Characterization of sequence determinants of enhancer function using natural genetic variation

Marty G. Yang,1,2,* Emi Ling,1,3,* Christopher J. Cowley,1,4 Michael E. Greenberg,1,# and Thomas Vierbuchen5,6,#,‡

1 Department of Neurobiology, Harvard Medical School, Boston, MA, USA
2 Program in Neuroscience, Harvard Medical School, Boston, MA, USA
3 Present Address: Department of Genetics, Harvard Medical School, Boston, MA, USA
4 Present Address: Laboratory of Mammalian Cell Biology and Development, Howard Hughes Medical Institute, The Rockefeller University, New York, NY, USA
5 Developmental Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA
6 Center for Stem Cell Biology, Memorial Sloan Kettering Cancer Center, New York, NY, USA

* These authors contributed equally
# Correspondence: meg@hms.harvard.edu and vierbuct@mskcc.org
‡ Lead Contact

ABSTRACT

Sequence variation in enhancers, a class of cis-regulatory elements that control cell type-specific gene transcription, contributes significantly to phenotypic variation within human populations. Enhancers are short DNA sequences (~200 bp) composed of multiple binding sites (4-10 bp) for transcription factors (TFs). The transcriptional regulatory activity of an enhancer is encoded by the type, number, and distribution of TF binding sites that it contains. However, the sequence determinants of TF binding to enhancers and the relationship between TF binding and enhancer activity are complex, and thus it remains difficult to predict the effect of any given sequence variant on enhancer function. Here, we generate allele-specific maps of TF binding and enhancer activity in fibroblasts from a panel of F1 hybrid mice that have a high frequency of sequence variants. We identified thousands of enhancers that exhibit differences in TF binding and/or activity between alleles and use these data to define features of sequence variants that are most likely to impact enhancer function. Our data demonstrate a critical role for AP-1 TFs at many fibroblast enhancers, reveal a hierarchical relationship between AP-1 and TEAD TF binding at enhancers, and delineate the nature of sequence variants that contribute to AP-1 TF binding. These data represent one of the most comprehensive assessments to date of the impact of sequence variation on enhancer function in chromatin, with implications for identifying functional cis-regulatory variation in human populations.
INTRODUCTION

How genetic variation within and across species gives rise to phenotypic diversity is a central question in genetics. Genetic variants within ~2-3 x 10^6 CRE sequences in the human genome contribute significantly to phenotypic variation in complex traits, including the risk of developing a number of common diseases (Carroll 2008; Maurano et al., 2012; Pickrell 2014; Li et al., 2016; Boyle et al., 2017). The majority of CREs are gene-distal enhancers that regulate gene expression by potentiating transcription, often in a cell type- or cell state-specific manner (Keilwagen et al., 2019). Importantly, we lack the capacity to reliably distinguish functional sequence variants among a large excess of neutral variants at CREs in silico. As a result, it remains necessary to perform direct experimental measurements of enhancer activity to determine the functional consequences of sequence variants (Farh et al., 2015; Nasser et al., 2021), which can be difficult to scale to a large number of variants and is not readily possible in many human cell types (Kasowski et al., 2010; Ding et al., 2014; Tehranchi et al., 2016). Thus, there is a critical need to better define sequence-to-function relationships for enhancers to facilitate the large-scale identification of putative, functional enhancer sequence variants in human populations (Levo and Segal, 2014).

A typical enhancer is bound by ~4-5 sequence-specific TFs that recognize short sequence motifs (~6-12 consecutive nucleotides; Bilu and Barkrai, 2005; Meuleman et al., 2020). TFs effectively function as adaptor proteins to recruit transcriptional regulatory complexes, leading to potentiation of transcription at associated gene promoters. Enhancer activity is highly cell type-specific and this specificity of function is encoded by the combination of TF-binding motifs present within the enhancer. As a result, enhancers that control transcription in distinct cell types are often bound by combinations of TFs unique to each cell type (Spitz and Furlong, 2012). Thus, different sequence features determine TF binding and enhancer activity in each cell type, complicating efforts to identify generalizable features of enhancer sequences that can be used to prioritize enhancer sequence variants in silico. Conversely, highly divergent enhancer sequences can exhibit similar patterns of gene regulation between species, underscoring the complex relationships between genetic variation and enhancer function (Hare et al., 2008).

Although enhancers cannot be defined by a singular set of sequence features, they do, however, exhibit stereotyped chromatin features, such as chromatin accessibility, histone post-translational modifications (e.g. H3K4me1/2 and H3K27ac), and bi-directional transcription of short enhancer RNAs (Heintzman et al., 2007; Boyle et al., 2008; Creyghton et al., 2010; Kim et al., 2010; Rada-Iglesias et al., 2011). These chromatin signatures have been used extensively to map millions of putative enhancers across a wide range of cell types and developmental stages, but a deeper mechanistic understanding of the sequence motifs that control TF binding at enhancers and enhancer activity remains elusive (Kundaje et al., 2015).

Our lab and others have utilized the extensive genetic variation present among inbred mouse strains as an efficient model system to identify sequence variants that impact cis-regulatory function (Heinz et al., 2013; Vierbuchen et al., 2017; Wong et al., 2017; van der Veeken et al., 2019). Furthermore, it is possible to cross highly distinct inbred mouse strains to generate F1 hybrids, allowing for a direct comparison of two alleles for each enhancer locus within the same cellular environment and a highly controlled and quantitative assessment of how SNPs affect their activity. Given these features, the use of naturally occurring genetic variation is a powerful approach for deciphering sequence features of TF binding and enhancer activity in the native genomic context (Wittkopp and Kalay, 2011; Albert and Kruglyak, 2015; Lappalainen 2015; Pai et al., 2015).

Using mouse embryonic fibroblasts (MEFs) derived from two inbred mouse strains, we recently found that the binding of AP-1 TFs is required for chromatin accessibility and activity at the majority of active enhancers at which it binds across the genome (Vierbuchen et al., 2017). However, we observed that many instances of allele-specific AP-1 binding at enhancers could not be readily explained by sequence variants within AP-1 motif(s). These data indicate that, in some instances, sequence features outside AP-1 TF-binding sites play an important role in determining AP-1 binding at enhancers. We observed an enrichment of variants in motifs for putative collaborating TFs (e.g. TEAD) at these sites that lack AP-1 motif mutations, but the nature of this collaborative relationship is
complex and remains poorly defined. Both AP-1 and TEAD are broadly expressed and have critical roles in mediating signal-dependent transcription downstream of the Ras/MAPK and Hippo/YAP/TAZ pathways, respectively. Thus, delineating the sequence features that determine the binding of AP-1 and TEAD TFs has significant implications for enhancer function across a broad range of cell types.

In this study, we performed extensive allele-specific mapping of chromatin state (ATAC-seq, H3K27ac, H3K4me1/2, and H3K4me3) and TF binding (FOS, TEAD1, and CTCF) in F1-hybrid MEFs derived from crosses between C57BL/6J mice and a panel of nine inbred mouse lines. These analyses included several wild-derived inbred strains that contain a high frequency of SNPs/indels (1 in every ~85-170 bp) compared to C57BL/6J mice, which enabled us to efficiently examine the frequency and distribution of SNPs/indels at thousands of enhancers with allele-specific chromatin features and/or binding of TFs. We found that genetic variants in the core region of enhancer sequences (central ~100 bp) were most predictive of an allele-specific change in enhancer activity. AP-1 is often required for TEAD binding and the activity of most enhancers where it is bound, whereas in most cases, TEAD is not required for AP-1 binding, suggesting a model in which AP-1 functions as a pioneer TF and recruits co-factors and chromatin remodelers (e.g. mSWI/SNF) necessary for enhancer function in fibroblasts. Non-canonical sequence features, such as partial AP-1 motifs and nucleotide sequences flanking core AP-1 binding sites, are often missed by motif searches from AP-1-bound regions in a single genome, but they appear to be important for AP-1 binding affinity in chromatin when examined in F1-hybrid cells. Our findings provide new insight into how enhancer function is encoded within DNA sequences and will serve as an important framework for identifying functional cis-regulatory variation in evolution and disease. In addition, our data provide new insight into the crosstalk between Ras/MAPK and Hippo/YAP/TAZ signal-dependent gene expression and suggest that Ras/MAPK-induced AP-1 can play an instructive role in determining the output of Hippo/YAP/TAZ-dependent transcriptional programs.

RESULTS

Allele-specific mapping of sequencing reads in F1-hybrid MEFs

To identify genetic variants that modulate TF binding and/or chromatin state at CREs, we isolated MEFs from male F1-hybrid embryos derived from crosses between C57BL/6J females and males from nine distinct inbred mouse strains, including four wild-derived inbred strains (CAST/EiJ, MOLF/EiJ, PWK/PhJ, and SPRET/EiJ) that have a high frequency of SNPs/indels compared to C57BL/6J mice (Figure 1A; Table 1). Genome sequencing data is available for each inbred strain, meaning that we can query up to ten distinct alleles at each CRE sequence for differences in TF binding and/or CRE activity. To identify potential differences in CRE function that result from sequence variants between maternal (C57BL/6J) and paternal chromosomes, we generated the following allele-specific datasets from the four wild-derived inbred F1-hybrid strains: chromatin features associated with cis-regulatory function (ATAC, H3K4me1/2, H3K4me3, and H3K27ac), occupancy of TFs that bind many CREs in fibroblasts (FOS and TEAD1), putative insulator elements (CTCF), and nascent gene transcription levels (chromatin-associated RNA-seq; Figure 1B-F, S1A). Additional CUT&RUN experiments for FOS and H3K27ac were performed in remaining F1-hybrid lines to include data for more AP-1 bound enhancer alleles, which also contain on average fewer SNPs/indels than those in wild-derived mice. H3K27ac Hi-ChIP was also performed in C57BL/6J MEFs to link active enhancers to putative target genes and to other active CREs (Table 2). All experiments were conducted under two distinct conditions: (1) MEFs that were growth arrested in G0 by serum starvation and (2) serum-starved MEFs that were re-stimulated with 15% serum for 90 minutes. These defined conditions reduce technical variability between samples by synchronizing cells in the population at a specific stage of the cell cycle and allow us to measure the binding of stimulus-dependent TFs (e.g. AP-1 factors) at the peak of their activity (Vierbuchen et al., 2017).

From each F1-hybrid line, we identified primed (ATAC-seq summits that lack H3K27ac, but have H3K4me1) and active CREs (ATAC-seq summits overlapping H3K27ac peaks) on both alleles (Figure S2A). For all active CREs, we classified sites as either gene-proximal (promoters) and gene-distal (putative enhancers) based on their distance to the nearest annotated TSS (Figure S2C). In total, we found 76,517 unique genomic loci defined as
active CREs from the nine F1 hybrids surveyed, and 50.4% of allele pairs at these sites harbor SNP(s) within +/- 60 bp of the ATAC-seq summit used to define each enhancer locus.

**Identification of allele-specific CREs in F1-hybrid MEFs**

In aggregate, 24.4% of pairs of active enhancer alleles on autosomes show a statistically significant difference in H3K27ac levels between maternal (C57BL/6J) and paternal alleles (Figure 1C, S2B). Among these allele-specific sites, 56.2% and 15.3% exhibit a >2-fold and >4-fold difference in H3K27ac signal, respectively. To determine whether differences in H3K27ac between alleles are associated with changes in transcription of the gene that they regulate, we first identified high-confidence enhancer-TSS interactions using H3K27ac Hi-ChIP data, and then examined whether transcription of the linked gene was higher on the chromosome with the active enhancer allele. We found that allele-specific enhancers are more likely to interact with genes that exhibit allele-specific transcriptional differences than enhancers that have similar levels of H3K27ac on each allele (14.5% and 9.1% of active enhancers with detectable H3K27ac Hi-ChIP loops with an active promoter, respectively; Figure S2D). This suggests that allele-specific differences in H3K27ac are indicative of functional differences in the transcriptional regulatory activity of enhancers, consistent with findings from previous studies (Creyghton et al., 2010; Arnold et al., 2013; Fulco et al., 2019).

Compared to H3K27ac levels at enhancers, levels of promoter-associated histone modification H3K4me3 (3.6%) and gene-distal binding of CTCF (2.6%) are less likely to exhibit significant differences between alleles (Figure 1D-E). These data are consistent with previous studies suggesting that promoters and CTCF-binding sites are more likely to be functionally conserved than enhancers among groups of distantly related mammals (Schmidt et al., 2012; Villar et al., 2015; Fudenberg and Pollard, 2019). Furthermore, for each class of active CREs, we found that the frequency of sites with allele-specific H3K27ac signal is proportional to the frequency of SNPs between maternal and paternal alleles (Table 3). We noted that the number of genes with an allele-specific skew in expression level per strain also scaled with the total number of SNPs/indels relative to C57BL/6J in the given strain (Figure S1B; Table 4).

**Inferring mechanisms of enhancer selection and activation from genetic variants**

Several mechanisms have been proposed for how TFs initially bind enhancers to activate expression of genes that were previously silent. It remains unclear whether TF binding is sufficient to displace histone octamers at nucleosomal enhancers or if TF-mediated recruitment of additional co-regulatory proteins, such as chromatin remodeling complexes, is also required (Lidor-Nili et al., 2011; Paakinaho et al., 2017; Johnson et al., 2018). For instance, it is thought that H3K4me1/2 deposition is indicative of enhancers that have been partially activated or primed (Heintzman et al., 2007). However, it is not known whether the majority of these sites only become active later in development (i.e. subsequently gain H3K27ac; Creyghton et al., 2010; Rada-Iglesias et al., 2011; Bonn et al., 2012; Bogdanovic et al., 2012), or if they typically are fully activated in a single step (i.e. concurrently gain H3K4me1/2 and H3K27ac), such as upon the binding of signal-dependent TFs or during cellular differentiation (Kaikkonen et al., 2013; Ostuni et al., 2013). To address these hypotheses, we examined our allele-specific H3K4me1 and H3K4me2 ChIP-seq datasets, which contain thousands of allele pairs that have significant differences in these histone modifications (Figure 1F). Our previous work suggested that disruption of AP-1 TF binding results in the loss of histone marks associated with both primed and active enhancers (Vierbuchen et al., 2017), but whether this feature is generally applicable for all enhancers (independent of AP-1 binding) and whether there are mutations that inactivate enhancers without affecting H3K4me1/2 levels were unresolved. To assess whether the priming and activation of enhancers are genetically separable processes, we focused on H3K4me1/2 levels at enhancers with the greatest difference in H3K27ac levels between alleles. We observed that 70.1% of enhancers in the top decile of allele-specific enhancers have a significant and >2-fold concordant loss of H3K4me1 on the inactive allele, compared to 0.6% of the bottom decile of allele-specific enhancers (peaks with the smallest, statistically significant fold changes in H3K27ac levels between alleles). Chromatin accessibility and AP-1 binding exhibit similar changes to H3K4me1 at enhancers with strongly allele-specific H3K27ac (Figure 1G). Together,
these data reveal that few if any SNPs/indels cause a significant loss of enhancer H3K27ac and maintain strong enrichment of H3K4me1/2 and chromatin accessibility, suggesting that enhancer priming/selection and activation are not separable steps mediated by distinct TF-binding events at enhancers in MEFs. Instead, our data are consistent with a model in which the persistence of H3K4me1/2 (in the absence of H3K27ac) might reflect the kinetics at which histone modifications decay or are removed, and that H3K4me1/2 enrichment might indicate enhancers where low levels of transcription have occurred in a previous stage of development (Jadhav et al., 2019).

### Contribution of cis- and trans-acting effects on enhancer activity

In F1 hybrids, both enhancer alleles are exposed to the same nucleoplasmic environment, and thus observed differences between the two alleles are generally considered to be caused directly by local SNPs (i.e. SNPs within the ~200 bp sequence of chromatin accessibility at the CRE in question). However, each enhancer allele is also located within a cis-regulatory unit or topologically associated domain (TAD), which contains additional genetic changes outside of the enhancer itself that could potentially impact TF binding or chromatin state at the enhancer in an allele-specific manner (Kilpinen et al., 2013; Grubert et al., 2015). These “locus-scale” cis-acting mechanisms could include: (1) sequence variants in other CREs at the same locus that interact with an enhancer, (2) gains or losses in CTCF-binding sites that influence 3D interactions between CREs within the cis-regulatory unit associated with that enhancer, (3) structural variants that rearrange CREs at the locus such that the enhancer is subject to different 3D interactions, and (4) variation in repeat elements (e.g. LINEs, SINEs) within the locus that are not generally well annotated in genomic datasets (Ou et al., 2019). Another possible explanation for allele-specific activity of CREs is parent-of-origin specific imprinting. We excluded CREs at known imprinted loci from subsequent analyses due to the differing nature of this type of allele-specific transcriptional regulation.

To quantify the relative impact of these cis-acting, locus-level mechanisms, we analyzed sequencing reads from our allele-specific histone modification datasets, which typically flank the functional CRE sequence and can thus be mapped to one allele or the other even when there are no SNPs present in the accessible chromatin window at enhancers (Figure 2A, S3B). We reasoned that enhancers lacking SNPs/indels should only show an allelic skew in H3K27ac levels when these aforementioned non-local mechanisms significantly contribute to the function of those enhancers. Only 9.1% of enhancers that have no SNPs/indels in their central 150 bp (centered on the ATAC-seq summit used to initially define the CRE) exhibit a significant, allele-specific, >2-fold skew in H3K27ac levels on flanking nucleosomes, compared to 22.1% of enhancers with SNPs/indels (Figure 2F). This result suggests that it is relatively rare for SNPs outside the enhancer sequence itself to influence the function of the enhancer in question. In addition, allele-specific 0-SNP enhancers are not situated significantly close (than H3K27ac-matched shared 0-SNP enhancers) to an allele-specific CTCF peak (Figure S3A, S3C), implying that rearrangement of CTCF-dependent TAD boundaries is not a major contributor to allele-specific differences in enhancer activity. However, allele-specific 0-SNP enhancers are more likely than shared 0-SNP enhancers to be located near another allele-specific enhancer (median = 48,623 bp and 75,664 bp, respectively), but they are not significantly closer to the nearest active CRE (Figure S3D-E). Consistent with this finding, we observed that allele-specific 0-SNP enhancers are frequently located in enhancer clusters with another allele-specific enhancer (within ~1-2 kb apart). In such cases, it is difficult to rule out the possibility that the quantification of H3K27ac-marked nucleosomes flanking the 0-SNP enhancer is not simply detecting diffuse signal from other enhancer(s) in the cluster (Figure 2B). Furthermore, based on H3K27ac Hi-ChIP, we rarely observed 0-SNP allele-specific enhancers connected via a long-range loop (e.g. >10 kb) with another allele-specific SNP/indel-containing enhancer in the same TAD. Thus, while previous studies have observed that allele- or species-specific CREs tend to be highly interconnected with one another (Prescott et al., 2015; Link et al., 2018), which has been interpreted to suggest that CREs within the same topological domain can modulate each other’s function, our data indicate that locus-scale, cis-acting mechanisms exert limited effects on enhancer activity.

Next, we examined the extent to which trans-acting effects contribute to changes in cis-regulatory function and gene expression between each of the F1-hybrid MEF lines. Trans-acting mechanisms should, in principle, affect each allele within the same F1 hybrid equally, but genetic variation between the distinct F1-hybrid MEF lines could
confound quantitative comparisons of allele-specific enhancer function across F1 hybrids. To measure inter-F1
trans-acting differences, we examined chromatin state and gene expression on the C57BL/6J X-chromosome,
which is present across all F1-hybrid lines. Applying the same criteria that we had used for defining allele-specific
CREs on autosomes, we did not observe any CREs with significantly different H3K27ac levels on the C57BL/6J X-
chromosome between F1 hybrids (Figure S4A). A similar analysis of chromatin-associated RNA-seq data revealed
that expression of a small subset of C57BL/6J X-chromosome genes differed significantly between different F1-
hybrid strains. For example, 9.3% of expressed genes on the C57BL/6J X-chromosome differed by >2-fold in
expression between C57BL/6J x CAST/EiJ and C57BL/6J x SPRET/EiJ MEFs (Figure S4B). This includes a number
of genes critical for transcriptional regulation (e.g. Smarca1, a critical unit of the mSWI/SNF chromatin remodeling
complex, is expressed at ~2-fold lower levels in C57BL/6J x SPRET/EiJ hybrid MEFs compared to all other F1-hybrid
MEFs). Taken together, we observe a limited degree of trans-acting effects on histone modification levels at CREs
and on gene transcription. Therefore, for some subsequent analyses, we merged chromatin and TF-binding data
from allele pairs across different F1-hybrid lines.

**Distribution of genetic variants that influence cis-regulatory function**

To characterize the types of variants that occur within CREs that cause changes in enhancer activity, we started by
examining active enhancers with the largest differences in H3K27ac between alleles. We reasoned that these
enhancers would contain large-effect, loss-of-function mutations on one allele, which would identify TF-binding
sites likely required for enhancer function. Across nine F1-hybrid strains, we identified a total of 29,185 pairs of
enhancer alleles with a significant and >2-fold difference in H3K27ac levels between alleles. Allele-specific
enhancers have a significantly higher frequency of SNPs/indels than H3K27ac signal-matched shared enhancers
(Figure 2C). Moreover, enhancers with a greater number of genetic variants are more likely to have allele-specific
H3K27ac signal and to show larger quantitative differences in H3K27ac levels (Figure 2F-G). Across many
mammalian species, allele-specific enhancers in F1-hybrid MEFs exhibit slightly less evolutionary conservation than
shared enhancers (Figure S3F). Active promoters and gene-distal CTCF peaks tolerate, on average, a greater
number of SNPs/indels than enhancers (Figure 2D-E). Altogether, these data suggest that enhancer sequences
are less functionally conserved than other classes of CREs (Villar et al., 2015), and that quantitative differences in
enhancer activity correlate with the degree of sequence divergence between alleles.

We also hypothesized that the location of genetic variants within the enhancer is likely to impact whether a
given SNP/indel affects enhancer function. To explore this further, we examined the distribution of SNPs/indels
relative to the center of the accessible chromatin region at pairs of enhancer alleles with allele-specific or shared
H3K27ac levels. This revealed an enrichment of SNPs/indels within a 150 bp window centered on the ATAC-seq
summit at allele-specific enhancers (Figure 2H). In contrast, there was not a similar enrichment of SNPs/indels in
allele-specific H3K4me3 peaks at promoters (Figure 2J). We also examined enhancer loci with a single SNP/indel
present, since these sites can inform us about genetic variants that are sufficient to block enhancer function. We
found that 14.6% of 1-SNP/indel enhancers show an allele-specific and >2-fold skew in H3K27ac levels, and we
observed a more focal enrichment directly at the core region of the enhancer (which we define as the central ~100
bp of the <200 bp accessible chromatin window) of SNPs/indels at allele-specific enhancers (Figure 2I). Together,
these data suggest that the core region is most likely to harbor SNPs/indels that significantly modulate chromatin
state at enhancers. Interestingly, previous studies have proposed that a subset of TFs might bind preferentially
away from the enhancer core, including the well-studied pioneer TF FOXA (Grossman et al., 2017). Recent data on
pioneer TF occupancy suggests that these motifs near the edges of an accessible chromatin region could represent
sites of direct DNA binding upon unwrapping from the nucleosome core or regions where pioneer TFs directly
bind flanking nucleosomes (Meers et al., 2019). However, our data indicate that most SNPs impacting enhancer
function in fibroblasts are centrally located within enhancers, suggesting that these mode(s) of pioneer TF binding
are not strictly required for enhancer activity in our model system.
Identification of TF-binding motifs required for enhancer activity in MEFs

We next sought to determine candidate TF motifs that are responsible for driving enhancer activity in MEFs. Analysis of the top decile of active MEF enhancers (based on relative H3K27ac levels) in the C57BL/6J genome using the KMAC algorithm (Guo et al., 2018) identified k-mers that matched putative binding motifs for several TF families (AP-1, TEAD, and ETS; Figure S5A; Table 5). For AP-1, the k-mer identified by KMAC (5’-VTGACTCAB-3’; V indicates A/C/G and B indicates C/G/T) includes the known core AP-1 site (known as a TRE; TGASTCA), which is bound by heterodimers of FOS and JUN family TFs or homodimers of JUN family members (Risse et al., 1989; Eferl and Wagner, 2003). VTGACTCAB is the most enriched k-mer at active fibroblast enhancers (30.8%, versus 1.6% of control sequences; AUC = 0.450). AP-1 TFs bind DNA as dimers, with the basic leucine zipper (bZIP) DNA-binding domain of each FOS/JUN monomer recognizing half of a palindromic consensus motif (TGASTCA). The enriched AP-1 k-mer we observed indicates that certain flanking nucleotides on either side of the TRE were strongly disfavored (i.e. T and A were depleted from the nucleotide on the 5’ and 3’ ends of the AP-1/TRE motif, respectively).

KMAC also identified an enriched k-mer (5’-GGAATK-3’; K indicates G/T) that matched the known core binding motif for the TEAD family of TFs (GGAAT; Farrance et al., 1992) and includes an additional restricted nucleotide on the 3’ end (10.9%, versus 0.6% of control sequences; AUC = 0.289). TEAD TFs are broadly expressed in developing and adult cell types and function as transcriptional effectors of the Hippo/YAP/TAZ signaling pathway that regulates cell growth and proliferation (Chen et al., 2010).

We observed a similar enrichment of AP-1 and TEAD k-mer clusters at both constitutive enhancers and enhancers that regulate late-response genes activated by serum stimulation (Vierbuchen et al., 2017; Figure S5B). This finding suggests that the specific dynamics of enhancer activation cannot be readily distinguished by the presence or absence of specific k-mers alone, and suggests that the sequence features that determine whether an enhancer is constitutively active or signal-responsive (e.g. enhancers that regulate late-response genes) are more complex (Bevington et al., 2016; Comoglio et al., 2019).

Since these AP-1 and TEAD k-mers were defined by their enrichment at enhancers in the C57BL/6J genome (in comparison to GC-matched control regions), we next sought to determine the impact of SNPs within these k-mers on AP-1 and TEAD binding at active enhancers using allele-specific TF-binding data (Figure 3A-B). We performed a series of validations to ensure that distinct methods (ChIP-seq and CUT&RUN) were providing similar quantitative information on TF-binding levels (Figure S6A-G). We reasoned that an enrichment of SNPs would occur only at nucleotide positions required for sequence-dependent binding of these TFs and not at neighboring positions flanking these nucleotides (Maurano et al., 2015). For active enhancer loci with allele-specific FOS binding, an enrichment of SNPs is observed at the core AP-1 motif (Figure 3C). Within the core motif, the lowest enrichment of SNPs was observed at the central nucleotide, consistent with in vitro experiments suggesting that this nucleotide does not strongly influence AP-1-binding affinity (Risse et al., 1988). Recent in vitro studies of AP-1 binding affinity to AP-1/TRE motifs suggests that the three nucleotides flanking the core AP-1 motif (TGASTCA) can strongly modulate AP-1 TF binding by altering the shape of the AP-1/TRE motif (Leonard and Kerppola, 1998; Rohs et al., 2010; Yella et al., 2018). Given these data, we assessed whether these flanking sequences play a role in determining AP-1 binding site selection in chromatin. Our data in Figure 3C suggests SNPs at the 5’ and 3’ flanking nucleotides of the AP-1/TRE motif (VTGACTCAB) can affect AP-1 binding at active FOS-bound enhancers. More broadly, there is an enrichment of SNPs in the three nucleotides flanking each side of the core AP-1 binding site when considering all allele-specific FOS-bound sites from our data (NNVTGACTCABNN; 9.9% and 5.6% at allele-specific and shared FOS peaks, respectively; Figure S5D). These results provide further evidence that sequences immediately flanking core AP-1 motifs should be considered in future assessments of AP-1 binding motif preferences.

Next, we found that an additional nucleotide beyond the core TEAD-binding site (GGAAT) was restricted to G/T in the KMAC output (GGAATK), and SNPs were enriched at all positions within this k-mer at allele-specific
TEAD1-bound enhancers (Figure 3D). For allele-specific CTCF sites with >2-fold difference in signal, we found a ~14 bp window of enriched SNPs (i.e. broader than typical DNA-binding TFs, like AP-1 or TEAD) that disrupt CTCF binding (Figure 3E), closely mirroring the ~15-20 bp sequence that CTCF is predicted to bind in vivo (Kim et al., 2007). These data define that AP-1 and TEAD consensus k-mers are most enriched in active enhancers in fibroblasts and functionally validate the importance of individual nucleotides for TF binding in the native chromatin context.

Having observed the importance of these k-mers for AP-1 and TEAD binding at active enhancers, we next performed a targeted analysis to assess the impact of AP-1 and TEAD SNPs on enhancer chromatin state. For AP-1 and TEAD, we separately identified a set of active enhancers that have a single instance of their respective TF k-mers and have a SNP/indel within the k-mer sequence. This allowed us to directly assess the effect of core motif SNPs on TF binding and enhancer chromatin state. For both AP-1 and TEAD, core motif SNPs are correlated with a significant loss in binding of each TF at enhancers, as expected (Figure 3F, 3J). Loss of AP-1 binding is associated with a substantial decrease in chromatin accessibility, H3K4me1/2, and H3K27ac levels on the allele with the mutated AP-1 site (Figure 3G-I), similar to our previous observations from a smaller set of enhancers (Vierbuchen et al., 2017). This finding suggests that at enhancers that contain a single consensus AP-1 site and are bound by FOS/JUN, a variant that changes a nucleotide in the core AP-1 motif is likely to result in a complete loss of enhancer function, consistent with data from plasmid-based reporters that suggest AP-1 motifs are required for transcriptional activation (Malik et al., 2014; Liu et al., 2016). Similar analysis of TEAD1-bound enhancers revealed a more modest decrease in ATAC-seq, H3K4me1/2, and H3K27ac signal associated with the allele containing the predicted inactive TEAD motif (Figure 3K-M), suggesting that loss in TEAD occupancy has less severe consequences on enhancer function than loss of AP-1 binding. Taken together, these data suggest that AP-1 and TEAD motifs play a central role in enhancer function in fibroblasts.

Hierarchical relationships between AP-1 and TEAD TFs

Although SNPs that disrupt core TF-binding motifs (AP-1, TEAD, and ETS) are enriched at enhancers with allele-specific TF binding, our data also indicate that these are not sufficient to explain all regulatory variation between alleles. For example, among all enhancers in the top decile of allele-specific H3K27ac signal, only 13.3% had a SNP/indel overlapping a core AP-1, TEAD, and/or ETS motif in their central region. In contrast, we found that 21.5% of allele-specific insulators (with >2-fold difference in CTCF signal) that contain a CTCF position weight matrix (PWM) had at least one SNP/indel overlapping the CTCF binding site. These data favor a model in which other types of SNPs outside core TF-binding motifs can collectively modulate enhancer activity, whereas the sequence determinants of CTCF binding at insulators can more readily be predicted by the presence or absence of SNPs in CTCF motifs.

In previous work, we found that strain-specific AP-1 peaks in MEFs (in a comparison of two inbred mouse strains) that lack a mutation in a core AP-1 site were enriched for SNPs in TEAD motifs, suggesting a model in which AP-1 binding was dependent, at least in part, on the presence of TEAD binding sites (Vierbuchen et al., 2017). However, we lacked TEAD binding data, which prevented us from examining in depth the sequence determinants and functional relationship of AP-1 and TEAD binding at enhancers across the genome. Other data have suggested that AP-1 and TEAD TFs coordinate to regulate transcriptional programs critical for cell growth and proliferation during normal development and in the context of cancer (Liu et al., 2016; Zanconato et al. 2018). Since multiple AP-1 and TEAD TFs are also often co-expressed in the same cell types and can play functionally redundant roles with one another (Seo et al., 2021), it has been difficult to examine how these two TFs that exhibit extensive co-occupancy (~70-80% shared binding) work together at enhancers to regulate gene transcription. With our newly generated AP-1 and TEAD binding data across four wild-derived F1-hybrid lines, we could more systematically examine a larger number of loci to define the functional relationship between AP-1 and TEAD.

We first quantified how often consensus TF motifs are mutated at allele-specific versus shared AP-1 and TEAD peaks. If the binding of a given TF was entirely dependent on the occupancy of another TF, we would expect to observe a similar loss in binding of the dependent TF, regardless of which TF motif was mutated. For these
analyses, we included all distal FOS and TEAD1 peaks in our dataset, irrespective of whether occur at active enhancers. We observed that AP-1 motif mutations are frequently associated with a loss of TEAD binding, whereas AP-1 binding is more weakly affected by TEAD motif mutations (Figure 4A-B). Strikingly, AP-1 motif mutations were as enriched at allele-specific TEAD peaks as TEAD mutations were (compared to sites with shared TEAD binding). Analysis of AP-1 and TEAD co-bound sites (independent of whether they contained consensus AP-1 or TEAD motifs) further supported a hierarchical binding relationship between these TFs. For example, 50.2% (n = 821/1,635 allele pairs) of allele-specific FOS-bound sites also exhibit an allele-specific loss of TEAD1 binding, whereas only 8.7% (n = 821/9,416 allele pairs) of allele-specific TEAD1 peaks showed significant allele-specific FOS signal. In summary, these data are consistent with previous studies that suggest that AP-1 can serve as a pioneer TF to facilitate the binding of other TFs, such as the glucocorticoid receptor, PU.1, and C/EBP (Biddie et al., 2011; Heinz et al., 2013), and that AP-1 binding is required for inducible chromatin remodeling and nucleosome displacement at late-response gene enhancers (Vierbuchen et al., 2017).

Data from in vitro studies examining TF-binding specificity have shown that TFs that bind to composite motifs often prefer sequences that are distinct from their consensus individual motifs (Jolma et al., 2015). This led us to consider the possibility that sites at which AP-1 and TEAD bind together might exhibit differential motif requirements from sites where only one of these two TFs bind. We observed that AP-1 only peaks contain at least one AP-1 k-mer (65.9%; VTGACTCAB, VTGAATCAB, or VTTAGTCAY), whereas AP-1/TEAD co-bound peaks were less likely to contain a consensus motif (53.8%; Figure 4C). Similarly, TEAD-only peaks (44.0%) had a higher frequency of TEAD k-mers (GGAATK) than AP-1/TEAD peaks (36.0%; Figure 4D). These data suggest that the motif requirements for AP-1/TEAD co-bound regions are more flexible than sites at which only one of the TFs bind.

**Distribution of genetic variants outside TF motifs suggests collaborative competition model for AP-1/TEAD binding**

Thus far, our data suggest that many instances of AP-1 and TEAD binding cannot be explained solely by mutations in consensus, core motifs for these TFs. This lack of enriched TF motif mutations has been observed for other classes of TFs and in a variety of model systems, suggesting that this is a general, unresolved problem in genetic studies of TF binding (Deplancke et al., 2016). Our dataset allowed us to systematically look for recurrent features of SNPs/indels associated with allele-specific AP-1 or TEAD binding. These analyses can help reveal the prevalence of various proposed mechanisms by which sequences outside of core motif occurrences influence TF binding. In particular, SNPs outside known TF-binding sites allow us to dissect the role of motif spacing on the ability of TFs to cooperate with one another to bind enhancers. Subtle changes in motif syntax have been shown to alter enhancer function (Erceg et al., 2014; Farley et al., 2016; Shen et al., 2021), and conversely, the arrangement of TF-binding motifs can also be highly flexible in other contexts (Amost and Kulkarni, 2005; Junion et al., 2012; King et al., 2020; Jindal and Farley, 2021).

First, we identified allele-specific gene-distal binding sites for FOS, TEAD1, and CTCF, and then examined the frequency of SNPs/indels (relative to the ATAC-seq peak center) at these sites compared to sites with shared binding on both alleles. In Figure 2I, we plotted SNP/indel distributions at enhancers with allele-specific histone acetylation, whereas these analyses focus instead on TF-binding sites independent of H3K27ac levels. When comparing allele-specific and shared FOS peaks, we found an increased frequency of SNPs/indels within an ~100 bp window centered on the ATAC-seq peak summit, which is similar to the pattern observed at enhancer loci with allele-specific H3K27ac levels (Figure 5A).

Next, given that AP-1 motif SNPs likely contribute to the distribution observed in Figure 5A, we repeated this analysis, but excluded allele-specific AP-1 peaks that have a SNP/indel in their consensus AP-1 motif (VTGACTCTCAB) and plotted the SNP/indel frequency relative to this k-mer instead of the ATAC-seq peak summit. This revealed an enrichment of SNPs/indels within +/- ~50 bp of the AP-1 motif (Figure 5D). This distribution of SNPs/indels is consistent with a collaborative competition model for AP-1 binding: motifs for AP-1 and collaborating TFs occur at non-fixed distances to one another, and they can collaboratively bind as long as all
motifs are situated on the same side of the nucleosomal dyad (<75 bp of one another). Similar results were observed at TEAD peaks (Figure 5B, 5E), but a narrower distribution of enriched SNPs/indels was found at allele-specific CTCF sites (+/− ~10 bp) when the CTCF PWM was present on both the C57BL/6J and paternal allele (Figure 5C, 5F). This result indicates a more restricted length scale at which genetic variants can disrupt CTCF binding than those that we observed for more typical DNA-binding TFs like FOS and TEAD1, and suggests that CTCF binding is less dependent on collaborative binding with other TFs. Overall, we observed that genetic variants associated with an allele-specific loss in enhancer activity or AP-1 binding are enriched at the core of enhancers and that these SNPs/indels tend not to occur at highly stereotyped positions, but instead block collaborative competition with AP-1 via likely disruption of additional TF motifs within +/− ~50 bp of AP-1 sites.

**Contribution of partial or degenerate AP-1 motifs to AP-1 binding affinity**

Low-affinity TF-binding sites have been shown to play an important role in the function of enhancers, but they are difficult to identify from PWM or k-mer searches because they often bear little resemblance to predicted core TF motifs (Kriebelbauer et al., 2019). Binding of TFs to their cognate motifs on nucleosomes is often restricted by steric hindrance between TFs and histone octamers. Similarly, some pioneer TFs are thought to bind partial motifs on nucleosomes (Soufi et al., 2015; Roberts et al., 2021). We considered the possibility that additional low-affinity sites, such as half motifs, make an important contribution to AP-1 binding of enhancers in fibroblasts, and that some instances of allele-specific AP-1 binding lacking core motif mutations could be explained by SNPs in degenerate recognition motifs not readily detected by traditional searches.

To examine whether binding of AP-1 half sites contributes to AP-1 recognition at enhancers, we chose to examine the occurrence of AP-1 half site frequency at AP-1 bound sites. Allele-specific and shared AP-1 peaks contained, on average, a similar number of AP-1 half sites (in the context of TGASVDB k-mers; mean = 1.17 and 1.20 k-mers per peak in the central 150 bp, respectively). However, we observed a ~2-fold greater frequency of AP-1 half sites containing SNPs (TGASVDB) in allele-specific versus shared FOS peaks (mean = 0.25 and 0.13 k-mers per peak, respectively), suggesting that AP-1 half sites contribute to AP-1 TF binding in chromatin. However, based on our prior analysis of mutations in full AP-1 sites, it is clear that disruption of one of two half sites within an AP-1 consensus motif has a strong effect on AP-1 binding in most cases, which suggests that AP-1 half sites alone might not be sufficient for binding in the absence of another intact AP-1 motif at the same enhancer (Figure 3C, 3F). Importantly, we observed no difference in the frequency of TGASVDB k-mers across different deciles of FOS-binding signal, suggesting the occurrence of half AP-1 sites is not correlated with regions where AP-1 is highly or lowly bound (Figure S5C).

**Identification of k-mers predictive of AP-1 binding and/or activity using machine learning**

Support vector machine (SVM) algorithms have been utilized in a variety of contexts to perform classification of DNA sequences in a supervised manner (Barozzi et al., 2014; Ghandi et al., 2016). We sought to apply a gapped k-mer SVM approach (gkm-SVM) to our datasets that has been optimized to detect k-mers of similar length to typical TF-binding motifs (Ghandi et al., 2016). Our initial goal was to determine whether additional k-mers could contribute to our ability to distinguish TF-bound sites versus non-bound sites and whether these k-mers matched binding sites that are frequently disrupted in our allele-specific TF-binding dataset.

We first compared 60 bp of DNA sequences from AP-1 peaks in C57BL/6J MEFs (positive set) to GC- and length-matched, randomly sampled background DNA sequences from the C57BL/6J genome (negative set). The area under the receiver operating characteristic curve (AUROC = 0.872) from this gkm-SVM analysis is highly similar to the corresponding value obtained from a control analysis of CTCF peaks from human cells, suggesting that the gkm-SVM is able to classify FOS-bound and unbound regions with a low rate of detecting false positives while correctly assigning true negatives. Similarly, a relatively high value for the area under the precision-recall curve (AUPRC = 0.881) indicates that the gkm-SVM is able to reliably distinguish true and false positives. Together, these results suggest that the information within the central 60 bp sequences (+/− 30 bp relative to the ATAC-seq summit)
at FOS-bound peaks is sufficient to train a model to reliably distinguish FOS-bound sites from control non-coding regions of the same genome (Figure S7A). Inclusion of additional sequence beyond this central 60 bp (up to a total of 300 bp) had only a slight positive effect on the performance of the model (Figure 6A-B). Conversely, shortening of DNA sequence below 60 bp resulted in a drop-off in performance of the model. Next, to determine whether sequences outside of the AP-1 motif contribute to the performance of the model, we repeated this same analysis, but we computationally masked all occurrences of core AP-1 sites. This revealed a slight drop in AUROC and AUPRC values, suggesting that the model retains some predictive capacity when core AP-1 motif sequences are excluded (Figure 6A-B, S7E).

k-mers containing AP-1 sites were the largest contributors to the performance of the model, as expected (Figure S7D). Interestingly, k-mers containing AP-1 sites also contributed the most to model performance when gkm-SVM was applied to TEAD1-bound sites (Figure S7B, S7F), consistent with our observations that AP-1 binding is required for TEAD binding at many enhancers (Figure 5). When we ran the gkm-SVM on CTCF peaks, we observed the expected enrichment of k-mers matching critical nucleotides of the well-documented CTCF site (Figure S7C, S7G).

We next sought to apply this gkm-SVM approach to attempt to identify k-mers that distinguish between AP-1 binding sites with and without H3K27ac. Our data indicate that the AP-1 binding is required for the function of many of the active enhancers at which they bind in MEFs. However, AP-1 binding alone is clearly not sufficient for enhancer activity. For example, 34.9% of gene-distal FOS peaks do not overlap H3K27ac peaks. This suggests that the sequence features that are permissive for AP-1 binding in MEFs might be separable from those that confer activity. For this gkm-SVM analysis, we selected a curated set of FOS-bound allele pairs (n = 2,697) that (1) have equivalent levels of FOS binding, (2) contain a consensus AP-1 site on both alleles, and (3) exhibit allele-specific H3K27ac levels. We input 60 bp DNA sequences (centered on the shared AP-1 consensus motif) from the active (positive set) and inactive (negative set) alleles at FOS-bound enhancers (Figure S7H). In contrast to the results above, the gkm-SVM failed to discriminate between these two sets of sites (AUROC = 0.086 and AUPRC = 0.318), suggesting that sequence features predictive of H3K27ac are more complex and cannot be readily captured by this k-mer based SVM approach.

**Generalizability of sequence determinants of AP-1 binding across species**

Having defined some features of sequences that determine AP-1 binding to CREs in MEFs, we next wanted to examine TF occupancy data across a large number of cell types to understand more about TF-binding events that occur together with AP-1 at CREs. To do this, we used DNase-seq footprinting data generated from a large panel of human tissues and cell types (Vierstra et al., 2020). These data provide an unbiased view of individual TF-DNA binding interactions within CREs.

First, we identified TF footprints that overlap an extended AP-1 k-mer (VTGACTCAB) within CREs, which we interpret as individual instances of AP-1 binding. We found a total of 164,705 TF footprints (from among >4 x 10⁶ total footprinted regions) that contain VTGACTCAB k-mers. These AP-1 footprints were centrally enriched within CREs (Figure 6C), consistent with the distribution of AP-1 k-mers observed at AP-1 bound peaks in MEFs, as well as previous data examining AP-1 motif frequency within human DNase-seq peaks (Grossman et al., 2017). The majority of AP-1 footprints were <30 bp in width (83.3%; median = 17 bp), which suggests that they represent the footprint caused by binding of a single AP-1 homo/heterodimer (Figure 6D). CREs with AP-1 footprints typically have a total of ~3-4 additional TF footprints (Figure 6E), and the median distance between AP-1 footprints and the nearest other TF footprint is ~24 bp (Figure 6F). gkm-SVM analysis of sequences flanking AP-1 footprints (60 bp windows) revealed an enrichment of TEAD and ETS k-mers, consistent with our observations at AP-1 bound sites in MEFs (Figure S7I). Together, these data are consistent with a model in which AP-1 is critical for CRE function across many cell types and provide further insight into the nature of TF-binding events that occur with AP-1 binding at CREs. These data will be valuable for disentangling the complex sequence features that control AP-1 binding and enhancer function across diverse cell types and tissues.
DISCUSSION

In this study, we leverage natural genetic variation across inbred mouse strains to identify sequence variants that impact TF binding and/or enhancer activity in the native chromatin context. To systematically assess the effect of many genetic variants on CRE function, we mapped TF binding (AP-1, TEAD, CTCF) and multiple chromatin features (ATAC-seq, H3K27ac, H3K4me1/2/3) in up to ten distinct alleles for each CRE locus. By assessing the frequency and distribution of genetic variants at large numbers of CREs with shared or allele-specific TF binding and/or cis-regulatory activity, we define features of cis-acting genetic variants that are most predictive of differences in chromatin state and/or TF binding.

In this study, we find that loss of the active enhancer histone modification H3K27ac is not genetically separable from loss of H3K4me1/2. We observed few, if any, instances in which SNP(s) within an enhancer cause a loss of H3K27ac without a concurrent loss in H3K4me1/2. This is interesting to consider given a number of observations about the relative contribution of these histone modifications to enhancer function: (1) we found that enhancers that regulate late-response genes exhibit H3K4me1 enrichment in serum-starved MEFs, but have low chromatin accessibility and lack H3K27ac, and upon serum-stimulation, gain H3K27 acetylation and inducibly bind AP-1 TFs (Vierbuchen et al., 2017), (2) a recent study suggests that enhancers with H3K4me1 enrichment that lack H3K27ac are not, in fact, primed or poised for future activity, but instead that this chromatin state is a remnant of activity in a prior developmental stage (Kim and Shiekhattar, 2015; Jadhav et al., 2019), and (3) catalytic mutants of Mll3/4, enzymes responsible for H3K4me1/2 deposition do not appear to affect recruitment of RNA polymerase II to enhancers (Dorighi et al., 2017; Jang et al., 2017). Taken together, these observations suggest that H3K4me1/2-only enhancers might exhibit different cis-regulatory features compared to enhancers that have H3K27ac enrichment because they represent enhancers that were active in a previous developmental stage characterized by a distinct complement of TFs expressed.

The genetic variants between inbred mouse strains that we examine provide insight into normal enhancer turnover across evolutionary timescales (Villar et al., 2015). However, using F1 hybrids, we can more directly quantify the relative contributions of local cis-acting variants versus differences in enhancer function due to trans-acting divergence. We find that greater numbers of SNPs/indels at enhancers are correlated with higher probabilities of allele-specific activity between mouse strains, and that allele-specific enhancers in our study are also less conserved across greater evolutionary timescales than shared enhancers. This is consistent with the observation that many loci contain multiple enhancers with overlapping cis-regulatory activity, such that loss of any individual single enhancer is likely insufficient to cause a large change in gene expression and/or organismal phenotype (Osterwalder et al., 2018). SNPs/indels that are sufficient to block enhancer H3K27ac tend to occur within ~50 bp of the center of the accessible chromatin region used to define enhancer sequences. This underscores that genetic variants within the core region of enhancers should be prioritized in genome-wide association studies for human traits and/or disease risk. Interestingly, our data indicate lower levels of enriched SNPs towards the edges of ATAC-seq peaks at enhancers. A previous study proposed that TFs might bind preferentially to specific regions within accessible chromatin at enhancers and that pioneer TF-binding sites (e.g. FOXA factors) tend to be situated near the outer regions of accessible chromatin, compared to other TF families that exhibit more central enrichment, such as AP-1 (Grossman et al., 2017).

The distribution of SNPs that impact AP-1 binding relative to the AP-1 site at enhancers is interesting to consider in the context of a recent paper that looked at sequence features that distinguish AP-1 bound enhancers with high versus low activity in reporter assays (Chaudhari and Cohen, 2018). Analysis of their data with a supervised machine learning approach (gkm-SVM) suggested that most variation in AP-1 bound enhancer activity can be predicted from sequences within +/- 10 bp of and including the core AP-1 motif. The proximity of these sequences to AP-1 sites contrasts with the broader window (+/- ~50 bp relative to AP-1 motifs) that we observe to determine AP-1 binding in chromatin. More generally, whether massively parallel reporter assays (MPRAs) can fully recapitulate the cis-regulatory activity of different alleles in the genomic context has important implications for studying the effects of disease variants on enhancer activity.
We define sequence determinants for binding of AP-1 and TEAD family TFs at enhancers. We find that AP-1 TF binding is frequently required for TEAD occupancy at a large number of enhancers across the genome. This striking dependency of TEAD binding on AP-1 has been unclear due to the difficulty of performing AP-1 or TEAD loss-of-function studies because both AP-1 and TEAD TFs are encoded by multiple family members that are broadly co-expressed and have extensive functional redundancy. At a subset of enhancers, we observed that AP-1 binding can also be dependent on TEAD. These results are generally consistent with a collaborative competition model for AP-1 and TEAD binding at enhancers (Miller and Widom, 2003; Long et al., 2016). Furthermore, we find that TEAD can potentially bind to more degenerate k-mers at sites at which AP-1 is also bound, suggesting that the sequence determinants of TEAD binding differ depending on whether AP-1 binds at the enhancer or not. These findings highlight the difficulty of understanding the determinants of TF binding at cis-regulatory elements.

The requirement of AP-1 for TEAD to bind many enhancers is interesting to consider given that these TFs have been previously shown to co-regulate gene expression programs associated with cell proliferation and tissue growth. Both AP-1 and TEAD are transcriptional effectors of intercellular signaling pathways. AP-1 TFs are activated by Ras/MAPK signaling and TEAD TFs are required for the binding of the transcriptional co-activators YAP and TAZ, whose nuclear localization is directly regulated by Hippo signaling. Our data suggests a mechanism for crosstalk between these two signaling pathways in which the transcriptional output of the Hippo pathway can be modulated depending on whether Ras/MAPK is active or not. This instructive function of AP-1 in selecting the enhancers at which TEAD TFs can bind is similar to the role for AP-1 in facilitating the binding of transcriptional effectors from several other signaling pathways, including the glucocorticoid receptor, NF-κB, and SMAD (Biddie et al., 2011; Heinz et al., 2013; Li et al., 2019).

Enhancer sequences tend to be occluded by histone octamers prior to TF binding, suggesting that a subset of TFs must access their cognate motifs at enhancers on nucleosomes during the process of enhancer selection (Tillo et al., 2010; Barozzi et al., 2014). This feature complicates efforts to determine sequence features required for TF binding because the affinity and type of TF motifs that TFs bind to can often differ between nucleosomal context and naked DNA. Furthermore, different classes of TFs utilize distinct mechanisms to engage with histone octamers (Michael and Thoma, 2021). Based on nucleosome-binding studies, bZIP TFs, including AP-1 factors, appear to bind their cognate motifs on nucleosomes (He et al., 2013) only when they are present on outer edges of nucleosomes, which are thought to permit TF binding when the DNA sequences at these flanking regions intermittently unwind from the histone octamer (known as nucleosome breathing; Zhu et al., 2018; Zhou et al., 2019). This is consistent with structural data suggesting that AP-1 TFs cannot bind their full cognate motif (TGASTCA) on nucleosomes due to steric constraints (Michael and Thoma, 2021). We observe an enrichment of SNPs in AP-1 half sites (TGA SVDB) at allele-specific AP-1 peaks, raising the possibility that partial AP-1 motifs contribute to AP-1 binding to nucleosomal enhancers. The binding to partial motifs has been observed for other nucleosome-binding TFs, such as OCT4 (Soufi et al., 2015). Thus, we favor a model in which (1) AP-1 recognizes partial motifs towards the edges of nucleosomes, (2) AP-1 dimers can then bind both halves of the core AP-1 site when it is accessible upon nucleosome breathing, and (3) the initial binding of AP-1 facilitates the binding of other dependent TFs (e.g. TEAD) via collaborative competition to evict the histone octamer and/or recruit co-regulatory proteins and chromatin remodelers to enhancers. However, we cannot exclude the possibility that there is a pioneer TF in fibroblasts that binds highly degenerate motifs that cannot be detected with our analyses that facilitates AP-1 binding to nucleosomes.

Our F1-hybrid dataset provides new insights into how DNA sequences within CREs contribute to TF binding and enhancer function. We believe that this F1-hybrid approach for examining TF function is a powerful tool to uncover sequence determinants of TF binding that cannot be easily detected from motif searches or motif enrichment analysis alone. Our F1-hybrid datasets identify thousands of enhancer allele pairs that differ subtly in their DNA sequences and yet have strongly allele-specific functional properties. In the future, using deep learning or other sophisticated approaches to analyze these data could provide further insight into more complex sequence features (i.e. beyond the presence or absence of specific k-mers) that underlie TF binding and enhancer activity.
Similarly, incorporating F₁-hybrid data from additional cell types can further reveal both context-specific and broadly applicable mechanisms of TF binding and enhancer activity (Halow et al., 2021). More broadly, this F₁-hybrid approach represents a powerful tool for understanding complex cis-regulatory processes and can accelerate efforts to identify functional non-coding variants that contribute to human disease and complex traits.

MATERIALS AND METHODS

Mice
All animal experiments were approved by the National Institutes of Health and the Harvard Medical School Institutional Animal Care and Use Committee and were conducted in compliance with the relevant ethical regulations. 6-week-old female C57BL/6J mice were obtained from Jackson Labs (Bar Harbor, ME, USA) (Stock No. 000664) for all breeding pairs. 4- to 8-week-old male mice from the following strains were also obtained from Jackson Labs: CAST/EiJ (Stock No. 000928), MOLF/EiJ (Stock No. 000550), PWK/PhJ (Stock No. 003715), SPRET/EiJ (Stock No. 001146), 129S1/SvlmJ (Stock No. 002448), A/J (Stock No. 000646), BALB/cJ (Stock No. 000651), DBA/2J (Stock No. 000671), NOD/ShiLtJ (Stock No. 001976). No new mouse strains were generated in this study.

The study of inbred mice that are more genetically divergent from C57BL/6J in combination with the use of longer sequencing reads increases the proportion of informative allele-specific reads. However, higher frequencies of SNPs/indels per strain results in a greater percentage of CREs with multiple genetic variants, making it difficult to assign which specific SNP/indel is likely responsible for observed changes in TF binding or chromatin state. Therefore, to balance these considerations, we included wild-derived strains with a relatively high frequency of SNPs/indels compared to C57BL/6J mice (1 SNP/indel per ~85-170 bp), as well as more commonly used inbred strains that are less genetically divergent from C57BL/6J mice (1 SNP/indel per ~1,000 bp; Table 1).

Generation of MEF lines
Embryos were harvested on embryonic day 13.5-14.5 and washed in room-temperature PBS. The heads and internal organs were removed, and the dissected tissue was re-washed in PBS. Individual embryos were placed at the center of 15-cm plates and incubated for 45 min in 1 mL trypsin-EDTA 0.25% (Life Technologies 25200072). Excess trypsin was carefully aspirated, and the dissected tissue was manually dissociated with scissors for ~1 min. Dissociated cells were then incubated in ~1 mL trypsin-EDTA at 37°C in 5% CO₂ for 30 min. Complete media was prepared by supplementing DMEM (Life Technologies (Carlsbad, CA, USA) 12430062) with 10% CCS (Thermo Fisher 21003175), Penicillin-Streptomycin (Thermo Fisher 15140148), MEM non-essential amino acids (Thermo Fisher 11140050), and 1 mM sodium pyruvate (Thermo Fisher 11360070). Trypsin was quenched with 10 mL complete media, and cells were rapidly triturated up/down 10 times with a 10 mL serological pipette to generate a single-cell suspension. An additional 10 mL complete media was added per plate, and cells were grown at 37°C in 5% CO₂.

When cells became fully confluent in ~2-3 days, MEFs were washed in PBS and trypsinized in 3 mL trypsin-EDTA. A small aliquot of cells from each embryo were frozen for genotyping (see below). Cells were pelleted by spinning at 300 g and expanded onto five 15-cm plates with 20 mL complete media per plate. When fully confluent once again, MEFs were trypsinized and frozen down in freeze media (50% complete media, 40% CCS, and 10% DMSO) in aliquots of 1 plate per cryogenic vial. Cells were placed at -80°C for ~24 hr in a cell freezing container and then transferred to liquid N₂ for long-term storage.

For genotyping, cells were processed with the DNeasy Blood and Tissue kit (QIAGEN (Hilden, Germany) 69506). All MEF lines were tested for mycoplasm contamination with the following primer pairs: 5’-CTTCWTGACTTYYCAGACCCAGGTAT-3’ (Myco2(cb)) with 5’-ACACCATGGGAGYTGGTAAT-3’ (Myco11(cb)) and
5'-GGTGTGGGTGAGTATTACAAARTCAATT-3' (Myc05(cb)) with 5'-GGAGTGAGTGGATCCATAATTGTGA-3' (Myc06(cb)). Genotyping for the sex of each MEF line was performed with the following primer pair: 5'-CTGAAGCTTTTGCTTTTGAG-3' with 5'-CCACTGCCAAATCTTTG-3'. A single 340 bp product was expected for female cells, and an additional 310 bp product was present in male cells.

**Generation of FOS antibody**

We generated an in-house antibody against the full-length mouse protein for c-Fos (NCBI Reference Sequence: NP_034364.1). Briefly, we purified GST-c-Fos-His as detailed in Sharma et al., 2019 and injected the recombinant protein into immunocompromised rabbits. Serum was collected and affinity purified using a protein A column before use in ChIP-seq and CUT&RUN experiments.

**Cell culture**

Cells were thawed onto one 15-cm plate per MEF line and grown in complete media until fully confluent. For ChIP-seq and Hi-ChIP experiments, MEFs were split onto five 15-cm plates and grown in complete media until ~70-80% confluent. Cells were washed in 10 mL room-temperature PBS and switched to 20 mL warmed starve media (0.5% CCS, with the same supplement concentrations as complete media). After 26+ hours in starve media, samples to be serum stimulated were incubated with 20 mL warmed stimulation media (30% CCS, with the same supplement concentrations as complete media) for 0, 10, or 90 min.

For ATAC-seq, RNA-seq, and CUT&RUN experiments, MEFs were thawed as above and were split into 6-well dishes at a concentration of 5 x 10^5 cells per well in 2 mL warmed starve media. Cells were grown for 26+ hours, and appropriate wells were serum stimulated with 2 mL warmed stimulation media.

**Crosslinking cells**

Media was aspirated from MEFs, and 2 mL or 15 mL crosslinking buffer (10 mM HEPES pH 7.5, 100 mM NaCl, 1 mM EDTA, 1 mM EGTA) with 1% formaldehyde was added for 6-well or 15 cm dishes, respectively. Cells were crosslinked by shaking gently for 10 min at room temperature. Crosslinking was quenched by adding glycine to a final concentration of 125mM and incubating for 5 min at room temperature while shaking. Cells were washed once in 2 mL or 15 mL PBS for 6-well or 15 cm dishes, respectively. Cells were scraped and collected in 1 mL or 5 mL cold PBS for 6-well or 15 cm dishes, respectively, and pelleted by spinning at 1,000 g for 5 min at 4°C.

**ATAC-seq libraries**

MEFs from a 6-well dish were washed twice in 1 mL cold PBS and pelleted each time by spinning at 300 g for 5 min at 4°C. 50,000 MEFs were resuspended in 50 μL cold ATAC lysis buffer (10 mM Tris-HCl pH 7.5, 10 mM NaCl, 3 mM MgCl₂, 0.1% NP-40 0.1% Tween 20, 0.01% digitonin) and incubated for 3 min on ice. Lysed cells were washed once in 1 mL ATAC wash buffer (10 mM Tris-HCl pH 7.5, 10 mM NaCl, 3 mM MgCl₂, 0.1% Tween 20) by gently inverting the tube 3 times and pelleted by spinning at 500 g for 10 min at 4°C. Pelleted nuclei were resuspended in 50 μL transposition mix (25 μL 2x TD Buffer (Illumina (San Diego, CA, USA) 20034197), 2.5 μL TDE1 transposase (Illumina 20034197), 0.5 μL 10% Tween 20, 0.5% 1% digitonin, 16.5 μL PBS, 5 μL NF-H₂O) and incubated for 30 min at 37°C with a Thermomixer set to 1,000 rpm. Samples were purified with MinElute PCR Purification Kit (QIAGEN 28004) per manufacturer’s instructions and eluted in 13 μL NF-H₂O. Libraries were amplified by adding the following to 10 μL purified DNA: 2.5 μL 25 μM Ad1 universal primer, 2.5 μL 25 μM Ad2.* indexing primer, 25 μL NEBNext Hi-Fi 2x PCR Master Mix (NEB (Ipswich, MA, USA) M0541S), 10 μL NF-H₂O. After an initial 5 PCR cycles, libraries were quantified by qPCR by adding the following to 5 μL partially amplified DNA: 0.5 μL 25 uM Ad1 universal primer, 0.5 μL 25 μM Ad2.* indexing primer, 5 μL NEBNext Hi-Fi 2x PCR Master Mix, 0.15 μL 1x SYBR Green I (Thermo Fisher S7563), 3.85 μL NF-H₂O. All primer sequences referenced are described in Buenrostro et al., 2015. The number of additional PCR cycles required for amplifying remaining libraries was determined by the number of qPCR cycles needed to reach 1/3 of the maximum SYBR green signal. Libraries were purified with AMPure XP beads (0.5x volume; Beckman Coulter (Indianapolis, IN, USA) A63881), and the supernatant was retained to remove large fragments. Primer dimers were removed by a subsequent cleanup with AMPure XP beads (1.3x initial volume), and
libraries were eluted in 20 μL NF-H2O. Libraries were sequenced on an Illumina NextSeq 500 with 40 bp paired-end reads.

**ChIP-seq libraries**

Crosslinked MEFs per protocol above from 15 cm dishes were resuspended in 1 mL L1 buffer (50 mM HEPES pH 7.5, 140 mM NaCl, 1 mM EDTA, 1 mM EGTA, 0.25% Triton X-100, 0.5% NP-40, 10% glycerol, 10 mM sodium butyrate) per 15 cm dish starting material and rotated for 10 min at 4°C to lyse cells. Nuclei were pelleted by spinning at 1,350 g for 5 min at 4°C and resuspended in 1 mL L2 buffer (10 mM Tris-HCl pH 8.0, 200 mM NaCl, 10 mM sodium butyrate) per 15 cm dish starting material and rotated for 10 min at room temperature. Nuclei were pelleted by spinning at 1,350 g for 5 min at 4°C and resuspended in 300 μL LB3 buffer (10 mM Tris-HCl pH 8.0, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 0.1% sodium deoxycholate, 0.5% N-lauroylsarcosine, 10 mM sodium butyrate) per 15 cm dish starting material. Chromatin was sonicated with a Bioruptor Plus (Diagenode (Denville, NJ, USA)) on “high” power setting with an “on” interval of 30 sec and “off” interval of 45 sec for 36 cycles). DNA concentration was determined by taking 100 μL aliquot of sonicated chromatin, decrosslinking at 95°C for 15 min, and purifying with QIAquick PCR Purification Kit (QIAGEN 28104) and quantifying by Nanodrop. 1 μg of purified chromatin in 10% glycerol was run on a 2% agarose gel and stained with ethidium bromide for 30 min to validate fragment size (typically within ~200-1,000 bp). The remainder of the sonicated chromatin was transferred to 1.5 mL tubes and centrifuged at 16,000 g for 10 min at 4°C to pellet insoluble debris. Triton X-100 was added to soluble chromatin to a final 1% concentration. Protein A Dynabeads (Thermo Fisher 10008D) were washed twice in 1 mL cold block solution (0.5% BSA (w/v), 1% Triton X-100, diluted in LB3 buffer). For coupling antibodies to beads, 15 μL bead slurry per IP were resuspended in 1.5 mL cold block solution, and the appropriate amount of antibody (0.5 μg for anti-H3K27ac (Abcam (Waltham, MA, USA) ab4729), 0.5 μg for anti-H3K4me1 (Abcam ab8895), 0.5 μg for anti-H3K4me2 (Abcam ab7766), 0.5μg for anti-H3K4me3 (Abcam ab5850), 2 μg for anti-FOS (in-house generated antibody and Santa Cruz Biotechnology (Dallas, TX, USA) sc-7202X), 2 μg for anti-TEAD1 (Abcam ab135353), 2 μg for anti-CTCF (Active Motif (Carlsbad, CA, USA) 61312), and 2 μg for anti-JUND (Santa Cruz Biotechnology sc74)) was added before rotating beads for >2 hrs at 4°C. For pre-clearing chromatin, 15 μL bead slurry was added to appropriate amount of chromatin (40 μg for histone modifications, 80 μg for transcription factors), and additional cold LB3 buffer with 1% Triton X-100 was added such that all samples had a final volume of 1.5 mL before rotating samples for > 2 hrs at 4°C. Pre-cleared chromatin was added to antibody-coupled beads, and additional cold LB3 buffer with 1% Triton X-100 was added such that all samples had a final volume of 1.8 mL before rotating samples overnight at 4°C. 50 μL of pre-cleared chromatin was stored at -20°C for making input libraries. For all wash steps listed below, samples were rinsed with 1 mL cold wash buffer and rotated for 5 min at 4°C before separating beads with a magnet and discarding supernatant. Samples were washed sequentially twice in low salt buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCl pH 8.0, 150 mM NaCl), twice in high salt buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCl pH 8.0, 500 mM NaCl), twice in LiCl buffer (250 mM LiCl, 1% NP-40, 0.5% sodium deoxycholate, 1 mM EDTA, 10 mM Tris-HCl pH 8.0), and once in TE buffer (50 mM Tris-HCl pH 8.0, 10 mM EDTA). Samples were eluted from beads by addition of 200 μL TE buffer with 1% SDS and incubating at 65°C for 30 min, with brief vortexing every 10 min to mix. IP samples were placed on magnet, and supernatant was transferred to new tubes. Input samples were also thawed, and 150 μL TE buffer with 1% SDS was added. Both IP and input samples were decrosslinked by incubating at 65°C overnight. 10 μg RNase A (Sigma Aldrich (St. Louis, MO, USA) R6513) was added and samples were incubated at 37°C for 1 hr to digest RNA. 7 μL 20 μg/μL proteinase K (New England Biolabs P8107) was added, and samples were incubated at 55°C for 2 hr to digest protein. DNA was extracted with 1 volume of 25:24:1 phenol-chloroform-isoamyl alcohol and purified with QIAquick PCR Purification Kit (QIAGEN 28104). Libraries were prepared with the Ovation Ultralow V2 DNA-Seq Library Preparation Kit (NuGEN (Redwood City, CA, USA) 0344NB-32) per manufacturer’s instructions. Libraries were sequenced on an Illumina NextSeq 500 with 150 bp paired-end reads.

**CUT&RUN libraries**

Crosslinked MEFs per protocol above from 6-well dishes were washed once in 2 mL PBS and collected in 1 mL cold NE1 buffer (20 mM HEPES pH 7.5, 10 mM KCl, 1 mM MgCl2, 1 mM DTT, 0.1% Triton X-100, Roche Protease Inhibitor Cocktail (Millipore (Burlington, MA, USA) 11873580001)). Cells were permeabilized to isolate nuclei by...
rotating for 10 min at 4°C. Nuclei were pelleted by spinning at 500 g for 5 min at 4°C and resuspended in 1 mL cold CUT&RUN wash buffer (20 mM HEPES pH 7.5, 0.2% Tween-20, 150 mM NaCl, 0.1% BSA, 0.5 mM spermidine, 10 mM sodium butyrate, Roche Protease Inhibitor Cocktail). 10 µL concanavalin-coated bead slurry (Bangs Laboratories (Fishers, IN, USA) BPS31) per sample was washed twice in 1.5 mL CUT&RUN binding buffer (20 mM HEPES pH 7.5, 10 mM KCl, 1 mM CaCl$_2$, 20 mM MnCl$_2$) and resuspended in a final volume of 10 µL binding buffer per sample. After adding 10 µL bead slurry to each sample, tubes were inverted 10 times and incubated for 10 min at room temperature to bind nuclei. Beads were separated from wash buffer by placing on magnet for >30 sec and were resuspended in 50 µL antibody buffer (0.1% Triton X-100, 2 mM EDTA, diluted in CUT&RUN wash buffer). Antibodies (in-house anti-FOS, anti-H3K27ac (Abcam ab4927), or rabbit IgG (Cell Signaling Technology (Danvers, MA, USA) 2729S)) were added at 1:50 dilution, and samples were incubated overnight at 4°C. Beads were washed twice in 1 mL Triton-wash buffer (0.1% Triton X-100, diluted in CUT&RUN wash buffer) and resuspended in 50 µL antibody buffer. Protein-A MNase (Skene and Henikoff, 2017) was added to a final concentration of 700 ng/mL, and samples were incubated for 1 hr at 4°C. Beads were washed twice in 1 mL Triton-wash buffer and resuspended in 100 µL Triton-wash buffer. 2 µL 100 mM CaCl$_2$ was added per sample to activate the MNase and each sample incubated on ice for 30 min. 100 µL 2x STOP buffer (340 mM NaCl, 20 mM EDTA, 4 mM EGTA, 0.1% Triton X-100, 50 µg/mL RNase A (Sigma Aldrich R6513), 2 pg/mL yeast spike-in DNA (provided by S. Henikoff)) was added, and samples were incubated for 20 min at 37°C to release CUT&RUN fragments from nuclei. Samples were placed on magnet, and supernatant was transferred to a new tube and added to 2 µL 10% SDS and 2 µL 20 mg/mL proteinase K (New England Biolabs P8107). Samples were incubated overnight at 65°C to reverse crosslinks. DNA was extracted with 1 volume of 25:24:1 phenol-chloroform-isoamyl alcohol and precipitated in 2.5 volumes of 100% ethanol with 2 µL glycogen (Sigma Aldrich 10901393001). Pellet was washed once in 1 mL 100% ethanol and dissolved in 40 µL 10 mM Tris-HCl pH 8.5. Libraries were prepared as described in (Skene and Henikoff, 2017), with two subsequent AMPure XP bead cleanups (1.1x volume) to fully remove contaminating adapter dimers from libraries. Libraries were sequenced on an Illumina NextSeq 500 with 40 bp paired-end reads.

**Hi-ChIP libraries**

Hi-ChIP was performed as previously described in (Mumbach et al., 2017) with the following modifications. 15 µL of Mbol restriction enzyme (New England Biolabs R0147) was used for digesting chromatin from 15 million MEFs. Sonication was performed with a Covaris M220 with the following settings: duty cycle = 5, PIP = 70, cycles/burst = 200, and time = 8 min. 75 µL of Protein A Dynabeads (Thermo Fisher 10008D) was used for IP and 1 µg of anti-H3K27ac (Abcam ab4927) antibody was used per sample to typically yield 12.5 ng DNA. Accordingly, 0.6725 µL of transposase enzyme (Nextera 20034197) was used to insert adapters, and libraries were amplified for 8 PCR cycles.

**RNA-seq libraries**

MEFs from 15 cm dish were washed once in 15 mL cold PBS and pelleted by spinning at 300 g for 5 min at 4°C. Cell pellet was resuspended in 200 µL cold cytoplasmic lysis buffer (10 mM Tris-HCl pH 7.5, 150 mM NaCl, 0.15% NP-40) and rotated for 5 min at 4°C. Lysate was layered on top of 500 µL cold sucrose buffer (10 mM Tris-HCl pH 7.5, 150 mM NaCl, 24% sucrose (w/v)) and centrifuged at 16,000 g for 10 min at 4°C. These steps were repeated once more to achieve higher purity in the nucleoplasmic fraction. Pelleted nuclei were resuspended in 200 µL glycogen buffer (20 mM Tris-HCl pH 7.9, 75 mM NaCl, 0.5 mM EDTA, 50% glycogen, 0.85 mM DTT), and an equal volume of cold nuclear lysis buffer was added (20 mM HEPES pH 7.5, 7.5 mM MgCl$_2$, 0.2 mM EDTA, 300 mM NaCl, 1 M urea, 1% NP-40, 1 mM DTT). Tubes were gently vortexed twice for 2 sec, incubated for 1 min on ice, and centrifuged at 18,000 g for 2 min at 4°C. These steps were repeated once more to achieve higher purity in the chromatin fraction. The remaining chromatin pellet was resuspended in 50 µL cold PBS and vortexed briefly. 500 µL of TRIzol (Thermo Fisher 15596026) was added to the pellet and vortexed for several minutes until fully resuspended. Chromatin-associated RNA was isolated with RNAeasy Mini Kit (QIAGEN 74104) per manufacturer’s instructions, and libraries were generated from 250 ng starting material with NEBNext Ultra Directional RNA Library Prep Kit for Illumina (New England Biolabs E7765). Libraries were sequenced on an Illumina NextSeq 500 with 150 bp paired-end reads.

**Pseudogenome generation**
SNPs occurring in the CAST/EiJ, MOLF/EiJ, PWK/PhJ, and SPRET/EiJ genomes relative to the mm10 reference genome were obtained from SNP release version 5 of the Mouse Genomes Project (Keane et al., 2011). Only high-confidence SNPs annotated with the PASS filter, filtered using VCFtools (version 0.1.12; Danecek et al., 2011), were used in all analyses. A separate pseudogenome for each wild-derived strain was constructed from these SNPs using Modtools (version 1.0.2; Huang et al., 2014).

Allele-specific read mapping
Reads were trimmed with the paired-end option and with SLIDINGWINDOW:5:30 using Trimmomatic (Bolger et al., 2014). Paired-end reads that survived trimming were re-paired using the bmap utility (Bushnell 2014). Both unpaired and paired reads were concurrently mapped to the C57BL/6J and appropriate pseudogenome with bowtie2 using default parameters (Langmead and Salzberg, 2012). Mapped reads were converted to .bam format with samtools view (Li et al., 2009) using the following options -h -b -F 3844 -q 10 and sorted by coordinate. Reads initially mapped to each pseudogenome were converted back to C57BL/6J coordinates by running Lapels (Huang et al., 2014). All unpaired reads were then resorted by query name with samtools view -n and their flags were fixed with samtools fixmate. Informative reads (i.e. those that overlapped SNPs) were subsetted with the extractasReads.R utility from asSeq (Sun 2012) and remapped to the reciprocal genome using the same commands as above. To retrieve our final set of allele-specific reads, we inputted the informative reads into the WASP pipeline (van de Geijn et al., 2015) to retain only those reads that map to a single locus in only one genome. Tag directories for both alleles were generated with HOMER’s makeTagDirectory command with total mapped reads (i.e. before running WASP pipeline) and allele-specific reads.

For visualization purposes, mapped reads in .bam format were also converted to .bed format, and unique reads were retained (with sort -k1,1 -k2,2n -k3,3n -u) and extended by 150 bp with bedtools slop -l 0 -r 150. All samples were normalized to a depth of 10 million reads, and read coverage was calculated by bedtools genomeCoverageBed. The output .bedgraph file was then converted with UCSC’s bedGraphToBigWig utility, and all tracks displayed were visualized with the UCSC Genome Browser (GRCm38/mm10).

ATAC-seq peak calling
Reads from individual bioreplicates were pooled with samtools merge. Two pseudoreplicates consisting of a random subset (50%) of total reads were generated by samtools view -h -b -s 1.5 and samtools view -h -b -s 2.5. Peaks were called from pooled reads and from two pseudoreplicates using macs2 (Liu 2014) with the following options: -p 1e-1 --nomodel --extsize 200. Peaks were also called using spp (Kharchenko et al., 2008) with -npeak=500000 to include a large set of putative peaks. For both macs2 and spp, reads from input DNA pooled from all ChIP-seq experiments were used as the control sample. To analyze consistency of peak calling across pseudoreplicates, we employed an Irreproducible Discovery Rate (IDR) approach (Landt et al., 2012) by running batch-consistency-analysis.R and ranking peaks by p.value for macs2 and signal.value for spp. Peaks with a global IDR score of 0.0025 or less were retained for downstream analyses.

Allele-specific CUT&RUN peak calling
Peak calling was performed as detailed above for ATAC-seq data, except reads mapping to the C57BL/6J and corresponding pseudogenome for each F1-hybrid line were inputted separately into macs2 and spp. CUT&RUN peaks were then intersected with all ATAC summits detected across all genotypes and were recentered on the summit of ATAC-seq signal. Peaks from both the C57BL/6J and pseudogenome were concatenated, and only peaks with at least one SNP/indel within +/- 60 bp of the ATAC summit were retained for allele-specific analysis (as highly “mappable” sites). To determine whether the CUT&RUN signal is significantly skewed towards one allele, we used HOMER (Heinz et al., 2010) to annotate read coverage with -noadj -size -250,250 for AP-1 and -noadj -size -500,500 for H3K27ac. These counts were inputted in DESeq2 (Love et al., 2014), and all peaks with an FDR < 0.1 were considered allele-specific. Both allele-specific and shared peaks were then filtered by the following criteria: (1) when peak summits occurred within 1 kb of one another, only the summit with the highest pooled ATAC-seq signal was retained for downstream analyses, (2) peaks within the bottom quintile of pooled ATAC-seq signal per condition per F1-hybrid line were also excluded as low-signal sites, and (3) peaks that fell within 100 kb of gene
bodies of known imprinted genes were filtered out of our remaining dataset to rule out differences in CRE activity that result from parent-of-origin effects (Shen et al., 2014).

**Validation of CUT&RUN experiments**

Since we modified existing protocols for CUT&RUN (Skene and Henikoff, 2017) to decrease the number of cells and sequencing reads compared to those typically required for generating ChIP-seq data, we performed a series of analyses to ensure that we were still able to quantitatively measure TF binding. While other MNase-based methods have reported sequence-dependent biases that could result in preferential cutting at open chromatin regions (Chung et al., 2010), we noted a similar fraction of reads in peaks from CUT&RUN and ChIP-seq when using identical antibodies (Figure S6D), suggesting that we observe a minimal open chromatin bias with our modified CUT&RUN protocol. We also noted similar levels of binding at FOS peaks with ChIP-seq using our newly generated FOS antibody in comparison with a previously available commercial antibody, and we confirmed the specificity of our antibody by comparing peaks found in our FOS-binding data with HA ChIP-seq data from wild-type C57BL/6J MEFs and C57BL/6J MEFs that express FOS-FLAG-HA (Figure S6A-C). When we computationally separate shorter (<120 bp) from nucleosomal (>150 bp) FOS CUT&RUN reads, we found that sub-nucleosomal reads were more likely to be enriched at the core of FOS-bound enhancers, showed greater signal-to-noise relative to an IgG control antibody (Figure S6F), and could be used to detect footprints containing AP-1-binding motifs (as a proxy for the detection for sequence-specific AP-1 binding; Figure S6G).

**Motif footprinting with CUT&RUN reads**

Since CUT&RUN utilizes an antibody-targeted MNase for cleaving DNA fragments at TF-bound regions, individual cut sites derived from both ends of paired-end sequencing reads can be used for higher resolution mapping of specific nucleotides bound by TFs within peaks. DNA motifs that are recurrently protected (termed “footprints”) from MNase by chromatin-associated protein binding were identified from FOS CUT&RUN experiments performed in serum-stimulated C57BL/6J MEFs. Peak calling, motif identification, and footprinting analysis were performed using CUT&RUNtools (Zhu et al., 2019) with default parameters. Shown in Figure S6G is the aggregated cut site probability within +/- 100 bp of all identified MTGAGTCA sites at FOS CUT&RUN peaks, suggesting that our FOS CUT&RUN experiments are able to detect direct binding sites for AP-1.

**Allele-specific ChIP-seq peak calling**

Peak calling was performed as detailed above for CUT&RUN data, except we considered all ChIP-seq peaks that overlapped ATAC-seq summits.

**Detection of significant Hi-ChIP interactions**

H3K27ac Hi-ChIP reads were aligned with HiC-Pro (Servant et al., 2015) using an MboI-digested mm10 genome as the reference genome. Significant H3K27ac loops were determined by running hichipper (Larea et al., 2018), inputting 1 kb windows centered on all previously identified distal active enhancers from C57BL/6J MEFs (Vierbuchen et al., 2017) as possible loop anchors. Only loops that were supported by at least 10 paired-end tags per replicate and had a p-value < 1e-4 from hichipper were retained from each timepoint (0m and 90m). Using these criteria, we noted similar numbers of H3K27ac Hi-ChIP loops in our dataset as those from other cell types (Mumbach et al., 2017). We generated tracks for visualization by retaining the midpoint of all significant loops.

**Analysis of allele-specific gene expression**

Reads were mapped with STAR 2.7.3 (Dobin et al., 2013) with the following options to enable WASP filtering of allele-specific reads: --outSAMattributes NH HI AS nM vG vA --waspOutputMode SAMtag. Genome-specific reads were extracted and converted into .bam format with samtools view -h -b -F 3844 -q 10. The featureCounts command from Subread (Liao et al., 2013) was used to quantify the number of allele-specific reads per genome that overlapped each mm10 Refseq gene bodies. Genes with an average expression per sample < 1 were filtered out, and counts from individual genotypes and timepoints were inputted into edgeR. Genes with an FDR < 0.05 by glmQLFTest were considered allele-specific in their expression.
Scatterplots for quantifying TF binding or chromatin state across alleles

Allele-specific reads are converted from .bam files into tag directories for HOMER (Heinz et al., 2010). Single bp coordinates, typically from ATAC-seq summits, are annotated with separate tag directories for the C57BL/6J and pseudogenome-specific reads with the following options: mm10 -noadj -size -250,250 for TFs and mm10 -noadj -size -500,500 for histone modifications. The resulting read coverage values are log2 transformed and plotted with geom_point in ggplot2 against one another.

Aggregate plots for averaging TF binding or chromatin state across peaks

Allele-specific reads were prepared as described above with HOMER. ATAC-seq peak centers or TF motif k-mers were then annotated with allele-specific read tag directories with the following options: mm10 -noadj -noann -nogene -size -1000,1000 -hist 10. Individual coverage values across 10 bp bins are plotted with geom_line in ggplot2.

Number and position of SNPs/indels at cis-regulatory elements

To determine the total number of SNPs/indels within the central 150 bp of enhancers as in Figure 2I, we used bedtools window -w 75 -c and centered on the ATAC-seq summit present at each putative CRE. We also mapped the positioning of SNPs/indels relative to the ATAC-seq summit by using bedtools window -w 200, and computed the difference in coordinates between the ATAC-seq summit and the closest nucleotide present in the SNP/indel. For defining the locations of putative CTCF motifs, we inputted the MA0139.1 profile from the JASPAR database into FIMO and limited the maximum stored scores to $10^6$ per genome.

Mammalian conservation scores at cis-regulatory elements

To more directly compare allele-specific and shared CREs with similar levels of transcriptional activity, we subsampled the pool of shared CREs such that the distribution of H3K27ac levels (+/- 500 bp from ATAC-seq summit) on the active allele matched that of the allele-specific CREs. Bigwig files with phastCons scores for 60 vertebrate species for each mouse mm10 chromosome were obtained from UCSC. For each CRE, we computed a phastCons score for a 150 bp window centered on the ATAC-seq summit using the bigWigAverageOverBed script from UCSC Tools (Version 3.6.3).

Identifying recurrent k-mer clusters at cis-regulatory elements with KMAC

Nucleotide sequences present at the central 60 bp of enhancers were extracted using bedtools getfasta (Quinlan and Hall, 2010). These .fasta files were inputted into KMAC (Guo et al., 2018) as the positive sequences (using the appropriate --k_seq[n] and enriched k-mer clusters are determined relative to random GC-matched control regions of equal length from the C57BL/6J genome (using --gc -1) with the following additional options: --k_min 5 -k_max 10 --k_top 10.

Identifying k-mers that distinguish classes of AP-1 bound sites with gkm-SVM

Coordinates for AP-1 peaks were converted to appropriate pseudogenome coordinates with modmap (Huang et al., 2014) with -f and -d options. Nucleotide sequences in .fasta format for both alleles of each locus (60 bp window) were obtained with bedtools getfasta and were concatenated across different F1-hybrid lines. For performing the active versus inactive FOS-bound site comparison, we used the gkm-SVM package developed by Dr. Michael Beer’s lab (Ghandi et al., 2016) and generated the kernel matrix by inputting the active allele DNA sequence (with higher H3K27ac levels) as the positive set and the corresponding inactive allele DNA sequence as the negative set. SVM training was done with the gkmsvm_trainCV command using default parameters and k-mer weights were calculated for all possible 10-mers with gkmsvm_classify.

ACKNOWLEDGEMENTS

We would like to thank members of the Greenberg Lab for their scientific advice and input throughout this project, L. Hu for assistance in generating the anti-FOS antibody used in this study, and S. Bhunia for help with data.
visualization. We are grateful to L. Boxer, A. Carter, C. Davis, E. Duffy, E. Griffith, and E. Li for their helpful feedback on this manuscript.

M.G.Y. was supported by National Institutes of Health under training grants T32EY00711030 and T32AG000222. E.L. was supported by the National Science Foundation Graduate Research Fellowship under grant numbers DGE0946799 and DGE1144152. This work was funded by the NIH (R01 NS115965 to M.E.G.).

**AUTHOR CONTRIBUTIONS**

Conceptualization: M.G.Y., E.L., M.E.G., and T.V.
Methodology: M.G.Y., E.L., M.E.G, and T.V.
Investigation: M.G.Y., E.L., C.J.C., and T.V.
Formal Analysis: M.G.Y., E.L., and T.V.
Writing - Original Draft: M.G.Y., E.L., M.E.G., and T.V.
Writing - Review & Editing: M.G.Y., E.L., M.E.G., and T.V.
Supervision: M.E.G. and T.V.
Funding Acquisition: M.E.G. and T.V.

**DATA AVAILABILITY**

All genomic data reported in this study have been deposited in the NCBI Gene Expression Omnibus (pending accession number).

**DECLARATION OF INTERESTS**

The authors declare no competing interests.

**FIGURE LEGENDS**

**Figure 1. Identification of allele-specific TF binding and chromatin features in F1-hybrid fibroblasts.** (A) F1-hybrid male MEFs were derived from crosses between female C57BL/6J mice and male mice from a panel of inbred mouse strains. Experiments were performed in quiescent (0 min) and serum-stimulated (90 min) MEFs from at least two independent male embryos as biological replicates for each assay. Reads were mapped to either the maternal or paternal allele to quantify chromatin state and TF binding at CREs in an allele-specific manner. For wild-derived strains, ATAC-seq data was generated using MEFs from corresponding parental lines and compared with chromatin accessibility in C57BL/6J MEFs. Similarly, H3K27ac Hi-ChiP data was obtained only from starved and serum-stimulated MEFs from C57BL/6J mice. All other genomic data indicated herein were obtained using MEFs derived from male F1-hybrid embryos. (B) Example genome browser track of a locus (chr5:147,587,473-147,599,697 in mm10 genome) with an allele-specific enhancer (indicated in gray, on the right) and a shared enhancer (indicated in gray, on the left) in C57BL/6J x SPRET/EiJ F1-hybrid MEFs. Normalized read densities for ATAC-seq and H3K27ac ChIP-seq for each allele are shown. (C-F) Scatterplots of maternal (C57BL/6J) and paternal allele-specific signal for histone modifications and CTCF binding (n = 61,288 proximal H3K27ac, n = 138,662 distal H3K27ac, n = 47,485 distal CTCF, n = 46,853 proximal H3K4me3, n = 127,077 distal H3K4me1, and n = 97,084 distal H3K4me2 allele pairs, respectively). Points indicated in light and dark colors represent peaks with and without a significant skew in signal between alleles, respectively (FDR < 0.1 with DESeq2). CTCF and H3K4me3 levels were less likely to show an allele-specific skew in signal, in comparison with H3K27ac levels at active enhancers (Fisher’s exact test, p < 2.2 x 10^{-16} for CTCF, p < 2.2 x 10^{-16} for H3K4me3). (G) Scatterplot of allele-specific H3K4me1, ATAC-seq, and FOS binding signal at top decile of allele-specific enhancers, comparing signal from the active and inactive alleles (defined based on relative H3K27ac levels) to one another.

**Figure 2. Frequency and distribution of sequence variants at allele-specific CREs.** (A) Schematic depicting TF-bound enhancer with zero SNPs/indels (denoted by red X’s) in the transposase-accessible CRE region (indicated in orange). Nucleosomes flanking both ends of the accessible region at active enhancers are marked by histone post-translational modifications (e.g. H3K27ac), which are used as proxies for the transcriptional state of
each enhancer. DNA sequences in these flanking regions tend to also be less conserved that sequences within enhancers themselves, thus often allowing sequencing reads to be correctly assigned to one of two genomes in F1 hybrid cells in the absence of SNPs/indels within enhancer sequences. (B) Example genome browser track of a locus (chr11:113,290,106-113,416,149 (top) and chr11:113,350,775-113,356,042 (bottom) in mm10 genome) with an allele-specific 0-SNP/indel enhancer (indicated in gray, on the left) within an enhancer “cluster” (indicated in box) in C57BL/6J x CAST/EiJ F1-hybrid MEFs. The 0-SNP enhancer is situated <2 kb from a SNP/indel-containing enhancer (indicated in gray, on the right) within the same cluster. Normalized read densities for ATAC-seq and H3K27ac ChIP-seq for each allele are shown. Tick marks indicate positions of annotated SNPs/indels that distinguish the C57BL/6J and CAST/EiJ genomes. (C-E) Histogram of number of SNPs/indels present within the central 150 bp of allele-specific and signal-matched, shared active enhancers, promoters, and putative insulators (mean = 2.36 and 1.57 SNPs/indels for enhancers, respectively, two-tailed unpaired t-test, p < 2.2 x 10^{-16} for enhancers; mean = 1.84 and 1.09 SNPs/indels for promoters, respectively; two-tailed unpaired t-test, p = 7.9 x 10^{-4} for promoters; mean = 2.95 and 2.03 SNPs/indels for insulators, respectively; two-tailed unpaired t-test, p < 2.2 x 10^{-16} for insulators). (F) Proportion of enhancers that show allele-specific and >2-fold difference in signal, plotted as a function of the number of SNPs/indels present within their central 150 bp. (H-J) Frequency of SNPs/indels at positions relative to ATAC-seq summits for allele-specific (and >2-fold), 1-SNP active enhancers, and H3K4me3-marked promoters. In all plots, SNP/indel frequency was compared to corresponding signal-matched shared sites.

Figure 3. Characterization of TF-binding motifs and changes in chromatin state at allele-specific binding sites for AP-1, TEAD, and CTCF. (A-B) Scatterplots of maternal (C57BL/6J) and paternal allele-specific signal for AP-1 and TEAD binding (n = 85,198 distal FOS, and 75,350 distal TEAD1 allele pairs, respectively). Points indicated in light and dark colors represent peaks with and without a significant skew in signal between alleles, respectively (FDR < 0.1 with DESeq2). (C-D) Distribution of SNPs centered on respective k-mers (denoted by dashed lines) at allele-specific, active, and gene-distal FOS and TEAD1 peaks with >2-fold difference in binding signal across alleles (n = 263 and n = 1,035 allele pairs, respectively). (E) Distribution of SNPs centered on CTCF PWM (MA0139.1; denoted by dashed lines) at allele-specific, gene-distal CTCF peaks with >2-fold difference in binding signal across alleles (n = 1,663 allele pairs). (F-I) Aggregate plot of allele-specific FOS, ATAC-seq, H3K4me1, and H3K27ac reads centered on ATAC-seq summits at active FOS peaks. These sites have been selected because they contain a single SNP/indel-containing AP-1 site and no shared AP-1 sites within 75 bp of the ATAC-seq summit. Signal is compared between alleles with intact versus mutated core AP-1 motifs (TGASTCA). (J-M) Aggregate plot of allele-specific TEAD1, ATAC-seq, H3K4me1, and H3K27ac reads centered on ATAC-seq summits at active TEAD1 peaks. These sites have been selected because they contain a single SNP/indel-containing TEAD site and no shared TEAD sites within 75 bp of the ATAC-seq summit. Signal is compared between alleles with intact versus mutated extended TEAD motifs (GGAATK).

Figure 4. Impact of genetic variants on AP-1 and TEAD binding. (A) Percentage of allele-specific versus shared gene-distal FOS peaks that contain strain-specific core AP-1 (TGASTCA; Fisher’s exact test, p < 2.2 x 10^{-16}) or extended TEAD (GGAATK; Fisher’s exact test, p = 2.123 x 10^{-9}) k-mer(s) within 75 bp of their respective ATAC-seq summits. (B) Percentage of allele-specific versus shared gene-distal TEAD1 peaks that contain strain-specific core AP-1 (TGASTCA; Fisher’s exact test, p < 2.2 x 10^{-16}) or extended TEAD (GGAATK; Fisher’s exact test, p < 2.2 x 10^{-16}) k-mer(s) within 75 bp of their respective ATAC-seq summits. (C) Percentage of FOS-only peaks versus FOS/TEAD1 co-bound peaks in the C57BL/6J genome with at least one bindable AP-1 k-mer (VTGACTCB, VTGAACTCAB, or VTAGTCCAY) present within 50 bp of their respective ATAC-seq summits (Fisher’s exact test, p < 2.2 x 10^{-16}). (D) Percentage of TEAD1-only peaks versus FOS/TEAD1 co-bound peaks in the C57BL/6J genome with at least one extended TEAD k-mer (GGAATK) present within 50 bp of their respective ATAC-seq summits (Fisher’s exact test, p = 2.406 x 10^{-9}).
Figure 5. Distribution of sequence variants that influence AP-1, TEAD, and CTCF binding. (A-C) Frequency of SNPs/indels at positions relative to ATAC-seq summits at allele-specific (with >2-fold difference in signal between alleles) versus shared gene-distal FOS, TEAD1, and CTCF peaks (Pearson’s Chi-squared test, p = 9.7 x 10^-8 for AP-1, p < 2.2 x 10^-16 for TEAD, p < 2.2 x 10^-16 for CTCF, 100 bp window centered on ATAC-seq summit in all cases). (D-F) Frequency of SNPs/indels at positions relative to shared VTGACTCAB and GGAATK k-mers within 75 bp of the ATAC-seq summit at allele-specific (with >2-fold difference in signal between alleles) versus shared gene-distal FOS and TEAD1 peaks, respectively. Sites have been filtered to exclude any peaks that include SNPs/indels that overlap their cognate k-mers. (F) Frequency of SNPs/indels at positions relative to shared CTCF PWM (MA0139.1) within 75 bp of the ATAC-seq summit at allele-specific (with >2-fold difference in signal between alleles) versus shared gene-distal CTCF peaks. Sites have been filtered to exclude any peaks that include SNPs/indels at disrupt the CTCF PWM in a strain-specific manner.

Figure 6. Examination of sequence features of AP-1 bound sites with machine learning and TF footprint data from diverse human cell types. (A-B) Area under ROC and P-R curves for gkm-SVM comparison of FOS peaks in the C57BL/6J genome (positive set) and length-matched, random genomic regions (negative set). Shown are AUC values for different lengths of input DNA sequence, ranging from 10 to 300 bp, indicated in black. The same analysis was repeated after masking all instances of core AP-1 motifs (TGASTCA), indicated in green. (C) Frequency of consensus, human DNase footprints (from Vierstra et al., 2020) containing an extended AP-1 k-mer (VTGACTCAB) at positions relative to DNase-seq summits (n = 164,705 footprints). (D) Width of DNase footprints that contain an extended AP-1 k-mer (VTGACTCAB). (E) Number of additional DNase footprints within 100 bp of DNase-seq summits at DNase peaks with a VTGACTCAB-containing footprint. (F) Distance between VTGACTCAB-containing footprints and nearest neighboring DNase footprint.

SUPPLEMENTAL FIGURE LEGENDS

Supplemental Figure 1. Identification of allele-specific transcripts from chromatin-associated RNA-seq. (A) Volcano plots of chromatin-associated RNA-seq read density at annotated gene bodies. Indicated on the horizontal axis is the ratio of allele-specific signal between paternal and maternal (C57BL/6J) alleles for each F1-hybrid strain. Transcripts whose expression levels are significantly allele-specific in DEseq2 (FDR < 0.1) and edgeR (FDR < 0.05) are highlighted in blue. (B) Number of SNPs/indels annotated in each wild-derived inbred mouse strain, relative to the C57BL/6J genome. Number of transcripts per F1-hybrid line with significant allele-specific expression from reads pooled across all timepoints (0, 20, and 90 min).

Supplemental Figure 2. Properties of allele-specific enhancers and association with gene expression. (A) Percentage of primed CREs (ATAC-seq summits overlapping H3K4me1/2 peaks) identified in any F1-hybrid strain that are active regulatory elements (overlap H3K27ac peaks). In total, we found n = 283,339 pairs of active CRE alleles across all F1-hybrids in our dataset, including n = 142,898 pairs that contained SNP(s) within the central 120 bp relative to their respective ATAC-seq summits, which we considered to be highly mappable sites. (B) Percentage of active CREs with a significant skew in H3K27ac levels between alleles, which we denoted as allele-specific enhancers or promoters (based on distance to their nearest annotated TSS) for subsequent analyses (n = 31,627/138,622 allele pairs; FDR < 0.1 with DEseq2). (C) Distance to nearest annotated TSS for active CREs identified in any F1-hybrid strain (median = 19,786 bp). CREs were considered gene-distal if they occurred >1 kb from the nearest annotated TSS (cutoff denoted by dashed line). (D) Percentage of active enhancers that exhibit an allele-specific skew in nascent transcription levels for putative target genes, which were identified by statistically significant H3K27ac Hi-ChIP loops formed with active promoters that overlap an annotated TSS. Allele-specific enhancers are more likely than enhancers with shared H3K27ac signal to be linked with allele-specific transcripts (n = 1,002/6,907 and n = 3,403/37,390 detectable Hi-ChIP loops for allele-specific and shared enhancers, respectively; Fisher’s exact test, p = 2.2 x 10^-16).

Supplemental Figure 3. Contribution of locus-scale, cis-acting mechanisms to enhancer activity. (A) Distribution of H3K27ac signal from the active allele at (1) allele-specific enhancers, (2) all shared enhancers,
and (3) a randomly subsampled set of shared enhancers that were signal-matched to the active allele of allele-specific sites. (B) Frequency of SNPs/indels for enhancers in our dataset assigned as 0-SNP/indel enhancers, which we defined as those having no annotated genetic variants within 75 bp of the ATAC-seq summit. No notable differences in the pattern of SNPs/indels were observed between allele-specific and shared 0-SNP enhancers in regions flanking the central 150 bp window that contains no genetic variants. (C-E) Cumulative distribution of distances to nearest allele-specific CTCF peak, nearest allele-specific CRE, and nearest active CRE. Shown is a comparison between allele-specific and signal-matched, shared enhancers, and both sets of sites contain zero SNPs/indels in their central 150 bp (Kolmogorov-Smirnov, \( p = 0.0142 \) for nearest allele-specific CTCF peak; Kolmogorov-Smirnov D = 0.02788 for nearest allele-specific CTCF peak; \( p < 0.0001 \) for nearest allele-specific CRE; Kolmogorov-Smirnov D = 0.1313 for nearest allele-specific CRE; \( p < 0.0001 \) for nearest active CRE; Kolmogorov-Smirnov D = 0.05563, for nearest active CRE). (F) Cumulative distribution of phastCons scores for allele-specific and signal-matched, shared active enhancers and promoters (Kolmogorov-Smirnov, \( p < 0.0001 \) and Kolmogorov-Smirnov D = 0.1274 for enhancers; Kolmogorov-Smirnov, \( p < 0.0001 \) and Kolmogorov-Smirnov D = 0.1137 for promoters).

Supplemental Figure 4. Quantification of trans-acting effects on enhancer chromatin state and gene expression. (A) Scatterplot of H3K27ac CUT&RUN signal at active CREs identified on the X-chromosome of the C57BL/6J x 129/SvImJ F1 hybrid (n = 216 peaks). Shown are pairwise comparisons of H3K27ac read densities from the C57BL/6J x 129/SvImJ line and normalized H3K27ac CUT&RUN signal from other F1-hybrid lines. Since we setup our breeding scheme such that each F1-hybrid MEF line was derived only from male embryos that contain a single X-chromosome inherited from their C57BL/6J mother, this enabled us to quantify chromatin state without the need for allele-specific mapping at this particular set of CREs. (B) Scatterplot of nascent RNA-seq read counts overlapping gene bodies of expressed transcripts on the C57BL/6J X-chromosome (with an average of at least 100 reads per F1-hybrid line). For example, amongst 427 expressed genes on the X-chromosome, we found that n = 39 (9.3%) were expressed with >2-fold difference in signal between alleles, when comparing the C57BL/6J x CAST/EiJ and the C57BL/6J x SPRET/EiJ F1 hybrids.

Supplemental Figure 5. Sequence determinants of AP-1 binding. (A) Top enriched k-mer clusters identified with the KMAC algorithm present in the top decile of active C57BL/6J enhancers (Guo et al., 2018). DNA sequence from central 60 bp of enhancer regions were compared with random GC- and length-matched regions in the C57BL/6J genome. AP-1 k-mers were present in n = 1,410/4,579 active enhancers and n = 74/4,579 control regions, respectively. TEAD k-mers were present in n = 499/4,579 active enhancers and n = 29/4,579 control regions, respectively. (B) Top enriched k-mer clusters identified with the KMAC algorithm present in late-response gene enhancers in the C57BL/6J genome. Control regions used for comparison were constitutively active enhancers in the C57BL/6J genome. (C) Mean number of TGASVDB k-mers (corresponding to putative AP-1 half sites and/or some degenerate AP-1 binding motifs) present within 75 bp of the ATAC-seq summit at FOS peaks identified in the C57BL/6J genome. Data are separated into deciles based on aggregate FOS-binding signal within +/- 250 bp of their respective ATAC-seq summits. (D) Percentage of allele-specific versus shared FOS peaks that contain an extended AP-1 motif (VTGACTCAB) with a SNP in the 3 bp flanking the core motif (TGACTCA; Fisher’s exact test, \( p = 5.2 \times 10^{-5} \)).

Supplemental Figure 6. Comparison of ChIP-seq and CUT&RUN datasets for quantifying AP-1 binding. (A) Scatterplot of FOS ChIP-seq read density at FOS peaks originally identified in C57BL/6J MEFs, comparing signal from commercial (Santa Cruz Biotechnology sc-7202X) and in-house generated FOS (1096AE) antibodies. (B) Scatterplot of read density at FOS peaks originally identified in C57BL/6J MEFs, comparing signal from FOS and HA ChIP-seq experiments performed in MEFs expressing an epitope-tagged form of FOS (FOS-FLAG-HA). (C) Scatterplot of HA ChIP-seq read density at FOS peaks originally identified in C57BL/6J MEFs, comparing signal in wild-type C57BL/6J MEFs and FOS-FLAG-HA MEFs. (D) Fraction of mapped reads in C57BL/6J MEFs that fall within FOS and H3K27ac peaks (500 bp and 1 kb windows, respectively). Identical antibodies were used for ChIP-seq and CUT&RUN experiments. (E) Fragment lengths for mappable paired-end FOS and H3K27ac CUT&RUN reads (median = 139 and 163 bp, respectively). (F) Aggregate plot of FOS CUT&RUN reads mapped to
the C57BL/6J genome, pooled from data across five F1-hybrid lines, and centered on ATAC-seq summits at FOS peaks originally identified in C57BL/6J MEFs. Paired-end reads are categorized based on fragment size as either sub-nucleosomal (<120 bp) or nucleosomal (>150 bp). (G) MNase cut site probability for FOS CUT&RUN reads defined using CUT&RUNTools (Zhu et al., 2019) identifies a TF footprint centered on MTGAGTCA k-mer at FOS peaks, suggesting that our CUT&RUN data can reliably identify instances of direct AP-1 binding.

**Supplemental Figure 7. Application of gkm-SVM algorithm to identify k-mers required for AP-1, TEAD, and CTCF binding.** (A-C) ROC and P-R curves for gkm-SVM comparison of FOS, TEAD1, and CTCF peaks compared to randomly sampled regions in the C57BL/6J genome. (D-G) Top enriched 10-mers from gkm-SVM comparing the central 60 bp from n = 4,000 randomly selected FOS (with and without masking TGASTCA k-mers), TEAD1, and CTCF peaks in C57BL/6J MEFs (positive set) with n = 4,000 randomly sampled 60 bp windows across the C57BL/6J genome (negative set). Core motif sequences are indicated in bold and flanking nucleotides are underlined. (H) Top enriched 10-mers from gkm-SVM comparing the central 60 bp from active (positive set) and inactive allele (negative set) at FOS peaks with significantly allele-specific H3K27ac levels. (I) Top enriched 10-mers from gkm-SVM comparing the central 60 bp from n = 4,000 DNase footprints containing VTGACTCAB k-mers (positive set) and n = 4,000 random gene-distal genomic windows in the human genome (hg38) centered on VTGACTCAB k-mers (negative set).

**TABLE LEGENDS**

**Table 1. Total numbers of SNPs/indels per inbred mouse strain relative to the C57BL/6J reference strain.** ATAC-seq peaks were considered highly mappable if they contained a SNP/indel within a 120 bp window centered on their respective ATAC-seq summits.

**Table 2. Identification of significant interactions between active CREs.** Hi-ChIP experiments were performed after serum-starved and restimulated (90 min) C57BL/6J MEFs. Only intra-chromosomal loops (p < 1e-4) with at least 10 paired-end reads connecting them per bioreplicate were retained for analysis.

**Table 3. Number of allele pairs with allele-specific and shared signal for each histone modification and transcription factor examined in our dataset.** For FOS and H3K27ac experiments, the data from ChIP-seq (wild-derived strains; CAST/EiJ, MOLF/EiJ, PWK/PhJ, SPRET/EiJ) and CUT&RUN (less divergent inbred strains; 129S1/SvImJ, A/J, BALB/cJ, DBA/2J, NOD/ShiLtJ) were merged. Only CUT&RUN peaks with a SNP/indel present within 60 bp of the ATAC-seq summit were included for allele-specific analyses for non-wild-derived (i.e. less genetically divergent) strains.

**Table 4. Significant allele-specific transcripts from chromatin-associated RNA-seq data.** Reads were pooled across replicates and across timepoints of serum stimulation (0, 20, and 90 min). Positive and negative fold-changes indicate genes expressed at higher levels on the paternal, wild-derived allele and maternal, C57BL/6J allele, respectively.

**Table 5. Location, allele-specific H3K27ac values, and DNA sequences for top decile of allele-specific enhancers.** Enhancer loci were selected based on those with greatest fold-change in H3K27ac signal between active and inactive alleles (+/- 500 bp window centered on their respective ATAC-seq summits).
REFERENCES

Albert, F. W., & Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. Nature Reviews. Genetics, 16(4), 197-212. https://doi.org/10.1038/nrg3891

Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, Ł. M., Rath, M., & Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science (New York, N.Y.), 339(6123), 1074-1077. https://doi.org/10.1126/science.1232542

Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, Ł. M., Rath, M., & Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science (New York, N.Y.), 339(6123), 1074–1077. https://doi.org/10.1126/science.1232542

Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., Kundaje, A., & Zeitlinger, J. (2021). Base-resolution models of transcription-factor binding reveal soft motif syntax. Nature Genetics, 53(3), 354-366. https://doi.org/10.1038/s41588-021-00782-6

Barozzi, I., Simonatto, M., Bonifacio, S., Yang, L., Rohs, R., Ghisletti, S., & Natoli, G. (2014). Coregulation of transcription factor binding and nucleosome occupancy through DNA features of mammalian enhancers. Molecular Cell, 54(5), 844-857. https://doi.org/10.1016/j.molcel.2014.04.006

Bevington, S. L., Cauchy, P., Piper, J., Bertrand, E., Lalli, N., Jarvis, R. C., Gilding, L. N., Ott, S., Bonifer, C., & Cockerill, P. N. (2016). Inducible chromatin priming is associated with the establishment of immunological memory in T cells. The EMBO Journal, 35(5), 515-535. https://doi.org/10.15252/embj.201592534

Biddie, S. C., John, S., Sabo, P. J., Thurman, R. E., Johnson, T. A., Schiltz, R. L., Miranda, T. B., Sung, M. H., Trump, S., Lightman, S. L., Vinson, C., Stamatoyanopoulos, J. A., & Hager, G. L. (2011). Transcription factor AP1 potentiates chromatin accessibility and glucocorticoid receptor binding. Molecular Cell, 43(1), 145-155. https://doi.org/10.1016/j.molcel.2011.06.016

Bilu, Y., & Barkai, N. (2005). The design of transcription-factor binding sites is affected by combinatorial regulation. Genome Biology, 6(12), R103. https://doi.org/10.1186/gb-2005-6-12-r103

Bogdanovic, O., Fernandez-Miñán, A., Tena, J. J., de la Calle-Mustienes, E., Hidalgo, C., van Kruysbergen, I., van Heeringen, S. J., Veenstra, G. J., & Gómez-Skarmeta, J. L. (2012). Dynamics of enhancer chromatin signatures mark the transition from pluripotency to cell specification during embryogenesis. Genome Research, 22(10), 2043-2053. https://doi.org/10.1101/gr.134833.111

Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S., & Crawford, G. E. (2008). High-resolution mapping and characterization of open chromatin across the genome. Cell, 132(2), 311-322. https://doi.org/10.1016/j.cell.2007.12.014

Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell, 169(7), 1177-1186. https://doi.org/10.1016/j.cell.2017.05.038
Buenrostro, J. D., Wu, B., Chang, H. Y., & Greenleaf, W. J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Current Protocols in Molecular Biology, 109, 21.29.1-21.29.9. https://doi.org/10.1002/0471142727.mb2129s109

Bushnell, B. (2014). BBMap: A Fast, Accurate, Splice-Aware Aligner. Lawrence Berkeley National Laboratory. LBNL Report #: LBNL-7065E. Retrieved from https://escholarship.org/uc/item/1h3515gn

Carroll S. B. (2008). Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell, 134(1), 25-36. https://doi.org/10.1016/j.cell.2008.06.030

Chen, L., Loh, P. G., & Song, H. (2010). Structural and functional insights into the TEAD-YAP complex in the Hippo signaling pathway. Protein & Cell, 1(12), 1073-1083. https://doi.org/10.1007/s13238-010-0138-3

Chung, H. R., Dunkel, I., Heise, F., Linke, C., Krobitsch, S., Ehrenohofer-Murray, A. E., Sperling, S. R., & Vingron, M. (2010). The effect of micrococcal nuclease digestion on nucleosome positioning data. PloS One, 5(12), e15754. https://doi.org/10.1371/journal.pone.0015754

Comoglio, F., Simonatto, M., Polletti, S., Liu, X., Smale, S. T., Barozzi, I., & Natoli, G. (2019). Dissection of acute stimulus-inducible nucleosome remodeling in mammalian cells. Genes & Development, 33(17-18), 1159-1174. https://doi.org/10.1101/gad.326348.119

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frapton, G. M., Sharp, P. A., Boyer, L. A., Young, R. A., & Jaenisch, R. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences of the United States of America, 107(50), 21931-21936. https://doi.org/10.1073/pnas.1016071107

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. Bioinformatics (Oxford, England), 27(15), 2156-2158. https://doi.org/10.1093/bioinformatics/btr330

Dorighi, K. M., Swigut, T., Henriques, T., Bhanu, N. V., Scruggs, B. S., Nady, N., Still, C. D., 2nd, Garcia, B. A., Adelman, K., & Wysocka, J. (2017). Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation. Molecular Cell, 66(4), 568-576.e4. https://doi.org/10.1016/j.molcel.2017.04.018

Ercg, J., Saunders, T. E., Girardot, C., Devos, D. P., Hufnagel, L., & Furlong, E. E. (2014). Subtle changes in motif positioning cause tissue-specific effects on robustness of an enhancer’s activity. PLoS Genetics, 10(1), e1004060. https://doi.org/10.1371/journal.pgen.1004060

Eferl, R., & Wagner, E. F. (2003). AP-1: a double-edged sword in tumorigenesis. Nature Reviews. Cancer, 3(11), 859-868. https://doi.org/10.1038/nrc1209
Farh, K. K., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., Shoresh, N., Whitton, H., Ryan, R. J., Shishkin, A. A., Hatan, M., Carrasco-Alfonso, M. J., Mayer, D., Luckey, C. J., Patsopoulos, N. A., De Jager, P. L., Kuchroo, V. K., Epstein, C. B., Daly, M. J., Hafler, D. A., … Bernstein, B. E. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature, 518(7539), 337-343. https://doi.org/10.1038/nature13835

Farley, E. K., Olson, K. M., Zhang, W., Rokhsar, D. S., & Levine, M. S. (2016). Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers. Proceedings of the National Academy of Sciences of the United States of America, 113(23), 6508-6513. https://doi.org/10.1073/pnas.1605085113

Farrance, I. K., Mar, J. H., & Ordahl, C. P. (1992). M-CAT binding factor is related to the SV40 enhancer binding factor, TEF-1. The Journal of Biological Chemistry, 267(24), 17234-17240.

Fudenberg, G., & Pollard, K. S. (2019). Chromatin features constrain structural variation across evolutionary timescales. Proceedings of the National Academy of Sciences of the United States of America, 116(6), 2175-2180. https://doi.org/10.1073/pnas.1808631116

Fulco, C. P., Nasser, J., Jones, T. R., Munson, G., Bergman, D. T., Subramanian, V., Grossman, S. R., Anyoha, R., Doughty, B. R., Patwardhan, T. A., Nguyen, T. H., Kane, M., Perez, E. M., Durand, N. C., Lareau, C. A., Stamenova, E. K., Aiden, E. L., Lander, E. S., & Engreitz, J. M. (2019). Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. Nature Genetics, 51(12), 1664-1669. https://doi.org/10.1038/s41588-019-0538-0

Ghandi, M., Mohammad-Noori, M., Ghareghani, N., Lee, D., Garraway, L., & Beer, M. A. (2016). gkmSVM: an R package for gapped-kmer SVM. Bioinformatics (Oxford, England), 32(14), 2205-2207. https://doi.org/10.1093/bioinformatics/btw203

Grossman, S. R., Zhang, X., Wang, L., Engreitz, J., Melnikov, A., Rogov, P., Tewhey, R., Isaakova, A., Deplancke, B., Bernstein, B. E., Mikkelson, T. S., & Lander, E. S. (2017). Systematic dissection of genomic features determining transcription factor binding and enhancer function. Proceedings of the National Academy of Sciences of the United States of America, 114(7), E1291-E1300. https://doi.org/10.1073/pnas.1621150114

Grubert, F., Zaugg, J. B., Kasowski, M., Ursu, O., Spacek, D. V., Martin, A. R., Greenside, P., Srivas, R., Phanstiel, D. H., Pekowska, A., Heidari, N., Euskirchen, G., Huber, W., Pritchard, J. K., Bustamante, C. D., Steinmetz, L. M., Kundaje, A., & Snyder, M. (2015). Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. Cell, 162(5), 1051-1065. https://doi.org/10.1016/j.cell.2015.07.048

Guo, Y., Tian, K., Zeng, H., Guo, X., & Gifford, D. K. (2018). A novel k-mer set memory (KSM) motif representation improves regulatory variant prediction. Genome Research, 28(6), 891-900. https://doi.org/10.1101/gr.226852.117

Halow, J. M., Byron, R., Hogan, M. S., Ordoñez, R., Groudine, M., Bender, M. A., Stamatoyannopoulos, J. A., & Maurano, M. T. (2021). Tissue context determines the penetrance of regulatory DNA variation. Nature Communications, 12(1), 2850. https://doi.org/10.1038/s41467-021-23139-3

Hare, E. E., Peterson, B. K., Iyer, V. N., Meier, R., & Eisen, M. B. (2008). Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. PLoS Genetics, 4(6), e1000106. https://doi.org/10.1371/journal.pgen.1000106

He, X., Chatterjee, R., John, S., Bravo, H., Sathyarayana, B. K., Biddie, S. C., FitzGerald, P. C., Stamatoyannopoulos, J. A., Hager, G. L., & Vinson, C. (2013). Contribution of nucleosome binding preferences and
co-occurring DNA sequences to transcription factor binding. BMC Genomics, 14, 428. https://doi.org/10.1186/1471-2164-14-428

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A., Wang, W., Weng, Z., Green, R. D., Crawford, G. E., & Ren, B. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature Genetics, 39(3), 311-318. https://doi.org/10.1038/ng1966

Heintz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., & Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular Cell, 38(4), 576-589. https://doi.org/10.1016/j.molcel.2010.05.004

Heinz, S., Romanoski, C. E., Benner, C., Allison, K. A., Kaikkonen, M. U., Orozco, L. D., & Glass, C. K. (2013). Effect of natural genetic variation on enhancer selection and function. Nature, 503(7477), 487-492. https://doi.org/10.1038/nature12615

Huang, S., Holt, J., Kao, C. Y., McMillan, L., & Wang, W. (2014). A novel multi-alignment pipeline for high-throughput sequencing data. Database : The Journal of Biological Databases and Curation, 2014, bau057. https://doi.org/10.1093/database/bau057

Jadhav, U., Cavazza, A., Banerjee, K. K., Xie, H., O’Neill, N. K., Saenz-Vash, V., Herbert, Z., Madha, S., Orkin, S. H., Zhai, H., & Shivdasani, R. A. (2019). Extensive Recovery of Embryonic Enhancer and Gene Memory Stored in Hypomethylated Enhancer DNA. Molecular Cell, 74(3), 542-554.e5. https://doi.org/10.1016/j.molcel.2019.02.024

Jang, Y., Wang, C., Zhuang, L., Liu, C., & Ge, K. (2017). H3K4 Methyltransferase Activity Is Required for MLL4 Protein Stability. Journal of Molecular Biology, 429(13), 2046-2054. https://doi.org/10.1016/j.jmb.2016.12.016

Jindal, G. A., & Farley, E. K. (2021). Enhancer grammar in development, evolution, and disease: dependencies and interplay. Developmental Cell, 56(5), 575-587. https://doi.org/10.1016/j.devcel.2021.02.016

Johnson, T. A., Chereji, R. V., Stavreva, D. A., Morris, S. A., Hager, G. L., & Clark, D. J. (2018). Conventional and pioneer modes of glucocorticoid receptor interaction with enhancer chromatin in vivo. Nucleic Acids Research, 46(1), 203-214. https://doi.org/10.1093/nar/gkx1044

Jolma, A., Yin, Y., Nitta, K. R., Dave, K., Popov, A., Taipale, M., Enge, M., Kivioja, T., Morgunova, E., & Taipale, J. (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature, 527(7578), 384-388. https://doi.org/10.1038/nature15518

Junion, G., Spivakov, M., Girardot, C., Braun, M., Gustafson, E. H., Birney, E., & Furlong, E. E. (2012). A transcription factor collective defines cardiac cell fate and reflects lineage history. Cell, 148(3), 473-486. https://doi.org/10.1016/j.cell.2012.01.030

Kaikkonen, M. U., Spann, N. J., Heintz, S., Romanoski, C. E., Allison, K. A., Stender, J. D., Chun, H. B., Tough, D. F., Prinjha, R. K., Benner, C., & Glass, C. K. (2013). Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. Molecular Cell, 51(3), 310-325. https://doi.org/10.1016/j.molcel.2013.07.010

Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S. M., Habegger, L., Rozowsky, J., Shi, M., Urban, A. E., Hong, M. Y., Karczewski, K. J., Huber, W., Weissman, S. M., Gerstein, M. B., Korfel, J. O., & Snyder, M. (2010). Variation in transcription factor binding among humans. Science (New York, N.Y.), 328(5975), 232-235. https://doi.org/10.1126/science.1183621
Keane, T. M., Goodstadt, L., Daneczek, P., White, M. A., Wong, K., Yalcin, B., Heger, A., Agam, A., Slater, G., Goodson, M., Furlotte, N. A., Eskin, E., Nellåker, C., Whitley, H., Cleak, J., Janowitz, D., Hernandez-Pliego, P., Edwards, A., Belgard, T. G., Oliver, P. L., ... Adams, D. J. (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. Nature, 477(7364), 289-294. https://doi.org/10.1038/nature10413

Keilwagen, J., Posch, S., & Grau, J. (2019). Accurate prediction of cell type-specific transcription factor binding. Genome Biology, 20(1), 9. https://doi.org/10.1186/s13059-018-1614-y

Kilpinen, H., Waszak, S. M., Gschwind, A. R., Raghav, S. K., Witwicki, R. M., Orioli, A., Migliavacca, E., Wiederkehr, M., Gutierrez-Arcelus, M., Panousis, N. I., Yurovsky, A., Lappalainen, T., Romano-Palumbo, L., Planchnon, A., Bielser, D., Broyis, J., Padioleau, I., Udin, G., Thurnheer, S., Hacker, D., ... Dermitzakis, E. T. (2013). Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. Science (New York, N.Y.), 342(6159), 744-747. https://doi.org/10.1126/science.1242463

Kim, T. H., Abdullaev, Z. K., Smith, A. D., Ching, K. A., Loukinov, D. I., Green, R. D., Zhang, M. Q., Lobanenkov, V. V., & Ren, B. (2007). Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell, 128(6), 1231-1245. https://doi.org/10.1016/j.cell.2006.12.048

Kim, T. K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., Harmin, D. A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., Markenscoff-Papadimitriou, E., Kuhl, D., Bito, H., Worley, P. F., Kreiman, G., & Greenberg, M. E. (2010). Widespread transcription at neuronal activity-regulated enhancers. Nature, 465(7295), 182-187. https://doi.org/10.1038/nature09033

Kim, T. K., & Shiekhattar, R. (2015). Architectural and Functional Commonalities between Enhancers and Promoters. Cell, 162(5), 948-959. https://doi.org/10.1016/j.cell.2015.08.008

King, D. M., Hong, C., Shepherdson, J. L., Granas, D. M., Maricque, B. B., & Cohen, B. A. (2020). Synthetic and genomic regulatory elements reveal aspects of cis-regulatory grammar in mouse embryonic stem cells. eLife, 9, e41279. https://doi.org/10.7554/eLife.41279

Kharchenko, P. V., Tolstorukov, M. Y., & Park, P. J. (2008). Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nature Biotechnology, 26(12), 1351-1359. https://doi.org/10.1038/nbt.1508

Kribelbauer, J. F., Rastogi, C., Bussemaker, H. J., & Mann, R. S. (2019). Low-Affinity Binding Sites and the Transcription Factor Specificity Paradox in Eukaryotes. Annual Review of Cell and Developmental Biology, 35, 357-379. https://doi.org/10.1146/annurev-cellbio-100617-062719

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y. C., ... Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. Nature, 518(7539), 317-330. https://doi.org/10.1038/nature14248

Lappalainen T. (2015). Functional genomics bridges the gap between quantitative genetics and molecular biology. Genome Research, 25(10), 1427-1431. https://doi.org/10.1101/gr.190983.115

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., Cayting, P., Chen, Y., DeSalvo, G., Epstein, C., Fisher-Aylor, K. I., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A. J., Hoffman, M. M., Iyer, V. R., ... Snyder, M. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Research, 22(9), 1813-1831. https://doi.org/10.1101/gr.136184.111
Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods, 9(4), 357-359. https://doi.org/10.1038/nmeth.1923

Lareau, C. A., & Aryee, M. J. (2018). HiChiper: a preprocessing pipeline for calling DNA loops from HiChIP data. Nature Methods, 15(3), 155-156. https://doi.org/10.1038/nmeth.4583

Leonard, D. A., & Kerppola, T. K. (1998). DNA bending determines Fos-Jun heterodimer orientation. Nature Structural Biology, 5(10), 877-881. https://doi.org/10.1038/2316

Levo, M., & Segal, E. (2014). In pursuit of design principles of regulatory sequences. Nature Reviews. Genetics, 15(7), 453-468. https://doi.org/10.1038/nrg3684

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England), 25(16), 2078-2079. https://doi.org/10.1093/bioinformatics/btp352

Li, Q. V., Dixon, G., Verma, N., Rosen, B. P., Gordillo, M., Luo, R., Xu, C., Wang, Q., Soh, C. L., Yang, D., Crespo, M., Shukla, A., Xiang, Q., Dündar, F., Zumbo, P., Witkin, M., Koche, R., Betel, D., Chen, S., Massagué, J., … Huangfu, D. (2019). Genome-scale screens identify JNK-JUN signaling as a barrier for pluripotency exit and endoderm differentiation. Nature Genetics, 51(6), 999-1010. https://doi.org/10.1038/s41588-019-0408-9

Li, Y. I., van de Geijn, B., Raj, A., Knowles, D. A., Petti, A. A., Golan, D., Gilad, Y., & Pritchard, J. K. (2016). RNA splicing is a primary link between genetic variation and disease. Science (New York, N.Y.), 352(6285), 600-604. https://doi.org/10.1126/science.aad9417

Lidor Nili, E., Field, Y., Lubling, Y., Widom, J., Oren, M., & Segal, E. (2010). p53 binds preferentially to genomic regions with high DNA-encoded nucleosome occupancy. Genome Research, 20(10), 1361-1368. https://doi.org/10.1101/gr.103945.109

Liu T. (2014). Use model-based Analysis of ChIP-Seq (MACS) to analyze short reads generated by sequencing protein-DNA interactions in embryonic stem cells. Methods in Molecular Biology (Clifton, N.J.), 1150, 81-95. https://doi.org/10.1007/978-1-4939-0512-6_4

Liu, X., Li, H., Rajurkar, M., Li, Q., Cotton, J. L., Ou, J., Zhu, L. J., Goel, H. L., Mercurio, A. M., Park, J. S., Davis, R. J., & Mao, J. (2016). Tead and AP1 Coordinate Transcription and Motility. Cell Reports, 14(5), 1169-1180. https://doi.org/10.1016/j.celrep.2015.12.104

Liao, Y., Smyth, G. K., & Shi, W. (2013). The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Research, 41(10), e108. https://doi.org/10.1093/nar/gkt214

Link, V. M., Duttke, S. H., Chun, H. B., Holtman, I. R., Westin, E., Hoeksema, M. A., Abe, Y., Skola, D., Romanoski, C. E., Tao, J., Fonseca, G. J., Troutman, T. D., Spann, N. J., Strid, T., Sakai, M., Yu, M., Hu, R., Fang, R., Metzler, D., Ren, B., … Glass, C. K. (2018). Analysis of Genetically Diverse Macrophages Reveals Local and Domain-wide Mechanisms that Control Transcription Factor Binding and Function. Cell, 173(7), 1796-1809.e17. https://doi.org/10.1016/j.cell.2018.04.018

Long, H. K., Prescott, S. L., & Wysocka, J. (2016). Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. Cell, 167(5), 1170-1187. https://doi.org/10.1016/j.cell.2016.09.018
Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Malik, A. N., Vierbuchen, T., Hemberg, M., Rubin, A. A., Ling, E., Couch, C. H., Stroud, H., Spiegel, I., Farh, K. K., Harmin, D. A., & Greenberg, M. E. (2014). Genome-wide identification and characterization of functional neuronal activity-dependent enhancers. Nature Neuroscience, 17(10), 1330-1339. https://doi.org/10.1038/nn.3808

Maurano, M. T., Humert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A. K., Canfield, T. K., Giste, E., Diegel, M., Bates, D., ... Stamatoyannopoulos, J. A. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science (New York, N.Y.), 337(6099), 1190-1195. https://doi.org/10.1126/science.1222794

Maurano, M. T., Haugen, E., Sandstrom, R., Vierstra, J., Shafer, A., Kaul, R., & Stamatoyannopoulos, J. A. (2015). Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. Nature Genetics, 47(12), 1393-1401. https://doi.org/10.1038/ng.3432

Meuleman, W., Muratov, A., Rynes, E., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Teodosiadis, A., Reynolds, A., Haugen, E., Nelson, J., Johnson, A., Frerker, M., Buckley, M., Sandstrom, R., Vierstra, J., Kaul, R., & Stamatoyannopoulos, J. (2020). Index and biological spectrum of human DNase I hypersensitive sites. Nature, 584(7820), 244-251. https://doi.org/10.1038/s41586-020-2559-3

Michael, A. K., & Thomä, N. H. (2021). Reading the chromatinized genome. Cell, 184(14), 3599-3611. https://doi.org/10.1016/j.cell.2021.05.029

Miller, J. A., & Widom, J. (2003). Collaborative competition mechanism for gene activation in vivo. Molecular and Cellular Biology, 23(5), 1623-1632. https://doi.org/10.1128/MCB.23.5.1623-1632.2003

Mumbach, M. R., Satpathy, A. T., Boyle, E. A., Dai, C., Gowen, B. G., Cho, S. W., Nguyen, M. L., Rubin, A. J., Granja, J. M., Kazane, K. R., Wei, Y., Nguyen, T., Greenside, P. G., Corces, M. R., Tycko, J., Simeonov, D. R., Suliman, N., Li, R., Xu, J., Flynn, R. A., ... Chang, H. Y. (2017). Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. Nature Genetics, 49(11), 1602-1612. https://doi.org/10.1038/ng.3963

Nassar, J., Bergman, D. T., Fulco, C. P., Guckelberger, P., Doughty, B. R., Patwardhan, T. A., Jones, T. R., Nguyen, T. H., Ulirsch, J. C., Lekschas, F., Mualim, K., Natri, H. M., Weeks, E. M., Munson, G., Kane, M., Kang, H. Y., Cui, A., Ray, J. P., Eisenhaure, T. M., Collins, R. L., ... Engreitz, J. M. (2021). Genome-wide enhancer maps link risk variants to disease genes. Nature, 593(7858), 238-243. https://doi.org/10.1038/s41586-021-03446-x

Osterwalder, M., Barozzi, I., Tissières, V., Fukuda-Yuzawa, Y., Mannion, B. J., Afzal, S. Y., Lee, E. A., Zhu, Y., Plajzer-Frick, I., Pickle, C. S., Kato, M., Garvin, T. H., Pham, Q. T., Harrington, A. N., Akiyama, J. A., Afzal, V., Lopez-Rios, J., Dickel, D. E., Visel, A., & Pennacchio, L. A. (2018). Enhancer redundancy provides phenotypic robustness in mammalian development. Nature, 554(7691), 239-243. https://doi.org/10.1038/nature25461

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J., Hellinga, A. J., Lugo, C., Elliott, T. A., Ware, D., Peterson, T., Jiang, N., Hirsch, C. N., & Hufford, M. B. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biology, 20(1), 275. https://doi.org/10.1186/s13059-019-1905-y

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics (Oxford, England), 26(6), 841-842. https://doi.org/10.1093/bioinformatics/btq033
Paakinaho, V., Presman, D. M., Ball, D. A., Johnson, T. A., Schiltz, R. L., Levitt, P., Mazza, D., Morisaki, T., Karpova, T. S., & Hager, G. L. (2017). Single-molecule analysis of steroid receptor and cofactor action in living cells. Nature Communications, 8, 15896. https://doi.org/10.1038/ncomms15896

Pai, A. A., Pritchard, J. K., & Gilad, Y. (2015). The genetic and mechanistic basis for variation in gene regulation. PLoS Genetics, 11(1), e1004857. https://doi.org/10.1371/journal.pgen.1004857

Pickrell J. K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. American Journal of Human Genetics, 94(4), 559-573. https://doi.org/10.1016/j.ajhg.2014.03.004

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S. A., Flynn, R. A., & Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. Nature, 470(7333), 279-283. https://doi.org/10.1038/nature09692

Risse, G., Jooss, K., Neuberg, M., Brüller, H. J., & Müller, R. (1989). Asymmetrical recognition of the palindromic AP1 binding site (TRE) by Fos protein complexes. The EMBO Journal, 8(12), 3825-3832.

Roberts, G. A., Ozkan, B., Gachulincová, I., O’Dwyer, M. R., Hall-Ponsele, E., Saxena, M., Robinson, P. J., & Soufi, A. (2021). Dissecting OCT4 defines the role of nucleosome binding in pluripotency. Nature Cell Biology, 23(8), 834-845. https://doi.org/10.1038/s41556-021-00727-5

Rohs, R., Jin, X., West, S. M., Joshi, R., Honig, B., & Mann, R. S. (2010). Origins of specificity in protein-DNA recognition. Annual Review of Biochemistry, 79, 233-269. https://doi.org/10.1146/annurev-biochem-060408-091030

Schmidt, D., Schwalie, P. C., Wilson, M. D., Ballester, B., Gonçalves, A., Kutter, C., Brown, G. D., Marshall, A., Flicek, P., & Odom, D. T. (2012). Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. Cell, 148(1-2), 335-348. https://doi.org/10.1016/j.cell.2011.11.058

Seo, J., Koçak, D. D., Bartelt, L. C., Williams, C. A., Barrera, A., Gersbach, C. A., & Reddy, T. E. (2021). AP-1 subunits converge promiscuously at enhancers to potentiate transcription. Genome Research, 31(4), 538-550. https://doi.org/10.1101/gr.267898.120

Skene, P. J., & Henikoff, S. (2017). An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. eLife, 6, e21856. https://doi.org/10.7554/eLife.21856
Soufi, A., Garcia, M. F., Jaroszewicz, A., Osman, N., Pellegrini, M., & Zaret, K. S. (2015). Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. Cell, 161(3), 555-568. https://doi.org/10.1016/j.cell.2015.03.017

Spitz, F., & Furlong, E. E. (2012). Transcription factors: from enhancer binding to developmental control. Nature Reviews. Genetics, 13(9), 613-626. https://doi.org/10.1038/nrg3207

Sun W. (2012). A statistical framework for eQTL mapping using RNA-seq data. Biometrics, 68(1), 1-11. https://doi.org/10.1111/j.1541-0420.2011.01654.x

Tehranchi, A. K., Myrthil, M., Martin, T., Hie, B. L., Golan, D., & Fraser, H. B. (2016). Pooled ChIP-Seq Links Variation in Transcription Factor Binding to Complex Disease Risk. Cell, 165(3), 730-741. https://doi.org/10.1016/j.cell.2016.03.041

Tillo, D., Kaplan, N., Moore, I. K., Fondufe-Mittendorf, Y., Gossett, A. J., Field, Y., Lieb, J. D., Widom, J., Segal, E., & Hughes, T. R. (2010). High nucleosome occupancy is encoded at human regulatory sequences. PloS One, 5(2), e9129. https://doi.org/10.1371/journal.pone.0009129

van de Geijn, B., McVicker, G., Gilad, Y., & Pritchard, J. K. (2015). WASP: allele-specific software for robust molecular quantitative trait locus discovery. Nature Methods, 12(11), 1061-1063. https://doi.org/10.1038/nmeth.3582

van der Veeken, J., Zhong, Y., Sharma, R., Mazutis, L., Dao, P., Pe’er, D., Leslie, C. S., & Rudensky, A. Y. (2019). Natural Genetic Variation Reveals Key Features of Epigenetic and Transcriptional Memory in Virus-Specific CD8 T Cells. Immunity, 50(5), 1202-1217.e7. https://doi.org/10.1016/j.immuni.2019.03.031

Vierbuchen, T., Ling, E., Cowley, C. J., Couch, C. H., Wang, X., Harmin, D. A., Roberts, C., & Greenberg, M. E. (2017). AP-1 Transcription Factors and the BAF Complex Mediate Signal-Dependent Enhancer Selection. Molecular Cell, 68(6), 1067-1082.e12. https://doi.org/10.1016/j.molcel.2017.11.026

Vierstra, J., Lazar, J., Sandstrom, R., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Haugen, E., Rynes, E., Reynolds, A., Nelson, J., Johnson, A., Frerker, M., Buckley, M., Kaul, R., Meuleman, W., & Stamatoyannopoulos, J. A. (2020). Global reference mapping of human transcription factor footprints. Nature, 583(7818), 729-736. https://doi.org/10.1038/s41586-020-2528-x

Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., Lukk, M., Pignatelli, M., Park, T. J., Deaville, R., Erichsen, J. T., Jasinska, A. J., Turner, J. M., Bertelsen, M. F., Murchison, E. P., Flice, P., & Odom, D. T. (2015). Enhancer evolution across 20 mammalian species. Cell, 160(3), 554-566. https://doi.org/10.1016/j.cell.2015.01.006

Wittkopp, P. J., & Kalay, G. (2011). Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nature Reviews. Genetics, 13(1), 59-69. https://doi.org/10.1038/nrg3095

Wong, E. S., Schmitt, B. M., Kazachenka, A., Thybert, D., Redmond, A., Connor, F., Rayner, T. F., Feig, C., Ferguson-Smith, A. C., Marioni, J. C., Odom, D. T., & Flice, P. (2017). Interplay of cis and trans mechanisms driving transcription factor binding and gene expression evolution. Nature Communications, 8(1), 1092. https://doi.org/10.1038/s41467-017-01037-x

Yella, V. R., Bhimsaria, D., Ghoshdastidar, D., Rodríguez-Martínez, J. A., Ansari, A. Z., & Bansal, M. (2018). Flexibility and structure of flanking DNA impact transcription factor affinity for its core motif. Nucleic Acids Research, 46(22), 11883-11897. https://doi.org/10.1093/nar/gky1057
Zanconato, F., Forcato, M., Battilana, G., Azzolin, L., Quaranta, E., Bodega, B., Rosato, A., Bicciato, S., Cordenonsi, M., & Piccolo, S. (2015). Genome-wide association between YAP/TAZ/TEAD and AP-1 at enhancers drives oncogenic growth. Nature Cell Biology, 17(9), 1218-1227. https://doi.org/10.1038/ncb3216

Zhou, K., Gaullier, G., & Luger, K. (2019). Nucleosome structure and dynamics are coming of age. Nature Structural & Molecular Biology, 26(1), 3-13. https://doi.org/10.1038/s41594-018-0166-x

Zhu, F., Farnung, L., Kaasinen, E., Sahu, B., Yin, Y., Wei, B., Dodonova, S. O., Nitta, K. R., Morgunova, E., Taipale, M., Cramer, P., & Taipale, J. (2018). The interaction landscape between transcription factors and the nucleosome. Nature, 562(7725), 76-81. https://doi.org/10.1038/s41586-018-0549-5

Zhu, Q., Liu, N., Orkin, S. H., & Yuan, G. C. (2019). CUT&RUNTools: a flexible pipeline for CUT&RUN processing and footprint analysis. Genome Biology, 20(1), 192. https://doi.org/10.1186/s13059-019-1802-4
Figure 1. Identification of allele-specific TF binding and chromatin features in F1-hybrid fibroblasts
Figure 2. Frequency and distribution of sequence variants at allele-specific CREs
Figure 3. Characterization of TF-binding motifs and changes in chromatin state at allele-specific binding sites for AP-1, TEAD, and CTCF.
Figure 4. Impact of genetic variants on AP-1 and TEAD binding
Figure 5. Distribution of sequence variants that influence AP-1, TEAD, and CTCF binding
Figure 6. Examination of sequence features of AP-1 bound sites with machine learning and TF footprint data from diverse human cell types
Supplemental Figure 1. Identification of allele-specific transcripts using chromatin-associated RNA-seq
Supplemental Figure 2. Properties of allele-specific enhancers and association with gene expression
Supplemental Figure 3. Contribution of locus-scale, cis-acting mechanisms to enhancer activity
Supplemental Figure 4. Quantification of trans-acting effects on enhancer chromatin state and gene expression
Supplemental Figure 5. Sequence determinants of AP-1 binding
Supplemental Figure 6. Comparison of ChIP-seq and CUT&RUN datasets quantifying AP-1 binding
Supplemental Figure 7. Application of the gkm-svm algorithm to identify k-mers required for AP-1, TEAD, and CTCF binding