 binding regions with significant QTLs. Using a

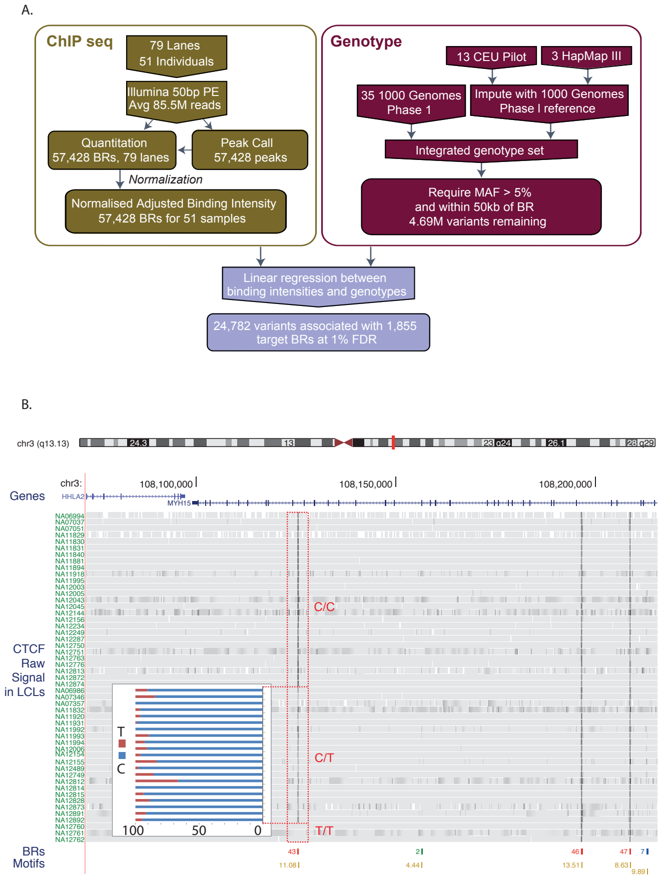

Figure 2.7: A. Flow chart indicating the overall design of the experiment. B. Overview of the binding intensities of a binding site across samples in three genotype groups of the associated SNP. ChIP-seq signal from the samples is aligned as tracks for this region of chromosome 3. The greyness is proportional to fragments mapped at the position, indicating binding intensity, with dark grey indicating high fragment count. Samples are grouped by their genotype at SNP rs936266, C/C, C/T or T/T, respectively. Binding sites were identified, as shown in the binding region track along with the number of samples passing the peak calling threshold. The colours of the binding regions represent the consistency of identifying the binding region across samples. Specifically, red binding regions were identified in 10 or more cell lines, blue binding regions in 5-9 cell lines and green binding regions in 2-4 cell lines. Finally the bottom track shows the corresponding CTCF motifs, with quality score attached to each site. The binding intensity decreases for T heterozygotes and further for T homozygotes. The inset panel shows allele-specific binding for the C and T allele (blue and red, respectively) in the heterozygous individuals (C/T) as percentage of the total count. Binding intensities consistently favour the C allele over the T allele.

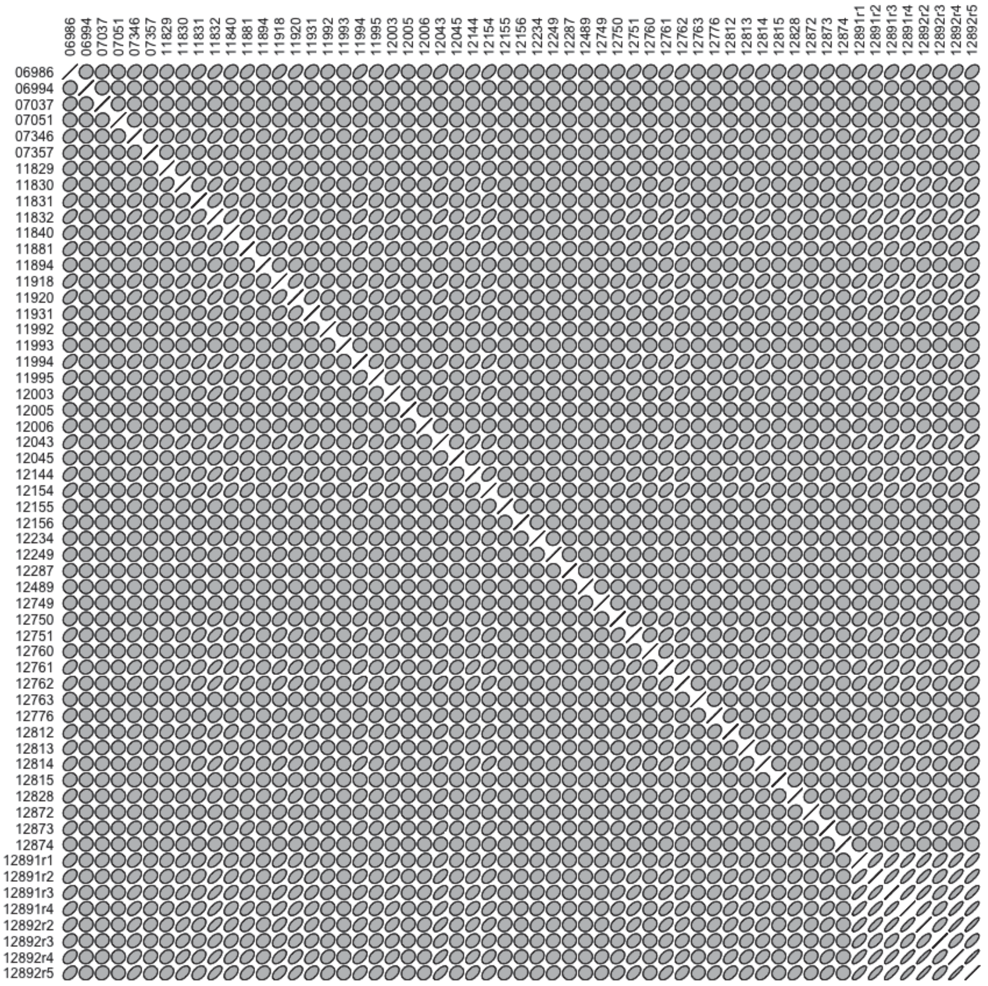

Figure 2.8: Higher correlation within day replicates compared to between different samples. We calculate the pair-wise Spearman correlation among all samples, including the two day-replicates, 12891 and 12892, shown as the last two sets of four samples. A diagonal line in each cell represents perfect correlation whereas a full circle represents no correlation. Increasingly flattened ellipses indicate a greater degree of correlation. When comparing among the day replicates, we obtained a correlation coefficient of 0.83 and 0.82 for GM12891 and GM12892, respectively. We also looked at the mean correlation of all the other samples and found a correlation of 0.17. Therefore we see much higher correlation within day replicates than that of all other samples.

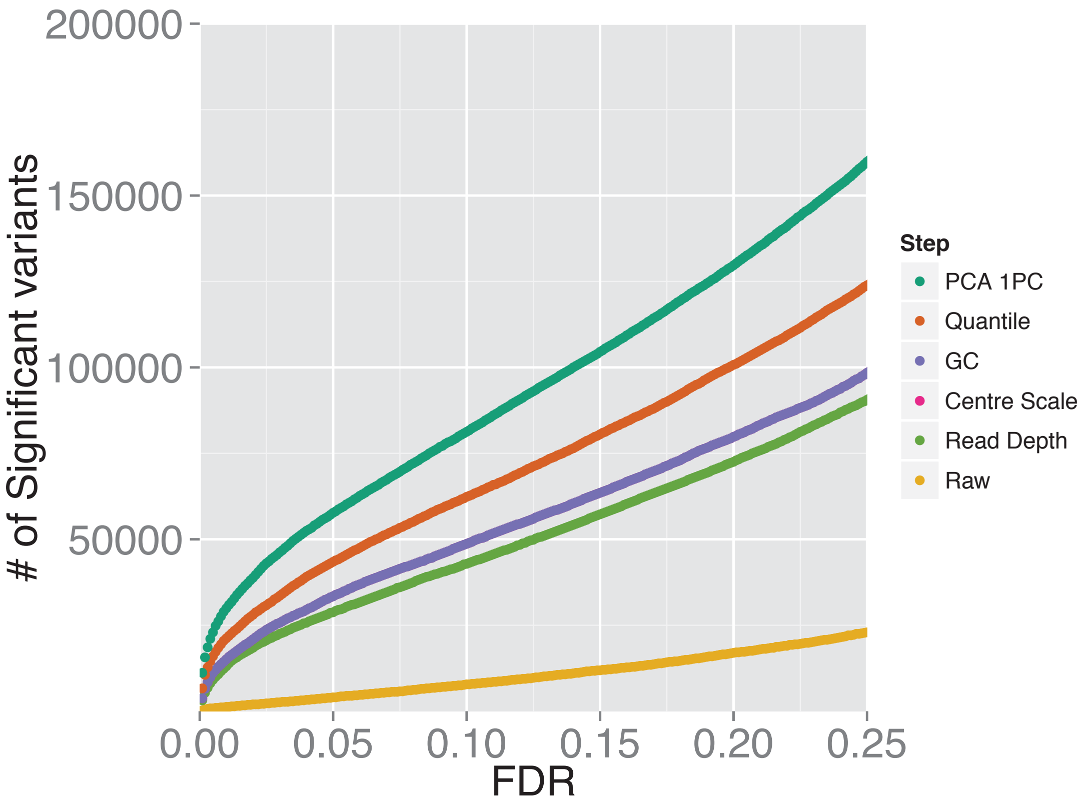

Figure 2.9: The number of significant QTLs found as a function of false discovery rate (FDR), plotted for the raw data and after each stage of the data normalization procedure that we used. We first normalised the binding intensities for each sample by the total read depth for that sample. We then corrected for GC composition by removing the median count of binding regions in the same GC bin (100 bins in total) from each binding region. The measures for each binding region were then centre-scaled by removing the mean and then dividing by the standard deviation (track hidden behind GC as center scale does not affect regression). This was followed by a quantile normalization, which maps the measures of each sample to normal quantiles across all binding regions. Lastly, we removed the first principal component that explains the most global phenotypic variation.

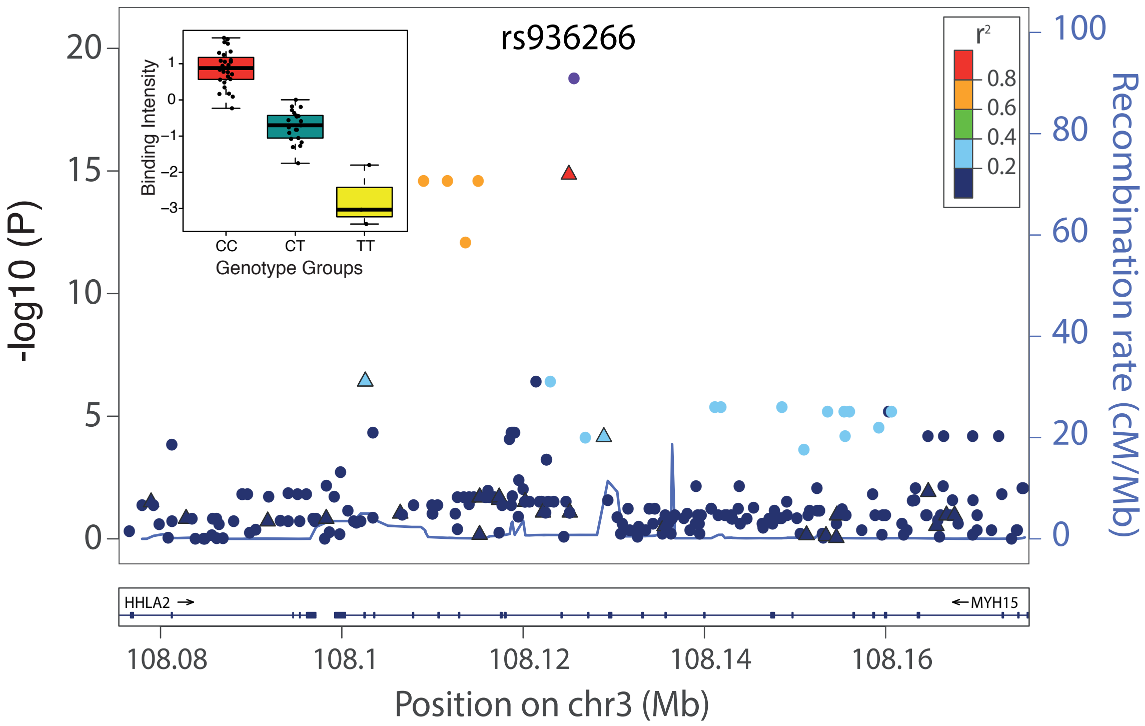

Figure 2.10: An example CTCF QTL. Here shows all associations for all variants in the region of the binding region at chr3:108125397-108125829. SNPs are shown as solid circles and INDELs are shown as triangles, colored by R^2. Inset is boxplot showing the normalized adjusted binding intensity (NABI) for the different possible genotypes of SNP rs936266. Genotype is strongly associated with the binding intensity of the binding region (P=1.69E-19), with the C allele favoring binding.

more liberal False Discovery Rate (FDR) (Storey and Tibshirani, 2003) approach to take advantage of the smaller number of effectively independent tests occurring in these limited *cis*-regions, we discovered 1,837 binding regions (3% of total binding regions) with at least one significant variant at the 1% FDR level; relaxing the threshold to 10% FDR we discover 6,747 binding regions (12% of the total) (Table 2.2).

We chose to focus further analysis on the 1% FDR threshold as this provided ample QTLs from which to derive insights. We only considered one association per binding region, because the small number of samples meant that there was insufficient power

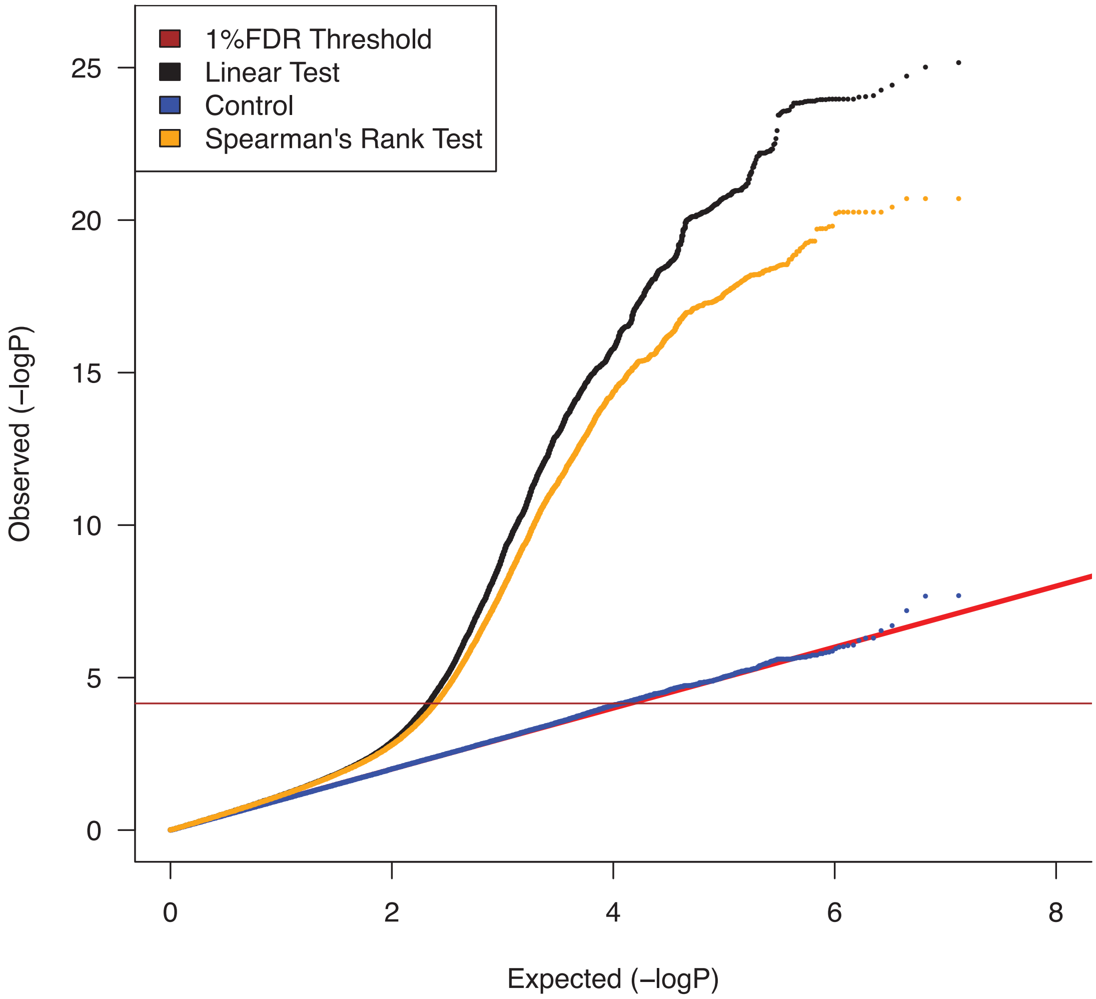

Figure 2.11: A Quantile-Quantile plot showing the distribution of the observed (y-axis) compared to the expected P values(x-axis). The red line is the distribution of the P values from the null model. The brown line on the y-axis shows the 1% FDR level determined by the q value method (Storey and Tibshirani, 2003). Black and blue dots indicate P values from the linear tests and permutation controls, where sample labels are randomly permuted. Association test by linear methods can be inappropriate and give spurious signal if the normality assumption is not met. Although in our normalization procedure the binding measures are mapped to normal quantiles sample-wise, it is still possible that the normality assumption does not hold binding region-wise. To test if this would bias the QTL mapping we performed the same tests using the Spearman's rank method (orange line). We see a slight elevation of the black line, suggesting the rank test is more conservative but would give similar results (1476 out of 1837 QTL binding regions overlap between two tests), and our linear test is mostly appropriate. The subsequent analysis is based on the discovery set from the linear test at a 1% FDR (brown line) threshold.

Study Parameters	
Traits (Binding Regions)	57,428
Variants	4,687,317
SNPs	4,250,881
INDELs	436,436
Study Results	
Binding Regions	1,837
Variants	24,534
SNPs	22,954
INDELs	1,580
GWAS overlaps	61
eQTL overlaps	366

Table 2.2: Summary statistics of the CTCF QTL scan.

for a conditional analysis for secondary associations in almost all cases. Within this set of associations, the genetic variant accounted for a substantial fraction of the variation in CTCF binding (median R^2 0.38, Figure 2.12). When comparing the effect sizes and the proportion of variance explained between QTL at 1% FDR and 10%FDR, the 1% FDR set has higher values (the average absolute value of beta = 1.1 (0.37-3.39)) than the 10% FDR set (average absolute value of beta = 0.80 (0.26-2.83), Figure 2.13).

We summarized the collective set of variants which might be involved in each binding region association as being the cluster of SNPs within one order of magnitude of the P-value of the lead variant. 24,534 variants were identified in at least one cluster at the 1% FDR level, 13.4 variants on average per binding region (Table 2.2). As expected, these variants were mainly clustered around the target binding region, and when a CTCF binding motif could be identified (1341 of the 1837 cases) and a cluster QTL variant was present in the motif, the frequency was correlated with the information content and the GERP score (Cooper et al., 2005) of the motif (Figure 2.14B), as seen previously (Maurano et al., 2012). This is not driven by any biases

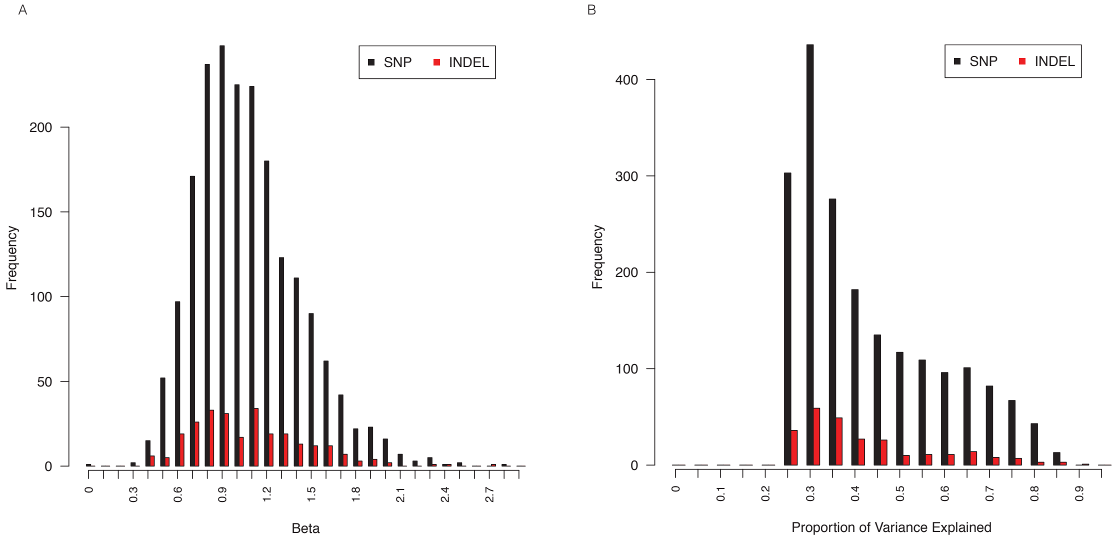

Figure 2.12: The distributions of the effect size (β) and proportion of phenotypic variance explained of the QTL variants.

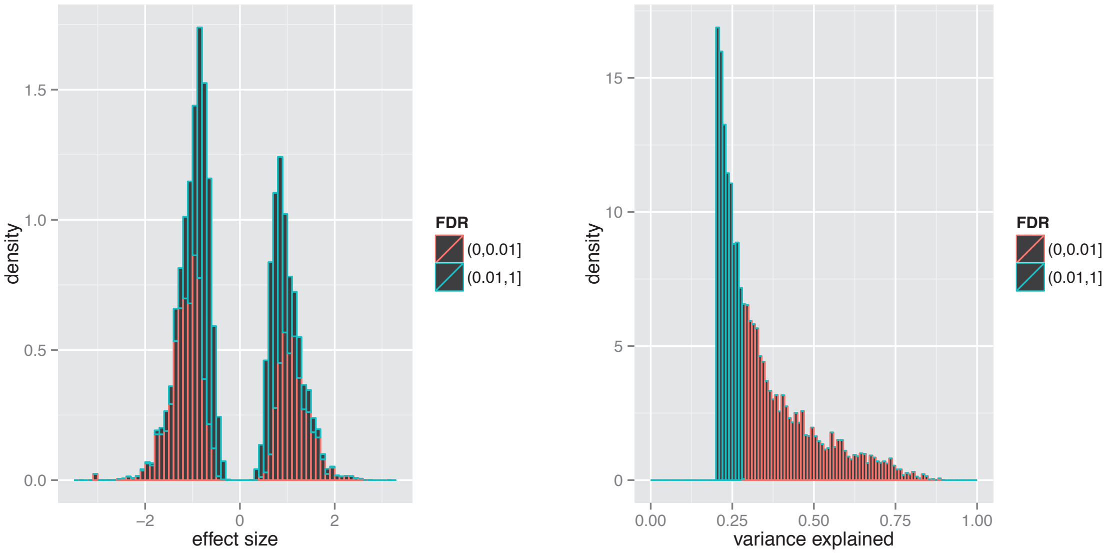

Figure 2.13: Effect sizes and proportion of variance explained of QTLs discovered at 1%FDR and 10%FDR.

Significance	Binding region count	Binding regions with motif	QTL in motif	≤1kb	≤10kb	≤30kb
10% FDR	6,747	5,260	453	1,386	2,583	4,057
1% FDR	1,837	1,341	344	747	1,023	1,199
Bonferroni	509	360	164	258	322	341

Table 2.3: CTCF QTLs with associated variants in different distance ranges.

in the distribution of the variants around the CTCF binding interface (Figure 2.14). However, only a minority of significant binding regions had a QTL candidate within the motif (344/1341), and in only a small majority of cases was there a QTL within 1kb (747/1341), of the binding region (Table 2.3). Considering that out of 45,668 binding regions that contain at least one motif, 2,090 (4.6%) binding regions have at least one variants on its binding motif. The QTL set shows strong enrichment of functional variants that are on motif: out of 1,341 that are motif containing, 344 (25.6%) have QTL variants within the motif.

We explored further the cases where there was no proximal variant in the cluster. There was not a substantial difference in genotype quality around the associated binding regions in these cases compared to binding regions with proximal effects, suggesting that there is not a large missing data problem. When considering all 1000 Genomes Project variants including those with allele frequency below 5%, in 95.5% of these cases, there was a proximal variant within 1kb of the binding region in linkage disequilibrium (LD) with the distal lead variant, where LD was defined as the absolute value of D' $>$ 0.5. In approximately half of these cases the P-value of the proximal association either fell just outside the one order of magnitude threshold to fall in the cluster, or was just under the FDR threshold (Figure 2.16). In the 99 such cases where such a proximal variant was within the CTCF binding motif, the position of the variant was correlated with the information content of the position in the motif (Figure 2.17). Therefore a substantial fraction of the apparently distal cases appear to be explained by proximal cases. However still only a minority can be explained by

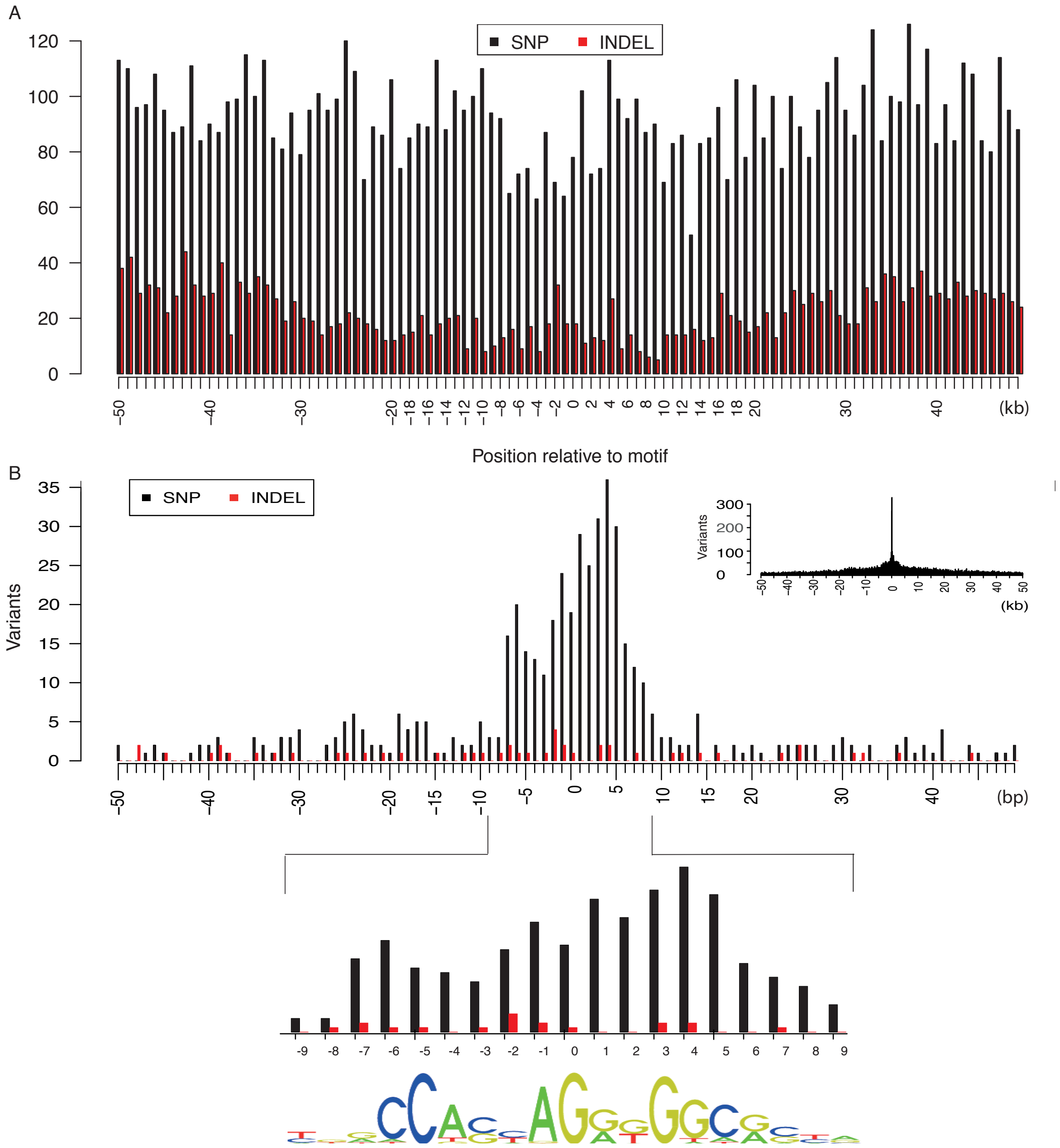

Figure 2.14: A. The distribution of variants (SNPs and INDELs) within a 50kb window over all binding regions that contain a motif. Y axis indicates the number of variants at a given position indicated by X axis with respect to the binding motif. Variants are uniformly distributed throughout the window, except a small reduction at the center corresponding to the high information content of the motif. B. The density of QTL variants with respect to distance from the motif of the associated binding regions. Density plots are shown at kb (inset) and base pair resolution (main plot). SNP and INDEL are shown as black and red bars respectively. For these cases the QTL density correlates with the information content of the motif (Spearman's rank ρ=0.63) shown at the bottom.

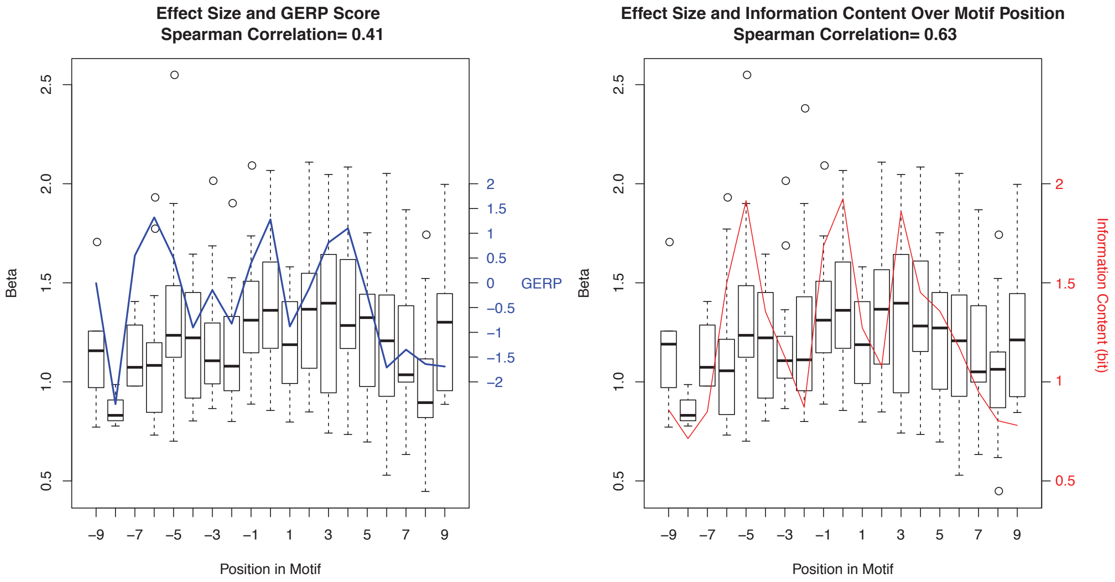

Figure 2.15: The QTL effect size correlates with the information content and the GERP score for the variants present in the motif .

variants in the binding motif itself. We looked at overlapping between the associated variants and other markers including DNase I hypersensitivity sites and transcription factor binding sites. We did not see a clear enrichment towards a particular motif. We also conducted the analysis excluding short INDELs to replicate the more commonplace association analysis using only SNPs. In an INDEL-free analysis we would have missed QTLs in 67 binding regions entirely (~5% of significant binding regions), and for 56 additional binding regions the closest observed explanatory SNP would have been over 1 kb away from the motif inside the peak. For these SNPs, there is usually a short indel with similar direct P-value inside the binding region. We further explored whether another cause for distal QTL effects could be due to the distal variant affecting a second neighbouring binding region of CTCF, which in turn influenced the primary binding region, but there was only one case where we could find any evidence for this model (Figure 2.18).

We additionally investigated the cases where there exist binding interactions between the QTL binding region and the neighboring region. For each of the four categories with sufficient abundance (model 1, 2, 3 and 4 in Figure 2.18), we compared the average signals between the QTL binding region (B1) and the neighboring binding region (B2) for a number of molecular markers using data obtained from the ENCODE project (?). We observed distinct patterns of regulatory signals between model 1,2 and model 3,4 (Figure 2.19). We saw that when there exists interactions between two binding regions (model 3,4), active transcription factors, enhancers and active histone markers tend to be more enriched in the QTL binding regions, as shown in red in Figure 2.19. This change is not driven by their distances being closer to the transcription start site (TSS) by chance, measured as the distance to the closest TSS, because the neighboring binding regions have similar distance to the TSS as the QTL binding regions (red and green lines in the density plots, Figure 2.19). We observed corresponding changes in histone modifications (H2AZ, H3k27ac, H3k4me1, H3k4me2 and H3k4me3) depending on the direction of the interactions between two binding regions (Figures 2.19 and Figure 2.20).

The effect size distribution with respect to allele frequency shows increased effect sizes for lower frequency SNPs, with a clear absence of large effects of common alleles (Figure 2.21). There is no statistical difference in effect size distribution between SNP and indel variants (Figure 2.21).

The dual-end sequencing of the ChIP-seq fragments provides the resolution to discover specific binding modes that influence the spatial distribution of the recovered fragments. To analyse this, we characterised ChIP-seq binding regions by metrics that summarised the extent of the peak and the position of the summit on a per individual basis, and used these additional metrics as phenotypes in a quantitative trait analysis using the methods described above. In detail, for each binding region in each sample, we measure the average left end, middle point and right end of sequencing fragments. The variation of the average positions across samples thus reflect variations in binding shapes across samples. Out of all 57,428 binding regions that were tested, we found 25 shifts in peak shape driven by a genetic locus at the 1% FDR. Ten cases were also associated with a change in peak height. An example is shown in Figure 2.22, with the two homozygous genotypes showing the creation of a new associated peak, and merging of a double peak, and from visual inspection the other cases also look as if

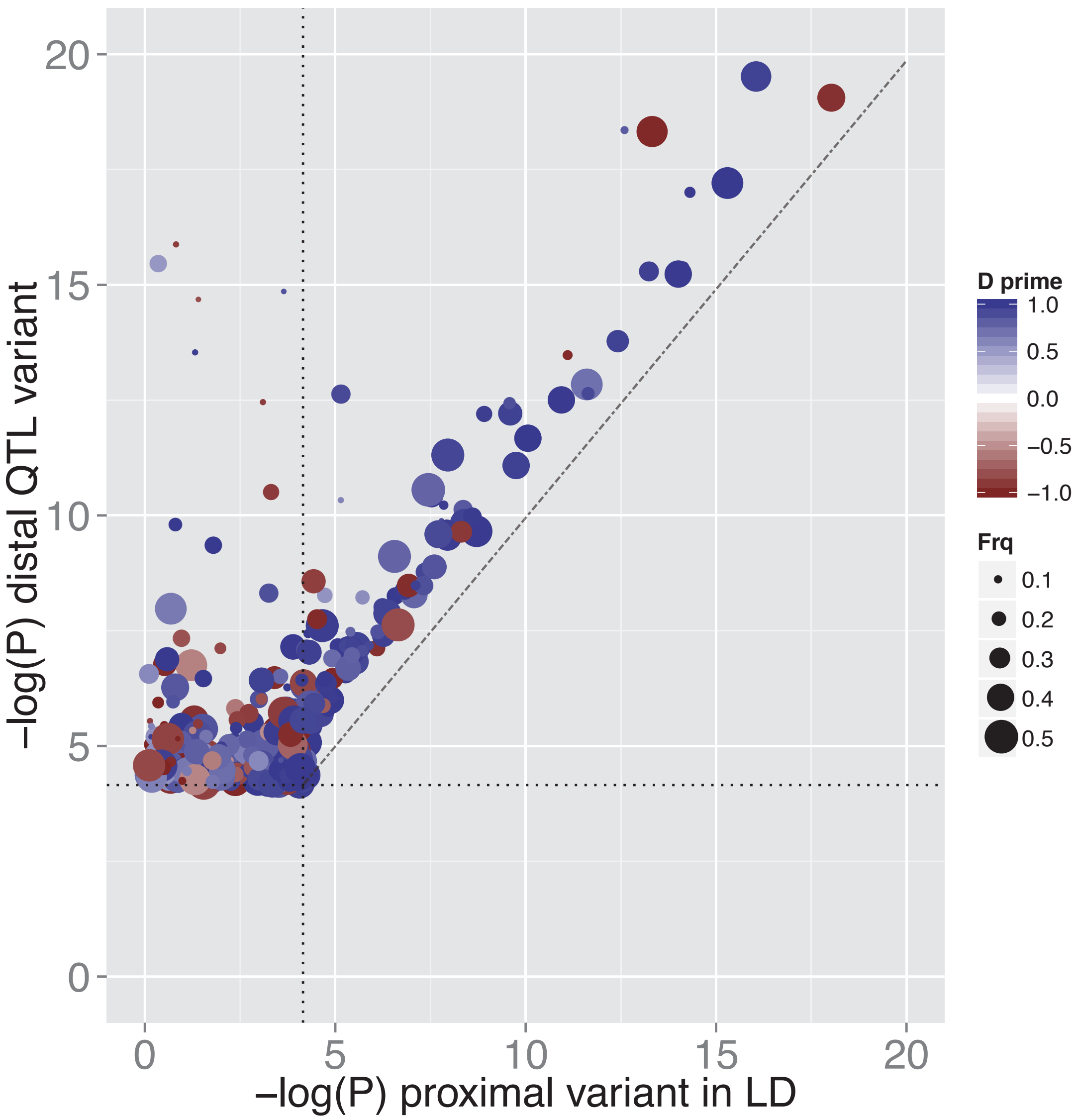

Figure 2.16: P value distribution of the proximal variants. Here the P values from the association between the CTCF binding and the lead distal QTL variants are plotted against that of the proximal variants, which are in LD with the distal QTL variants. The horizontal and vertical dashed lines are the 1% genome wide FDR threshold established in the main analysis. The diagonal line assists to indicate same P values. Each dot is colored by its D' value of LD with its size scaled by the allele frequency of the proximal variant.

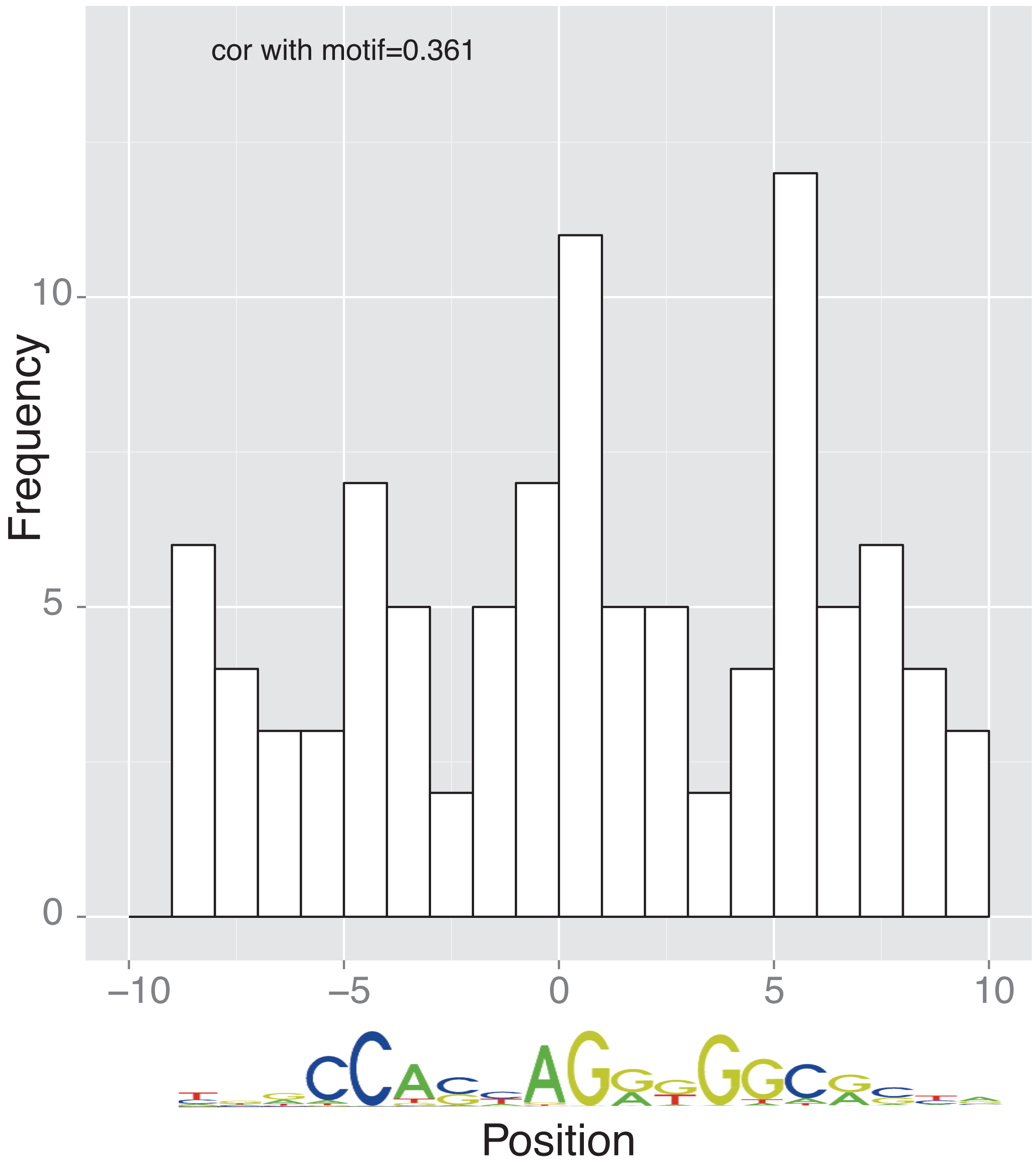

Figure 2.17: Distribution of the proximal variants that are on motif and in LD with the distal lead QTL variants. Here the proximal variants were aligned to the motif positions. We saw a correlation between their distribution and the information content of the motif at $\rho = 0.36$.

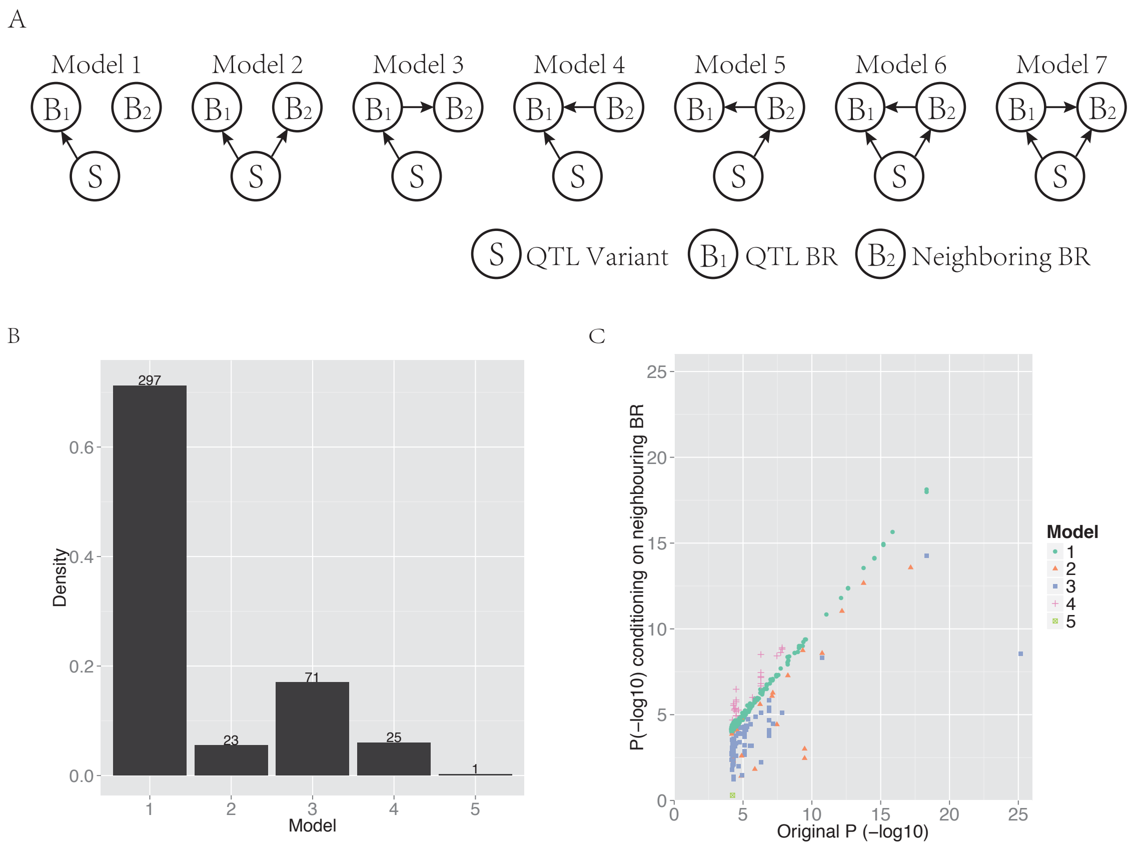

Figure 2.18: Evidence for indirect effects when a second binding region is present in the distal QTL window. Many (75.5%) of our distal QTLs contain a second CTCF binding region in their 50kb *cis*-window. To explore possible causal relationships between the lead variant, the associated binding region(BR1) and the second binding region(BR2) we constructed seven graphical models (A) and compared them using the Bayesian Information Criterion (BIC). In each case we assign the most likely model, chosen as having the lowest BIC (AIC showed same results). The frequency of the chosen models (B) suggests that there is almost never evidence for the association effect of the distal variant being mediated via a secondary binding region. The most frequently preferred model (1) did not involve BR2 at all; for the next most preferred models (3 and 4) there was some evidence of interactions between neighbouring CTCF binding sites, but we could not explain the variant association to BR1 binding via BR2. The only models which support mediation of binding at BR1 via BR2 are 5 and 6, and in only one case do we see one of these being selected. The P value of BR1 when conditioned on BR2 is plotted in (C). We further investigated the enrichment of a range of ENCODE (**?**) signals over the QTL binding region and the neighboring region. We found the association between two binding regions (model 3,4) tend to correlate with the active regulatory signals (Figure 2.19).

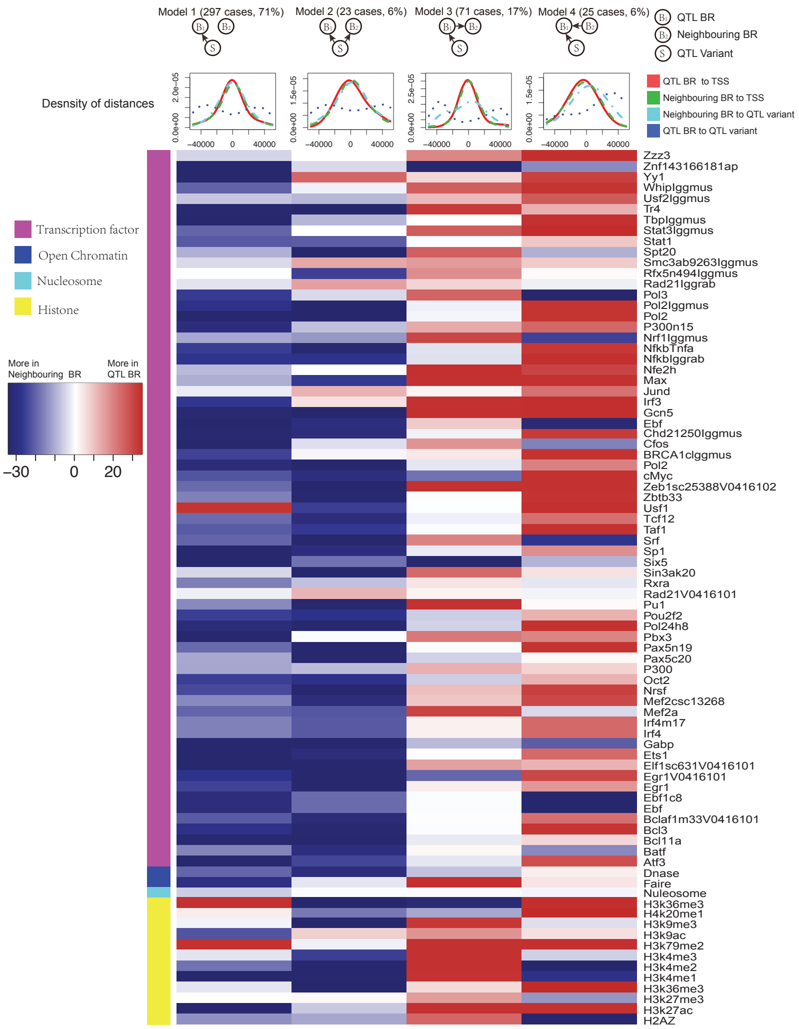

Figure 2.19: The interaction between QTL binding region and neighboring binding region correlates with regulatory events. The distal QTL set is as previously described (Figure 2.18). For each of the four categories with sufficient abundance (model 1, 2, 3 and 4), we compare the average signals between the QTL binding region (B1) and the neighboring binding region (B2) for a number of molecular markers using data obtained from the ENCODE project (Birney et al., 2007). We observed distinct patterns of regulatory signals between model 1,2 and model 3,4. We saw that when there exists interactions between two binding regions (model 3,4), active transcription factors, enhancers and active histone markers tend to be more enriched in the QTL binding regions, as shown in red. This change is not driven by their distances being closer to the transcription start site (TSS) by chance, measured as the distance to the closest TSS, because the neighboring binding regions have similar distance to the TSS as the QTL binding regions (red and green lines in the density plots). Some of the histone modifications (H2AZ, H3k27ac, H3k4me1, H3k4me2 and H3k4me3) swap enrichment direction between model 3 and model 4 depending on the direction of interaction between B1 and B2 (also see Figure 2.20 for more detailed enrichment signals).

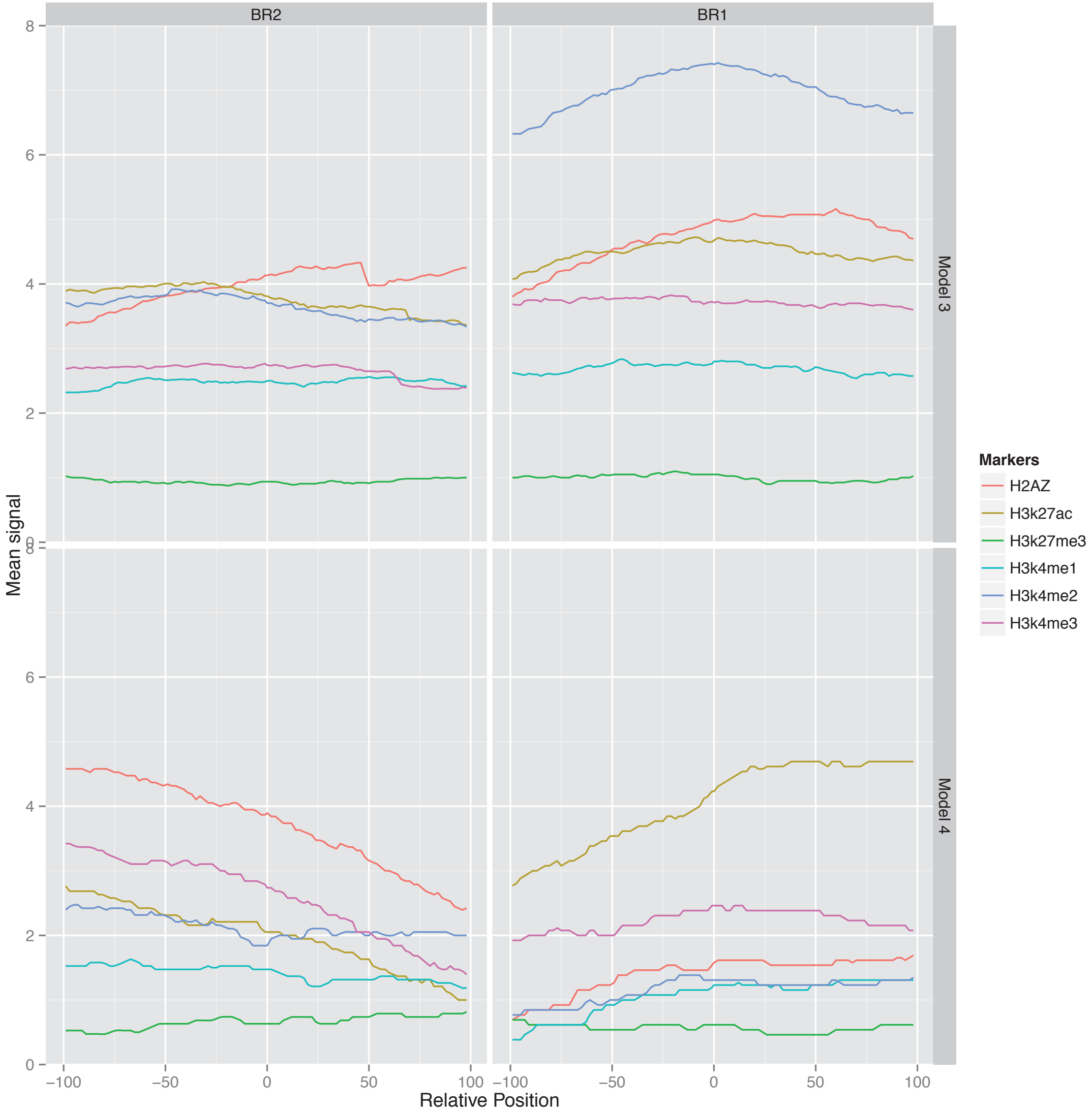

Figure 2.20: Change of histone modifications depending on the interaction models between the QTL binding region and the neighboring binding region (see Figure 2.18 and Figure 2.19 for explanations about the models).

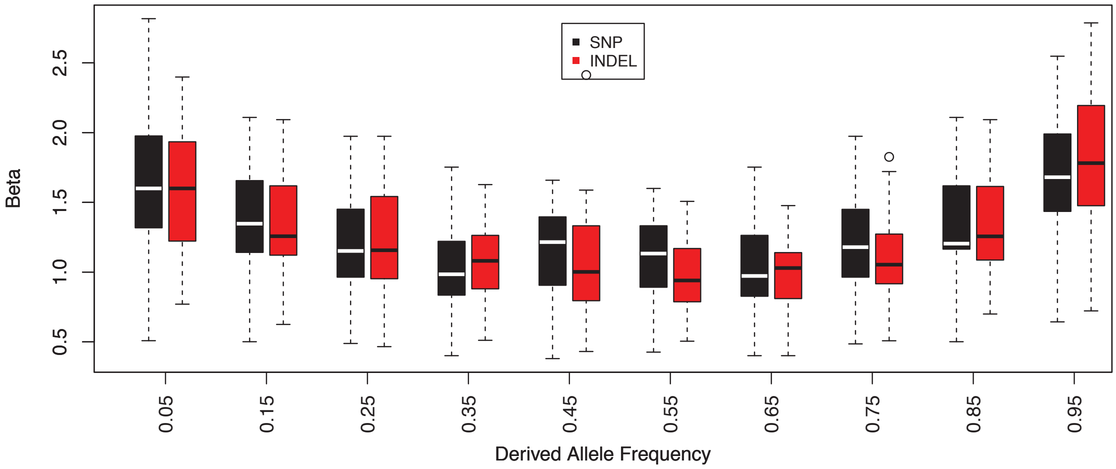

Figure 2.21: Effect size versus derived allele frequency for all CTCF QTLs identified at 1 % FDR.

they can be explained as two CTCF peaks in close proximity, one or both of which is under *cis*-genetic control.

There are 61 CTCF QTL variants that overlap with disease and trait associated variants from other studies (Table 2.4, GWAS Catalog Hindorff et al., 2009). In particular there is a disproportionate overlap with immune system related diseases (20 variants; χ^2 P-value 1.7E-9). This is consistent with the lymphocyte origin of LCLs, and may suggest roles of CTCF binding in the disease phenotypes.

In summary, these results are consistent with previous studies(Kasowski et al., 2010; Maurano et al., 2012; Reddy et al., 2012; Stefflova et al., 2013) that observed substantial variation in transcription factor binding within and between species, only a minority of which could be accounted for by genetic differences in the binding motif. We also found that only 25.7% of our QTLs could be explained by a genetic variant in the motif. The majority of the remainder can be explained by changes within 1kb of the motif, consistent with observations that transcription factor binding differences between mouse strains are more likely if there are genetic differences within 200bp of

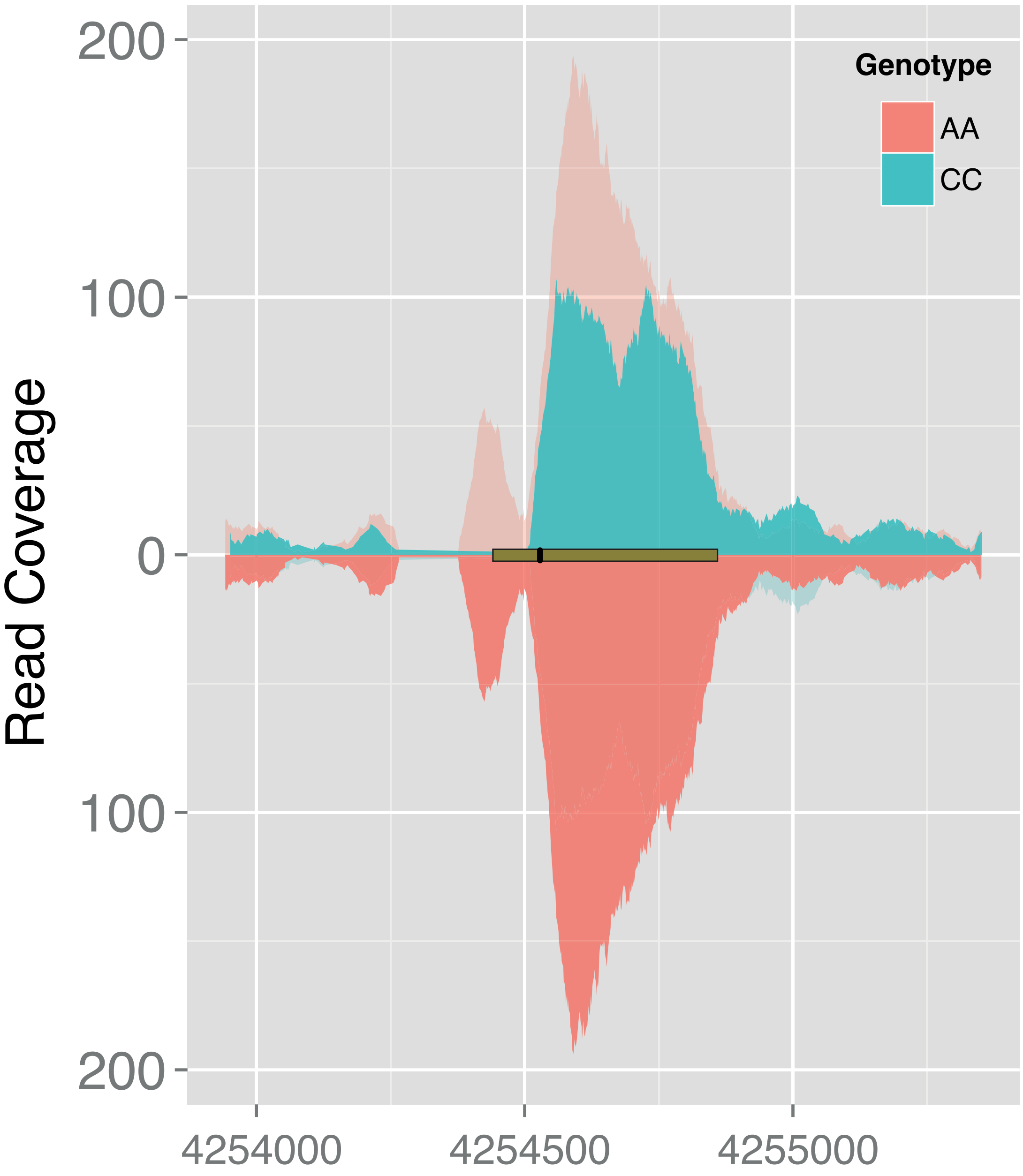

Figure 2.22: Example of CTCF peak shape QTL. Reads for samples in each homozygous genotype group at QTL rs11935835 were merged (AA and CC, respectively), and the average CC genotype profile is plotted above the main axis (in green), and the average AA genotype profile below (in red); each plot is reflected on the other axis in a lighter colour to allow visual comparison. The AA genotype has stronger overall binding, with a second peak to the left, whereas the CC genotype has a double peak. The heterozygote has intermediate profile between these two (not visualized in this figure). The binding region is marked as a brown box with the SNP position marked by a black vertical dash.

the binding site (Heinz et al., 2013). However there remain some genetic associations for which we are not able to identify any proximal candidate, suggesting that longer range influences can make some contribution to CTCF binding. Using published gene expression data for a subset of these samples, we looked at correlations between CTCF bindings and expression levels of nearby genes. We did not see strong correlations between the two, suggesting more complex role of CTCF in influencing genes.

Table 2.4: The overlap between CTCF QTL variants and GWAS variants.

PMID	Disease/Trait	CHR	SNP	-log(P_{GWAS})	Binding region start	-log(P_{QTL})
18204098	Systemic lupus erythematosus	8	rs13277113	10	11339579	4.891
17611496	Asthma	17	rs7216389	10.046	38028921	14.668
21627779	Alzheimer's disease	11	rs1562990	10.398	60018960	15.264
23263486	Urate levels	5	rs17632159	10.398	72431291	5.383
19079260	Body mass index	1	rs2568958	11	72796185	23.913
19079260	Body mass index	1	rs2568958	11	72808237	20.605
19079260	Body mass index	1	rs2568958	11	72794697	20.529
17554300	Type 1 diabetes	12	rs11171739	11	56435260	4.979
22396660	Nephrolithiasis	5	rs11746443	11.046	176797734	4.184
21460841	Alzheimer's disease (late onset)	11	rs4938933	11.097	60018960	14.791
21102463	Crohn's disease	6	rs415890	11.523	167411229	5.477
19430480	Type 1 diabetes	17	rs2290400	12.222	38028921	14.668
23222517	Red blood cell traits	22	rs5749446	12.523	32870620	4.932
21804548	Asthma	12	rs1701704	12.699	56435260	4.931
22561518	Vitiligo	12	rs2456973	13.523	56435260	4.931
22423221	Mean platelet volume	6	rs210134	14.699	33546527	5.42
22700719	Chronic lymphocytic leukemia	6	rs210142	15.046	33546527	5.42
21833088	Multiple sclerosis	16	rs7200786	16.046	11196016	4.233
21829393	Type 1 diabetes autoantibodies	12	rs1701704	17.301	56435260	4.931
17554260	Type 1 diabetes	12	rs2292239	19.699	56435260	5.796
23128233	Inflammatory bowel disease	6	rs1819333	20.155	167411229	5.477
23128233	Inflammatory bowel disease	21	rs7282490	25.699	45659281	4.193
21829393	Type 1 diabetes autoantibodies	12	rs2292239	26.523	56435260	5.796
21149283	Iron status biomarkers	11	rs236918	27	117051957	5.29
22139419	Platelet counts	6	rs210134	35.155	33546527	5.42

PMID	Disease/Trait	CHR	SNP	-log(P_{GWAS})	Binding region start	-log(P_{QTL})
22504420	Bone mineral density	7	rs6959212	37.398	38110179	8.869
21079607	Anorexia nervosa	3	rs6782029	5.046	11661145	4.572
23251661	Obesity-related traits	3	rs1044826	5.097	139072763	4.783
18839057	Attention deficit hyperactivity disorder	16	rs11646411	5.155	82772300	4.259
23192594	Body mass index (interaction)	18	rs11876941	5.301	50906413	9.637
22589738	Subcutaneous adipose tissue	1	rs990871	5.398	72794697	20.061
22589738	Subcutaneous adipose tissue	1	rs990871	5.398	72808237	19.855
22365631	Temperament (bipolar disorder)	21	rs2150410	5.398	40547111	14.838
23319000	Metabolite levels (HVA/MHPG ratio)	2	rs6750634	5.398	50763433	14.085
23251661	Obesity-related traits	17	rs1051424	5.523	57976434	11.83
21998595	Height	6	rs2224391	5.523	5261260	11.231
23319000	Metabolite levels (HVA-5-HIAA Factor score)	8	rs13251954	5.699	29034453	4.16
20195514	Primary tooth development (time to first tooth eruption)	17	rs9674544	6.097	47091576	4.54
20862305	Type 2 diabetes	15	rs1436955	6.155	62417944	4.854
22797727	Renal function-related traits (sCR)	5	rs12654812	6.301	176797734	4.943
21833088	Multiple sclerosis	5	rs4075958	6.301	176797734	4.344
21408207	Systemic lupus erythematosus	8	rs2736340	6.523	11339579	4.891
22451204	Parkinson's disease	2	rs6430538	6.699	135540345	15.085
22797727	Renal function-related traits (eGRFcrea)	5	rs12654812	6.699	176797734	4.943
20228799	Ulcerative colitis	17	rs8067378	7	38028921	14.668
19023125	Brain imaging in schizophrenia (interaction)	5	rs245201	7.046	127169342	5.004
21118971	Small-cell lung cancer	11	rs716274	7.046	103408608	4.285
19079261	Body mass index	1	rs2815752	7.222	72796185	23.913
19079261	Body mass index	1	rs2815752	7.222	72808237	20.605
19079261	Body mass index	1	rs2815752	7.222	72794697	20.529
20596022	Alopecia areata	12	rs1701704	7.523	56435260	4.931
19079260	Weight	1	rs2568958	7.699	72796185	23.913
19079260	Weight	1	rs2568958	7.699	72808237	20.605

PMID	Disease/Trait	CHR	SNP	-log(P_{GWAS})	Binding region start	-log(P_{QTL})
19079260	Weight	1	rs2568958	7.699	72794697	20.529
23128233	Inflammatory bowel disease	5	rs12654812	7.699	176797734	4.943
19165918	Systemic lupus erythematosus	8	rs2618476	7.699	11339579	4.891
20195514	Primary tooth development (number of teeth)	17	rs9674544	7.699	47091576	4.54
18464913	Protein quantitative trait loci	11	rs7112513	8.222	117051957	5.29
19503088	Rheumatoid arthritis	8	rs2736340	8.222	11339579	4.891
20881960	Height	7	rs6959212	8.699	38110179	8.869
19801982	Bone mineral density (spine)	7	rs1524058	9	38110179	9.083
23291587	Behcet's disease	12	rs2617170	9	10563751	5.679
21459883	Dilated cardiomyopathy	1	rs10927875	9	16321009	4.468
18198356	Type 1 diabetes	12	rs1701704	9.046	56435260	4.931
22446961	Kawasaki disease	8	rs2736340	9.046	11339579	4.891
22139419	Mean platelet volume	3	rs10512627	9.301	124339333	9.465
19820697	Hematological parameters	22	rs9609565	9.398	32870620	4.943
23128233	Inflammatory bowel disease	14	rs194749	9.523	69255227	4.287
21943158	Cardiovascular disease risk factors	11	rs508487	9.699	117051957	4.597
22001757	Liver enzyme levels (alkaline phosphatase)	8	rs6984305	9.699	9178038	4.199
23251661	Obesity-related traits	7	rs11976180	5.15490196	143760743	6.02128023
23064961	Dental caries	13	rs735539	5.397940009	21285708	7.303681407
22566498	Response to angiotensin II receptor blocker therapy	11	rs11020821	6.045757491	94234754	5.607190203
22566498	Response to angiotensin II receptor blocker therapy (opposite direction w/ diuretic therapy)	11	rs11020821	5.397940009	94234754	5.607190203
22561518	Vitiligo	11	rs4409785	12.69897	95311198	7.479585622
22001757	Liver enzyme levels (gamma-glutamyl transferase)	1	rs10908458	14.69897	155085124	5.091098739
21833088	Multiple sclerosis	11	rs4409785	6.22184875	95311198	7.479585622
21037568	Hodgkin's lymphoma	2	rs1432295	7.698970004	61066413	7.269570862

PMID	Disease/Trait	CHR	SNP	-log(P_{GWAS})	Binding region start	-log(P_{QTL})
20972438	Bladder cancer	4	rs798766	12.39794001	1731408	5.210110236
20708005	Non-alcoholic fatty liver disease histology (other)	13	rs1305088	5.045757491	29252925	4.210516688
20395239	Optic disc size (cup)	12	rs10858945	5.22184875	90456729	4.268051964
20348956	Urinary bladder cancer	4	rs798766	11	1731408	5.210110236
19343178	Height	7	rs849141	10.52287875	28182178	4.157298252
19197348	Quantitative traits	7	rs2527866	5.522878745	157091071	7.170945424
18759275	Uric acid levels	3	rs6442522	5.301029996	15440342	17.31606037

Allele-specific bias analysis of CTCF binding provides independent confirmation of QTLs This data set represents an excellent resource to directly examine allele-specific biases in TF binding at heterozygous sites in a larger set of individuals than previous studies (McDaniell et al., 2010). Allele-specific binding refers to statistically significant biases in binding to the two alleles in a diploid cell, at sites where a heterozygous polymorphism allows the two alleles to be distinguished. Allele-specific binding thus is an independent way of assessing how genetic variants at binding sites might affect binding variation. Although the two alleles at heterozygous SNPs are normally referred to as the reference or alternate allele (referring to which base is found in the reference genome sequence and which is the alternate base), here we chose to categorize the two alleles as ancestral (shared with chimp) or derived (human specific). This has two advantages. First, any residual effect of biases in aligning sequence reads to the reference allele will be minimized. Second, measuring allele-specific binding in terms of the ancestral and derived allele provides information about how evolutionary changes might affect CTCF binding.

After processing the reads, we identified allele-specific sites using a binomial null model of equal occupancy of both alleles at heterozygous sites, using a 5% FDR corrected threshold, similar as described previously(McDaniell et al., 2010). Allele specific variants were identified using reads pooled across all individuals for each allele. This process identified 589 SNPs that have replicated in at least two individuals showing significant allele-specific bias. We examined the allele counts of all heterozygous individuals at these 589 SNPs. For most sites (91.5%) the allele-specific biases were consistent between individuals, confirming the predominantly genetic basis of allele-specific binding (Figure 2.23). At such sites, the same ancestral or derived allele was preferred for binding across 2 or more individuals.

However, there were 50 (8.5%) sites which showed significant but opposite allele-specific biases between two or more individuals. Six of these 50 sites could potentially be explained by virtue of being close to loci known to be subject to allelic exclusion (the Immunoglobulin heavy chain), a process that affects one allele randomly (see Discussion). One site lies in the KCNQ1 imprinted locus, where the regulatory status depends on parent of origin rather than genotype. The 46 other sites at which the allele- specific binding bias switches between individuals (Appendix Table A.1) could represent new random allelic exclusion loci or imprinted sites, or could arise because the site at which we see allele specificity is incompletely linked with the causal variant (Lappalainen et al., 2013). We tested whether there was a SNP which specifically explained the allele specific switching site; for 28 cases this was the case. We are not able to directly test whether any of these sites could be due to imprinting because parent-of-origin information is not available for the heterozygous alleles of these individuals.

Interestingly, a significant majority (68%, P <1E-16) of the SNPs showed increased binding to the ancestral allele (Figure 2.23). Alignment bias towards the reference allele has been reported before (McDaniell et al., 2010) and because the ancestral allele is more likely to be the reference allele, the increased binding to the ancestral allele could be the result of the alignment bias. To rule out this possibility, we analyzed the cases where the ancestral allele is the alternate allele and found that the binding bias remained towards the ancestral allele (Figure 2.25). Additionally, we repeated the allele-specific analysis after using a variant-aware aligner. The results were largely identical to what we observed as described above, indicating that the preference for the ancestral allele is not a trivial outcome of any alignment bias (Figure 2.26).

The allele-specific signal at binding regions (intra-individual measurements) mostly correlated linearly with the QTL effect size (inter-individual measurements) (Figure 2.24). There were however exceptions to this, and these were mainly cases in which there was an allele-specific signal but not inter-individual QTL. We observed QTLs

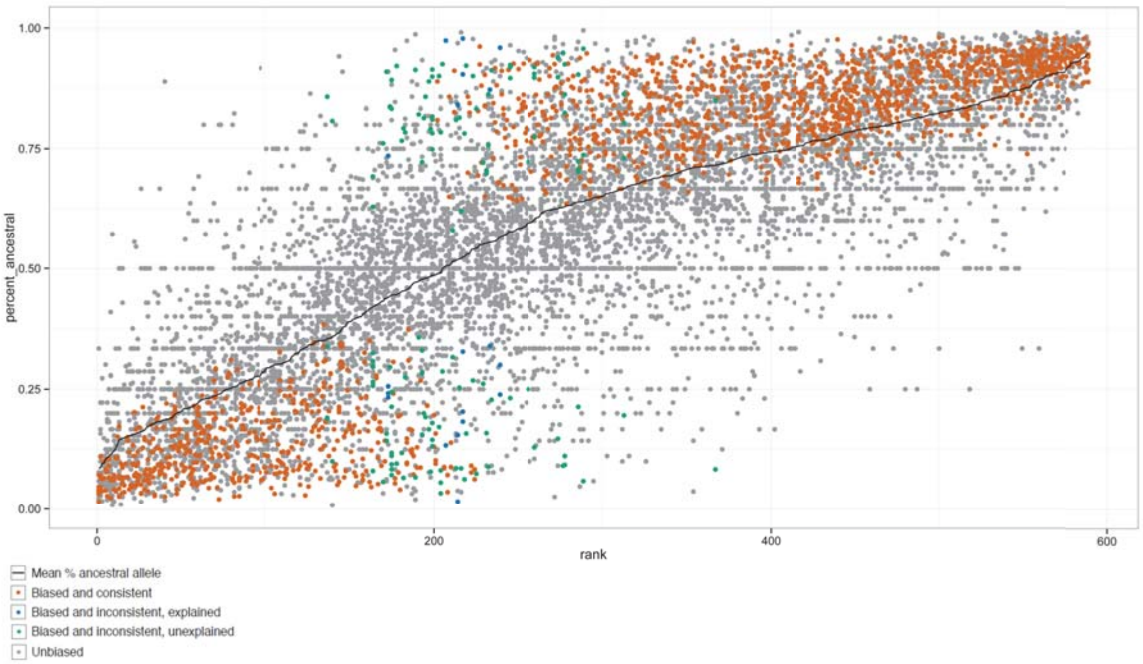

Figure 2.23: Summary of allele-specific analysis. SNP loci that show significant allele-specific CTCF binding in at least 2 samples are included. The y-axis represents the proportion of the total read counts from the ancestral allele. The 589 SNP loci are ordered by mean proportion ancestral allele for all heterozygous samples (black line). Heterozygous samples that do not pass the allele-specificity threshold are shown as gray points. Significant and consistent allele-specific samples (ie. the binding bias is toward the same allele) are represented by orange points. Significant but inconsistent samples are either blue (inconsistency explained by the nature of the site) or green (inconsistency unexplained).

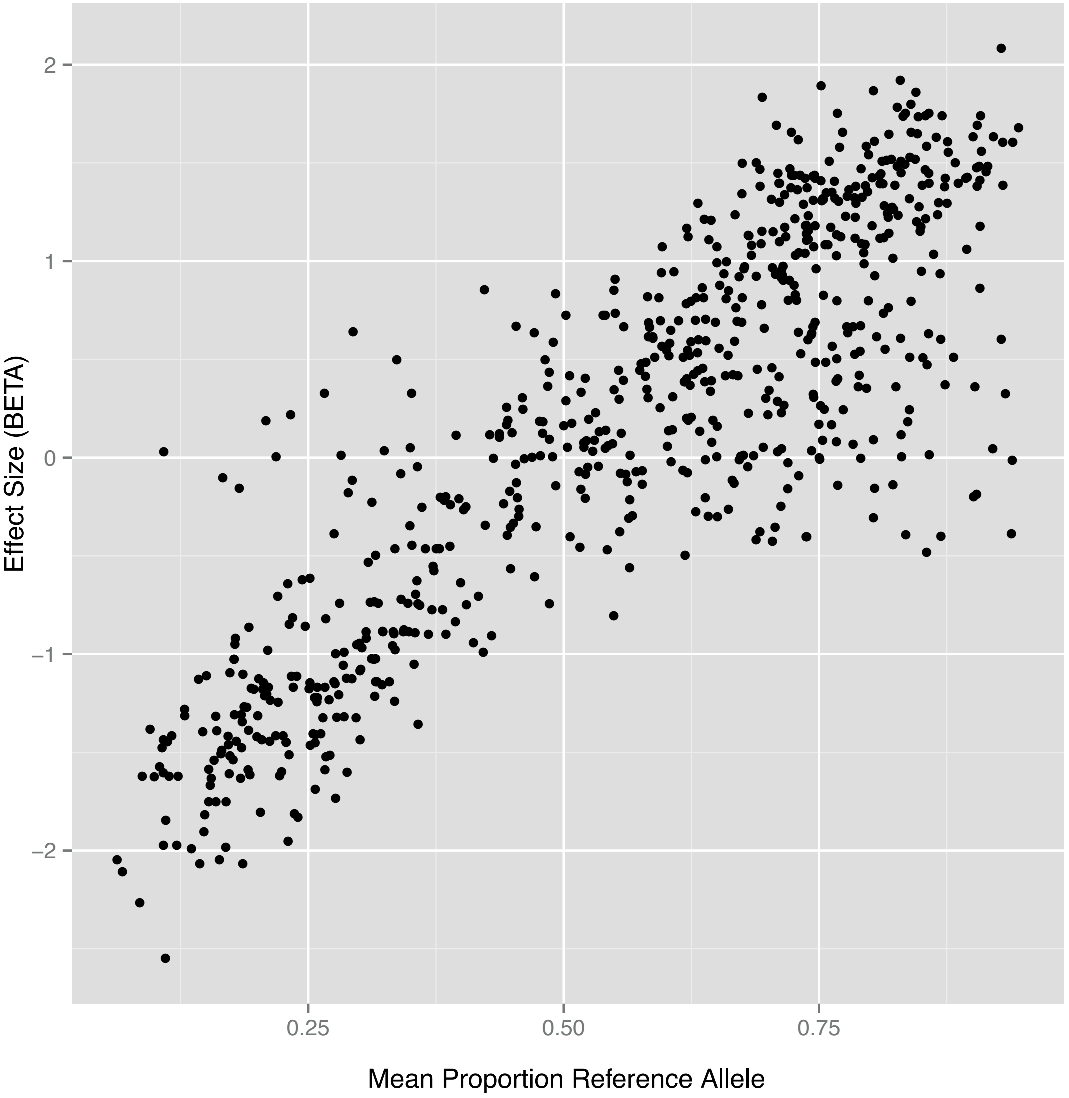

Figure 2.24: Allele-specificity correlates with QTL effect size (β). The mean proportion reference allele count for all heterozygous samples at SNP loci that show significant allele-specificity in at least 2 samples are plotted against the QTL effect size (β) at that locus. Only the β values from associations where the SNP is located within the associated binding region are shown.

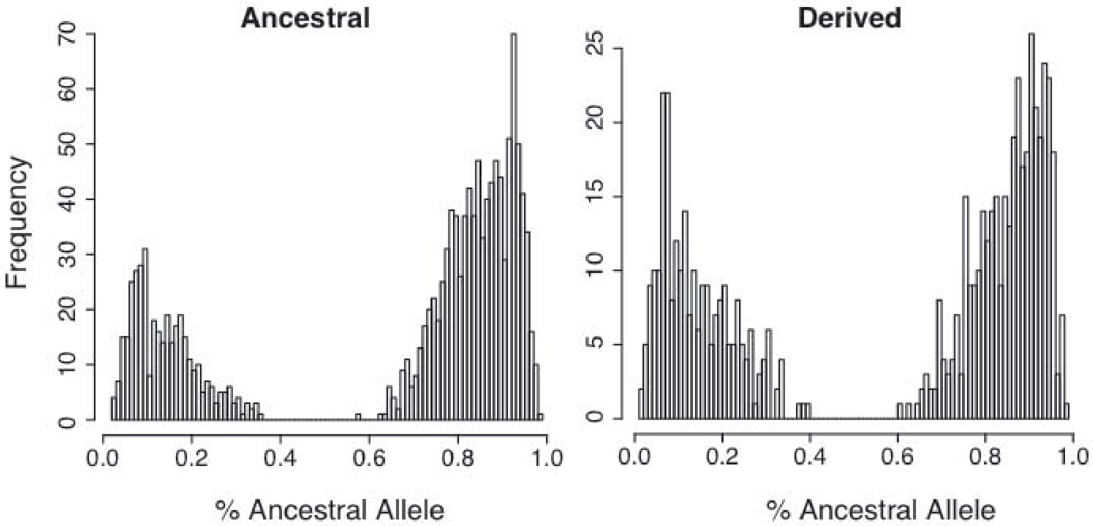

Figure 2.25: Effect of the reference allele. Even when the reference allele is the derived allele (Derived), the binding bias remained towards the ancestral allele.

with strong effect size in binding regions tend to show strong allele-specificity (Figure 2.27).

2.8 Discussion

This study is the first association QTL study performed on transcription factor binding in humans to our knowledge. The single site QTL properties are consistent with and extend other studies such as the family based studies (McDaniell et al., 2010; Maurano et al., 2012), the DNase I QTL (Degner et al., 2012). We find a large number of QTLs, with the majority being within or close to the binding region, and approximately a quarter inside the bound CTCF motif. By using the 1000 Genomes Project cell lines, we can be reasonably confident that we have a full catalog of common variation of which some subset are the causal variants. Using this information we could show that for a large fraction of the associations where the initial analysis suggested a distal variant more than 1kb away, there was a plausible causal candidate also within 1kb of the binding motif. Overall this suggests that, at least for CTCF, the substantial majority (~75%) of common genetic variants in the region with a reasonably strong effect on transcription factor binding lie within 1kb of the binding motif, although only a minority are actually within the motif. This clarifies previous observations that genetic variants contributing to transcription factor binding (CTCF and many others) were typically not in the motif itself (Kasowski et al., 2010; Stefflova et al., 2013) but there was enrichment nearby (Heinz et al., 2013).

These results suggest that the regulatory mechanism is not readily explainable by a simple regulation model. When we overlap CTCF QTLs with binding of other transcription factors that are measured by the ENCODE project, CTCF QTL variants that are not within the canonical motif are characterized by a modest enrichment (approximately 2 fold compared to random) of H3K4me3 and other transcription factors, such as PU1, Rad21, Pol II, ZNF1, YY1, and USF1. In these cases it is possible that the effect may be mediated via collaborations between these factors and CTCF. There is a small fraction (1.5%) of CTCF QTLs overlap *cis* eQTLs discovered in previous studies, indicating limited *cis* genetic effect targeting both CTCF binding

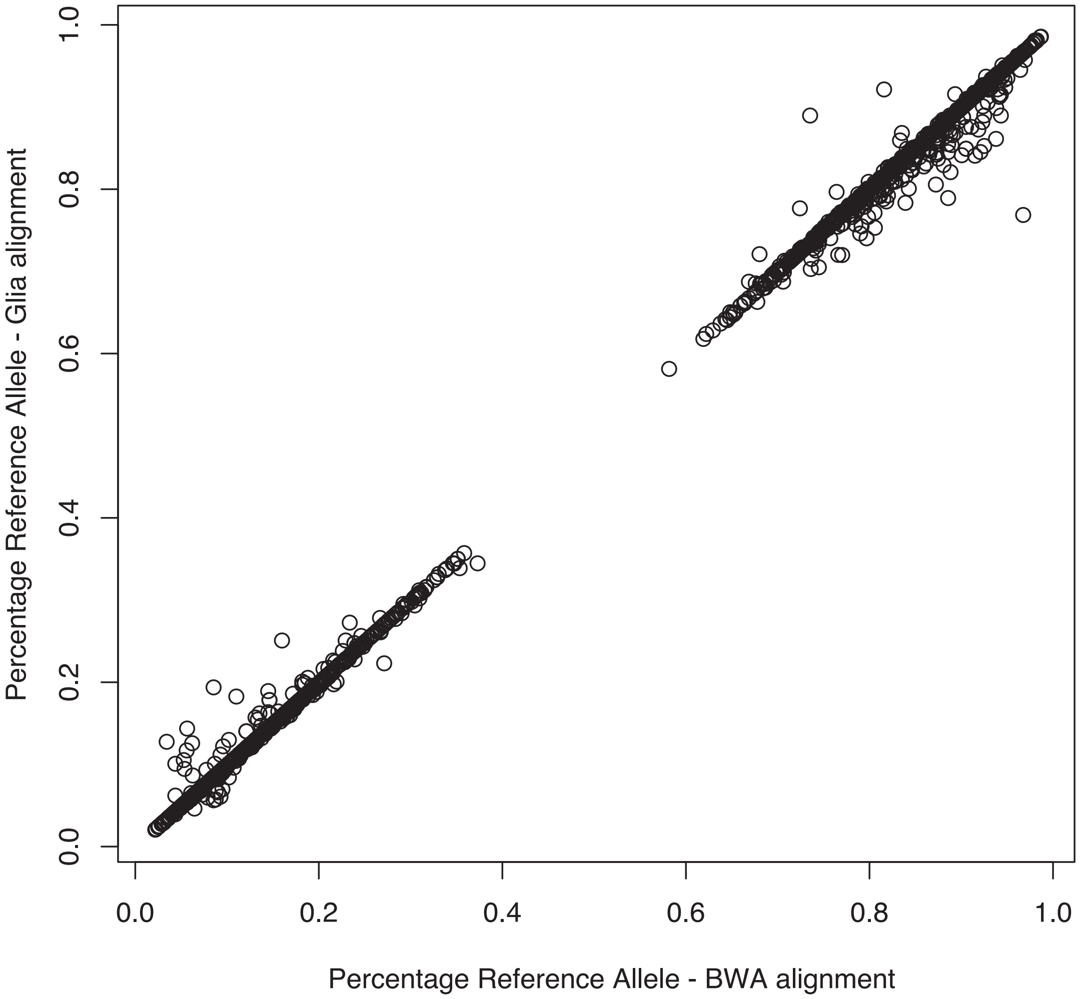

Figure 2.26: Effect of alignment to allele specific analysis. We performed local realignment using a variant aware aligner glia (https://github.com/ekg/glia), which align reads to a variant graph (Lee et al., 2002) built using supplied variants, and compared the allelic bias in our significant allele specific sites between the two alignments. We saw that the effect of local realignment is minimum.

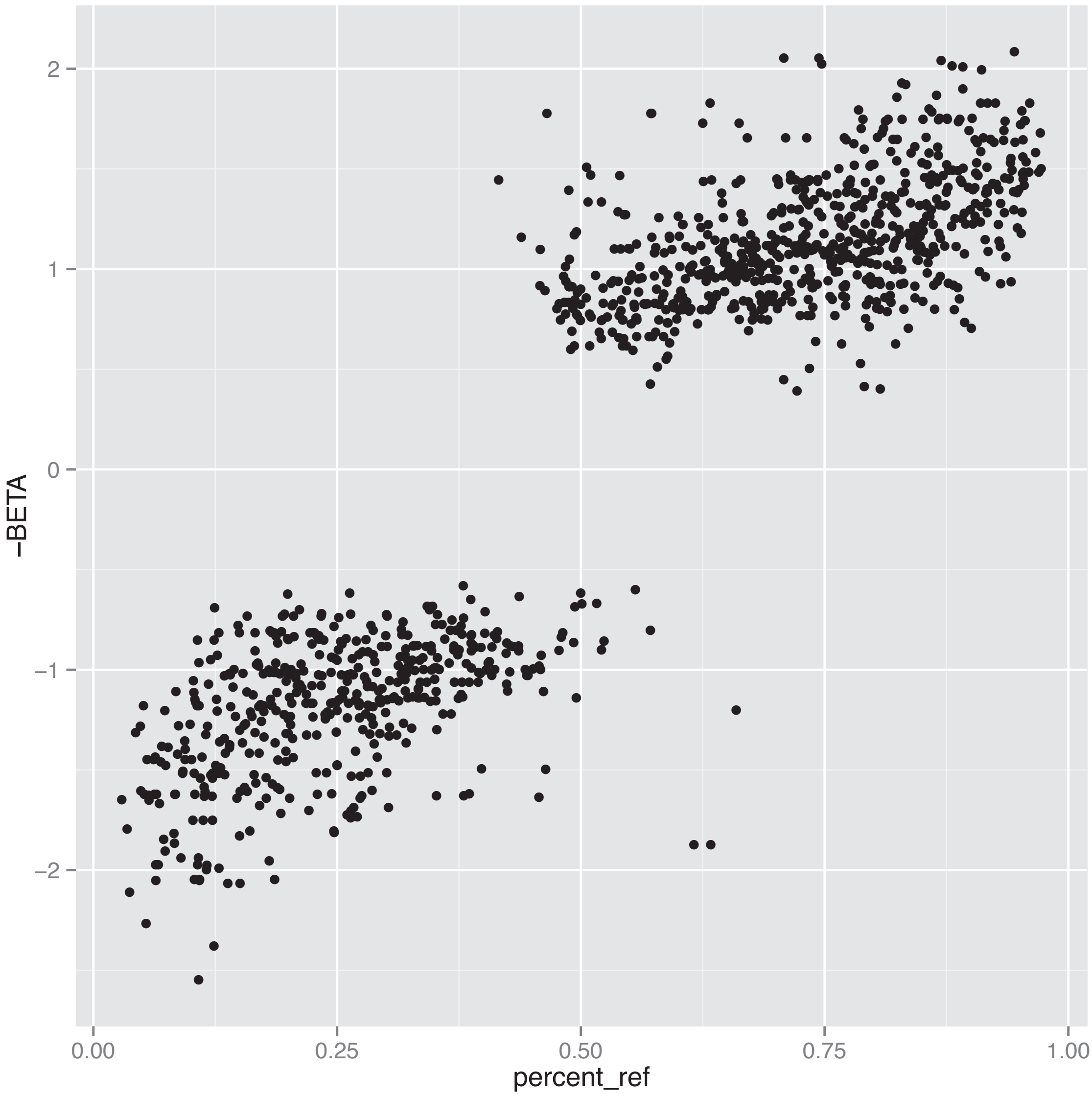

Figure 2.27: No QTLs with strong effect size in binding regions that do not show strong allele specificity. The x-axis shows allele specificity (measured as % reference), and the y-axis shows between-individual effect (β) orientated such that positive is towards reference.

and gene expression simultaneously. However, it is possible that CTCF affect gene expression distantly at a weaker level. Such effect could take place via higher order interactions via 3D chromatin structures such as looping (Bulger and Groudine, 2011). Much more samples may be needed to achieve sufficient power.

The results from this and many other studies suggest that motif adjacent sequences may influence transcription factor activities. This may exist when transcription factor binds to weaker secondary motifs that are close to the canonical motif or distribute around the canonical motif. Even those highly sequence specific transcription factors such as CTCF do not bind to their canonical motif in 100% cases, but instead show a distribution of binding occurring at different positions and alleles. Alternatively, it is also possible that variants that are on the canonical motif are removed from population due to their large effect. On the contrary, the ones in the close vicinity nearby may have weaker but significant phenotypic effects that is below selection (Farh et al., 2014).

We see hundreds of sites showing allele-specific binding. The idea that allele-specific events have similar effects inside one cell as genotypic effects do between individuals is commonplace (McVicker et al., 2013). Here we show that these two effects are well modeled by a linear relationship (at least for this assay), though there is also an interesting subset of allele-specific sites that show no QTL. In contrast there are few QTL loci that overlap binding regions without an allele-specific signal.

As expected, some of the allele specific sites switch specificity between the alleles in different samples, consistent with a nearby, incompletely linked causal allele, random allelic inactivation or parent-of-origin imprinting. Many of these sites can be explained by an incompletely linked nearby locus, highlighting that the causal variant is often not co-incident with the binding region.

Finally with more confident mapping of reads from paired read ChIP-seq data we are able to show that a consistent signal towards reference alleles is in fact predominantly due to a biological effect favoring ancestral alleles (at least for the CTCF transcription factor). This suggests that base pair changes segregating in the population tend to reduce binding of existing sites (rather than create new sites), at least for CTCF, and this is consistent with CTCF motif creation occurring by non-base pair changes, e.g. repeat deposition, as suggested in Schmidt et al. (2012). Similar observations have been made on the allele effect of gene expression, where the new mutations tend to reduce gene expression levels (Chaix et al., 2008).

The understanding of the non-coding variants, which comprise the vast majority of the disease susceptibility variants discovered so far is remains challenging mostly due to our limited knowledge of the regulatory mechanisms in the non-coding regions. This catalog of CTCF QTL sites is part of a growing set of molecular assays that are being examined in outbred individuals (for example, see Kasowski et al., 2010; Degner et al., 2012; Maurano et al., 2012; Kilpinen et al., 2013; McVicker et al., 2013). It provides a specific hypothesis for the 63 disease related loci which overlap these QTLs, and for future overlaps with other molecular, cellular and disease related phenotypes. The gradual unraveling of the different variant effects on different molecular behavior will provide a growing understanding of molecular and physiological processes in health and disease.

Systematic survey using data of multiple cellular events for the same set of samples could offer hints for understanding the underlying biological mechanisms. Much work remains to be done to collect such information. Ideally it should be done *in vivo* such that the data genuinely reflect the biological effect independent of technical interventions. Additionally data of each event should be collected in a time series with events at each time point collected simultaneously . Although this may not be possible with the current technologies, it is most likely to offer the correct information. Such experiments may also pose computational challenges. Methods are needed to deal with very large volumes of data. Meanwhile, the number of cellular events are usually far more numerous than the number of samples, raising a “small n, large p” problem. It is challenging to resolve the causal relationships among these variables. Nevertheless, the regulatory mechanisms largely remain to be understood, only after which it becomes possible to choose the right candidates for therapeutic interventions.