Interplay of cis and trans mechanisms driving transcription factor binding, chromatin, and gene expression evolution

Emily S Wong[1*], Bianca M Schmitt[2*], Anastasiya Kazachenka[3], David Thybert[1], Aisling Redmond[2], Frances Connor[2], Tim F Rayner[2], Christine Feig[2], Anne C. Ferguson-Smith[3], John C Marioni[1,2], Paul Flicek[1,4#], Duncan T Odom[2,4#]

[*] These authors contributed equally

[#] Corresponding authors: PF (flicek@ebi.ac.uk), DTO (duncan.odom@cruk.cam.ac.uk)

[1] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK.

[2] University of Cambridge, Cancer Research UK - Cambridge Institute, Li Ka Shing Centre, Cambridge, CB2 0RE, UK.

[3] University of Cambridge, Department of Genetics, Cambridge, CB2 3EH, UK.

[4] Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK.

SUMMARY

Noncoding regulatory variants play a central role in the genetics of human diseases and in evolution. We measured allele-specific TF binding affinity of three liver-specific TFs between crosses of two inbred mouse strains to elucidate the regulatory mechanisms underlying transcription factor (TF) binding variations in mammals. Our results highlight the pre-eminence of cis-acting variants on TF occupancy divergence. TF binding differences linked to cis-acting variants generally exhibit additive inheritance, while those linked to trans-acting variants are most often dominantly inherited. Cis-acting variants lead to local coordination of TF occupancies that decay with distance; distal coordination is also observed and may be modulated by long-range chromatin contacts. Our results reveal the regulatory mechanisms that interplay to drive TF occupancy, chromatin state, and gene expression in complex mammalian cell states.

INTRODUCTION

Understanding how genetic variation propagates into differences in complex traits and disease susceptibility is a major challenge. Evolutionary studies have revealed examples of regulatory variants linked to different organismal phenotypes[1]. Genome-wide studies have also found that many common disease-associated genetic variants lie in regulatory sequences[2–4] with genetic changes at local-regulatory elements leading to coordinated chromatin changes within constrained genomic domains[5,6].

A key determinant of transcriptional activation and spatiotemporal specificity is how strongly collections of transcription factors (TFs) bind to gene regulatory regions[7–9]. How transcription factor binding specificity and strength is shaped by cis- and trans-acting variation remains poorly understood[10], and understanding the interplay between TF binding and the surrounding chromatin state is critical for determining phenotypic diversity.

Cis-acting sequence changes substantially modulate TF occupancy[11,12], but direct disruption of TF-DNA binding motifs is relatively rare[13–19]. This seemingly conflicting observation may be potentially explained by changes to surrounding chromatin state, long range TF-TF connectivity[6] or cis-acting binding determinants near but outside the core binding motif[20].

Strategies used to dissect cis- and trans-acting mechanisms include QTL-based analyses and F1 crosses of genetically inbred organisms. QTL analysis correlates a measured trait (e.g. gene expression or TF binding intensity) with genetic variation. However, fully distinguishing between regulatory divergence in cis and in trans in eQTL and ChIP-QTL studies[26] requires large numbers of genetically diverse samples to achieve statistical power[27–30]. Alternatively, the regulatory mechanisms can be revealed by analysis of the patterns of divergence occurring in F1 genetic hybrids; this approach has been used to analyze gene expression in yeast[25,26], maize[27], fruit flies[21,22,28] and mouse[29,30].

To our knowledge, F1 hybrids have not been employed to comprehensively dissect TF binding differences in mammals. We created first-generation genetic hybrids from divergent mouse sub-species to dissect trans-acting mechanisms that affect both chromosomes equally due to a shared nuclear environment, from the allelic-specific differences caused by locally acting cis-directed mechanisms[21–24]. We further leveraged this strategy to interrogate the inheritance of TF binding occupancy, which reflects the selective pressures on TF binding[31,32].

By incorporating matched transcriptomic data from RNA-seq[29], our results provide a comprehensive and quantitative overview of how different layers of regulatory variation intertwine to create tissue-specific transcriptional regulation.

RESULTS

Identification of transcription factor binding events influenced by cis-acting variants using mouse reciprocal crosses

In order to dissect the extent of cis and trans variation in TF occupancy variation, transcription factor binding site (TFBS) occupancy was mapped with six biological replicates using chromatin immunoprecipitation followed by sequencing (ChIP-seq) against three liver TFs (HNF4A, FOXA1, CEBPA) in inbred mouse sub-species *C57*BL/6J (BL6) and CAST/*EiJ* (CAST) and their F1 crosses (BL6xCAST and CASTxBL6) (**Figure 1a, S1-3, Methods**); all data are in ArrayExpress (E-MTAB-4089). The large number (~19M) of single nucleotide variants (SNVs) between two parental strains is comparable to that found in human populations[38], and permits a substantial proportion of allele-specific TF binding to be measured.

Approximately 60-70,000 regions in the genome are bound by each TF (**Methods**), and approximately 20% had one or more SNVs with sufficient sequencing coverage to permit quantitative allelic resolution of TF binding (**Figure 1b**). Of these TFBS, in ~3-6% of these cases, SNVs directly disrupt a binding motif. Most (ca. 62%) SNVs are found in regions bound by only one TF, 34% are found in regions bound by exactly two TFs, and 5% by all three TFs, and are highly reproducible (**Figures 1c, S2**).

Cis and trans effects can be distinguished by the differences in binding affinities among F0 parents and their F1 offspring, as cis-acting variation must remain allele-specific[11,12,14,31,32] (**Figures S4a, S5**). TFBS that had informative SNVs for allelic resolution were classified into four regulatory categories – conserved (non-differential), cis, trans, and cistrans (affected by variants acting both in cis and in trans) (**Figure S4b**) (**Methods**).

Differences in TF binding occupancies between the two mouse strains were most frequently affected by cis-acting variation (44-49%), followed by cistrans (14-17%) and trans (8-13%); 23-30% of TF binding was conserved despite the presence of one or more variants near the site of binding (**Figure S4c**). Proportions of TFBSs belonging to each of the four categories were similar between all TFs. As expected, there are fewer conserved locations when SNVs directly disrupt the bound motif (**Figures S4c**)[21]. The substantial signal originating from trans effects in the trans and cistrans categories was visualised by subtracting the F1 BL6:CAST ratio from the corresponding F0 ratio (**Figure 1d)**. We validated our ChIP-seq measures of binding by independently performing allele-specific pyrosequencing (**Figure S6**). Approximately 40% of TFBS are regulated purely by cis-acting variation, compared with only 14% of liver-transcribed genes similarly regulated[14].

TF binding affinity is more strongly cis-driven than gene expression and is inherited additively

To quantitate the effect size of cis-acting variation on TF occupancy, we compared TF binding between F0 and F1 individuals using Pearson's correlation (**Figures 2, S7**, **Methods**). In the absence of noise, a correlation coefficient of zero indicates that cis and trans contributions are equal, whereas a correlation coefficient of one indicates the absence of trans effects. We find Pearson's r for TF binding to be significantly larger than that for gene

expression (TF binding: r=0.92, 95% CI (0.915, 0.919), P<2.2e-16; expression: r=0.62, 95% CI (0.607, 0.631), P<2.2e-16) (see also **SI S1**).

For lineage-specific TF binding locations, we constructed statistical models to test the extent of variation driven by cis versus cistrans regulation. If the divergence was purely due to variants regulated in cis, the binding strength in the F1 allele will be half that in the F0 mouse. If TF binding in the F1 mouse were also influenced by variants in trans, then these binding intensities would be either greater or less than half the level found in the parent (**Methods**). The vast majority (87%, 1056/1217) of lineage-specific TFBS were driven by cis variants (**Figure 2c-d**), while only 13% (161/1217) showed evidence of trans influence. Overall, lineage-specific sites are up to two times less likely to have contributions from trans variants

Binding intensities influenced by cistrans-acting mutations can either be balanced by compensatory mutations acting in trans (the difference in binding intensities in F1 < than for F0) or further diversified (the difference in binding intensities in F1 > than for F0). Under complete neutrality, both should be equally favoured[31]. The frequency of compensatory versus diversifying effects is not significantly different at lineage-specific TFBSs (binomial test, P=0.6) (**SI S2**, **Figure S8a**), suggesting many of these TF binding events are neutral. In comparison, of the 2,563 cistrans-regulated CEBPA binding sites shared between alleles, 64% are compensatory and 36% diversifying (binomial test, P<2.2e-16), suggesting non-lineage-specific TFBSs are more frequently subjected to selective forces. No lineage-specific TFBS affected by only trans-acting variants were observed (i.e. strain-specific in F0 but equally bound in F1). Our results strongly suggest that cis-directed variation either directly (e.g. modification of the binding motif) or indirectly (e.g. through remodelling of surrounding chromatin) play a required role in birth of TFBSs (**SI S3, Figure S8b**).

We evaluated the potential regulatory activity of the TF binding by mapping the genome-wide locations of H3K4me3 (marking transcription initiation sites) and H3K27ac (marking potential enhancer activity)[39] in F1 mouse livers (**Methods**). At promoters, TF occupancy changes driven by cis and cistrans variations were underrepresented (All TFs; binomial test; cis: P=1.1e-6, odds ratio (OR)=0.8; cistrans: P=1.2e-8, OR=0.6), and conserved sites were overrepresented (P<2.2e-16, OR=1.7) (**Figure 2b**). Regions showing enhancer activity were enriched for conserved, and depleted for TFBSs that were directed by cis and cistrans variants (cis: P=3.4e-3, OR=0.8; cistrans: P=1.6e-6, OR=0.6; conserved: P=3.3e-8, OR=1.5).

The stability of genomic occupancy at TFBSs was assessed by evaluating the TF occupancy in BL6 mice with a single allele deletion of *Cepba* or *Hnf4a*, which can reveal regulatory activity and gene expression with more direct TF dependency[40]. When TF expression was reduced, the change in TF occupancy level is greater for cis-directed variation compared to conserved binding (**Figure S9**). This suggests that TFBSs driven by cis variants are more prone to changes in TF expression while non-differentially bound TFBSs are buffered.

TFBSs can be inherited in an additive versus non-additive manner for cis-driven and trans-driven categories. Pure additive inheritance occurs when the combined occupancy of the F1 alleles is equal to the sum of the two parental (BL6 and CAST) F0 alleles[13,37,41]. Pure dominant inheritance occurs when the total allelic occupancy in the F1 offspring is equal to that of either parent (**Figure 2e**). We fitted statistical models for both scenarios and evaluated them using Bayesian Information Criteria (BIC) (**Methods**).

Of the 2,382 TFBSs driven by cis variants (**Methods**), 72% (1,720) showed additive inheritance (of which 1,215 had BIC>2), whereas 28% (662) appeared dominant, which may reflect assignment errors (see **Discussion**). In contrast, of 341 TFBSs driven in trans 74% (280) have dominant inheritance, whereas only 26% (61) were additive. Similar trends were observed for FOXA1 and HNF4A (**Supplementary Data File**). Under- or over-dominant TFBS inheritance appears rarely if at all (**SI S4)**.

In summary, variation in TF occupancy is strongly driven by cis-acting local variants, whereas TFBS affected by variation in trans are uncommon. Differing from *Drosophila* gene expression[37], mammalian TFBSs are largely inherited additively, and trans-driven TFBSs are mostly dominantly inherited (**Figure S10**).

Cis-influence on binding occupancy rapidly decays with distance

Chromatin state can depend on distal functional elements located tens to hundreds of kilobases away[5,6]; we therefore asked what impact cis-acting variation has on TF occupancy at varying distances.

We first confirmed that overlapping binding events from different TFs share cis-acting variants more often than expected by chance (**Figure S11**). We quantitated how strongly cis-acting variants influence distant TF binding occupancies using a complementary strategy to Waszak et al. (2015). Although the exact location of each causal variant is unknown, the genomic span (or effect distance) of a cis-acting variant can be inferred by examining co-variation in binding occupancies between neighbouring TFBSs (**Methods**, **Figure 3a, S12**).

The correspondence between TF binding occupancies decays at a logarithmic rate, with similar trends observed across all three TFs (**Figure 3b)**. The correspondence is 2-3 times lower at 50kb than at 3kb. Nevertheless, we detected cis-driven correspondence slightly above genomic background levels up to 400kb away (Rho=0.01–0.02, linear regression). These low incidences of long-range coordination may reflect an alternative mechanism to the local correspondence mediated by cis-acting variants. We determined the point of maximum curvature at which the correspondence between TFBSs began to more rapidly decay with distance (i.e. the elbow of the curve) using vector projection to estimate its location as 13kb (**Methods**). Our results were consistent across several bin sizes grouping nearby SNVs (**Supplementary Data File**). Different TF binding locations appear to be similarly correlated, as shown recently for chromatin domains[5,6].

Long-range coordination of TF occupancy could be affected by cis-variation via three-dimensional interactions, and we therefore searched for direct evidence that spatially distinct TFBSs interact (**Figure 3c**). We analyzed Hi-C data from BL6 mice[42] to identify the interaction endpoints that overlap CEBPA binding locations (**Methods**). As expected, conserved sites were more likely to overlap long-range interaction endpoints (logistic regression: P<0.05, OR=1.14–1.20) (**Table S1**). Chromatin interactions anchored on a cis-associated location were strongly enriched over the any-versus-any background (binomial test; P-value: cons versus cons=2.0e-8, cis versus trans=1.8e-9, cis versus cons=4.0e-10, cis versus cis=5.7e-6, cis versus cistrans=4.5e-4). Significant enrichment over the any-versus-any background set was observed for all categories of TFBS.

Our data support a model where the cis variants causal for differences in TF binding occupancy are mostly proximal to the TFBS they affect. However, regions with TF occupancy, including TFBSs affected by cis-variation, are disproportionately found at interaction endpoints for genomic domains, providing a possible mechanism for the observed long-range correlations.

Coordination of regulatory mechanisms underlying gene expression and TF binding intensity variations

The connection between genetic variation with TF binding, chromatin state and gene expression has recently been studied in human cell lines[22,24,25]. However, how genetic variants affect the interplay of these regulatory layers remains poorly understood.

As above, we classified the mechanisms of variation underlying the allelic changes in chromatin state and transcription based on whether these differences are cis-directed, conserved, trans-directed, or cistrans-directed (**Figure 4, Methods**). We then used logistic regression to establish whether the mechanisms of variation responsible for regulating TF binding differences are enriched or depleted within the corresponding chromatin and gene expression categories.

We found similar variant classes underlying TF binding occupancy, chromatin state, and gene expression at the same locus. For instance, promoters where allelic differences are caused by cis-acting variations are associated with TFBS where allelic differences are also caused by cis-acting variations (**Figure 4a-b**). This is compatible with models proposed by Kilpinen et al.[22]. Furthermore, there is a positive correlation between the directions of effect between allelic changes in TFBS occupancy and gene expression (**Figure S13**). In other words, when a TFBS increases its occupancy, then nearby gene transcription often increases (binomial test, P=2.9e-4) and with similar magnitude (Spearman's rank correlation, rho=0.29, P=6.3e-12). This effect is not caused by differences in expression levels (**SI S5**).

Promoter chromatin state and gene expression are correlated[43]. We identified a subtle but significant correspondence between the types of regulatory variation underlying promoter activity differences and gene expression differences (binomial test; cis P=0.04, cistrans P=0.03, conserved P=0.02, trans P=0.25) (**Figure 4c**).

Finally, TFBSs often act in concert with one another. Hence, we asked whether the collective effect of the cis- and trans-acting variants underlying changes to occupancy levels of multiple TFBSs propagate to gene expression. Using Shannon's entropy, we compared the mechanistic diversity of TF binding variants with the mechanisms of variation affecting nearby gene expression (**Methods**). Expression driven by cis- or cistrans-acting variants was significantly more likely to be associated proximally to TFBSs that are themselves driven by variation acting through diverse mechanisms (Mann Whitney U test) (**Figure 4d**). In contrast, conserved expression was likely to be associated with TFBSs directed by a similar type of variant (**SI S6**).

DISCUSSION

Directly connecting genome-wide observations of transcription factor binding with functional outputs in gene expression is challenging because of what appears to be two conflicting observations. On the one hand, most variation in the human genome associated with complex disease and other phenotypes is non-coding[2]. Even for Mendelian disorders, exome sequencing can suggest causative sequence changes in only a minority of cases (~25%)[44]. Both point to a major role for functional sequence changes in the regulatory regions of the genome, which subsequently lead to changes in gene expression. On the other hand, TF binding demonstrates both variability between even genetically identical individuals and such strikingly rapid evolutionary change[45] that it is tempting to conclude that the vast majority of TF binding is non-functional "biological noise"[46].

Here, we have undertaken a detailed and comprehensive dissection of the genetic mechanisms driving TF binding occupancy differences in mammals and integrated these results with chromatin and gene expression information. Our initial finding regarding how genetic sequence variation associates with TF binding differences between alleles is consistent with previous reports at a more limited set of locations in murine immune cells[48], human lymphoblast cells[23,25], and using computational simulations[49,50]. Specifically, almost three-quarters of assayed quantitative differences in TF binding occupancy appear to be purely the result of nearby cis-directed genetic differences.

However, our integrated analysis extending from TF binding to output gene expression using F1 inter-strain mouse crosses revealed a number of novel insights. First, the vast majority of trans-directed TF binding differences are dominantly inherited. Although most cis-driven binding is inherited additively as expected, a small proportion appears to show dominance/recessive variation. One plausible biological explanation is the presence of trans-directed elements that do not interact with cis-driven variation at each allele. Despite this, cis and trans-acting variants driving TF occupancy change show clear differences in their mode of inheritance. Second, allelic differences in TF binding are correlated at kilobase distances above the genomic background, likely driven by neighbouring cis variants. A minor fraction of TFBSs show long-range coordination, which may driven by high enrichment of TFBS at chromatin contacts. Such long-range correspondence is similar to recently described coordination of chromatin states within topological domains[6,51]. Third, we demonstrate interplay between the different mechanisms of variation that underlie transcription factor binding and tissue-specific gene expression *in vivo*. Aspects of the regulatory interplay between chromatin and gene expression has been reported in human cell lines[22,52–55].

The independently determined causal mechanisms of variation correspond well between TF occupancy and gene expression. This is potentially surprising given the difference in the overall patterns of regulatory mechanisms between TF binding and gene expression. Namely, protein-DNA interactions are shaped by a comparatively simple combination of DNA sequences, chromatin context, and (in some cases) noncoding RNA associations. In contrast, a multitude of regulatory processes influence gene expression, including TF binding as well as post-transcription processing, translation rate and mRNA degradation. Our results support a model whereby the variation underlying gene expression differences are a composite of the variation that modulate TF binding differences in multiple individual TFBSs.

Our analysis has specific limitations. Our approach cannot analyse the majority of TFBSs where no informative SNV is present, and these unclassified TFBSs are more likely to be

conserved. However, a change in the relative proportion of regulatory categories is not expected to influence our key findings, which were focused on the mechanism effect size. Our analysis ignores structural variants, and we cannot preclude that possibility that tissues other than liver may have greater trans-influenced TF binding differences. Although most tissue-specific gene expression appears to be driven by combinatorial TF binding of scores of TFs[10], we have profiled only a subset of three. However, analysis of the occupancy of over a hundred TFs in one tissue strongly suggest that our data will reflect the typical mechanistic contributions influencing the evolution of all tissue-specific TFs[56]. Finally, our technical definition of the cistrans-driven variation captures TFBSs with high biological and/or technical heterogeneity.

Our work builds upon previous findings of genomic coordination among TF binding, chromatin marks and transcription[5,6,22,57] and highlights the key role played by the basal variation that underlie TF binding in directing regulatory change. The cis- and trans-acting factors mechanistically driving changes in regulatory variation are likely to fundamentally contribute to the coordination of chromatin domains[5,6], which are themselves components of topologically associated domains[58].

ACKNOWLEDGEMENTS:

We thank the CRUK - CI Genomics, BRU, and Bioinformatics Cores for technical assistance and the EMBL-EBI systems team for management of computational resources. This research was supported by the European Molecular Biology Laboratory (E.S.W., D.T., J.M., P.F.), Cancer Research UK (B.S., T.R., F.C., C.F., A.R., D.T.O.), the BOLD ITN (B.S.), Darwin Fellowship (AK), the Wellcome Trust (WT095908) (P.F.), (WT095606) (AFS) and (WT098051) (P.F., D.T.O.), EMBO Long-term (ALTF1518-2012) and Advanced Fellowships (aALTF1672-2014) (E.S.W.), and by the European Research Council and EMBO Young Investigator Programme (D.T.O.).

AUTHOR CONTRIBUTIONS:

ESW, BS, DT, AFS, JCM, PF, DTO designed experiments; BS, AR, AK performed wet lab experiments; ESW performed computational analyses; BS, ESW, AK, TFR collated and quality controlled the data; FC, CF provided experimental assistance; ESW, BS, PF, DTO wrote the manuscript; PF, DTO oversaw the work.

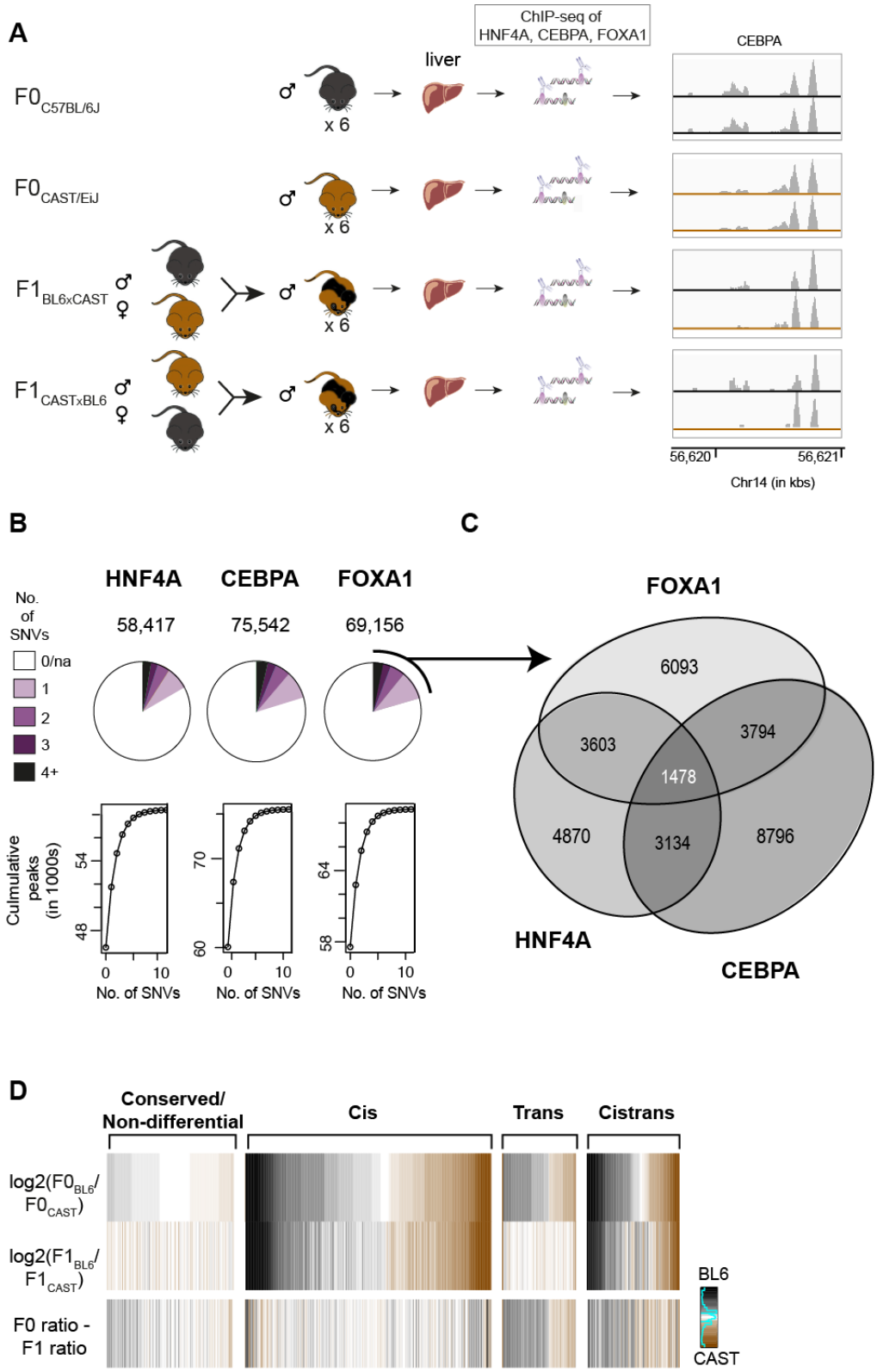

Figure 1. F1 mice were used to interrogate the regulation of TFBS variation
(A) *In vivo* binding of liver-specific TFs FOXA1, HNF4A and CEBPA were profiled in the livers of male mice from inbred strains C57BL/6J (BL6), CAST/EiJ (CAST) and their F1 crosses: C57BL/6J x CAST/EiJ (BL6xCAST) and CAST/EiJ x C57BL/6J (CASTxBL6). Six biological replicates were generated for each TF and genetic background combination. **(B)** The number of TFBS that could be classified were dependent on 1) the presence of one or

more SNVs under TF binding locations and 2) sequencing depth. TFBSs that did not overlap a SNV or where the loci did not meet our minimum read requirement (see **Methods**) are in white. For all others, the number of SNVs reflects those SNV loci which met our read count threshold and were classified to a regulatory category. Reference numbers for the representative libraries depicted are do3488, do3463 and do3483. **(C)** Venn diagram illustrates the numbers of classifiable SNVs that overlap between TFs. Each variant is at least 250bp from any other SNV. Numbers shown are the final numbers of regulatory loci used for downstream analyses. (**D)** Heatmap displaying BL6 (black) versus CAST (brown) binding intensity ratios for different regulatory categories for CEBPA. A subset of variants from each class was randomly sampled to match the overall distribution. Sparkline in key shows the number of observations at each colour category where density is increasing from left to right.

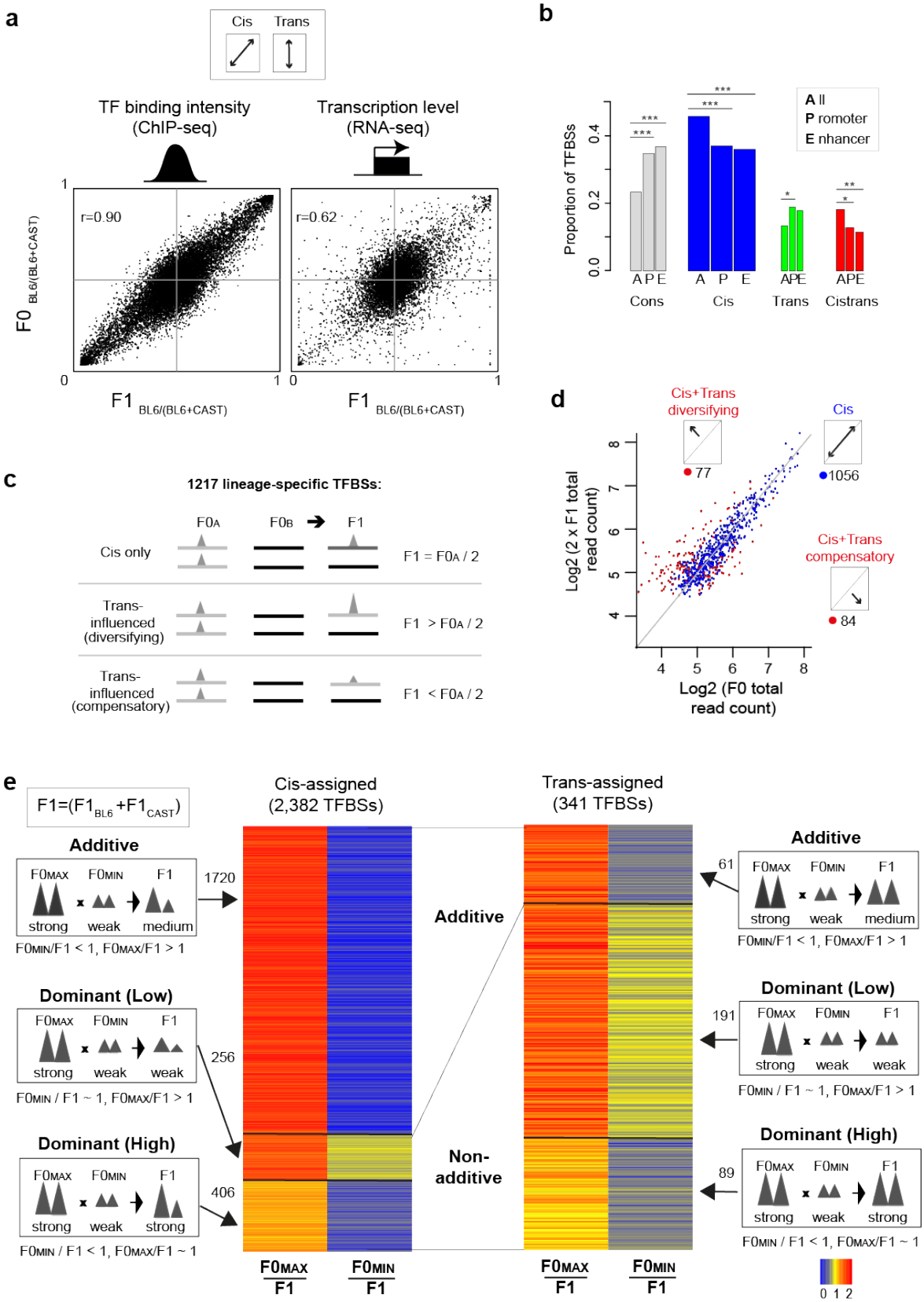

Figure 2. Differences in TF binding intensities strongly driven by cis directed variation and are additively inherited

(A) Mean F0 versus F1 TF binding intensity ratios (BL6 versus CAST) for CEBPA are plotted in the left panel. The right panel shows mean F0 versus F1 gene expression ratios for

liver-expressed protein-coding genes[14]. The correlation coefficient reflects the extent of cis-regulation. **(B)** Proportion in percentages of TF binding locations at promoters and enhancers. The width of the bar is proportional to the overall number of TFBSs in the 'All' category. Putative enhancers were annotated at regions containing H3K27ac with no evidence of H3K4me3; while regions containing H3k4me3, which largely also contain H3K27ac, denote promoter activity. Binomial tests were used to compare for enrichment at promoters and enhancers for each regulatory class based on the overall numbers of TFBSs ('All'). Data from CEBPA is shown. ***P<0.0001 **P<0.001* P<0.05. **(C)** Most highly allele-specific TFBSs are driven purely in cis. Lineage-specific TFBSs were defined as TFBSs where binding occurs either in BL6 or CAST in F0 individuals and in an allele-specific manner in F1 individuals based on a cut-off ($F0_{B6/(B6+CAST)}$>0.95, $F1_{B6/(B6+CAST)}$>0.95, $F0_{B6/(B6+CAST)}$ < 0.05, $F1_{B6/(B6+CAST)}$ < 0.05). These TFBSs can be sorted into the three categories described. Each category presents a testable model that is formally defined by the formulas on the right. We tested each of these scenarios using maximal likelihood estimation by assuming all counts follow negative binomial distributions (see **Methods**). **(D)** Mean log2 F0 total read counts were plotted against mean log2 F1 read count (BL6 + CAST allele) multiplied by 2. For the scatterplot, we used averages across biological replicates. Cis-driven TFBSs are thus expected to fall along the diagonal and these have been coloured blue (see **C**). Categories shown in the scatterplot were determined by maximal likelihood estimation. Data for CEBPA is shown. **(E)** The majority of cis-directed TFBSs are inherited additively. Trans- driven TFBSs may show additive or dominant inheritance patterns in TF binding intensities. We deciphered the different modes of inheritance by comparing overall peak binding intensities between F0 and F1 individuals. We constructed models to specifically test and partition trans-driven TFBSs into additive, dominant (high) and dominant (low) inheritance patterns, where high and low refer to the parental binding intensity that was inherited by the offspring. Trans-classed TFBSs that do not show sufficient difference between $F0_{MAX}$ and $F0_{MIN}$ were not considered (see Methods). The heatmap summarizes of the result of our classification process. Total F1 counts were individually scaled to 1 (yellow). Red indicates TFBSs where F1 > F0; blue indicates TFBSs where F1 < F0. CEBPA data is shown.

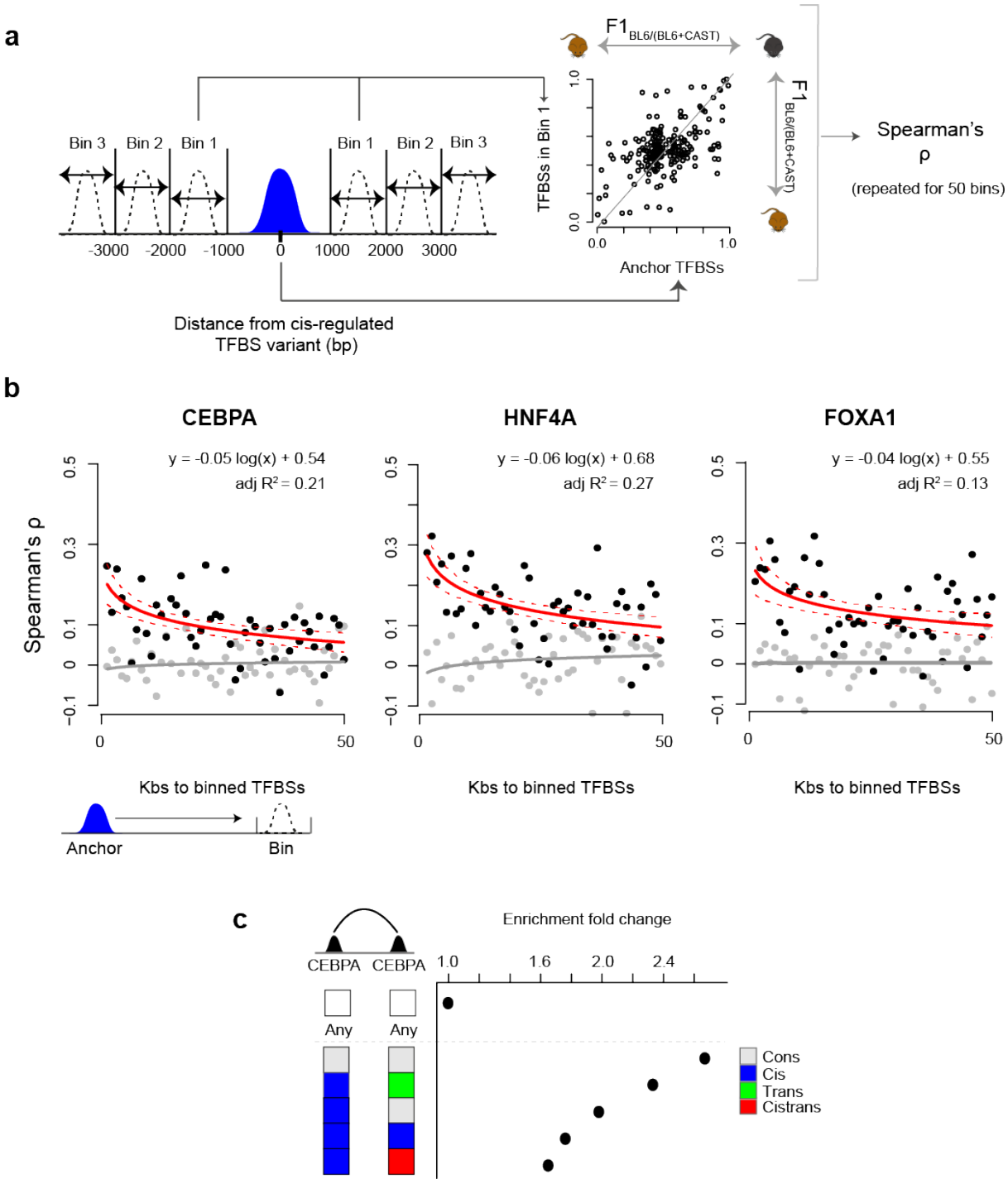

Figure 3. Rapid loss of cis-acting inter-peak correspondence with genomic distance
(A) Strategy for measuring the span of cis regulatory effect. Successive 1kb bins were taken from each cis-driven TFBS starting 400bps from the location of the SNV and extending in both directions. For each bin, Spearman's ρ was calculated using the BL6:CAST allelic ratio between queried TFBSs against TFBSs assigned as anchorages for the analysis. **(B)** Spearman's ρ values for each bin were plotted for each TF. The linear regression line calculated from these values is shown as a solid red line. Red dashed lines mark the 90% confidence intervals i.e. we have 90% confidence that the true slope of the line lies within the region bounded by the dashed lines. Adjusted R^2 of the regression line is indicated. Grey dots represent the null background distribution of correlation values. These data points were constructed by the random subsampling of TFBSs to anchor TFBSs (see **Methods**). The

numbers of TFBSs in each randomly sampled bin were matched to those in the observed bins. The grey line is the linear regression line for the correlation values derived from sampled points. **(C)** TFBSs are enriched at regions of chromatin contact. Enrichment values were calculated compared with expected rate of chromatin contact given the general enrichment for contact in each regulatory dataset (i.e. cons, trans, cis, cistrans). 'Any' denotes the null background set of randomly chosen locations in the genome

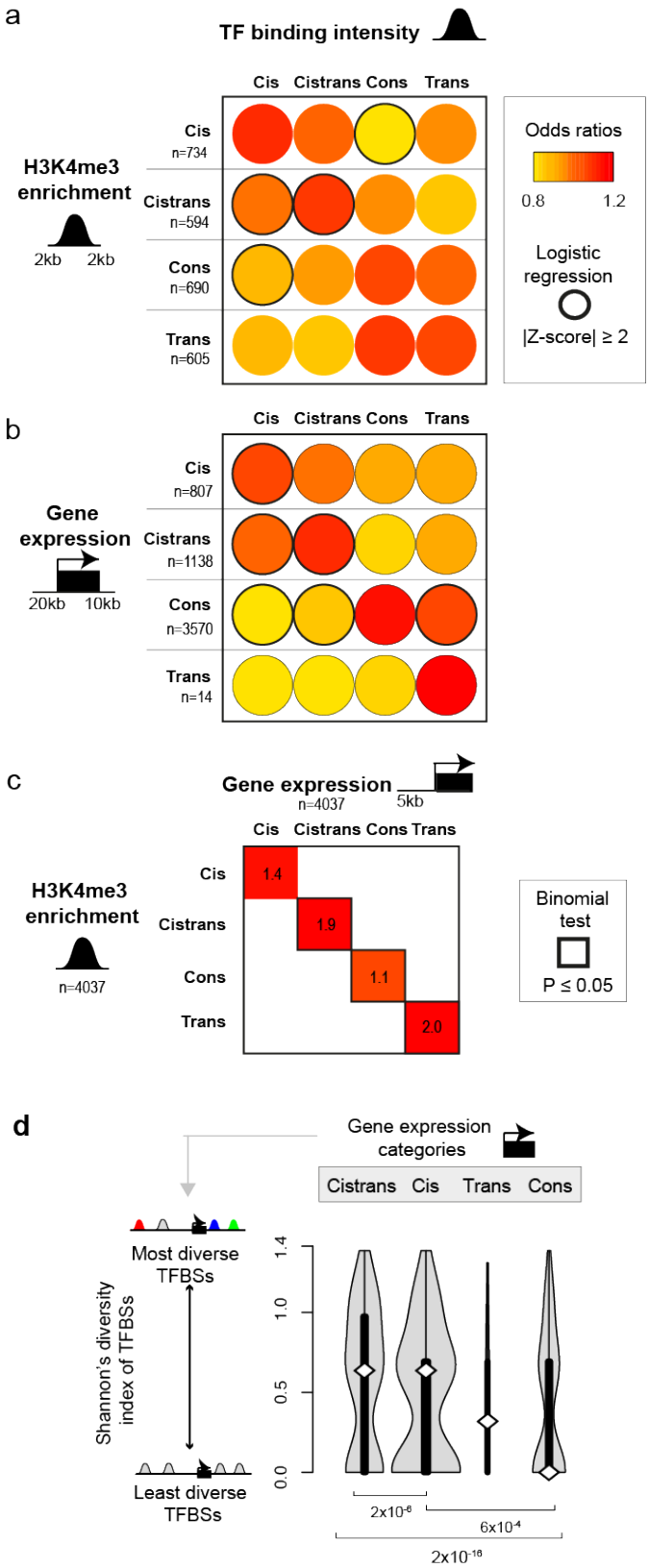

Figure 4. Genetic and epigenetic influences that change TF binding have parallel consequences for gene expression and chromatin

(A) Coordination between the regulation of variation in chromatin and TF binding occupancy variation is shown. TFBSs located 2kb upstream or downstream of the promoter mark, H3K4me3, were associated to that promoter mark. Separate logistic regressions were performed for each chromatin regulatory class (denoted by grey lines)(see Methods). Odds

ratios were mean-centred for comparison across chromatin regulation classes. Absolute values of Z-scores greater than two ($\alpha<0.05$) were denoted by a thick black border. **(B)** Coordination between the regulation of variation in gene expression levels and TF binding occupancy variation is shown. TFBSs located 20kb upstream or 10kb downstream of a TSS were associated to that gene. Separate logistic regressions were performed for genes of each expression regulation category (see Methods). Odds ratios were mean-centred for comparison across expression classes. Absolute values of Z-scores greater than two ($\alpha<0.05$) were denoted by a thick black border. **(C)** Association between chromatin and gene expression. Genes were linked to H3K4me3 modifications if the mark was located within 5kb upstream of the TSS. Binomial tests were performed based on the expected background probability of observing the same regulatory mechanism underlying both expression and histone enrichment change. **(D)** High diversity in mechanisms regulating TF binding variation is associated with gene expression that is affected by cistrans-acting variation. Diversity estimates were obtained using Shannon's diversity index. These were calculated on a gene-by-gene basis for TFBSs 20kb upstream and 10kb downstream of TSSs. These scores were compared between genes grouped by transcriptional regulatory class. Significant P-values for Mann-Whitney U tests are shown. The surface area of the violin plot is proportional to the number of genes in each class.

REFERENCES

1. Wray, G. A. The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.* **8,** 206–216 (2007).
2. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **337,** 1190–1195 (2012).
3. Fairfax, B. P. *et al.* Genetics of gene expression in primary immune cells identifies cell type–specific master regulators and roles of HLA alleles. *Nat. Genet.* **44,** 502–510 (2012).
4. Ballester, B. *et al.* Multi-species, multi-transcription factor binding highlights conserved control of tissue-specific biological pathways. *eLife* **3,** e02626 (2014).
5. Grubert, F. *et al.* Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell* **162,** 1051–1065 (2015).
6. Waszak, S. M. *et al.* Population Variation and Genetic Control of Modular Chromatin Architecture in Humans. *Cell* **162,** 1039–1050 (2015).
7. Biggin, M. D. Animal transcription networks as highly connected, quantitative continua. *Dev. Cell* **21,** 611–626 (2011).
8. Farley, E. K. *et al.* Suboptimization of developmental enhancers. *Science* **350,** 325–328 (2015).
9. Crocker, J. *et al.* Low Affinity Binding Site Clusters Confer Hox Specificity and Regulatory Robustness. *Cell* **160,** 191–203 (2015).
10. Jolma, A. *et al.* DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **advance online publication,** (2015).
11. Wittkopp, P. J., Haerum, B. K. & Clark, A. G. Evolutionary changes in cis and trans gene regulation. *Nature* **430,** 85–88 (2004).
12. Wittkopp, P. J., Haerum, B. K. & Clark, A. G. Regulatory changes underlying expression differences within and between Drosophila species. *Nat. Genet.* **40,** 346–350 (2008).
13. Lemos, B., Araripe, L. O., Fontanillas, P. & Hartl, D. L. Dominance and the evolutionary accumulation of cis- and trans-effects on gene expression. *Proc. Natl. Acad. Sci.* **105,** 14471–14476 (2008).
14. Goncalves, A. *et al.* Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. *Genome Res.* **22,** 2376–2384 (2012).
15. DeVeale, B., van der Kooy, D. & Babak, T. Critical Evaluation of Imprinted Gene Expression by RNA–Seq: A New Perspective. *PLoS Genet* **8,** e1002600 (2012).
16. Strogantsev, R. *et al.* Allele-specific binding of ZFP57 in the epigenetic regulation of imprinted and non-imprinted monoallelic expression. *Genome Biol.* **16,** (2015).
17. White, M. A., Myers, C. A., Corbo, J. C. & Cohen, B. A. Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. *Proc. Natl. Acad. Sci.* (2013). doi:10.1073/pnas.1307449110
18. Wilson, M. D. *et al.* Species-specific transcription in mice carrying human chromosome 21. *Science* **322,** 434–438 (2008).
19. Maurano, M. T. *et al.* Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nat. Genet.* (2015). doi:10.1038/ng.3432
20. Ding, Z. *et al.* Quantitative genetics of CTCF binding reveal local sequence effects and different modes of X-chromosome association. *PLoS Genet.* **10,** e1004798 (2014).
21. Stefflova, K. *et al.* Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. *Cell* **154,** 530–540 (2013).

22. Kilpinen, H. *et al.* Coordinated Effects of Sequence Variation on DNA Binding, Chromatin Structure, and Transcription. *Science* (2013). doi:10.1126/science.1242463
23. Reddy, T. E. *et al.* Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome Res.* **22,** 860–869 (2012).
24. McVicker, G. *et al.* Identification of Genetic Variants That Affect Histone Modifications in Human Cells. *Science* **342,** 747–749 (2013).
25. Kasowski, M. *et al.* Variation in Transcription Factor Binding Among Humans. *Science* **328,** 232–235 (2010).
26. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464,** 768–772 (2010).
27. Doss, S., Schadt, E. E., Drake, T. A. & Lusis, A. J. Cis-acting expression quantitative trait loci in mice. *Genome Res.* **15,** 681–691 (2005).
28. Hasin-Brumshtein, Y. *et al.* Allele-specific expression and eQTL analysis in mouse adipose tissue. *BMC Genomics* **15,** 471 (2014).
29. Lagarrigue, S. *et al.* Analysis of allele-specific expression in mouse liver by RNA-Seq: a comparison with Cis-eQTL identified using genetic linkage. *Genetics* **195,** 1157–1166 (2013).
30. Almlöf, J. C. *et al.* Powerful identification of cis-regulatory SNPs in human primary monocytes using allele-specific gene expression. *PloS One* **7,** e52260 (2012).
31. Tirosh, I., Reikhav, S., Levy, A. A. & Barkai, N. A Yeast Hybrid Provides Insight into the Evolution of Gene Expression Regulation. *Science* **324,** 659–662 (2009).
32. Wang, D. *et al.* Expression evolution in yeast genes of single-input modules is mainly due to changes in trans-acting factors. *Genome Res.* **17,** 1161–1169 (2007).
33. Springer, N. M. & Stupar, R. M. Allele-specific expression patterns reveal biases and embryo-specific parent-of-origin effects in hybrid maize. *Plant Cell* **19,** 2391–2402 (2007).
34. Emerson, J. J. *et al.* Natural selection on cis and trans regulation in yeasts. *Genome Res.* **20,** 826–836 (2010).
35. Sladek, R. & Hudson, T. J. Elucidating cis- and trans-regulatory variation using genetical genomics. *Trends Genet. TIG* **22,** 245–250 (2006).
36. Gibson, G. & Weir, B. The quantitative genetics of transcription. *Trends Genet.* **21,** 616–623 (2005).
37. McManus, C. J. *et al.* Regulatory divergence in Drosophila revealed by mRNA-seq. *Genome Res.* **20,** 816–825 (2010).
38. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526,** 75–81 (2015).
39. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 21931–21936 (2010).
40. Boj, S. F. *et al.* Functional Targets of the Monogenic Diabetes Transcription Factors HNF-1α and HNF-4α Are Highly Conserved Between Mice and Humans. *Diabetes* **58,** 1245–1253 (2009).
41. Gibson, G. *et al.* Extensive sex-specific nonadditivity of gene expression in Drosophila melanogaster. *Genetics* **167,** 1791–1799 (2004).
42. Vietri Rudan, M. *et al.* Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. *Cell Rep.* **10,** 1297–1309 (2015).
43. Consortium, T. E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

44. Yang, Y. *et al.* Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. *N. Engl. J. Med.* **369,** 1502–1511 (2013).
45. Schmidt, D. *et al.* Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328,** 1036–1040 (2010).
46. Spivakov, M. Spurious transcription factor binding: Non-functional or genetically redundant? *BioEssays* n/a–n/a (2014). doi:10.1002/bies.201400036
47. Bradley, R. K. *et al.* Binding Site Turnover Produces Pervasive Quantitative Changes in Transcription Factor Binding between Closely Related Drosophila Species. *PLoS Biol* **8,** e1000343 (2010).
48. Heinz, S. *et al.* Effect of natural genetic variation on enhancer selection and function. *Nature* **503,** 487–492 (2013).
49. Wray, G. A. The Evolution of Transcriptional Regulation in Eukaryotes. *Mol. Biol. Evol.* **20,** 1377–1419 (2003).
50. Stone, J. R. & Wray, G. A. Rapid Evolution of cis-Regulatory Sequences via Local Point Mutations. *Mol. Biol. Evol.* **18,** 1764–1770 (2001).
51. Gruber, J. D., Vogel, K., Kalay, G. & Wittkopp, P. J. Contrasting Properties of Gene-Specific Regulatory, Coding, and Copy Number Mutations in Saccharomyces cerevisiae: Frequency, Effects, and Dominance. *PLoS Genet* **8,** e1002497 (2012).
52. Cheng, C. *et al.* Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res.* **22,** 1658–1667 (2012).
53. Cheng, C. & Gerstein, M. Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. *Nucleic Acids Res.* (2011). doi:10.1093/nar/gkr752
54. Wong, E. S. *et al.* Decoupling of evolutionary changes in transcription factor binding and gene expression in mammals. *Genome Res.* **25,** 167–178 (2015).
55. Cusanovich, D. A., Pavlovic, B., Pritchard, J. K. & Gilad, Y. The Functional Consequences of Variation in Transcription Factor Binding. *PLoS Genet* **10,** e1004226 (2014).
56. Cheng, Y. *et al.* Principles of regulatory information conservation between mouse and human. *Nature* **515,** 371–375 (2014).
57. Ghanbarian, A. T. & Hurst, L. D. Neighboring Genes Show Correlated Evolution in Gene Expression. *Mol. Biol. Evol.* **32,** 1748–1766 (2015).
58. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485,** 376–380 (2012).

SUPPLEMENTARY MATERIALS

Interplay of cis and trans mechanisms driving TF binding, chromatin, and gene expression evolution

Emily S Wong[1*], Bianca Schmitt[2*], Anastasiya Kazachenka[3], David Thybert[1], Aisling Redmond[2], Frances Connor[2], Tim F Rayner[2], Christine Feig[2], Anne Ferguson-Smith[3], John C Marioni[1,2], Paul Flicek[1,4#], Duncan T Odom[2,4#]

[*] These authors contributed equally

[#] Corresponding authors: PF (flicek@ebi.ac.uk), DTO (duncan.odom@cruk.cam.ac.uk)

[1] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK.

[2] University of Cambridge, Cancer Research UK - Cambridge Institute, Li Ka Shing Centre, Cambridge, CB2 0RE, UK.

[3] University of Cambridge, Department of Genetics, Cambridge, CB2 3EH, UK.

[4] Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK.

METHODS:

Sample collection and preparation

All mice were housed in the same husbandry conditions within the Biological Resources Unit in the Cancer Research UK-Cambridge Institute under a Home Office Licence. C57BL/6J and CAST/EiJ mouse strains were used in experiments as parental strains (F0) as well as for breeding of reciprocal crosses of F1 mice. All mice used in the experiments were males between eight and 12 weeks of age, and harvested at the same time of day (between 8 and 11am). Liver perfusion was performed on mice post mortem, prior to tissue dissection. Harvested tissues were formaldehyde cross-linked for ChIP-seq experiments. Before cross-linking, dissected tissue was immediately chopped post mortem and added to a cross-linking solution containing 1% formaldehyde. Tissue was incubated for 20 min prior to quenching with 1/20th volume of 2.5 M glycine. Samples were incubated for a further 10 min before washing with PBS and flash-freezing and storage at -80°C.

Generation of HNF4A and CEBPA heterozygous mice

To create HNF4A and CEBPA heterozygous knockout mice, we acquired mice with targeted alleles from JAX (www.jax.com) (HNF4A stock number: 004665 [53]; CEBPA stock number: 006230 [54]). Heterozygous knockouts were generated via the Cre-loxP system [55] using the germline deleter strain PgkCre [56] and crossing it to *Cebpa*$^{FLOX/FLOX}$ and *Hnf4a*$^{FLOX/WT}$ mice. Ear biopsies were taken at the time of weaning for genotyping to confirm deletion via PCR (**Table S3**).

ChIP-seq experimental procedure

The ChIP-seq protocol was as described by Schmidt et al. (2009). Protein-bound DNA was immunoprecipitated with 10µg of an antibody against CEBPA (Santa Cruz, sc-9314), HNF4A (ARP 31946_P050), FOXA1 (ab5089, Abcam), H3K27ac (ab4729, Abcam), or H3K4me3 (Millipore 05-1339). Immunoprecipitated DNA was end-repaired, A-tailed, and Illumina sequencing adapters ligated before 16 cycles of PCR amplification. DNA fragments ranging from 200- to 300-bp in size were selected for 50-bp single-end read sequencing on an Illumina HiSeq 2000 according to the manufacturer's instructions.

Validation of allele-specific TF binding using pyrosequencing

We performed pyrosequencing to confirm the allele-specific occupancy of CEBPA in livers from F1 mice in both genetic cross directions. The assays and primers (**Table S2**) for pyrosequencing were designed using PyroMark Assay Design Software. The annealing temperature for PCR primers was optimized by gradient PCR. Primers' efficiency was confirmed using quality controls with different proportion of BL6 and CAST DNA (0/100%,

30/70%, 50/50%, 70/30%, 100/0%). PCR conditions: 1) 95°C – 5 min; 2) 94°C – 30 sec, optimized t°C – 30 sec, 72°C – 55sec, 40 cycles; 3) 72°C – 5 min. PCR product was mixed with streptavidin beads dissolved in binding buffer and shacked for 20 min. Sequencing primers were dissolved in annealing buffer and aliquoted into PSQ plate. DNA-Beads were cleaned on the PyroMark vacuum workstation and then mixed with PSQ Primer/Annealing Buffer. The samples were incubated at 85°C for 3 min, centrifuged for 3-4 minutes at 2500 rpm and then loaded to the pyrosequencer. PyroMark Gold Q96 SQA Reagents were used to load the pyrosequencer.

Read mapping, normalization and estimation of allele-specific binding level

We constructed the *Mus musculus castaneus* genome assembly using CAST/EiJ SNV calls (ENA accession: ERS076381) against the *Mus musculus* reference assembly (C57BL/6J) [57]. Single nucleotide variants (SNVs) were mapped from their original calls on NCBI37/mm9 to the latest version of the mouse assembly, GRCm38.p2/mm10, and nucleotides at each base position were changed to reflect point mutations in CAST. SNV calls were available for all autosomes and the X chromosome.

To assess allele-specific binding and histone enrichment, we aligned reads to an alignment index comprising of both GRCm38.p2/mm10 (BL6) and CAST assemblies. Indexing of the genomes was performed using BWA (Version 0.7.3a) [58]. Raw sequencing reads were first filtered and trimmed using Trimmomatic (Version 0.3) [59]. We required a minimum phred score of 30 using a sliding window of 20 bps, and only kept a read if it matched these criteria while maintaining a minimal overall length of 40bp. We aligned filtered reads using BWA with a maximum of 2 mismatches per read (-n 2). Reads that mapped equally well to multiple locations were discarded by filtering based on the 'XT:A:U' alignment tag. Our alignment statistics showed our approach aligned reads to each strain with high specificity (see **Figure S3**). The proportion of F1 reads aligning to the combined BL6 and CAST genomes were roughly 51:49, respectively. Proportions of BL6 TFBSs versus CAST TFBSs called from these alignments were also similar.

The mpileup program from the SAMtools package [60] was used to count the number of reads that overlapped each base of the joint assembly. We then filtered these counts to retain only those genomics positions where it was possible to distinguish between BL6 and CAST backgrounds. Further filtering was done to retain those sites where a minimum of 10 reads was mapped to either F0 CAST or F0 BL6 across replicates. For F1 crosses, we retained sites overlapping at least 10 reads for at least 10 allele-specific replicates. We repeated

these steps on a site-specific manner for each TF/histone mark irrespective of whether multiple SNVs existed at each ChIP-seq peak.

Prior to fitting statistical models and further downstream analyses, we normalized for sequencing depth by adjusting for differences in library sizes across biological replicates in F0 and F1 populations for each TF/histone mark. A constant scaling factor was estimated for each library based on the median of the ratio of reads at each SNV over its geometric mean across all libraries tested. This normalization constant was then applied to each library under the assumption that count differences attributable to biological effects only exists in a small proportion of the total number of sites. This procedure was performed using R Bioconductor package 'DESeq' [61].

To assess overall peak counts and determine the quality of each ChIP experiment, we also aligned reads from each library (F0 and F1) to the GRCm38.p2/mm10 genome using GSNAP [62] with a less stringent mapping criteria. We used a less conservative mismatch threshold (maximum mismatch of 3 bases per read) to allow F1 reads derived from the CAST allele to map against the BL6 genome. Based on overall SNV numbers between the strains, a rough estimation suggests that there are approximately 1 SNV every 100 bps, which distinguishes the strains. Regions bound by both TFs and covalently modified histones were called using MACS1.4 [63] using default parameters.

To mitigate the impact of potential batch effects, biological replicates for each TF for each genetic background were prepared and sequenced in three independent flowcells.

We estimated TF occupancy levels for the histone modification H3K4me3 by taking into account the fact that histone marks typically localise over a broader genomic region than do TFBSs. Wider regions cause a dilution in the number of reads overlapping SNVs, relative to binding site numbers and sequencing depth. Hence, to increase our ability to resolve binding differences at H3K4me3 loci, we summed the counts of all SNVs overlapping the same region. To ensure background-specific peaks were captured, we constructed a summary peak file comprising of the union of genomic intervals from peak calls from individuals of different genetic backgrounds (BL6, CAST and BL6xCAST) (library reference: do3342, do3337, do3411).

We identified between 6,000-8,000 TF bound regions per TF where two or more SNVs lie within close (<250 bp) proximity; ~85% of these co-located SNVs showed the same allelic direction of TF binding between BL6 and CAST. To avoid multiple counting of TF binding

events, we only used one SNV in any 250 bp region in further analyses. Our results were highly reproducible among replicates (**Figure S2**) with similar numbers of reads mapping to each genome (**Figure S3**).

Statistical models for identifying regulatory mechanisms

ChIP-seq read counts were used as a proxy for the binding intensities of a TF to the DNA [7]. Sites were classified into regulatory categories using the method of Gonclaves et al. [29].

We defined as conserved those regions with equal TF binding occupancy between BL6 and CAST in both F0 and F1 individuals, despite the presence of one or more variants near the site of binding; these types of sites could also be described as non-differentially bound [25]. We defined cis-driven TFBSs as sites where binding occupancy differences between strains were determined by locally acting genetic sequences; hence, the TF occupancy ratios between BL6 and CAST genomes found in the F0 parents is the same as that observed between alleles in the F1 offspring. TF binding affected in trans were defined based on TF binding occupancy differences between parents, but not between alleles in the F1 offspring. Finally, cistrans mechanisms show a complex mixture of cis and trans acting influences.

For each TF or histone mark, F0 counts from each strain were modelled as a negative binomial marginal distribution while F1 counts were modelled using a beta-binomial distribution where the parameters of the beta distribution were used to model the proportional contribution from each allele. For each TF and histone mark, there were 6 replicates (*i*) for each F0 strain and 12 replicates (*j*) for F1 samples. F0 counts for each strain (x_i, and y_i) were assumed to follow negative binomial distributions while F1 counts (n_j), were modeled on an allele-specific basis (z_j) using a beta-binomial distribution:

$$x_i \sim Po(\mu_i),\ y_i \sim Po(\upsilon_i), z_j \sim Bi(n_j, p_j)$$

$$\mu_i \sim Ga\left(r, \frac{p_\mu}{1-p_\mu}\right), \upsilon_i \sim Ga\left(r, \frac{p_\upsilon}{1-p_\upsilon}\right), p_j \sim Be(\alpha, \beta)$$

where x_i is formally defined as the binding intensity of the variant in the *i*th C57BL/6J F0 mouse, y_i is the binding intensity of the variant in the *i*th CAST/EiJ F0 mouse, n_j is the

number of reads mapping across both allelic variants in the *j*th F1 hybrid and z_j is the number of reads mapping to the C57BL/6J allele in the *j*th F1 hybrid.

We estimate the dispersion parameter *r* for F0 samples using the 'estimateDispersions' function within 'DESeq' with local regression fit. *r* was used as the reciprocal of the fitted dispersion value from 'DESeq'.

We constrained parameter estimation for each distribution based on four different regulatory scenarios and derived maximum likelihood values for each hypothetical case on a site-by-site basis. The four models are described below:

$$Conserved: p_\mu = p_\nu \; and \; \alpha = \beta$$

$$Cis: p_\mu \neq p_\nu \; and \; \frac{\alpha}{\alpha + \beta} = \frac{\frac{p_\mu}{1 - p_\mu}}{\frac{p_\mu}{1 - p_\mu} + \frac{p_\nu}{1 - p_\nu}}$$

$$Trans: p_\mu \neq p_\nu \; and \; \alpha = \beta$$

$$Cistrans: p_\mu \neq p_\nu \; and \; \alpha \neq \beta$$

To identify the most probable model at each variant we used the Bayesian information Criterion (BIC).

To avoid confounding results from the analysis of variants derived from the same binding site, downstream analyses only used variants spaced at least 250 bps apart. Hence, where two or more variants were found spaced within 250bps of one another, only one variant was chosen for subsequent analyses.

Identification of motif-disrupting variants

MEME [64] was used to perform *de novo* search for enriched motifs for all peaks called for three randomly chosen ChIP-seq samples from each TF (library identifiers do3488, do3463 and do3483). Sequences +/-50bp from all peak summits were extracted for analysis based

on the assumption of one motif per peak. Where multiple motifs exist in a peak, the motif sequence with the best score was retained.

Statistical test for regional enrichment of mechanisms driving TF occupancy

Enrichment for TF regulatory categories that overlapped with the location histone marks were assessed using the exact binomial test. Colocation was defined using an overlap of 1bp. The probability of Bernoulli success was defined for each TF based on their proportion of binding categories.

To assess whether collocating TFs (i.e. binding at the same SNV) share the same regulatory category (i.e. cis, cistrans, conserved, trans) more often than expected by chance, we calculated the expected probability of Bernoulli success as follows:

$$p_i = b_{FOXA1,i} \times b_{HNF4A,i} \times b_{CEBPA,i}$$

where *b* is the proportion of TFBSs in regulatory category *i* at TFBSs where all three TFs collocate.

Differential binding analysis between mice that are heterozygous versus wild-type for HNF4A and CEBPA

Cepba$^{FLOX/-}$ and *Hnf4a*$^{FLOX/-}$ mice were assayed by ChIP-seq using antibodies for CEBPA, HNF4A and FOXA1. Three biological replicates per condition (HET or WT) per antibody were compared for changes in binding intensity to their wildtype counterparts. We quantified the difference in TF binding intensity between heterozygous mice and wildtype mice, and then sorted the TFBSs based on whether their occupancy was conserved, or driven by cis- or cistrans-acting mechanisms. Binding intensities were considered as the number of reads at the summit of peaks that were called by MACS1.4 [63]. The same WT input libraries were used for peak calling in both HET and WT samples. We filtered out peaks with a read count cut-off of less than 11 reads in less than 5 libraries. Prior to differential binding comparisons, upper quantile normalization [65] was used to adjust for differences in sequencing depth between libraries. For each TF, 'edgeR' [66] was used to identify peaks with different binding intensities between HET and WT samples. A significance cut-off of FDR<0.1 was used.

Statistical models for assigning modes of TF occupancy inheritance

To identify the mode of TF binding intensity inheritance at non-conserved TFBSs, F0 and F1 libraries were first adjusted for differences in sequencing depth using the median of the ratio

of reads at each SNV over its geometric mean across all libraries as a constant normalization factor for each library [61]. Next, data from each SNV was fitted to statistical models reflecting either additive or dominant/recessive inheritance patterns. Models were constructed based the following premise: if offspring binding intensities were inherited via an additive mode of inheritance, we would expected the combined offspring binding intensity from both alleles to equal the summed binding intensity of parental alleles; on the other hand, if inherited through a dominant/recessive mode of inheritance, we would expect the combined binding intensity in the offspring across both alleles to equal the total intensity of one but not the other of its parents. We assumed read counts followed negative binomial distributions. Here, we formally define the models:

$$x_{max,i} \sim Po(p_{max,i}), x_{min,i} \sim Po(p_{min,i}), y_i \sim Po(o_i)$$

$x_{max,i}$ is defined as the normalized read count binding intensity of the variant in the *i*th F0 mouse from the parental strain showing the higher median binding intensity among replicates, $x_{min,i}$ is the normalized read count binding intensity of the variant in the *i*th F0 mouse from the parental strain with the lower median binding intensity among replicates. y_i is the binding intensity of the variant in the *i*th F1 mouse summed across both alleles.

$$p_{max,i} \sim Ga\left(r, \frac{S_{pmax}}{1 - S_{pmax}}\right), p_{min,i} \sim Ga\left(r, \frac{S_{pmin}}{1 - S_{pmin}}\right), o_i \sim Ga\left(r, \frac{S_o}{1 - S_o}\right)$$

As above, the dispersion parameter, *r*, was estimated using 'DESeq'. We used maximum likelihood estimation to fit the counts to the models below and used BIC to assess which of the following two models best fit counts from each cis- and trans-regulated site.

$$Dominant: S_{pmax} = S_o \ or \ S_{pmin} = S_o$$

$$Additive: S_{pmax} \neq S_o \ \ and \ \ S_{pmin} \neq S_o$$

We excluded those sites from our results where the parameter estimated for the offspring, S_o, was indistinguishable from the parameters estimated for both parent, i.e. if $S_o = S_{pmax}$ and $S_o = S_{pmin}$. Such sites were determined by comparing the dominant and additive models separately for $p_{max,i}$ and $p_{min,i}$ and excluding sites found to fit the dominant model in both. It is possible that additively inherited TFBSs may be misclassified if the difference in binding

intensities between the parental measurements is small enough that the F1 measurement is statistically indistinguishable from either parent due to measurement noise. To minimize this potential source of error, we restricted tested sites to those TFBSs where the difference between the means of $B6_{F0}$ and $CAST_{F0}$ across biological replicates was equal or greater than twice the standard deviation of the average binding intensity across biological replicates (this was set at 19 normalized counts or more). To further increase confidence in our results, we only used sites assigned to their regulatory category with BIC>1.

Over- and under-dominant TFBSs were identified by first restricting all TFBSs to those classified to a regulatory class with BIC>1. Normalised count data at each TFBS was fitted to the models described above. For each TFBS where the binding occupancy of each parent did not equalled to that of the offspring (i.e. $S_{pmax} \neq S_o$, $S_{pmin} \neq S_o$), TFBSs were classified as under-dominant if the mean F1 occupancy level among replicates was less than that of both parents, on the other hand, TFBSs where the mean F1 occupancy level was greater than that of both parents were termed over-dominant.

Statistical models to distinguish between cis and cistrans influences at lineage-specific TFBSs

Read counts were normalized between F0 and F1 libraries as described in the previous section [61]. Lineage-specific binding sites were defined as those sites meeting these criteria: ($ratio_{F0}$<0.05 and $ratio_{F1}$<0.05) or ($ratio_{F0}$>0.95 and $ratio_{F1}$>0.95). $ratio_{F0} = B6_{F0}/(B6_{F0}/CAST_{F0})$ and $ratio_{F1}=B6_{F1}/(B6_{F1}/CAST_{F1})$, where values were mean levels of binding among biological replicates. We expect that a lineage-specific site that is purely cis-regulated to possess F1 count levels that are half of that in F0. Significant deviation from this 2:1 ratio would indicate a trans effect. We constructed the following statistical models to test the likelihood of these scenarios for each lineage-specific site and used maximum likelihood estimation and BIC to choose the model of best fit. At each TFBS, reads across replicates were modelled using the negative binomial distribution.

$$x_i \sim Po(p_i), 2y_i \sim Po(o_i)$$

$$p_i \sim Ga\left(r, \frac{S_{pmax}}{1 - S_{pmax}}\right), o_i \sim Ga\left(r, \frac{S_o}{1 - S_o}\right)$$

x_i is defined as the normalized read count binding intensity of the variant in the *i*th F0 mouse from the strain of lineage-specific binding. y_i is the binding intensity of the variant in the *i*th

F1 mouse summed across both alleles. The dispersion parameter, *r*, was estimated using 'DESeq', as described above. We tested the two following scenarios:

$$Cis: S_{pmax} = S_o$$

$$Cistrans: S_{pmax} \neq S_o$$

Comparison of regulatory mechanisms underlying variation in gene expression and TF binding

Logistic regressions were used to examine the relationship between gene expression and TF binding. For each gene where expression variation is driven of each of the following mechanisms: cis, cistrans, conserved and trans, we represent the transcriptional context by taking all TFBSs 20kb upstream and 10kb downstream of the TSS and counting the numbers of each TFBSs in each TF regulatory category. Counts of TFBSs in each regulatory category (i.e. number of TFBSs where occupancy levels were driven in cis, etc) were then used as four independent predictive variables. Separate regressions were performed using each of the four expression regulatory classes in turn as the dependent variable. The binary nature of the dependent variable was defined using remaining regulatory categories. We used the same strategy to study the relationship between TF binding and chromatin state (H3k4me3), that is, the mechanistic relationship between TFBSs proximal to the histone mark was assessed using logistic regression. The size of the genomic regions used for the grouping of TFBSs was +/- 2kb from each histone mark location. To test for shared regulatory mechanisms between H3K4me3 and gene expression, the histone marks were first linked to genes (within 5kb upstream of a TSS). As H3K4me3 marks active promoters, we expect a 1:1 relationship between the histone modification and expressed genes. Binomial tests were then used to calculate the statistical enrichment of shared regulatory mechanisms at histone mark-gene pairs.

We computed the diversity of TF regulatory mechanisms for genes grouped by expression mechanisms using Shannon's diversity index (*H'*) [67], which was calculated for each gene as follows:

$$H' = -\sum_{i=1}^{4} a_i \ln a_i$$

where a_i is the proportion of binding sites belonging to the *i*th TF binding regulatory category within 20kb upstream or 10kb downstream of a liver-expressed protein-coding gene.

Gene expression levels show correlation with TFBS abundance. Thus, highly expressed genes are expected to be proximal to a more diverse set of mechanisms underlying TF occupancy change than by chance alone. Hence, to control for differences in expression levels, we subsampled genes to obtain matched gene expression levels between comparison sets. Gene expression levels were compared based on the average expression value among biological replicates of the more highly expressed parent. Mean expression levels were first log transformed then separated into 20 bins of equal consecutive intervals. Each cistrans-directed gene was then matched to a conserved regulated gene assigned to the same expression bin. In the same way, cis-driven genes were matched in expression values to conserved genes. All subsampling was done with replacement.

Measuring the coordination of TF binding occupancy

To determine the genomic region under the influence of any set of cis regulatory variants, we calculated correlation coefficients for binding intensities of TFBS pairs at successive genomic intervals away from each cis-directed TFBS. To capture the coordination of TF occupancies between TFBSs, we calculated Spearman's correlation coefficient of allelic proportions (BL6/(BL6+CAST)) between binding sites at consecutive distance bins centred upon cis-regulated variants. Spearman's Rho was calculated for each mutually exclusive bin with their 'anchor' peak. Interval width increased by 1kb at each succeeding bin extending from 1kb from the cis-driven variant. We performed linear regression using log-transformed distances as the predictor variable with Spearman's Rho estimates as the outcome variable to quantify the decay in correlation signal (**Methods**, **Figures 4a-b, S10**).

In order for meaningful inference, we generated a null distribution of the correlation of binding strengths by comparing occupancy levels of anchor TFBSs with the occupancies of other TFBS locations sampled randomly from across the genome. Null values were calculated using TFBSs that were randomly sampled from the total pool (without replacement) to simulate a set of binned peaks for each anchor peak (anchor peaks were kept constant). The total number of binned peak simulated was equal to the total number of anchored–binned peak pairings observed. Spearman's Rho was then calculated as described for the observed set.

To estimate the genomic distance at which the 'elbow' or maximum curvature of the curve occurs, we used a vector projection method on the fitted regression curve [68]. First, we drew a

line connecting the points from $x = 1$kb to where $x = 50000$. Next, for every point on this line at values of x we extended perpendicular lines to intersect with our regression line. We then measured the lengths of each of these lines and selected the point with the longest length as the estimate of the elbow.

Hi-C data processing and analysis

Hi-C libraries were generated from pooled liver samples from two 2-4 week old mice [37]. Raw data files were quality filtered using Trimmomatic [59] using identical parameters to those described above. We used the Homer Hi-C software (http://homer.salk.edu/homer/interactions/) to process Hi-C reads and to identify significant interactions. Restriction sites ('AAGCTT') were trimmed from our reads prior to mapping to the GRCm38.p2/mm10 genome using GSNAP [62] at a maximum of two mismatches per read. Only reads mapping to unique locations in the genome were retained. Paired reads that likely represent continuous genomic fragments or re-ligation events were removed if the reads are separated by less than 1.5x the sequencing insert fragment length (-removePEbg). Paired ends that originate from areas of unusually high read density were also removed by scanning 10kb regions in the genome and removing reads containing greater than five times the average number of reads (-removeSpikes 10000 5). Only reads where both ends of the paired read have a restriction site within the fragment length 3' to the read were kept (-both). We also filtered reads if their ends self-ligated with adjacent restriction sites (-removeSelfLigation).

To detect significant interactions between two genomic locations, it was necessary to create a background model that accounts for the primary sources of technical biases to count enrichment. Closely spaced loci are inevitably enriched for interactions due to their close proximity. We used Homer to normalize both for linear distance and read depth. We normalized our reads at 10 kb regions across the genome and examined the number of interactions occurring between these regions. Enrichment for significant interactions were identified using a binomial test against the expected number of interactions based on the background model that also accounts for the total number of reads mapping to each locus being tested. Briefly, the parameters for the binomial test are as follows: the probability of success is the expected interaction frequency (which vary depending on restriction site locations), the number of success is the number of reads mapping between the loci, and the number of trials is the overall number of significantly interacting reads.

SUPLEMENTARY RESULTS:

S1. TF binding affinity is more strongly regulated in cis than gene expression

In 80% of instances when we compared any randomly chosen TFBS to any randomly chosen expressed gene, the magnitude of cis effect was greater for TF occupancy than for gene expression (magnitude measured by the distance between F1 alleles over 10,000 random comparisons).

S2. Trans-influenced lineage-specific TFBS show little difference in compensatory versus diversifying effects

We next calculated the percentages of trans-influenced binding locations that have a decreased difference in the F1 mouse in binding intensities between the alleles (compensatory) versus increased differences (diversifying). Under complete neutrality, they should be equally favoured (Tirosh et al. 2009). The amount of compensatory versus diversifying effects is not significantly different at lineage-specific TFBSs (binomial test, P=0.6) (**Figure S8a**). In comparison, of the 2,563 non-lineage-specific cistrans-regulated CEBPA binding sites, 64% are compensatory and 36% diversifying (binomial test, P<2.2e-16) (**Figure S6**). These numbers closely mirror the proportion of compensatory versus diversifying effects reported for gene expression in liver (68% compensatory, 32% diverging) (Goncalves et al. 2012). No strain-specific TFBS that are regulated purely in trans were observed (i.e. strain-specific in F0 but equally bound in F1). In other words, our results strongly suggest that cis-directed mechanisms may either directly (e.g. modification of the binding motif) or indirectly (e.g. through the opening up of chromatin by altering the shape of the DNA) play a required role in birth of TFBSs.

S3. No difference in selective pressure was detected between strain-specific TFBS that are gained and those that are lost

The divergence time between BL6 and CAST is estimated to be less than a million years (Geraldes et al. 2008). Lineage-specific TFBS can be caused by: 1) lineage-specific loss of a TFBS that existed in the common ancestor of BL6 and CAST (plesiomorphic), or 2) lineage-specific gain since the most recent common ancestor in one strain (apomorphic) (**Figure 5c**). To identify gained versus lost TFBS, we compared our lineage-specific TFBS with matched TFBS data obtained from livers of *Mus Spretus* (SPR) (Stefflova et al. 2013), a mouse species of equal evolutionary distance (ca. ~1.5–2 MY) to both BL6 and CAST (Dejager et al. 2009). We distinguished between BL6 versus CAST lineage-specific binding sites that are apomorphic (present in BL6 not CAST or SPR and present in CAST not in BL6 or SPR) and plesiomorphic (shared between BL6 and SPR and between CAST and SPR but not BL6 and

CAST). Around 35% of TFBSs strain-specific between BL6 and CAST were also found in SPR, placing a lower bound on the number of plesiomorphic TFBSs. Proportions of cis and trans-influenced TF binding locations were evenly distributed between apomorphic and plesiomorphic (binomial test, P>0.01) suggesting that there is little difference in selection pressure between strain-specific TFBS that are gained and those that are lost.

S4. Over- and under- dominant patterns of TF occupancy inheritance

We searched for evidence of over- and under- dominant patterns of occupancy inheritance that correspond to respectively stronger or weaker F1 occupancy levels compared to parental measurements. In gene expression, this pattern of imbalance is thought to be associated with hybrid incompatibilities (Landry et al. 2005; McManus et al. 2010), and comprises approximately 27% and 8% (under- and over- , respectively) of expressed genes between two strains of fruit flies (McManus et al. 2010). In mice, we found that 6% and 11% of liver expressed genes showed under- and over- dominant modes of expression inheritance (**Figure S7**). In contrast, less than 1% of sites in mouse tissues were determined as under- or over- dominant across all TFBSs (where BIC>1) (**Supplementary Data File)**.

S5. Accounting for gene expression level when connecting the regulatory mechanisms underlying gene expression to TF occupancy

In addition to using logistic regression to connect the mechanism underlying TF binding with gene expression, we tested for similar associations using an alternative strategy which takes into account expression level differences between genes. As before, we took all protein-coding genes with at least one binding event in the region 20kb upstream and 10 kb downstream of the TSS. We then subsampled genes to match gene expression levels between regulatory classes. Again, we found that genes showing conserved expression levels were depleted for TFBSs with occupancy driven in cis (Mann-Whitney U test; on a per gene basis comparing the numbers of different TFBSs near conserved regulated genes against genes where expression is regulated in cis and cistrans, P=9.8e-11 and 2.2e-16, respectively). Hence, genes whose expression variation is regulated both in cis and cistrans possessed a higher than expected number of TFBSs driven in cis proximal to the TSS. Analysis of trans-regulated genes was generally non-informative due to the small number of genes (14) in this category.

S6. Accounting for gene expression level in TFBS diversity analysis

We considered the possibility that the association between diversity of TFBS and category of gene expression above might be due to differences in gene expression levels within each category. To control for this, we repeated the analysis by subsampling genes from each

regulatory category to generate subsets with matched expression levels. We observed little difference on our core results (Mann Whitney U test; cistrans versus conserved: P=1.4e-7, cis versus conserved: P=1.4e-3).

For the 24 libraries of each TF:

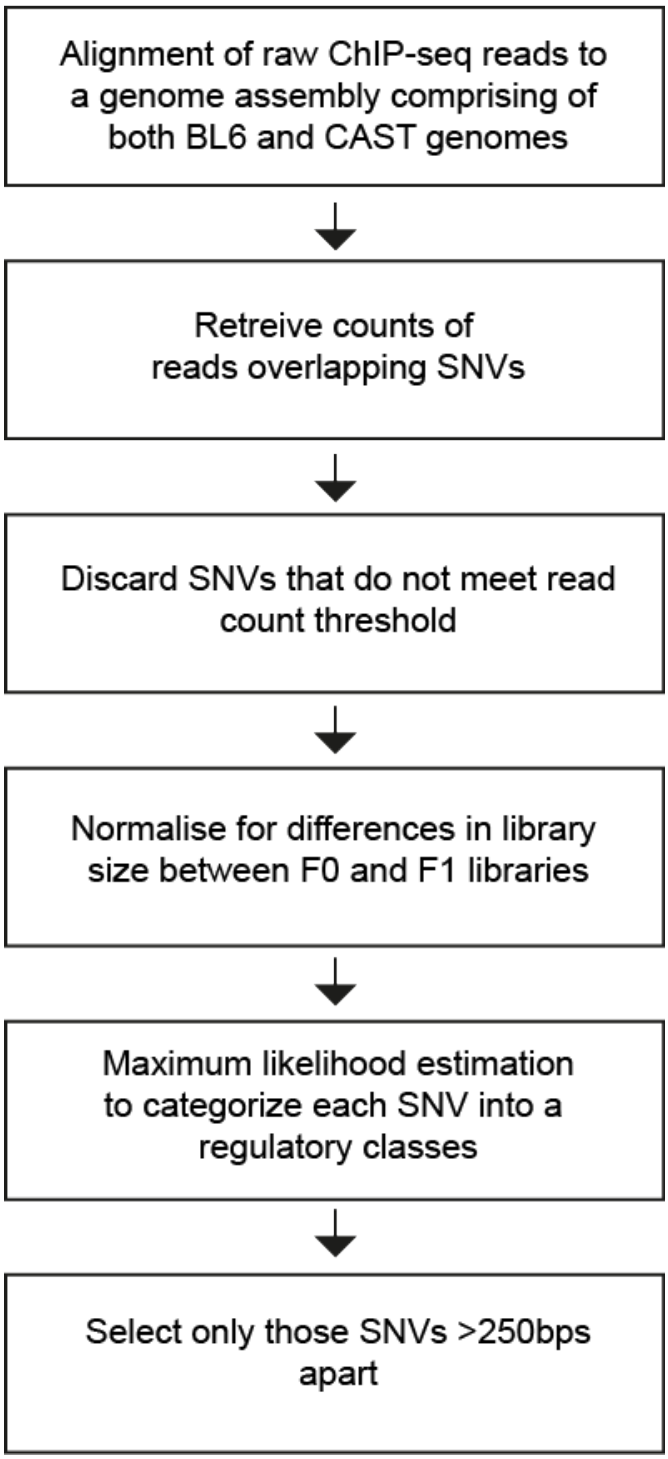

Figure S1. Overview of computational analysis pipeline

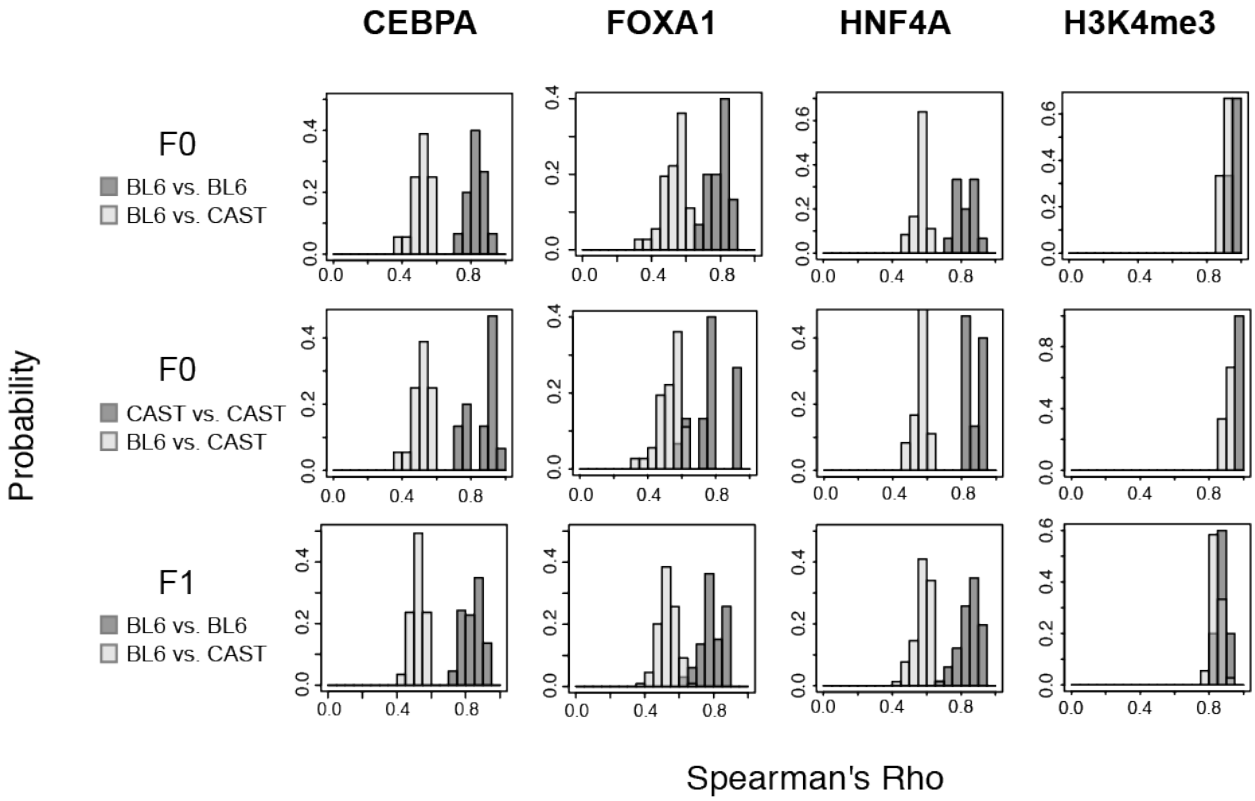

Figure S2. Correlation of ChIP-seq measurements at SNVs between libraries of the same genetic background are consistently greater than between libraries from different genetic backgrounds. The histograms show the frequency of Spearman's Rho for ChIP-seq measurements at SVNs between libraries. For each F0 library of the same TF or histone mark, Spearman's Rho was calculated using libraries generated from the same genetic background (either BL6 or CAST) (dark colour). The frequencies of these values were displayed in the same plot as Rho values generated from comparisons with libraries from individuals of the other strain (light colour). For F1 individuals, correlations were made on an allele-specific basis.

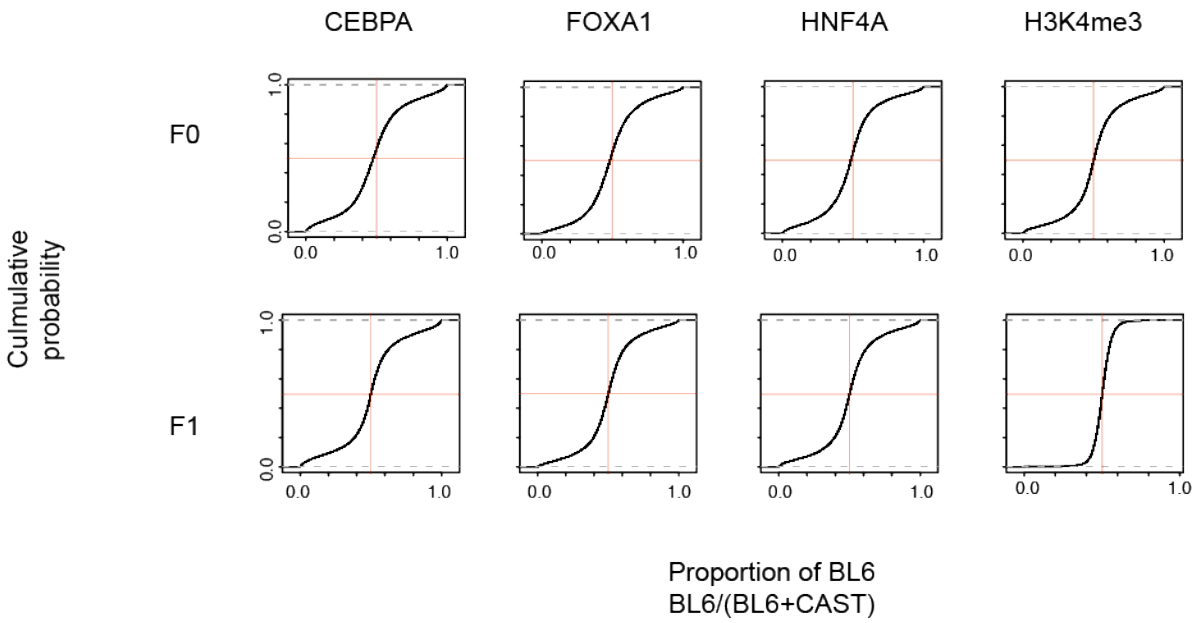

Figure S3. A similar proportion of BL6 and CAST reads were mapped. Culmulative probability based on the ratio of BL6:CAST ChIP-seq measurements at SNVs were plotted. The vertical red line indicates 0.5 culmulative probability and the horizontal red line indicates where BL6/(BL6+CAST) = 0.5 (i.e. an equal number of BL6 and CAST reads).

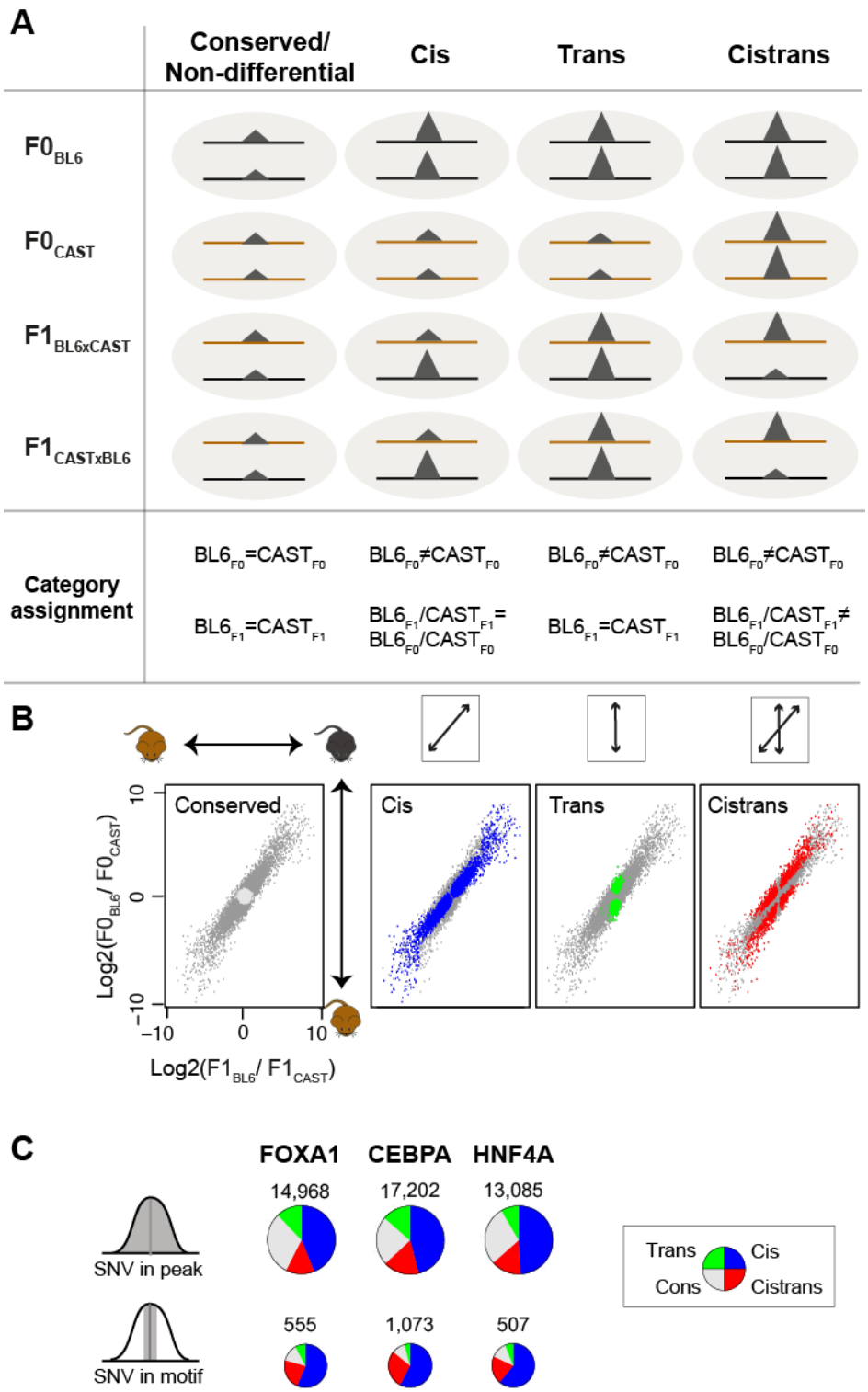

Figure S4. The assignment of regulatory status shows that TF occupancy levels are cis-driven for a large proportion of TFBSs

(A) Regulatory categories for variation in TF binding intensities were assigned based on comparison of normalized ChIP-seq read counts between BL6 and CAST at SNVs overlapping TFBSs. Due to a common nuclear environment, trans effects that are mediated by diffusible elements are expected to impact both alleles equally. Based on this, by

comparing BL6 and CAST ratios between F0 and F1 individuals one can classify TFBSs into various regulatory categories, namely – conserved, cis, trans, and cistrans. **(B)** Scatterplots of BL6 vs CAST ratios of TF binding intensities in F0 and F1 individuals. Each point represents a separate SNV. Regulatory categories are highlighted in separate scatterplots. Dark grey colour shows the remaining binding variants that do not belong in the highlighted category. CEBPA data is shown. **(C)** Pie charts depict the relative proportion of interrogated SNVs of each regulatory class for all SNVs overlapping TF bound locations, and only for SNVs positioned in the regulatory motif.

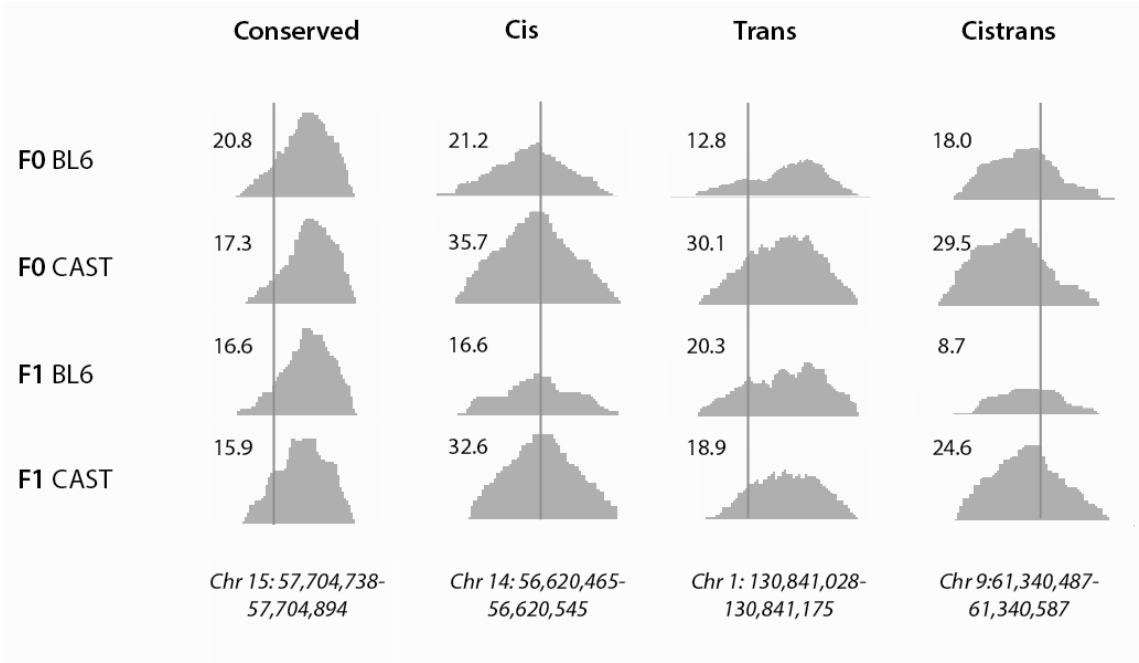

Figure S5. Examples of CEBPA binding sites classified into different regulatory modes. TFBSs were classified based on the statistical models described (see Methods). Numbers shown are normalized ChIP-seq counts at SNV locations. These sites are marked by a vertical line.

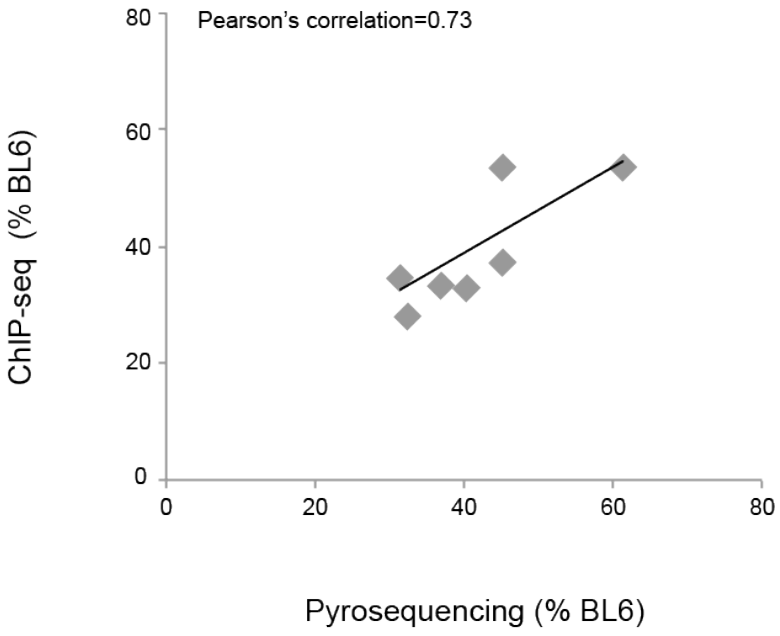

Figure S6. Validation of allele-specific measurements made using ChIP-seq. On the y-axis, a measure of allele-specific occupancy change (using the BL6 allele as the reference) determined from ChIP-seq is plotted for seven SNVs, each of which is located under a separate CEBPA peak. On the x-axis, the corresponding measure of allele-specific binding occupancy determined from pyrosequencing is shown. ChIP-seq measurements were derived from the average allelic ratio across all F1 individuals in the study. Pyrosequencing measurements were taken as the average measurement across six biological replicates (averaging across the three technical replicates). Importantly, the correspondence between the ChIP-seq and the pyrosequencing data was high for SNVs even at moderate levels of binding occupancy change in allelic ratio. SNVs for which primer allelic biases were detected were not included in plot (see **Table S2** for full list of primers).

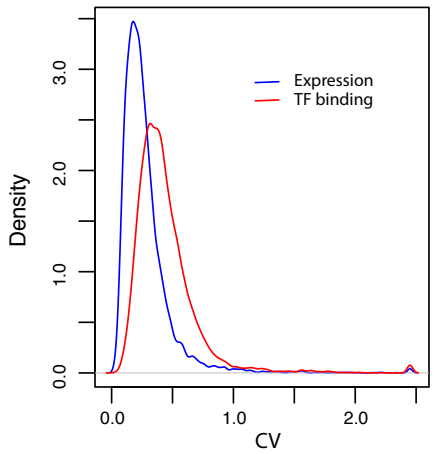

Figure S7. TF binding measurements are slightly noisier than gene expression. We use the coefficient of variance (SD/mean), a standardized measure of variance, to compare the level of noise across our replicates for TF binding to that of gene expression. The figure shows that expression measurements are more consistent across replicates than TF binding occupancy measurements. This is not surprising as the expression level of a gene can be estimated by many more SNVs than TF binding (thus reducing residual error). Therefore, the difference between TF binding and gene expression shown in Figure 2A cannot be explained by the difference in experimental error or biological noise.

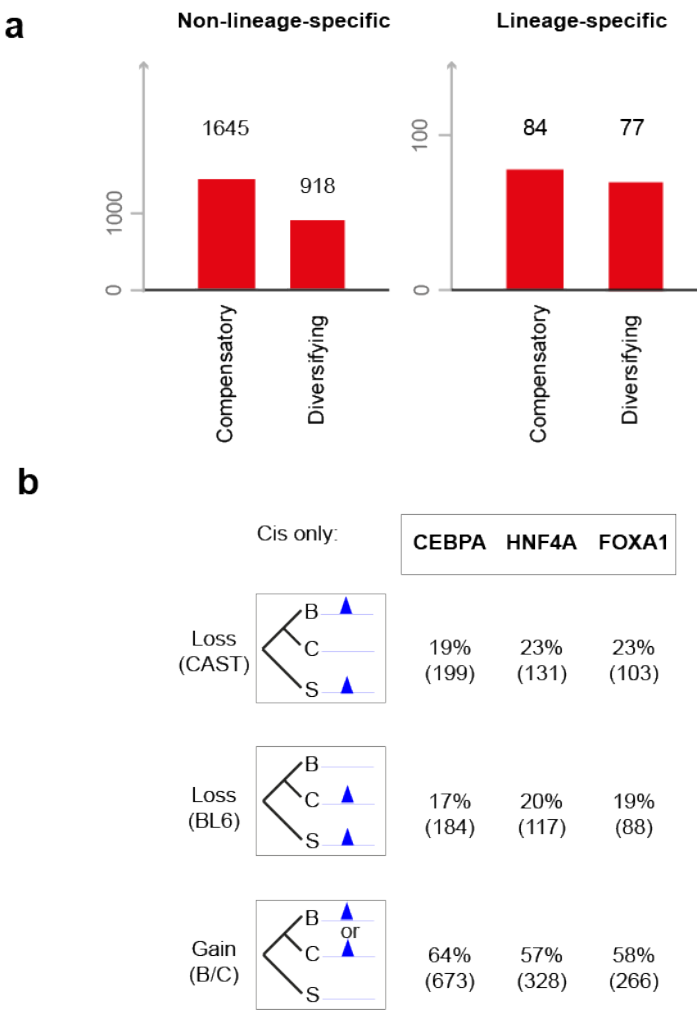

Figure S8. Compensatory versus diversifying cistrans modes of regulation. (**A)** Cistrans regulated sites can be further classified into those showing either diversifying or compensatory effects. These can then be compared to cistrans-regulated sites from all non-lineage-specific binding events. Data for CEBPA is shown. **(B)** Purely cis-driven highly allele-specific TFBSs were classified into those that were gained in BL6 or CAST and those which were lost in in BL6 or CAST based on parsimony by comparison with TF binding data in *Mus spretus* (counts in brackets).

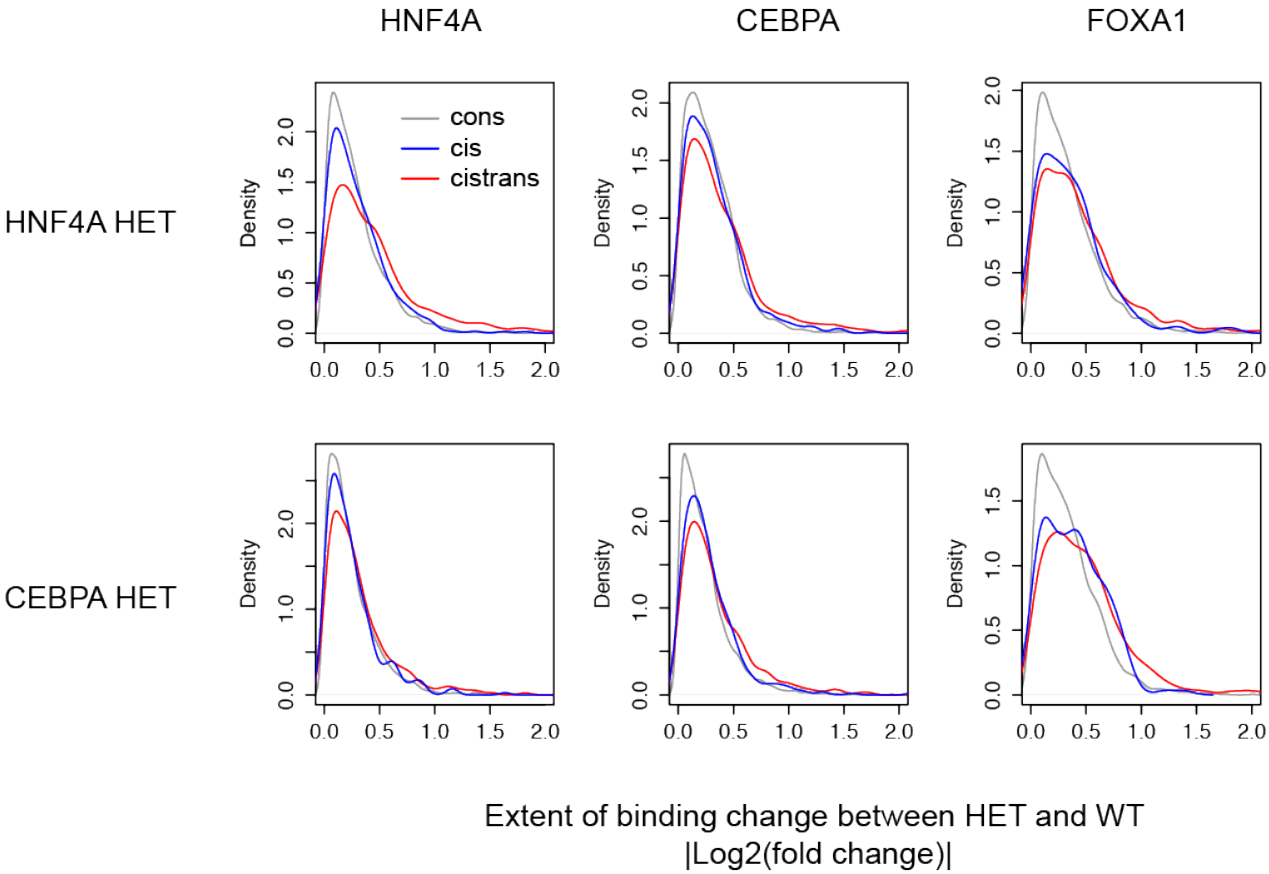

Figure S9. Perturbation of TF expression is more likely to cause changes in the binding occupancy of cis and cistrans driven TFBSs. To restrict analyses to confidently called regulatory categories, only cis and cistrans driven TFBSs classified with BIC>2 are shown.

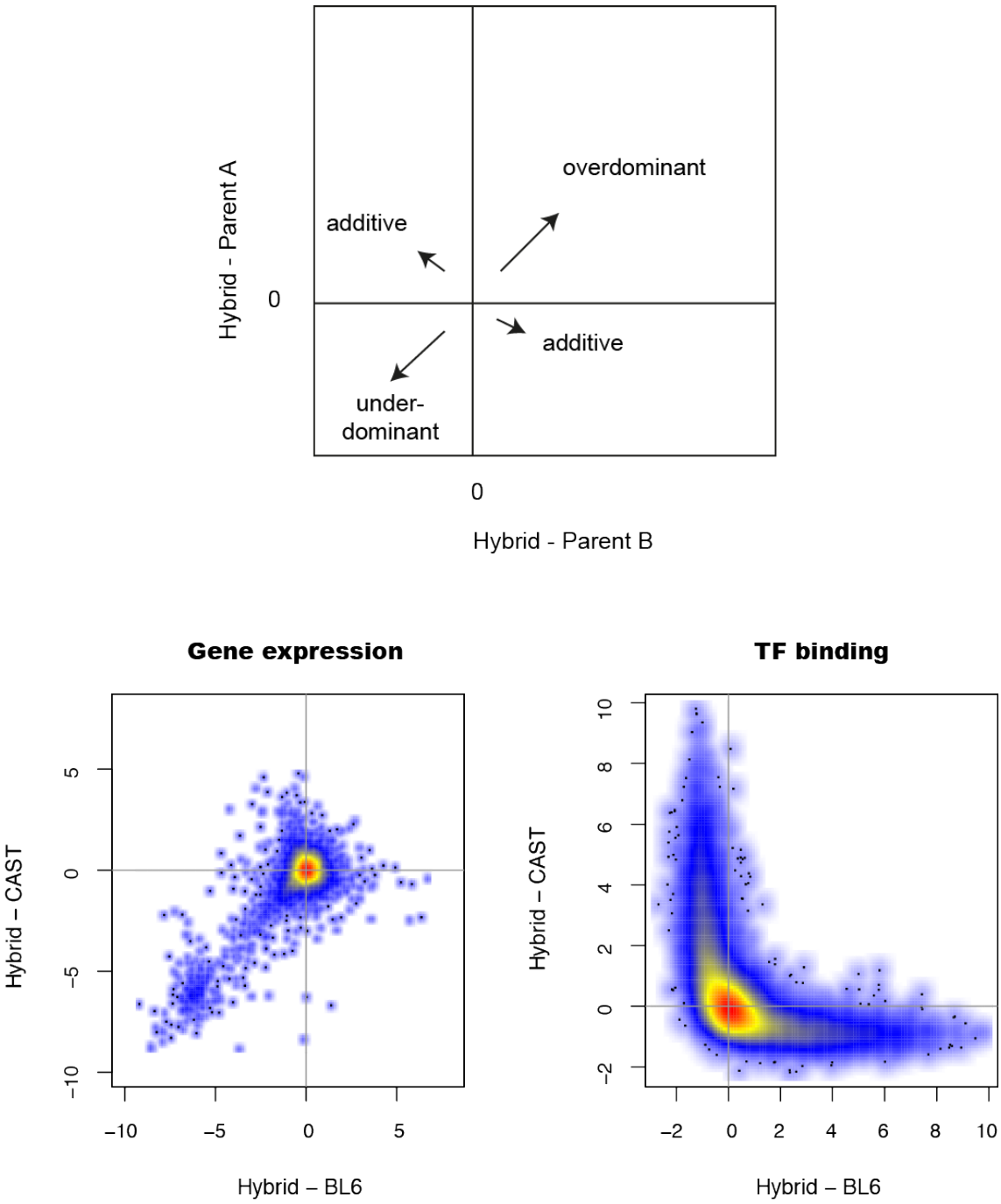

Figure S10. Patterns of inheritance of gene expression and TF occupancy levels are distinctly different. Smoothed density scatterplots showing the distribution of inheritance patterns for genes and TFBSs (CEBPA). Hybrid and parental values were summed across both alleles. Axes show log2 transformed parental subtracted from log2 transformed hybrid values (i.e. log2 (hybrid) - log2 (parental)). All values have been normalized for sequencing depth differences across F0 and F1 libraries.

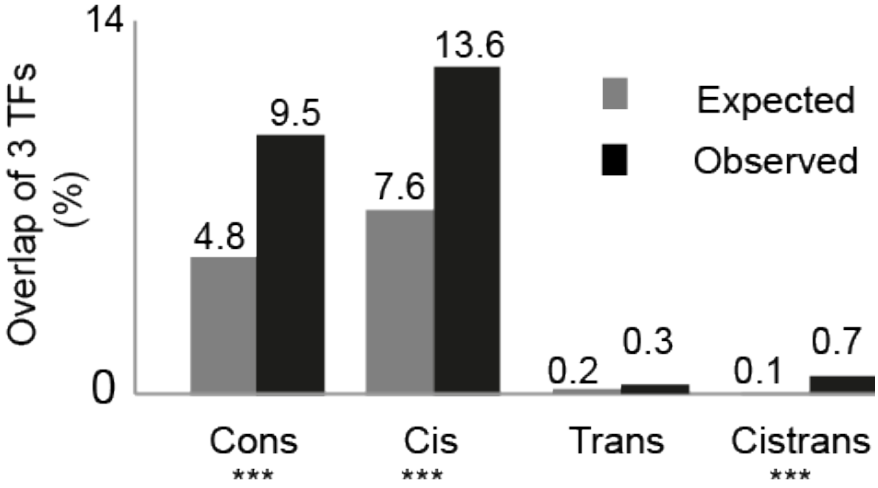

Figure S11. Colocating TFBSs show coordination in their mode of regulation. Percentage of expected and observed instances where all three colocating TFBSs (CEBPA, HNF4A, FOXA1) are regulated in cis, trans, conserved and cistrans. ***P<0.0001.

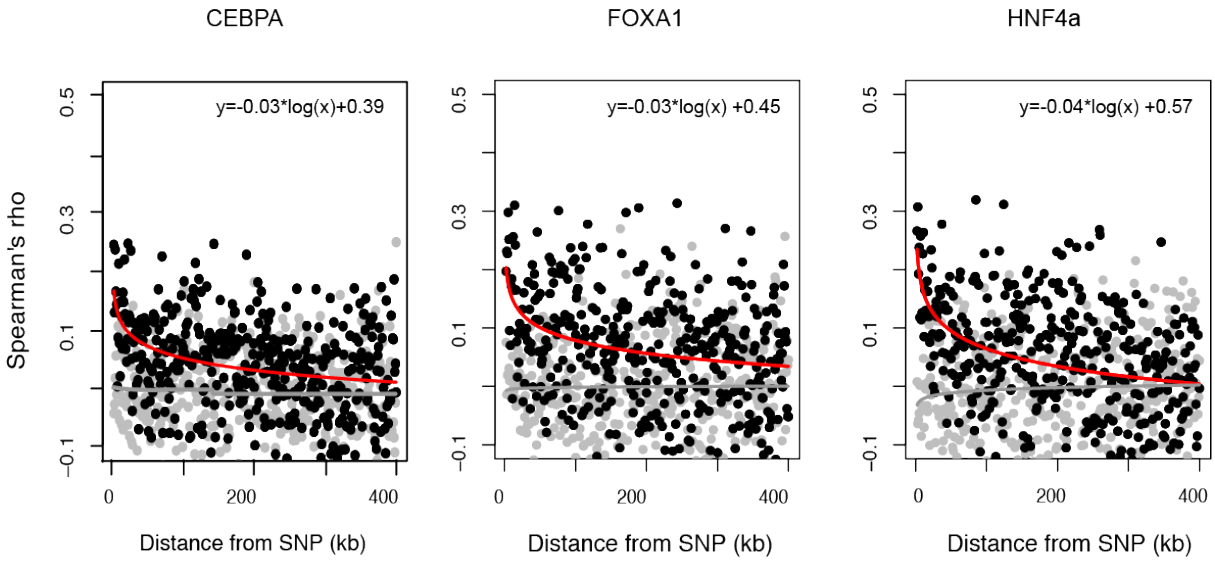

Figure S12. Rapid loss of cis-mediated inter-peak correspondence with genomic distance (0-400kb) Spearman's ρ values for each bin were plotted for each TF. Red solid line is the linear regression line. Grey dots represent the background distribution. These data points were constructed by random subsampling of TFBSs to anchor TFBSs (see Methods). The numbers of TFBSs in each randomly sampled bin were matched to those in the observed bins. The grey line is the linear regression line for the correlation values derived from sampled points.

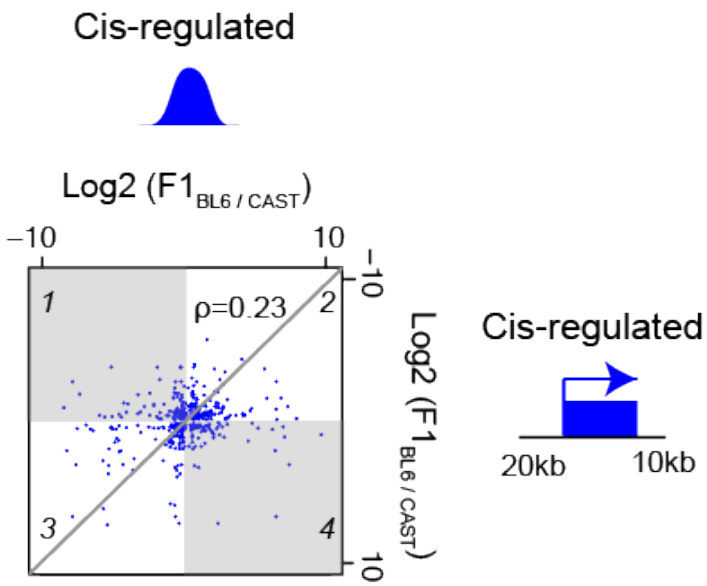

Figure S13. Comparison of allele-specific ratio for cis-regulated TF binding intensities and gene expression values for cis-regulated genes. TFBSs were associated to a gene based on their location either 20kb upstream or 10kb downstream of the TSS of an expressed cis-regulated protein-coding gene. Averaged values across biological replicates are plotted. The probability of points lying in quadrants 2 and 3 is 0.6.

Odds ratio of chromatin contact (relative to **Cis**)	**Cons**	**Trans**	**Cistrans**
CEBPA	**1.20** *** [1.08-1.34]	1.07 [0.94-1.22]	1.02 [0.90-1.14]
FOXA1	**1.16** ** [1.05-1.29]	**1.26** *** [1.10-1.44]	**0.82** *** [0.71-0.95]
HNF4A	**1.14** * [1.03-1.27]	1.10 [0.91-1.29]	0.97 [0.84-1.12]

Table S1. Odds ratios for chromatin contact enrichment at different regulatory categories. We derived odds ratios from the coefficients of a logistic regression analysis whereby the underlying regulatory mechanisms were regressed against whether a TFBS location overlapped a region displaying enrichment for long-range chromatin contact (as determined by Hi-C). 95% confident intervals are presented in brackets. ***P<0.0001 **P<0.001* P<0.05.

Primer name	Forward primer	Reverse primer	Sequencing primer
cis_1	TAAGCTGCGAGAACCTCTGAT	[Btn]CCTGCTTGGCTTGTGTGAAT	ACGTCCTCCCTGACC
cis_2	[Btn]ACTGGCAAGAGGCAATGAGC	TCTCATGGGGGACTTCGG	GTAGGGCCTGGGCGT
cis_3	TATGGGGATTACGGGGTCTG	[Btn]GTTTGGAAAGAACCCCGACAG	GAAAGTGGAAAGCCTC
cistrans_1	AAAGGGAGCCTGGAACCACAT	[Btn]GGCATCCATCTTGACAGGAGTT	GAACCACATCTGCCTC
cistrans_2	[Btn]TGCAAGGAGCCATCATTCT	CACATCCGTTTGTTGCTGAG	ACTGAACTACATACTTACCA
cistrans_3	AGCTACTCTGAAGCGGTTTGC	[Btn]CCTGGGGCTTCACATCAAT	GAGGAAAAAAGAAATGTAGA
conserved_1	[Btn]TGTGTGTGCTGCAACTGATGG	GGGAGGCTTAGGAAGAGGTCAATA	CGGGGAGGAGGTGTG
trans_1	ATGCTTTGAACTGTTGCACTGTCT	[Btn]CCCCTAAGCAAGTCTCAAAGTG	TCTCTTCAGCTCAATTCTC
trans_2	CTGCTTGGTGCCTGTGCT	[Btn]GGTAACCAGAGTAGCGGCTCAG	CTGCGACCGAGCCAG
trans_3	GCTGGCAAGTGACCCTGAGT	[Btn]AGTCCCTGAACAGACACCCACTTA	CAGTATAGTTAGGAATCCCC

Table S2. Primers used for pyrosequencing validation

Primer name	Sequence	PCR product
HNF4α ko mice		
HNF4a_pp2_F	CAGCCCAAGGGAGAGAAGTG	500bp on excised allele
HNF4a_exc_junc8_R	CTGTGAGCCCTGGGAATCAG	
HNF4a_exc_1_F	TACTACCCAGGCTCCCTTCC	129bp on WT allele
HNF4a_exc_1_R	AGTGTGTAGCACAGGGTTCG	
C/EBPα ko mice		
6071_F	TGGCCTGGAGACGCAATGA	269bp on targeted, 235bp on WT allele
6072_R	CGCAGAGATTGTGCGTCTTT	
Cebpa_excision_F	GCCTGGTAAGCCTAGCAATCCT	300bp on excised allele
Cebpa_excision_R	TGGAAACTTGGGTTGGGTGT	

Table S3. Primers used for the genotyping of heterozygous knockout mice