Systematic analysis of naturally occurring insertions and deletions that alter transcription factor spacing identifies tolerant and sensitive transcription factor pairs

Zeyang Shen[1,2], Rick Z. Li[1], Thomas A. Prohaska[3], Marten A. Hoeksema[1], Nathan J. Spann[1], Jenhan Tao[1], Gregory J. Fonseca[1,4], Thomas Le[5], Lindsey Stolze[6], Mashito Sakai[1,7], Casey E. Romanoski[6], Christopher K. Glass[1,3*]

[1]Department of Cellular and Molecular Medicine, School of Medicine, University of California San Diego, La Jolla, CA 92093, USA
[2]Department of Bioengineering, Jacobs School of Engineering, University of California San Diego, La Jolla, CA 92093, USA
[3]Department of Medicine, School of Medicine, University of California San Diego, La Jolla, CA 92093, USA
[4]Department of Medicine, McGill University, Montreal, Quebec, Canada
[5]Division of Biological Sciences, University of California San Diego, La Jolla, CA 92093, USA
[6]Department of Cellular and Molecular Medicine, College of Medicine, University of Arizona, Tucson, AZ 85721, USA
[7]Department of Biochemistry and Molecular Biology, Nippon Medical School, 1-1-5 Sendagi, Bunkyo-ku, Tokyo 113-8602, Japan
[*]Corresponding author. E-mail: ckg@ucsd.edu

Abstract

Regulation of gene expression requires the combinatorial binding of sequence-specific transcription factors (TFs) at promoters and enhancers. Prior studies showed that alterations in the spacing between TF binding sites can influence promoter and enhancer activity. However, the relative importance of TF spacing alterations resulting from naturally occurring insertions and deletions (InDels) has not been systematically analyzed. To address this question, we first characterized the genome-wide spacing relationships of 75 TFs in K562 cells as determined by ChIP-sequencing. We found a dominant pattern of a relaxed range of spacing between collaborative factors, including 46 TFs exclusively exhibiting relaxed spacing with their binding partners. Next, we exploited millions of InDels provided by genetically diverse mouse strains and human individuals to investigate the effects of altered spacing on TF binding and local

histone acetylation. Spacing alterations resulting from naturally occurring InDels are generally tolerated in comparison to genetic variants directly affecting TF binding sites. A remarkable range of tolerance was further established for PU.1 and C/EBPβ, which exhibit relaxed spacing, by introducing synthetic spacing alterations ranging from 5-bp increase to >30-bp decrease using CRISPR/Cas9 mutagenesis. These findings provide implications for understanding mechanisms underlying enhancer selection and for the interpretation of non-coding genetic variation.

Introduction

Genome-wide association studies (GWASs) have identified thousands of genetic variants associated with diseases and other traits (MacArthur et al., 2017; Visscher et al., 2017). Single nucleotide polymorphisms (SNPs) and short insertions and deletions (InDels) represent common forms of these variants. The majority of GWAS variants fall at non-protein-coding regions of the genome, implicating their effects on gene regulation (Farh et al., 2015; Ward & Kellis, 2012). Gene expression is regulated by transcription factors (TFs) in a cell-type-specific manner. TFs bind to short, degenerate sequences at promoters and enhancers, often referred to as TF binding motifs. Active promoters and enhancers are selected by combinations of sequence-specific TFs that bind in an inter-dependent manner to closely spaced motifs. SNPs and InDels can create or disrupt TF binding motifs and are a well-established mechanism for altering gene expression and biological function (Behera et al., 2018; Deplancke et al., 2016; Grossman et al., 2017; Heinz et al., 2013). InDels can additionally change spacing between motifs, but it remains unknown the extent to which altered spacing is relevant for interpreting genetic variation in human populations or between animal species.

Previous studies reported two major categories of motif spacing between inter-dependent TFs (Slattery et al., 2014). One category refers to the enhanceosome model (Slattery et al., 2014) that requires specific or "constrained" spacing. It is mainly provided by TFs that form ternary complexes recognizing composite binding motifs, exemplified by GATA, Ets and E-box transcription factors in mouse hematopoietic cells (Ng et al., 2014), MyoD and other cell-type-specific factors in muscle cells (Nandi et al., 2013), Sox2 and Oct4 in embryonic stem cells (Rodda et al., 2005). In vitro studies of the binding of pair-wise combinations of ~100 TFs to a diverse library of DNA sequences identified 315 out of 9400 possible interactive TF pairs that select composite elements with constrained positions of the respective recognition motifs (Jolma et al., 2015). Constrained spacing required for the optimal binding and function of interacting TFs can also occur between independent motifs, such as occurs at the interferon-β enhanceosome (Panne, 2008). In comparison to constrained spacing, another category of motif spacing allows TFs to interact over a relatively broad range (e.g., 100-200 bp), which we call "relaxed" spacing and is equivalent to the billboard model (Slattery et al., 2014). This type of spacing relationship is observed in collaborative or co-occupied TFs that do not target promoters or enhancers as a ternary complex (Heinz et al., 2010; Jiang & Singh, 2014; Sönmezer et al., 2021).

Substantial evidence indicates that the two categories of spacing requirement can experience a different level of impact from genetic variation. Reporter assays examining synthetic alterations of motif spacing revealed examples of TFs that require constrained spacing and have high sensitivity of transcription factor binding and gene expression on spacing (Farley et al., 2015; Ng et al., 2014; Panne, 2008). On the contrary, flexibility in motif spacing has been demonstrated using reporter assays in Drosophila (Menoret et al., 2013) and HepG2 cells (Smith et al., 2013).

80 However, these studies did not distinguish the impact of altered spacing on transcription factor
81 binding or subsequent recruitment of co-activators required for gene activation. Moreover, it
82 remains unknown the extent to which these findings are relevant to spacing alterations resulting
83 from naturally occurring genetic variation.
84
85 To investigate the effects of altered spacing on TF binding and function, we first characterized
86 the genome-wide binding patterns of seventy-five TFs based on their binding sites determined by
87 chromatin immuno-precipitation sequencing (ChIP-seq). We developed a computational
88 framework that assigned each spacing relationship to "constrained" or "relaxed" category and
89 associated spacings to the naturally occurring InDels observed in human populations to study the
90 selective constraints of different spacing relationships. As specific case studies, we leveraged
91 natural genetic variation in numerous human samples and from five strains of mice to study the
92 effect size of spacing alterations on TF binding activity and local histone acetylation. We find
93 that InDels altering spacing are generally less constrained and well tolerated when they occur
94 between TF pairs with relaxed spacing relationships. Finally, we established remarkable
95 tolerance in spacing for macrophage lineage determining TFs (LDTFs), PU.1 and C/EBPβ, by
96 introducing a wide range of InDels between their respective binding sites at representative
97 endogenous genomic loci using CRISPR/Cas9 mutagenesis in mouse macrophages.
98

99 ## Results

100 ### Transcription factors primarily co-bind with relaxed spacing

101 We characterized spacing relationships for 75 TFs of K562 cells covering diverse TF families
102 (Hu et al., 2019) based on the ChIP-seq data from ENCODE data portal (Davis et al., 2018).

After obtaining reproducible TF binding sites, we first used the corresponding position weight matrix (PWM) of each TF (Figure 1-table supplement 1) to scan through the sequence of every binding site and identified the locations of high-affinity motifs that are less than 50 bp from ChIP-seq peak centers (Fig. 1A; Figure 1-figure supplement 1). The binding sites of every pair of TFs were then merged to compute the edge-to-edge motif spacing at all the co-binding sites. Motif spacings were eventually aggregated to show a distribution within +/- 100 bp. To categorize spacing relationships, we used permutation tests on the gradients to test for specific spacing constraints and used Kolmogorov–Smirnov test (KS test) to test for a relaxed spacing relationship against random distribution.

We applied this computational framework to all possible pairs of TFs. By dissecting each TF's binding sites based on their spacing relationships with co-binding TFs, we found that 46 of the 75 TFs examined exclusively exhibited relaxed spacing relationships with other TFs (Fig. 1B). 26 factors could participate in either relaxed or constrained interactions, depending on the specific co-binding TFs. Only 3 TFs interacted with only constrained spacing, some of which might show additional relaxed spacing relationships by expanding the current set of TFs. The significant pairwise patterns of relaxed and constrained spacing relationships are illustrated in Figure 1C. Among 32 TF pairs with constrained spacing relationships, most bind closely to each other within 15 bp spacing (Figure 1-figure supplement 2; Figure 1-table supplement 2). Some of these TF pairs have been reported to recognize composite motifs such as GATA1-TAL1 and NFATC3-FOSL1 (Macián et al., 2001; Ng et al., 2014) (Fig. 1D; Figure 1-figure supplement 3), and some are novel constrained spacing patterns discovered by our analysis such as MEF2A-JUND and CEBPB-TEAD4 (Figure 1-figure supplement 3). There are also TF pairs, exemplified

126 by EGR1 and JUND, that bind relatively further away from each other but still require
127 constrained spacing (Fig. 1D). Previous studies demonstrated interactions between EGR1 and
128 AP-1 factors (Levkovitz & Baraban, 2002; Nakashima et al., 2003), but the underlying
129 mechanism for such constrained spacing at 29 bp needs to be further investigated. TFs exhibiting
130 relaxed spacing are exemplified by ETV1-TAL1 and JUND-KLF16, in which the frequency of
131 co-binding progressively declines with distance from the center of the reference TF (Fig. 1D). In
132 addition, the same type of spacing relationship is usually observed in different motif orientations
133 (Fig. 1D), consistent with previous findings (Lis & Walther, 2016). The same TF pairs can have
134 similar spacing relationships in different cell types, exemplified by CEBPB and JUND in K562
135 and HepG2 cells (Figure 1-figure supplement 4).

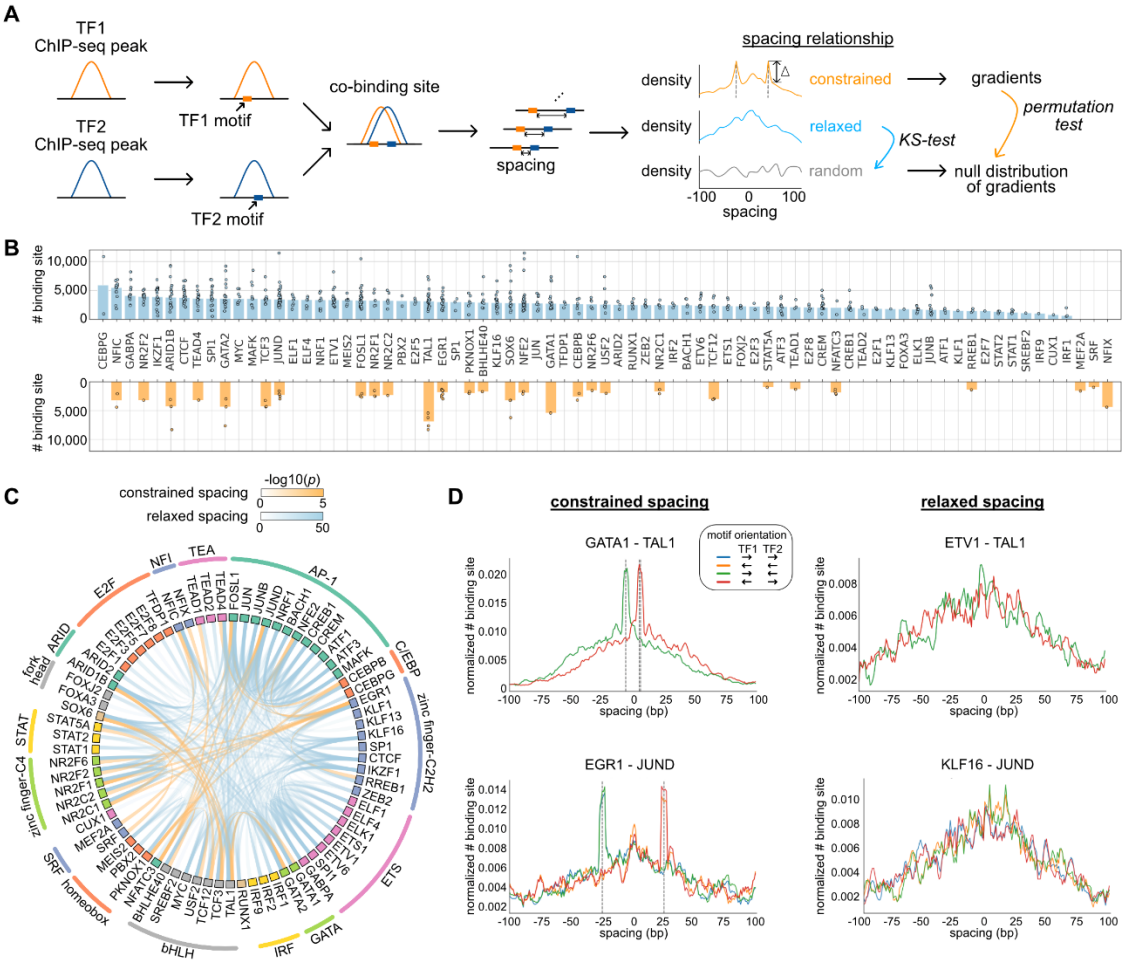


Figure 1. Characterization of spacing relationships for transcription factor pairs. (**A**) Schematic of data analysis pipeline for characterizing the spacing relationships based on TF ChIP-seq data. (**B**) Dissection of TF binding sites for TFs in K562 cells based on spacing relationships with co-binding TFs. Each dot represents the co-binding peak number of the corresponding TF and one of its co-binding TFs. Bar heights indicate means among the co-binding TFs. (**C**) Circos plot summarizing spacing relationships for all the TF pairs analyzed. Orange and blue bands represent significant constrained and relaxed spacing relationships, respectively. Color opacity indicates the level of significance. TFs are grouped and colored by TF family. (**D**) The spacing distributions of example TF pairs with constrained spacing or relaxed spacing relationships. Dashed lines indicate the significant constrained spacings. Since TAL1 motif is completely palindromic, the motif orientation is only differentiated by its co-binding partners.

Natural genetic variants altering spacing between relaxed transcription factors are associated with less deleteriousness in human populations

Based on a global view of the TF spacing relationships, we then studied whether these relationships associate with different levels of sensitivity to spacing alterations. Here, we leveraged more than 60 million InDels from gnomAD data (Karczewski et al., 2020), which were based on more than 75,000 genomes from unrelated individuals. We overlaid these InDels at motifs, between motifs, or within background regions of representative TF pairs of constrained and relaxed spacing relationships. We found that InDels at different regions have relatively similar distributions of InDel sizes with the majority being less than 5 bp (Fig. 2A). Next, we divided these InDels based on the allele frequency (AF) and the allele count (AC) into high-frequency variants (AF > 0.01%), rare variants (AF < 0.01%, AC > 1), and singletons (AC = 1). Most of the InDels at TF binding sites are singletons or rare variants (Figure 2-figure supplement 1). We compared the enrichment of these categories of InDels between TFs with different spacing relationships (Fig. 2B). The InDel compositions at motifs were not significantly different between constrained and relaxed spacing groups. On the contrary, singletons were significantly more enriched between motifs with constrained spacing, whereas high-frequency variants were significantly more depleted between these motifs. We also computed for several TF pairs with random spacing relationships as negative controls and found similar enrichments of InDels like those with relaxed spacing. Since common variants are associated with less deleteriousness and rare variants with more deleteriousness (Lek et al., 2016), our data suggest that InDels between motifs of TFs with constrained spacing could be just as damaging as those at motifs whereas InDels between motifs of TFs with relaxed spacing might have a much weaker effect.

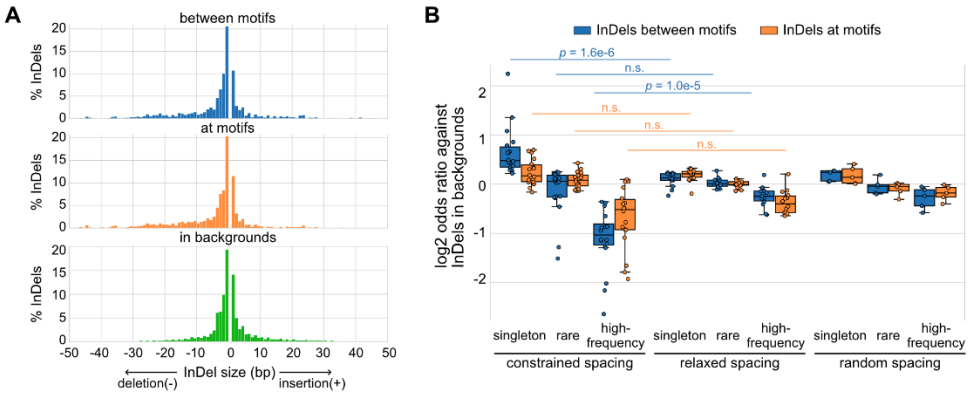
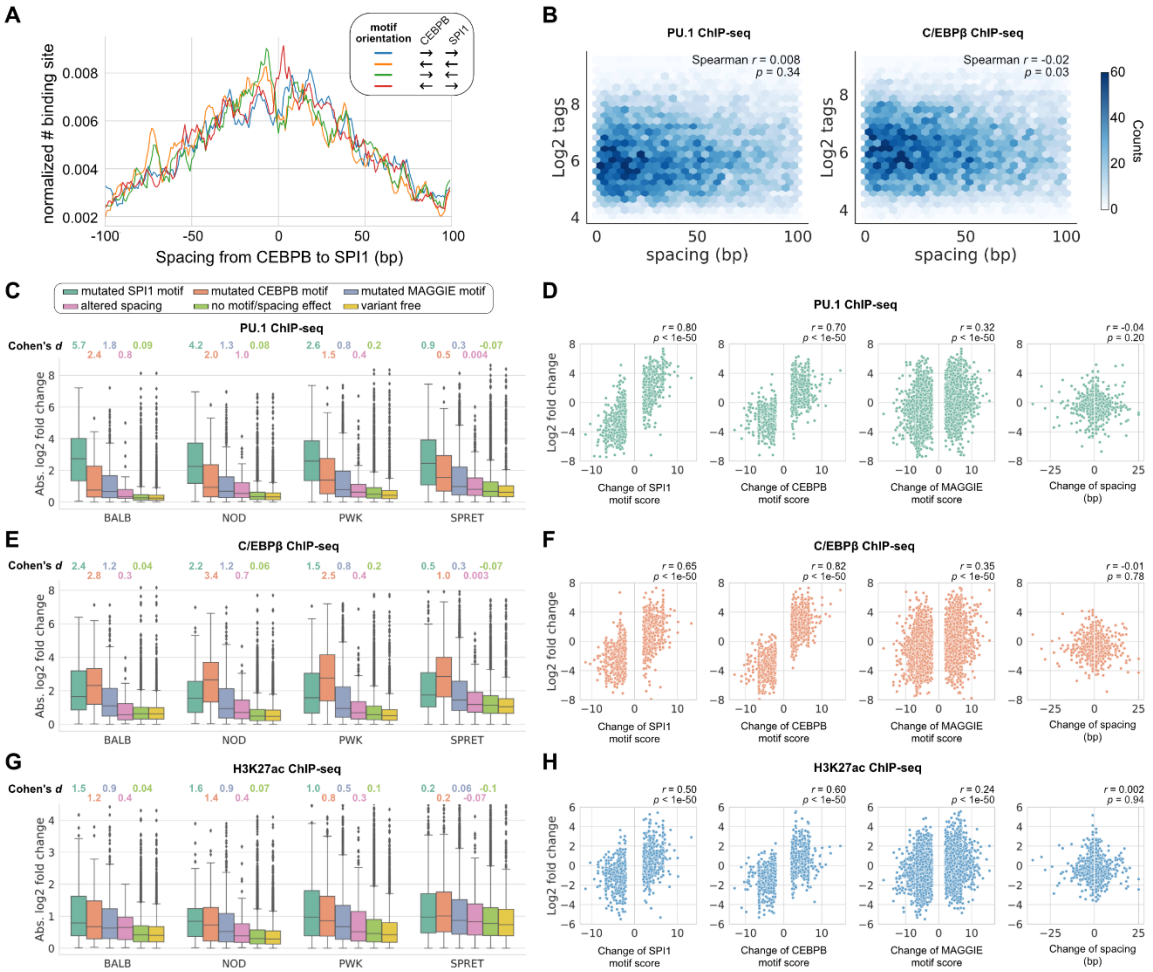


Figure 2. Naturally occurring InDels in human populations. (**A**) Size distributions of human InDels within different regions. (**B**) Log2 odds ratios for different categories of InDels. Each dot represents a TF pair with corresponding spacing relationship. Mann-Whitney U test was used to compare the odds ratios between different spacing relationships.

Spacing alterations across mouse strains are generally tolerated by relaxed transcription factor binding and promoter and enhancer function

To investigate the regulatory effects of naturally occurring InDels that alter motif spacing, we leveraged more than 50 million SNPs and 5 million InDels from five genetically diverse mouse strains and the ChIP-seq data of key TFs and histone acetylation for the bone marrow-derived macrophages (BMDMs) from every mouse strain (Link, Duttke, et al., 2018). The five mouse strains include C57BL/6J (C57), BALB/cJ (BALB), NOD/ShiLtJ (NOD), PWK/PhJ (PWK), and SPRET/EiJ (SPRET).We first characterized the spacing relationship between the macrophage lineage-determining TFs (LDTFs) PU.1 and C/EBPβ, which have been found to bind in a collaborative manner at regulatory regions of macrophage-specific genes (Heinz et al., 2010). Based on our computational framework for characterizing spacing relationships (Fig. 1A), these two TFs follow a relaxed spacing relationship independent of their motif orientations (Fig. 3A; KS p-value < 1e-6). Moreover, both PU.1 and C/EBPβ binding activities quantified by the ChIP-

188 seq tags were not correlated with motif spacing, suggesting no direct association between
189 spacing and TF binding (Fig. 3B).



190

191 **Figure 3. Effects of spacing alterations resulting from natural genetic variation across mouse strains.** (**A**)
192 Spacing distributions of PU.1 and C/EBPβ motif at co-binding sites. (**B**) Density plots showing the relationship
193 between TF binding activity and motif spacing for the co-binding sites. Log2 ChIP-seq tags were calculated within
194 300 bp to quantify the binding activity of PU.1 and C/EBPβ. The color gradients represent the number of sites.
195 Spearman correlation coefficients together with p-values are displayed. (**C, E, G**) Absolute log2 fold changes of
196 ChIP-seq tags between C57 and another strain for (**C**) PU.1 binding, (**E**) C/EBPβ binding, or (**G**) H3K27ac level.
197 Boxplots show the median and quartiles of every distribution. Cohen's d effect sizes comparing against variant-free
198 regions are displayed on top. (**D, F, H**) Correlations between change of motif spacing or motif score and change of
199 (**D**) PU.1 binding, (**F**) C/EBPβ binding, or (**H**) H3K27ac level. Spearman correlation coefficients together with p-
200 values are displayed.

We then conducted independent comparisons between C57 and one of the other four strains to investigate the effects of spacing alterations caused by natural genetic variation, which are mostly less than 5 bp like natural InDels in human populations (Figure 3-figure supplement 1). We first identified the co-binding sites of PU.1 and C/EBPβ for each strain and then, for each pairwise analysis, pooled the co-binding sites of C57 and the compared strain to obtain the testing set of regions. Based on the impacts of genetic variants on motif affinity and motif spacing, we categorized the testing regions into the following non-overlapping groups: 1) mutated PU.1 (i.e., SPI1) motif, 2) mutated C/EBPβ (i.e., CEBPB) motif, 3) mutated other functional motifs (i.e., MAGGIE motif), 4) altered spacing, 5) no motif affinity/spacing effect, and 6) variant free. Functional motifs were identified from PU.1 and C/EBPβ binding sites separately using MAGGIE (Shen et al., 2020), which is a computational tool that can identify motifs whose affinity changes are associated with TF binding changes (Figure 3-figure supplement 2). The effect of genetic variation was quantified by the log2 fold difference of ChIP-seq tag counts between strains at orthogonal sites (Fig. 3C). All the four independent comparisons showed that PU.1 binding is most strongly affected by PU.1 motif mutation, followed by C/EBPβ motif mutation and other functional motif mutation. Spacing alterations have a smaller effect size than any of these motif mutations, but still a relatively larger effect than variants affecting neither motif affinity nor spacing. Despite the moderate effect size of spacing alterations, we found such effect was independent of the size or direction of InDels (Fig. 3D). On the contrary, changes of PU.1 ChIP-seq tags are strongly correlated with changes of motif affinity (Fig. 3D). In addition, the effects of motif mutation and spacing alteration are not varied by the initial spacing between PU.1 and C/EBPβ motifs (Figure 3-figure supplement 3).

Similar findings were observed in C/EBPβ binding, except that expectedly C/EBPβ motif mutation had the largest effect size and the strongest correlation with C/EBPβ binding activity (Fig. 3E, F; Figure 3-figure supplement 3).

To investigate whether the effects of altered spacing on PU.1 and C/EBPβ binding can be generalized to hierarchical interactions with signal-dependent transcription factors (SDTFs), we leveraged the ChIP-seq data of PU.1, the NFκB subunit p65, and an AP-1 factor cJun for BMDMs treated with the TLR4-specific ligand Kdo2 lipid A (KLA) in the same five strains of mice (Link, Duttke, et al., 2018). Upon macrophage activation with KLA, p65 enters the nucleus and primarily binds to poised enhancer elements that are selected by LDTFs including PU.1 and AP-1 factors (Heinz et al., 2015). We observed a relaxed spacing relationship between PU.1 and p65 and between cJun and p65 (Figure 3-figure supplement 4). In addition, InDels altering motif spacing had a much smaller effect size on TF binding than motif mutations (Figure 3-figure supplement 5), consistent with our findings from PU.1 and C/EBPβ.

Although alterations in motif spacing had generally weak effects at the level of DNA binding, it remained possible that changes in motif spacing could influence subsequent steps in enhancer and promoter activation. To examine this, we extended our analysis to local acetylation of histone H3 lysine 27 (H3K27ac), which is a histone modification that is highly correlated with enhancer and promoter function (Creyghton et al., 2010). We leveraged the H3K27ac ChIP-seq data of untreated BMDMs in the five strains of mice (Link, Duttke, et al., 2018) and calculated the log fold changes of H3K27ac level within the extended 1,000-bp regions of the PU.1 and C/EBPβ co-binding sites. Like for TF binding, altered spacing demonstrated weaker effects on

histone acetylation than motif mutations (Fig. 3G; Figure 3-figure supplement 3), which is supported by the high consistency between change of TF binding and change of histone acetylation (Figure 3-figure supplement 6). The relative tolerance of spacing alteration was further reflected by a weak correlation between the change of acetylation level and the size of InDels, in comparison to a much stronger correlation with changes in motif affinity (Fig. 3H).

Human quantitative trait loci altering spacing between relaxed transcription factors have small effect sizes

To study the effects of spacing alteration on transcription factor binding and local histone acetylation in human cells, we leveraged the ChIP-seq data of ERG, p65, and H3K27ac in endothelial cells from dozens of individuals (Stolze et al., 2020). ERG is an ETS factor that functions as an LDTF in endothelial cells that selects poised enhancers where p65 binds in a hierarchical manner upon interleukin-1β (IL-1β) stimulation (Hogan et al., 2017). ERG and p65 follow a relaxed spacing relationship according to our method (Fig. 4A). Next, we obtained 557 TF binding quantitative trait loci (bQTLs) for ERG, 5,791 bQTLs for p65, 25,621 histone modification QTLs (hQTLs) for H3K27ac in untreated cells, and 21,635 hQTLs for H3K27ac in IL-1β-treated cells (Stolze et al., 2020). We further classified bQTLs and hQTLs based on their impacts on motif affinity and spacing: 1) mutated both ERG and p65 (i.e., RELA) motif, 2) mutated ERG motif only, 3) mutated p65 motif only, 4) mutated other functional motifs identified by MAGGIE (Shen et al., 2020), 5) altered spacing between ERG and p65 motif, 6) none of the above. To find functional motifs, we fed MAGGIE with 100-bp sequences around QTLs before and after swapping alleles at the center (Figure 4-figure supplement 1). As a result, only a small portion of bQTLs and hQTLs directly mutates an ERG or RELA motif (Fig. 4B;

Figure 4-figure supplement 2). However, such motif mutations are enriched in bQTLs compared to non-QTLs (Fisher's exact $p < 1e-4$). On the contrary, InDels that alter motif spacing are significantly depleted in p65 bQTLs (Fisher's exact $p = 1.3e-15$). These InDels from the dozens of individuals are predominantly shorter than 5 bp by following a similar size distribution of those in human populations (Figure 4-figure supplement 3). A large proportion of QTLs affect other functional motifs, implicating the complexity of TF interactions. More than a quarter of the QTLs affect neither motif affinity nor motif spacing, which can be explained by the high correlation of non-functional variants with functional variants due to linkage disequilibrium.

We further compared the effect sizes of different categories of QTLs. Despite being the minority among QTLs, variants that mutate both ERG and RELA motifs have the strongest effects on both p65 binding and histone acetylation in IL-1β-treated endothelial cells (Fig. 4C). In comparison, ERG binding and the basal level of histone acetylation are significantly affected by ERG motif mutations in untreated endothelial cells and not by p65 motif mutations, consistent with the hierarchical interaction of p65 only upon IL-1β stimulation (Figure 4-figure supplement 4). In both conditions of endothelial cells, spacing alterations have the smaller effect size than motif mutation categories and are not significantly different from likely non-functional variants in the "other" group. The examples showed a variant being both a p65 bQTL and a H3K27ac hQTL under the IL-1β state due to its impact on an ERG motif, and a 4-bp insertion between ERG and p65 motifs associated with no change in p65 binding or H3K27ac (Fig. 4D).

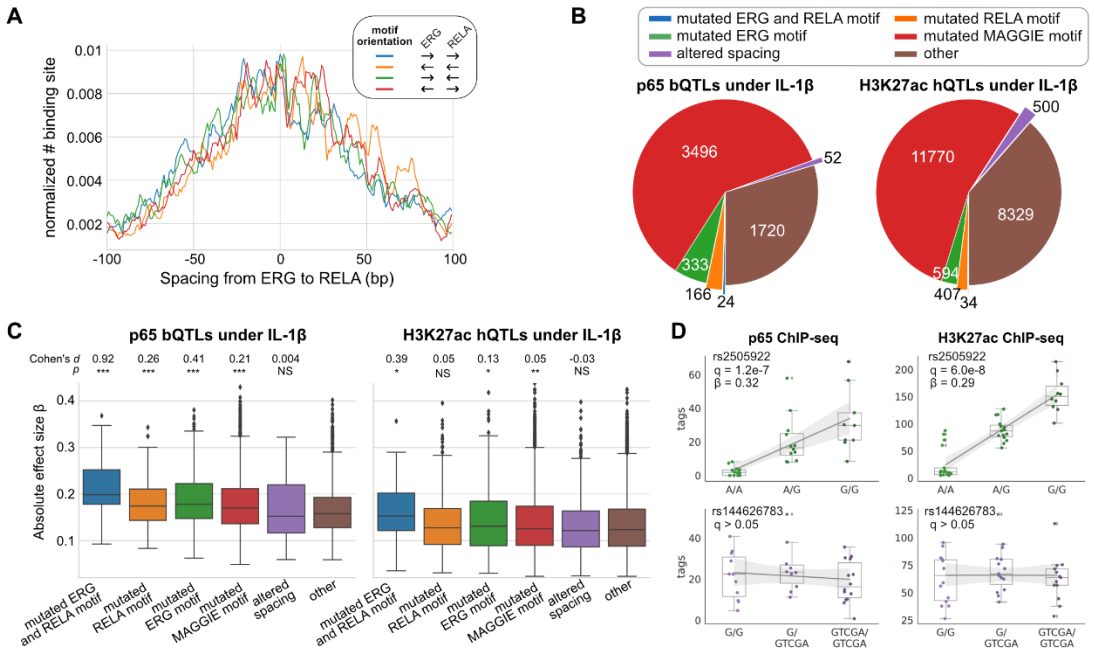
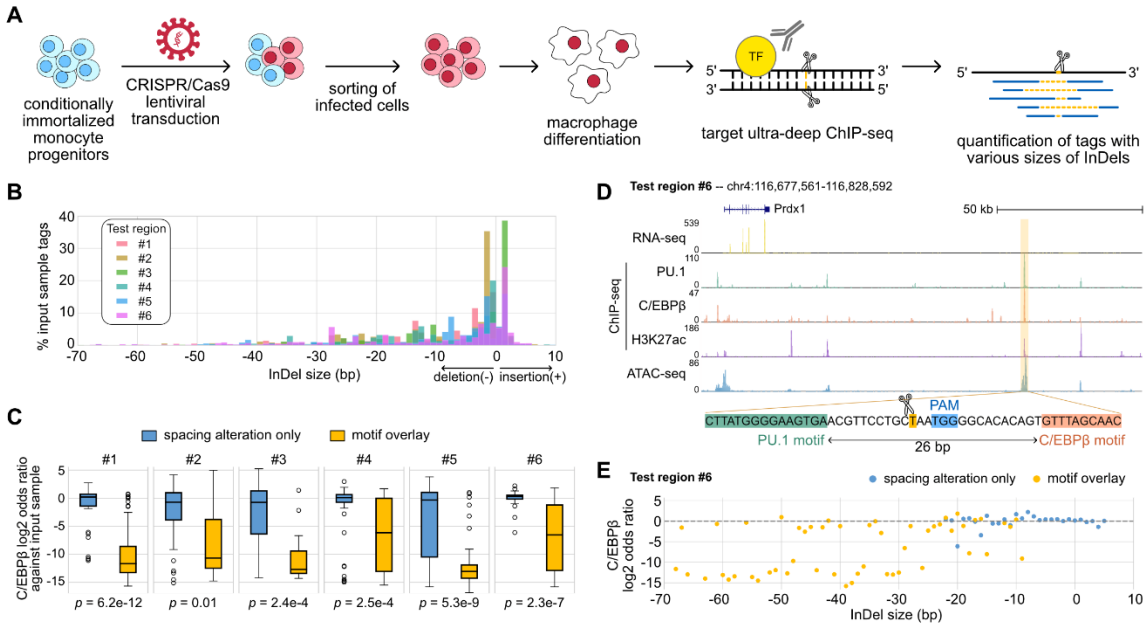


Figure 4. Effects of chromatin QTLs in human endothelial cells. (**A**) Spacing distributions of ERG and RELA motif at co-binding sites. (**B**) Classification of chromatin QTLs based on the impacts on motif and spacing. (**C**) Absolute correlation coefficients of different QTLs. Cohen's d and Mann-Whitney U test p-values comparing against the "other" group are displayed on top. *p<0.01, **p<0.001, ***p<0.0001. (**D**) Example QTLs for large effect size due to ERG motif mutation (upper) and trivial effect due to spacing alteration (lower).

Relaxed transcription factor binding is highly tolerant to synthetic spacing alterations

The generally small effects of InDels occurring between TF pairs exhibiting relaxed spacing relationships raised the question of the robustness and the extent of such tolerance at genomic locations lacking such variation. We addressed this question by using CRISPR/Cas9 editing to introduce synthetic InDels between binding sites observed for the LDTFs PU.1 and C/EBPβ in mouse macrophages (Fig. 5A). We used lentiviral transduction in Cas9-expressing ER-HoxB8 cells, which are conditionally immortalized monocyte progenitors, to introduce gRNAs targeting genomic sequences between the locations of PU.1 and C/EBPβ co-binding. The Cas9 nuclease activity at these sites resulted in non-homologous DNA repair that generated various sizes of

306 InDels in the populations of transduced cells. After sorting the successfully transduced ER-
307 HoxB8 cells and differentiating them into macrophages, we performed ChIP for C/EBPβ and
308 deeply sequenced amplicons of the target regions of Cas9 cleavage. Lastly, the reads were
309 mapped to the target regions by allowing various sizes of gaps at the cut sites and were
310 quantified by comparing to the input DNA samples.



311

312 **Figure 5. Effects of variable sizes of synthetic spacing alterations.** (**A**) Schematic for generating and analyzing
313 synthetic spacing alterations. (**B**) The distributions of valid read counts from the input sample based on the InDel
314 sizes of the reads. Negative InDel size indicates deletion, and positive size means insertion. (**C**) Log2 odds ratios by
315 comparing C/EBPβ ChIP-seq reads and input sample ChIP-seq reads. Y=0 indicates where TF binding has an
316 expected amount of activity. P-values were based on two-sample t-tests by comparing the InDel groups of each test
317 region. (**D**) Sequencing data of ER-HoxB8 cells at co-binding site of PU.1 and C/EBPβ. Highlighted is test region
318 #6 whose DNA sequence from PU.1 motif to C/EBPβ motif is shown. (**E**) Log2 odds ratios of test regions #6 as a
319 function of InDel size.

320

321 We tested six PU.1 and C/EBPβ co-binding sites with the original motif spacing ranging from 26
322 to 55 bp (Figure 5-table supplement 1) and quantified the effects of various InDels on C/EBPβ

323 binding. Among the six test regions, three of them have supportive evidence from naturally
324 occurring InDels of mouse strains (region #1, #3, #5) and the other three don't (region #2, #4,
325 #6). Based on the bioinformatic analysis of the ultra-deep sequencing reads from the input DNA
326 samples, we saw that the CRISPR/Cas9 system generated a wide range of InDels with most
327 deletions being < 30 bp and short insertions usually less than 5 bp (Fig. 5B). It provides longer
328 deletions than natural genetic variations found across mouse strains (Figure 3-figure supplement
329 1) and in human populations (Figure 2A). After classifying ChIP-seq reads based on the InDel
330 size and whether the InDel overlaps with any of the PU.1 and C/EBP motifs, we estimated the
331 effect size of InDels on C/EBPβ binding by calculating the odds ratio between C/EBPβ ChIP-seq
332 reads and input sample ChIP-seq reads for every InDel group. We found that InDels altering
333 spacing have significantly weaker effects on C/EBPβ binding in comparison to those overlapping
334 with at least one of the motifs (Fig. 5C). For some test regions, the effects of pure spacing
335 alterations are almost negligible, exemplified by test region #6 (Fig. 5D, E) and test region #1
336 (Figure 5-figure supplement 1). Test region #6 is located near a highly expressed gene *Prdx1* and
337 has strong binding of PU.1 and C/EBPβ binding and strong signals of H3K27ac and chromatin
338 accessibility indicated by ATAC-seq in ER-Hoxb8 cells, which all support its potential
339 regulatory function (Fig. 5D). The PU.1 and C/EBPβ motifs at this region are 26 bp apart. In
340 general, spacing alterations ranging from 5-bp increase to 22-bp decrease did not have a strong
341 effect on TF binding, indicated by a log2 odds ratio close to 0 (Fig. 5E). A small number of
342 outliers were observed at each region where specific InDels resulted in substantial loss of
343 binding (e.g., -20 bp, Fig. 5E). C/EBPβ binding at these specific InDels was generally
344 discontinuous with 1-bp increments (e.g., -19 bp and -21 bp, Fig. 5E). The basis for these highly
345 localized changes in the odds ratio in a small fraction of InDels that alter spacing is unclear. On

346 the contrary, deletions overlapping with motifs resulted in a general decrease in TF binding
347 activity. Similar results were found at test region #1 where PU.1 and C/EBPβ motifs are 41 bp
348 apart (Figure 5-figure supplement 1A). This *Ly9* enhancer also has a 5-bp insertion between
349 PU.1 and C/EBPβ motifs in BALB, NOD, and PWK mice, and shows unaffected binding of
350 PU.1 and C/EBPβ in the BMDMs of these strains (Figure 5-figure supplement 1B). As a result of
351 the synthetic InDels, the C/EBPβ binding activity was generally unaffected by spacing
352 alterations only whereas deletions affecting motifs substantially diminished TF binding (Figure
353 5-figure supplement 1C).
354
355

356 **Discussion**

357 By classifying the genome-wide spacing relationships of 75 co-binding TFs as "constrained" or
358 "relaxed", we revealed that relaxed spacing relationships were the dominant pattern of
359 interaction for majority of these factors. Among these factors, approximately half could also
360 participate in constrained spacing relationships with specific TF partners. We confirmed TF pairs
361 known to exhibit constrained relationships (e.g., GATA1-TAL1) and identified previously
362 unreported constrained relationships for additional pairs, including EGR1 and JUND. Overall,
363 this finding of a subset of constrained TF interactions on a genome wide level is consistent with
364 the locus-specific examples provided by functional and structural studies of the interferon-β
365 enhanceosome (Panne, 2008) and in vivo studies of synthetically modified enhancer elements in
366 Ciona (Farley et al., 2015). Each of these examples represents genomic regulatory elements in
367 which key TF motifs are tightly spaced in their native contexts (i.e., 0-9 bp between motifs).
368 Direct protein-protein interactions are observed between bound TFs at the interferon-β

enhanceosome, analogous to interactions defined for cooperative TFs that form ternary complexes (Morgunova & Taipale, 2017; Reményi et al., 2003). In the present studies, InDels between TF pairs exhibiting constrained spacings were under large selective constraints that were comparable to InDels at motifs, suggesting a deleterious effect of these spacing alterations on TF binding. However, the spacing analyses in this study did not directly consider the possible overlap or lack of spacing between TF binding sequences. Thus, we are not able to clearly distinguish effects of spacing alterations from effects of InDels on motifs at sites of tightly spaced composite motifs.

The observation that most TF pairs exhibited relaxed spacing relationships has intriguing implications for the mechanisms by which functional enhancers and promoters are selected from chromatinized DNA. In contrast to ternary complexes of TFs that cooperatively bind to composite elements as a unit, relaxed spacing relationships appear to not require specific protein-protein interactions between TFs for collaborative binding at most genomic locations. Although pioneering TFs necessary for selection of cell-specific enhancers have been reported to recognize their motifs within the context of nucleosomal DNA (Zaret & Carroll, 2011), the basis for collaborative binding interactions between TFs with relaxed spacings remains poorly understood.

While the current studies relying on natural genetic variation and mutagenesis experiments concluded clear tolerance of spacing alterations between motifs of TFs with relaxed spacings, the extent to which this set of binding sites is representative of all regulatory elements is unclear. For example, we observed outliers in which significant differences in TF binding between mouse strains were associated with InDels occurring between motifs. However, the proportion of

392 outliers was generally similar to that observed at genomic regions lacking such InDels, and such
393 strain differences may be driven by distal effects of genetic variation on interacting enhancer or
394 promoter regions (Hoeksema et al., 2021; Link, Duttke, et al., 2018). The remarkable tolerance
395 of synthetic InDels at two independent endogenous genomic locations between PU.1 and
396 C/EBPβ binding sites strongly support the generality of relaxed binding interactions for these
397 two proteins. Intriguingly, while the densities of C/EBP motifs increase with decreasing distance
398 to PU.1 motifs over a 100 bp range (Fig. 3A), deletions from 1 to >30 bp between PU.1-C/EBPβ
399 pairs did not result in improved binding. Instead, relatively constant binding was observed with
400 progressive deletions bringing two motifs close together until the deletions started to cause
401 mutations in one or both motifs. This is consistent with the lack of correlation between DNA
402 binding strengths and distances between these factors (Figure 3B). A limitation of these studies
403 is that few and relatively short insertions were obtained, preventing conclusions as to the extent
404 to which increases in spacing are tolerated.
405
406 In concert, the present studies provide a basis for estimation of the potential phenotypic
407 consequences of naturally occurring InDels in non-coding regions of the genome. The majority
408 of naturally occurring InDels are less than 5 bp in length. In nearly all cases, InDels of this size
409 range between motifs for TFs that have relaxed binding relationships are unlikely to alter TF
410 binding and function, and InDels of much greater length are frequently tolerated. In contrast,
411 InDels between motifs for TFs that have constrained binding relationships have the potential to
412 result in biological consequences. Application of these findings to the interpretation of non-
413 coding InDels that are associated with disease risk will require knowledge of the relevant cell

type in which the InDel exerts its phenotypic effect and the types of TF interactions driving the selection and function of the affected regulatory elements.

Methods

Sequencing data processing

We downloaded two replicates for each TF ChIP-seq data from ENCODE data portal (Davis et al., 2018). The mouse BMDM data and the human endothelial cell data were downloaded from the GEO database with accession number GSE109965 (Link, Duttke, et al., 2018) and GSE139377 (Stolze et al., 2020), respectively. We mapped ChIP-seq and ATAC-seq reads using Bowtie2 v2.3.5.1 with default parameters (Langmead & Salzberg, 2012) and mapped RNA-seq reads using STAR v2.5.3 (Dobin et al., 2013). All the human data downloaded from ENCODE were mapped to the hg38 genome. Data from C57BL/6J mice were mapped to the mm10 genome. Data from other mouse strains and endothelial cell data from different individuals were mapped to their respective genomes built by MMARGE v1.0 (Link, Romanoski, et al., 2018). More details are described below.

Based on the mapped ChIP-seq data, we called TF binding sites or peaks using HOMER v4.9.1 (Heinz et al., 2010). For data with replicates including ENCODE data and mouse data, we first called unfiltered 200-bp peaks using HOMER "findPeaks" function using parameters "-style factor -L 0 -C 0 -fdr 0.9 -size 200" and then ran IDR v2.0.3 with default parameters (Li et al., 2011) to obtain reproducible peaks. For data without replicates including human endothelial cell data and ER-HoxB8 ChIP-seq data, we called peaks using HOMER "findPeaks" with the default setting and parameters "-style factor -size 200".

Activity of TF binding was quantified by the ChIP-seq tag counts within 300-bp around peak centers and normalized by library size using HOMER "annotatePeaks.pl" script with parameters "-norm 1e7 -size -150,150". Activity of promoter and enhancer was quantified by normalized H3K27ac ChIP-seq tags within 1,000-bp regions around TF peak centers using parameters "-norm 1e7 -size -500,500".

Motif identification

Based on DNA sequences of the TF binding sites, we calculated motif scores by the dot products between PWMs and sequence vectors using Biopython package (Cock et al., 2009). The PWMs were obtained from either the JASPAR database (Fornes et al., 2020) or de novo motif analysis using HOMER "findMotifsGenome.pl" script (Heinz et al., 2010) if unavailable in the JASPAR database. The original PWMs were then trimmed to keep only the core motifs starting from the first position where information content greater than 0.3 to the last position of information content greater than 0.5 (Ng et al., 2014). The valid motifs were identified by a motif score passing a false positive rate (FPR) 0.1% and a location within 50 bp close to the peak center. The motif spacing is computed as the edge-to-edge distance between two core motifs at TF co-binding sites. If there are multiple valid motifs for one or both TFs, we computed the spacing between all possible combinations of valid motifs.

Characterization of spacing relationships

To test for the constrained spacing relationship between any two TFs, we developed a method to identify "spikes" in the spacing distribution. We first counted the TF pair distances at single-

base-pair resolution ranging from -100 bp to +100 bp. Next, we computed the slope at each position using the following formula:

$$S_i = \frac{\Delta_{i,i-1} + \Delta_{i,i+1}}{2}, i \in [-99,99]$$

$$\Delta_{i,i-1} = N_i - N_{i-1}$$

S_i is the average of single-step forward and backward slope at position i. N_i represents the number of TF pair at position i, and Δ is the difference in the number of TF pairs between two locations. We conducted permutation tests to compare each S_i to a simulated null distribution to determine a p-value based on the percentile rank. P-value smaller than 6.25e-05 (familywise error rate=0.05/200/4) is called significant, indicating a spike is found among motif spacing between the testing TF pair. The null distribution was generated by 1,000 iterations of 1,000 random spacing between 0 and 100 bp.

To test for the relaxed spacing relationship, we used Kolmogorov–Smirnov (KS) test to compare a spacing distribution to the random distribution. We randomly sampled integers between -100 and 100 to match the same size of the testing spacing distribution and then tested the spacing distribution against the distribution of the random integers to obtain a p-value. We repeated the above process 100 times and reported the average p-value.

Categorization of gnomAD variants based on allele frequency

We obtained InDels from gnomAD v3.1 (Karczewski et al., 2020). These gnomAD variants were overlapped with TF co-binding sites, specifically with two TF motifs and their intermediate sequences. For TF pairs with constrained spacing relationships, we only kept the co-bindings that have the significant constrained spacing +/- 2 bp. To account for region-by-region variation in

483 selective pressure, we also overlapped variants with 100-bp upstream and 100-bp downstream
484 background regions outside of TF co-binding sites. For each co-binding site, we categorized
485 InDels into high-frequency variants (AF > 0.01%), rare variants (AF < 0.01%, AC > 1), and
486 singletons (AC = 1) and computed the odds ratios of different categories of InDels between motif
487 or intermediate regions and background regions.
488
489 **Genetic variation processing and genome building**
490 Genetic variation of the five mouse strains was obtained from (Keane et al., 2011), and that of
491 the human individuals from which endothelial cell data were generated was derived from (Stolze
492 et al., 2020). We used MMARGE v1.0 with default variant filters (Link, Romanoski, et al., 2018)
493 to build separate genomes for each mouse strain and human individual. The sequencing data
494 from different samples were respectively mapped to the corresponding genomes and were then
495 shifted to a common reference genome using MMARGE "shift" function to facilitate comparison
496 at homologous regions. The reference genome is mm10 for mouse strains and hg19 for human
497 individuals.
498
499 **Motif mutation analysis**
500 We used MAGGIE (Shen et al., 2020) to identify functional motifs for different TF binding. To
501 prepare the inputs into MAGGIE based on the mouse strains data, we adapted a similar strategy
502 as described in (Shen et al., 2020). In brief, we conducted pairwise orthogonal comparisons of
503 TF peaks between each possible pair of the five strains to find strain-differential peaks. We then
504 extracted pairs of 200-bp sequences around the centers of the differential peaks from the
505 genomes of two comparative strains, the ones with TF binding as positive sequences paired with

506 those without TF binding as negative sequences. For the QTLs of human endothelial cells,
507 MAGGIE can directly work with a VCF file of QTLs with effect size and effect direction
508 indicated in a column of the file. We ran MAGGIE separately for each type of QTLs and
509 reported the significant motifs together with their p-values, which passed false discovery rate
510 (FDR) < 0.05 after the Benjamini–Hochberg controlling procedure.
511

512 **Categorization of genetic variation based on impacts on motif or spacing**

513 We categorized genetic variation based on its impact on motif affinity and motif spacing. Motif
514 mutations were defined by at least 2-bit difference in the motif score, which is equivalent to
515 approximately 4-fold difference in the binding likelihood. Mutations of other functional motifs
516 identified by MAGGIE required that at least one of the functional motifs had motif mutations.
517 InDels were first classified into motif mutation categories if eligible before being considered in
518 motif spacing. Therefore, spacing alterations were InDels between target motifs without any
519 motif mutations. Variants fitting neither motif mutation nor spacing alteration were gathered in a
520 separate group as a control. Another control category during analysis of mouse strains data was
521 defined by TF binding sites that have no genetic variation between strains.
522

523 **Statistical testing of effect size**

524 We conducted Mann-Whitney U tests between the control category and one of the testing
525 categories to test for significance. We also obtained the Cohen's *d* (Sullivan & Feinn, 2012)
526 between the control category and the testing categories and Spearman's correlation coefficients
527 as measures of effect size.
528

ER-HoxB8 cell-derived macrophage culture and CRISPR knockout

Bone marrow cells were isolated from femurs and tibias of a Cas9-expressing transgenic mouse (Jackson Laboratory, No.028555). Murine stem cell virus-based expression vector for ER-HoxB8 was gifted from Dr. David Sykes (Massachusetts General Hospital, Boston, MA). Cas9-expressing ER-HoxB8 conditionally immortalized myoid progenitor cells were generated following established protocols (Wang et al., 2006). In brief, bone marrow cells were purified with a Ficoll gradient (Ficoll-Paque-Plus, Sigma-Aldrich) and resuspended in RPMI 1640 containing 10% FBS, 1% penicillin/streptomycin and 10 ng/ml each of SCF, IL-3 and IL-6 (PeproTech). After 48 hours culture, $2.5x10^5$ cells in 1 ml were transduced with 2 ml of ER-HoxB8 retrovirus (in DMEM with 30% FBS) containing 0.5 μl/ml lentiblast A (OZ Biosciences), 2.5 μl/ml lentiblast B (OZ Biosciences) and 8 μg/ml polybrene (Sigma-Aldrich) in a well of fibronectin (Sigma-Aldrich)-coated 6-well culture plates and centrifuged at 1000g for 90 min at 22°C. After transduction, 6 ml of ER-HoxB8 cell culture media (RPMI 1640 supplemented with 10% FBS, 1% penicillin/streptomycin, 0.5 μM β-estradiol (Sigma-Aldrich), and 20 ng/ml GM-CSF (PeproTech)) were added and an additional half-media exchange with ER-Hoxb8 media performed the next day. Transduced cells were selected with G418 (Thermo Fisher) at 1 mg/ml for 48 hours. Thereafter, cells were maintained in ER-HoxB8 cell culture media. For baseline ATAC-seq and ChIP-seq of ER-HoxB8 cells prior to gRNA transduction, cells were washed twice with PBS, plated at a density of $3x10^6$ cells per 10 cm culture plate, and differentiated into macrophages in DMEM with 10% FBS, 1% pencillin/streptomycin, and 17 ng/ml M-CSF (Shenandoah) for 7 days with 2 culture media exchanges. Differentiated cells were washed twice with PBS and collected for sequencing experiments.

552 Guide RNA lentiviruses were prepared as previously described (Fonseca et al., 2019) with
553 modifications as follows. LentiGuide-mCherry was generated by modifying lentiGuide-puro
554 (Addgene) to remove a puromycin-resistant gene and replace it with mCherry. gRNA sequences
555 directed between the PU.1 and C/EBP motifs (and one each directed towards the motif itself)
556 were designed with CHOPCHOP web tool for genome engineering (Labun et al., 2019). One
557 CRISPR gRNA oligonucleotide was inserted for each target via PCR into a BsmBI cleavage site.
558 A list of gRNA targets used in this article is shown in Figure 5-table supplement 1. Lenti-X 293T
559 cells (Clontech) were seeded in poly-D-lysin (Sigma-Aldrich) coated 10 cm tissue culture plates
560 at a density of 3.5 million cells per plate in 10 ml of DMEM containing 10% FBS and 1%
561 penicillin/streptomycin, and then incubated overnight at 37°C. After replacement of the media to
562 6 ml of DMEM containing 30% FBS, plasmid DNAs (5 µg of lentiviral vector, 3.75 µg of
563 psPAX2 and 1.25 µg of pVSVG) were transfected into LentiX-293T cells using 20 µl of X-
564 tremeGENE™ HP DNA Transfection Reagent (Roche) at 37°C overnight. The media was
565 replaced with DMEM containing 30% FBS and 1% penicillin/streptomycin, and then cultured at
566 37°C overnight. The supernatant was filtrated with 0.45 µm syringe filters and used as lentivirus
567 media. Cell culture media was replaced, and virus was collected again after 24 hours. $1x10^6$
568 Cas9-expressing ER-HoxB8 cells were transduced with virus in 2 ml of lentivirus media and 1
569 ml of ER-HoxB8 cell media containing 0.5 µl/ml lentiblast A, 2.5 µl/ml lentiblast B, and 8 µg/ml
570 polybrene in a well of fibronectin-coated 6-well culture plates and centrifuged at 1000 g for 90
571 min at 22°C. After the transduction, 6 ml of ER-HoxB8 cell media was added to each well. Half
572 of the media was exchanged the next day and in the following days, cells were expanded and
573 passaged into T75 flasks. After 5 days, 250,000 successfully transduced cells (indicated by
574 mCherry fluorescence) for each gRNA were sorted by FACS using a Sony MA900. After FACS,

575 cells were expanded in ER-HoxB8 culture media. Differentiation into macrophages was carried
576 out as above in DMEM supplemented with M-CSF.
577
578 **RNA-seq library preparation**
579 Total RNA was isolated from cells and purified using Direct-zol RNA Microprep columns
580 according to the manufacturer's instructions (Zymo Research). 500 ng of total RNA were used to
581 prepare sequencing libraries from polyA enriched mRNA as previously described (Link, Duttke,
582 et al., 2018). Libraries were PCR-amplified for 14 cycles, size selected using Sera-Mag
583 Speedbeads (Thermo Fisher Scientific), quantified by Qubit dsDNA HS Assay Kit (Thermo
584 Fisher Scientific) and 75-bp single-end sequenced on a HiSeq 4000 (Illumina).
585
586 **ATAC-seq library preparation**
587 ATAC-seq libraries were prepared as previously described (Hoeksema et al., 2021). In brief,
588 $5x10^5$ cells were lysed at room temperature in 50 µl ATAC lysis buffer (10 mM Tris-HCl, pH
589 7.4, 10 mM NaCl, 3 mM MgCl2, 0.1% IGEPAL CA-630) and 2.5 µL DNA Tagmentation
590 Enzyme mix (Nextera DNA Library Preparation Kit, Illumina) was added. The mixture was
591 incubated at 37°C for 30 minutes and subsequently purified using the ChIP DNA purification kit
592 (Zymo Research) as described by the manufacturer. DNA was amplified using the Nextera
593 Primer Ad1 and a unique Ad2.n barcoding primers using NEBNext High-Fidelity 2X PCR MM
594 for 8-14 cycles. PCR reactions were size selected using TBE gels for 175 – 350 bp and DNA
595 eluted using gel diffusion buffer (500 mM ammonium acetate, pH 8.0, 0.1% SDS, 1 mM EDTA,
596 10 mM magnesium acetate) and purified using ChIP DNA Clean & Concentrator (Zymo

Research). Samples were quantified by Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific) and 75-bp single-end sequenced on HiSeq 4000 (Illumina).

Crosslinking for ChIP-seq

For PU.1, C/EBPβ, and H3K27ac ChIP-seq, culture media was removed, and plates were washed once with PBS and then fixed for 10 minutes with 1% formaldehyde (Thermo Fisher Scientific) in PBS at room temperature. Reaction was then quenched by adding glycine (Thermo Fisher Scientific) to 0.125M. After fixation, cells were washed once with cold PBS and then scraped into supernatant using a rubber policeman, pelleted for 5 minutes at 400x*g* at 4°C. Cells were transferred to Eppendorf DNA LoBind tubes and pelleted at 700x*g* for 5 minutes at 4°C, snap-frozen in liquid nitrogen and stored at -80°C until ready for ChIP-seq protocol preparation.

Chromatin immunoprecipitation

Chromatin immunoprecipitation (ChIP) was performed in biological replicates as described previously (Hoeksema et al., 2021). Samples were sonicated using a probe sonicator in 500 µl lysis buffer (10 mM Tris/HCl pH 7.5, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 0.1% deoxycholate, 0.5% sarkozyl, 1x protease inhibitor cocktail). After sonication, 10% Triton X-100 was added to 1% final concentration and lysates were spun at full speed for 10 minutes. 1% was taken as input DNA, and immunoprecipitation was carried out overnight with 20 µl Protein A Dynabeads (Invitrogen) and 2 µg specific antibodies for C/EBPβ (Santa Cruz, sc-150), PU.1 (Santa Cruz, sc-352X), and H3K27ac (Active Motif, 39135). Beads were washed three times each with wash buffer I (20 mM Tris/HCl, 150 mM NaCl, 0.1% SDS, 1% Triton X-100, 2 mM EDTA), wash buffer II (10 mM Tris/HCl, 250 mM LiCl, 1% IGEPAL CA-630, 0.7% Na-

620 deoxycholate, 1 mM EDTA), TE 0.2% Triton X-100 and TE 50 mM NaCl and subsequently
621 resuspended 25 µl 10 mM Tris/HCl pH 8.0 and 0.05% Tween-20. ChIP-seq libraries were
622 prepared on the Dynabeads as described below. For locus specific enrichment ChIP-seq, bead
623 complex was resuspended in 50 µl 1% SDS-TE. 4 µl ProtK, 4 µl RNase A, 3 µl 5 M NaCl was
624 added to these and the input samples and incubated at 50°C for 1 hour, reverse crosslinked at
625 65°C overnight and then eluted from the beads.
626
627 **ChIP-seq library preparation**
628 ChIP libraries were prepared while bound to Dynabeads using NEBNext Ultra II Library
629 preparation kit (NEB) using half reactions. DNA was polished, polyA-tailed and ligated after
630 which dual UDI (IDT) or single (Bioo Scientific) barcodes were ligated to it. Libraries were
631 eluted and crosslinks reversed by adding to the 46.5 µl NEB reaction 16 µl water, 4 µl 10% SDS,
632 4.5 µl 5 M NaCl, 3 µl 0.5 M EDTA, 4 µl 0.2 M EGTA, 1 µl RNAse (10 mg/ml) and 1 µl 20
633 mg/ml proteinase K, followed by incubation at 55°C for 1 hour and 75°C for 30 minutes in a
634 thermal cycler. Dynabeads were removed from the library using a magnet and libraries were
635 cleaned up by adding 2 µl SpeedBeads 3 EDAC (Thermo) in 124 µl 20% PEG 8000/1.5 M NaCl,
636 mixing well, then incubating at room temperature for 10 minutes. SpeedBeads were collected on
637 a magnet and washed two times with 150 µl 80% ethanol for 30 seconds. Beads were collected
638 and ethanol removed following each wash. After the second ethanol wash, beads were air dried
639 and DNA eluted in 12.25 µl 10 mM Tris/HCl pH 8.0 and 0.05% Tween-20. DNA was amplified
640 by PCR for 14 cycles in a 25 µl reaction volume using NEBNext Ultra II PCR master mix and
641 0.5 µM each Solexa 1GA and Solexa 1GB primers. Libraries were size selected using TBE gels
642 for 200 – 500 bp and DNA eluted using gel diffusion buffer (500 mM ammonium acetate, pH

8.0, 0.1% SDS, 1 mM EDTA, 10 mM magnesium acetate) and purified using ChIP DNA Clean & Concentrator (Zymo Research). Sample concentrations were quantified by Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific) and 75-bp single-end sequenced on HiSeq 4000.

Biotin-mediated locus specific enrichment ChIP-seq library preparation

After performing the target-specific ChIPs, we performed an initial PCR for locus-specific amplicon enrichment using NEBNext 2X High Fidelity PCR MM (NEB) and 5'-biotinylated stub adapter primers specific to appropriate genomic regions to be interrogated (Figure 5-table supplement 1). Initial hotstart/denaturation at 98℃ for 30 sec was followed by 10 cycles of amplification (98℃ for 15 sec, 65-67℃ for 15 sec, 72ºC for 30 sec) and then a final elongation at 72℃ for 5 min. After this, we performed a 0.7X AmpureXP clean-up and eluted in 20 µl 0.5x TT (5 mM Tris pH 8.0 + 0.025% Tween20). Dynabeads MyOne Streptavidin T1 beads were then washed in 1x Wash Binding Buffer (WBB, 2X WBB: 10 mM Tris-HCl (pH 7.5), 1 mM EDTA, 2 M NaCl, 0.1% Tween) and resuspended beads at 20 µl per sample in 2x WBB. 20 µl prepared Dynabeads MyOne Streptavidin T1 beads (in 2x WBB) were then added to cleaned up 20 µl 0.5x TT PCR fragments, mixed and incubated for 60min at RT with mild shaking. After this, beads were collected on a magnet and washed twice with 150 µl 1x WBB and once with 180 µl TET (TE + 0.05% Tween-20). Finally, beads were resuspended in 25 µl 0.5x TT and on bead PCR for addition of Illumina-specific adapters and 10-bp Unique Dual Indexes (UDIs) using NEBNext 2X High Fidelity PCR MM (NEB) and 25 PCR cycles was performed (Figure 5-table supplement 2). Libraries were size selected using TBE gels for 300-500 bp and DNA eluted using gel diffusion buffer and purified using ChIP DNA Clean & Concentrator (Zymo Research).

Samples were quantified by Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific) and 150-bp paired-end sequenced on NextSeq 500 (Illumina).

Analysis of variable InDels from CRISPR experiments

We mapped the reads to the target regions using the local alignment mode of Bowtie2 v2.3.5.1 (Langmead & Salzberg, 2012). To allow for InDels with tens of bases, we reduced the gap extend penalty and increased the gap open penalty so that the gaps could be long but not occur at multiple locations. Here are the adjusted parameters used in our mapping process: --local --rdg 10,1 --rfg 10,1. The mapped reads with gaps or InDels at unexpected locations rather than the Cas9 cut sites were removed. This step filtered out approximately 1% of the total reads (Figure 5-table supplement 2). The remaining reads were grouped based on the InDel size and whether the InDel overlaps with any of the PU.1 and C/EBP motifs. Tag counts were used as quantification of the signal intensity. InDel groups with less than 0.05% of the input reads were filtered out to reduce the low-intensity data. The effect of each InDel group on TF binding was computed by the odds ratio between TF ChIP-seq tags and input sample ChIP-seq tags: (# TF tags for an InDel group / # the rest of TF tags) / (# input tags for the same InDel group / # the rest of input tags).

Data and code availability

All sequencing data generated from this study have been made available by deposition in the GEO database: GSE178080. All raw tag counts are available in Figure 5-source data 1. The UCSC genome browser was used to visualize sequencing data. The codes for data analysis and

687 the processed files of ENCODE data are available on our Github repository:

688 https://github.com/zeyang-shen/spacing_pipeline.

689

690 **Acknowledgements**

691 The authors would like to thank J. Collier for technical assistance, the IGM core for library

692 sequencing, L. Van Ael for assistance with manuscript preparation. These studies were supported

693 by NIH grants DK091183 and HL147835 and a Leducq Transatlantic Network grant 16CVD01

694 to CKG. TAP was supported by NIH grant T32DK007044. MAH was supported by a Rubicon

695 grant from the Netherlands Organization for Scientific Research and postdoctoral grants from the

696 Amsterdam Cardiovascular Sciences Institute and the American Heart Association.

697

698 **Competing interests**

699 None declared

References

Behera, V., Evans, P., Face, C. J., Hamagami, N., Sankaranarayanan, L., Keller, C. A., Giardine, B., Tan, K., Hardison, R. C., Shi, J., & Blobel, G. A. (2018). Exploiting genetic variation to uncover rules of transcription factor binding and chromatin accessibility. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-03082-6

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & De Hoon, M. J. L. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, 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

Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., Hilton, J. A., Jain, K., Baymuradov, U. K., Narayanan, A. K., Onate, K. C., Graham, K., Miyasato, S. R., Dreszer, T. R., Strattan, J. S., Jolanki, O., Tanaka, F. Y., & Cherry, J. M. (2018). The Encyclopedia of DNA elements (ENCODE): Data portal update. *Nucleic Acids Research*, *46*(D1), D794–D801. https://doi.org/10.1093/nar/gkx1081

Deplancke, B., Alpern, D., & Gardeux, V. (2016). The Genetics of Transcription Factor DNA Binding Variation. *Cell*, *166*(3), 538–554. https://doi.org/10.1016/j.cell.2016.07.012

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Farh, K. K. H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., Shoresh, N., Whitton, H., Ryan, R. J. H., 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., … 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., Brandt, A. J., Rokhsar, D. S., & Levine, M. S. (2015).

731 Suboptimization of developmental enhancers. *Science*, *350*(6258), 325–328.
732 https://doi.org/10.1126/science.aac6948
733 Fonseca, G. J., Tao, J., Westin, E. M., Duttke, S. H., Spann, N. J., Strid, T., Shen, Z., Stender, J.
734 D., Sakai, M., Link, V. M., Benner, C., & Glass, C. K. (2019). Diverse motif ensembles
735 specify non-redundant DNA binding activities of AP-1 family members in macrophages.
736 *Nature Communications*, *10*(1), 414. https://doi.org/10.1038/s41467-018-08236-0
737 Fornes, O., Castro-Mondragon, J. A., Khan, A., Van Der Lee, R., Zhang, X., Richmond, P. A.,
738 Modi, B. P., Correard, S., Gheorghe, M., Baranašić, D., Santana-Garcia, W., Tan, G.,
739 Chèneby, J., Ballester, B., Parcy, F., Sandelin, A., Lenhard, B., Wasserman, W. W., &
740 Mathelier, A. (2020). JASPAR 2020: Update of the open-Access database of transcription
741 factor binding profiles. *Nucleic Acids Research*, *48*(D1), D87–D92.
742 https://doi.org/10.1093/nar/gkz1001
743 Grossman, S. R., Zhang, X., Wang, L., Engreitz, J., Melnikov, A., Rogov, P., Tewhey, R.,
744 Isakova, A., Deplancke, B., Bernstein, B. E., Mikkelsen, T. S., & Lander, E. S. (2017).
745 Systematic dissection of genomic features determining transcription factor binding and
746 enhancer function. *Proceedings of the National Academy of Sciences of the United States of*
747 *America*, *114*(7), E1291–E1300. https://doi.org/10.1073/pnas.1621150114
748 Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C.,
749 Singh, H., & Glass, C. K. (2010). Simple Combinations of Lineage-Determining
750 Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell
751 Identities. *Molecular Cell*, *38*(4), 576–589. https://doi.org/10.1016/j.molcel.2010.05.004
752 Heinz, S., Romanoski, C. E., Benner, C., Allison, K. A., Kaikkonen, M. U., Orozco, L. D., &
753 Glass, C. K. (2013). Effect of natural genetic variation on enhancer selection and function.
754 *Nature*, *503*, 487. https://doi.org/10.1038/nature12615
755 Heinz, S., Romanoski, C. E., Benner, C., & Glass, C. K. (2015). The selection and function of
756 cell type-specific enhancers. *Nature Reviews Molecular Cell Biology*, *16*(3), 144–154.
757 https://doi.org/10.1038/nrm3949
758 Hoeksema, M. A., Shen, Z., Holtman, I. R., Zheng, A., Spann, N. J., Cobo, I., Gymrek, M., &
759 Glass, C. K. (2021). Mechanisms underlying divergent responses of genetically distinct
760 macrophages to IL-4. *Science Advances*, *7*(25), eabf9808.
761 https://doi.org/10.1126/sciadv.abf9808

762 Hogan, N. T., Whalen, M. B., Stolze, L. K., Hadeli, N. K., Lam, M. T., Springstead, J. R., Glass,
763 C. K., & Romanoski, C. E. (2017). Transcriptional networks specifying homeostatic and
764 inflammatory programs of gene expression in human aortic endothelial cells. *ELife*, *6*(Cvd),
765 1–28. https://doi.org/10.7554/eLife.22536
766 Hu, H., Miao, Y.-R., Jia, L.-H., Yu, Q.-Y., Zhang, Q., & Guo, A.-Y. (2019). AnimalTFDB 3.0: a
767 comprehensive resource for annotation and prediction of animal transcription factors.
768 *Nucleic Acids Research*, *47*(D1), D33–D38. https://doi.org/10.1093/nar/gky822
769 Jiang, P., & Singh, M. (2014). CCAT: Combinatorial Code Analysis Tool for transcriptional
770 regulation. *Nucleic Acids Research*, *42*(5), 2833–2847. https://doi.org/10.1093/nar/gkt1302
771 Jolma, A., Yin, Y., Nitta, K. R., Dave, K., Popov, A., Taipale, M., Enge, M., Kivioja, T.,
772 Morgunova, E., & Taipale, J. (2015). DNA-dependent formation of transcription factor
773 pairs alters their binding specificity. *Nature*, *527*(7578), 384–388.
774 https://doi.org/10.1038/nature15518
775 Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R.
776 L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson,
777 M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J.
778 A., … Consortium, G. A. D. (2020). The mutational constraint spectrum quantified from
779 variation in 141,456 humans. *Nature*, *581*(7809), 434–443. https://doi.org/10.1038/s41586-
780 020-2308-7
781 Keane, T. M., Goodstadt, L., Danecek, P., White, M. A., Wong, K., Yalcin, B., Heger, A., Agam,
782 A., Slater, G., Goodson, M., Furlotte, N. A., Eskin, E., Nellåker, C., Whitley, H., Cleak, J.,
783 Janowitz, D., Hernandez-Pliego, P., Edwards, A., Belgard, T. G., … Adams, D. J. (2011).
784 Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*,
785 *477*(7364), 289–294. https://doi.org/10.1038/nature10413
786 Labun, K., Montague, T. G., Krause, M., Torres Cleuren, Y. N., Tjeldnes, H., & Valen, E.
787 (2019). CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing.
788 *Nucleic Acids Research*, *47*(W1), W171–W174. https://doi.org/10.1093/nar/gkz365
789 Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature*
790 *Methods*, *9*, 357. https://doi.org/10.1038/nmeth.1923
791 Lek, M., Karczewski, K. J., Minikel, E. V, Samocha, K. E., Banks, E., Fennell, T., O'Donnell-
792 Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P.,

Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., … Consortium, E. A. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291. https://doi.org/10.1038/nature19057

Levkovitz, Y., & Baraban, J. M. (2002). A Dominant Negative Egr Inhibitor Blocks Nerve Growth Factor-Induced Neurite Outgrowth by Suppressing c-Jun Activation: Role of an Egr/c-Jun Complex. *Journal of Neuroscience*, *22*(10), 3845–3854. https://doi.org/10.1523/jneurosci.22-10-03845.2002

Li, Q., Brown, J. B., Huang, H., & Bickel, P. J. (2011). Measuring reproducibility of high-throughput experiments. *Annals of Applied Statistics*, *5*(3), 1752–1779. https://doi.org/10.1214/11-AOAS466

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., … 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

Link, V. M., Romanoski, C. E., Metzler, D., & Glass, C. K. (2018). MMARGE: Motif mutation analysis for regulatory genomic elements. *Nucleic Acids Research*, *46*(14), 7006–7021. https://doi.org/10.1093/nar/gky491

Lis, M., & Walther, D. (2016). The orientation of transcription factor binding site motifs in gene promoter regions: Does it matter? *BMC Genomics*, *17*(1), 1–21. https://doi.org/10.1186/s12864-016-2549-x

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., MayPendlington, Z., Welter, D., Burdett, T., Hindorff, L., Flicek, P., Cunningham, F., & Parkinson, H. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, *45*(D1), D896–D901. https://doi.org/10.1093/nar/gkw1133

Macián, F., López-Rodríguez, C., & Rao, A. (2001). Partners in transcription: NFAT and AP-1. *Oncogene*, *20*(19), 2476–2489. https://doi.org/10.1038/sj.onc.1204386

Menoret, D., Santolini, M., Fernandes, I., Spokony, R., Zanet, J., Gonzalez, I., Latapie, Y., Ferrer, P., Rouault, H., White, K. P., Besse, P., Hakim, V., Aerts, S., Payre, F., & Plaza, S.

(2013). Genome-wide analyses of Shavenbaby target genes reveals distinct features of enhancer organization. *Genome Biology*, *14*(8), R86. https://doi.org/10.1186/gb-2013-14-8-r86

Morgunova, E., & Taipale, J. (2017). Structural perspective of cooperative transcription factor binding. *Current Opinion in Structural Biology*, *47*, 1–8. https://doi.org/10.1016/j.sbi.2017.03.006

Nakashima, A., Ota, A., & Sabban, E. L. (2003). Interactions between Egr1 and AP1 factors in regulation of tyrosine hydroxylase transcription. *Molecular Brain Research*, *112*(1–2), 61–69. https://doi.org/10.1016/S0169-328X(03)00047-0

Nandi, S., Blais, A., & Ioshikhes, I. (2013). Identification of cis-regulatory modules in promoters of human genes exploiting mutual positioning of transcription factors. *Nucleic Acids Research*, *41*(19), 8822–8841. https://doi.org/10.1093/nar/gkt578

Ng, F. S. L., Schütte, J., Ruau, D., Diamanti, E., Hannah, R., Kinston, S. J., & Göttgens, B. (2014). Constrained transcription factor spacing is prevalent and important for transcriptional control of mouse blood cells. *Nucleic Acids Research*, *42*(22), 13513–13524. https://doi.org/10.1093/nar/gku1254

Panne, D. (2008). The enhanceosome. *Current Opinion in Structural Biology*, *18*(2), 236–242. https://doi.org/10.1016/j.sbi.2007.12.002

Reményi, A., Lins, K., Nissen, L. J., Reinbold, R., Schöler, H. R., & Wilmanns, M. (2003). Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers. *Genes and Development*, *17*(16), 2048–2059. https://doi.org/10.1101/gad.269303

Rodda, D. J., Chew, J. L., Lim, L. H., Loh, Y. H., Wang, B., Ng, H. H., & Robson, P. (2005). Transcriptional regulation of Nanog by OCT4 and SOX2. *Journal of Biological Chemistry*, *280*(26), 24731–24737. https://doi.org/10.1074/jbc.M502573200

Shen, Z., Hoeksema, M. A., Ouyang, Z., Benner, C., & Glass, C. K. (2020). MAGGIE: leveraging genetic variation to identify DNA sequence motifs mediating transcription factor binding and function. *Bioinformatics (Oxford, England)*, *36*(1), i84–i92. https://doi.org/10.1093/bioinformatics/btaa476

Slattery, M., Zhou, T., Yang, L., Dantas Machado, A. C., Gordân, R., & Rohs, R. (2014). Absence of a simple code: How transcription factors read the genome. *Trends in*

Biochemical Sciences, *39*(9), 381–399. https://doi.org/10.1016/j.tibs.2014.07.002

Smith, R. P., Taher, L., Patwardhan, R. P., Kim, M. J., Inoue, F., Shendure, J., Ovcharenko, I., & Ahituv, N. (2013). Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nature Genetics*, *45*(9), 1021–1028. https://doi.org/10.1038/ng.2713

Sönmezer, C., Kleinendorst, R., Imanci, D., Barzaghi, G., Villacorta, L., Schübeler, D., Benes, V., Molina, N., & Krebs, A. R. (2021). Molecular Co-occupancy Identifies Transcription Factor Binding Cooperativity In Vivo. *Molecular Cell*, *81*(2), 255-267.e6. https://doi.org/10.1016/j.molcel.2020.11.015

Stolze, L. K., Conklin, A. C., Whalen, M. B., López Rodríguez, M., Õunap, K., Selvarajan, I., Toropainen, A., Örd, T., Li, J., Eshghi, A., Solomon, A. E., Fang, Y., Kaikkonen, M. U., & Romanoski, C. E. (2020). Systems Genetics in Human Endothelial Cells Identifies Non-coding Variants Modifying Enhancers, Expression, and Complex Disease Traits. *American Journal of Human Genetics*, *106*(6), 748–763. https://doi.org/10.1016/j.ajhg.2020.04.008

Sullivan, G. M., & Feinn, R. (2012). Using Effect Size—or Why the P Value Is Not Enough . *Journal of Graduate Medical Education*, *4*(3), 279–282. https://doi.org/10.4300/jgme-d-12-00156.1

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics*, *101*(1), 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005

Wang, G. G., Calvo, K. R., Pasillas, M. P., Sykes, D. B., Häcker, H., & Kamps, M. P. (2006). Quantitative production of macrophages or neutrophils ex vivo using conditional Hoxb8. *Nature Methods*, *3*(4), 287–293. https://doi.org/10.1038/nmeth865

Ward, L. D., & Kellis, M. (2012). Interpreting noncoding genetic variation in complex traits and human disease. *Nature Biotechnology*, *30*(11), 1095–1106. https://doi.org/10.1038/nbt.2422

Zaret, K. S., & Carroll, J. S. (2011). Pioneer transcription factors: Establishing competence for gene expression. *Genes and Development*, *25*(21), 2227–2241. https://doi.org/10.1101/gad.176826.111