Natural genetic variation affecting transcription factor spacing at regulatory regions is generally well tolerated

Zeyang Shen[1,2], Jenhan Tao[1], Gregory J. Fonseca[1] and Christopher K. Glass[1,3,4]

[1]Department of Cellular and Molecular Medicine, School of Medicine
[2]Department of Bioengineering, Jacobs School of Engineering
[3]Department of Medicine, School of Medicine, University of California, San Diego, La Jolla, CA, 92093, USA
[4]ckg@ucsd.edu

Abstract

Regulation of gene expression requires the combinatorial binding of sequence-specific transcription factors (TFs) at promoters and enhancers. Single nucleotide polymorphisms (SNPs) and short insertions and deletions (InDels) can influence gene expression by altering the sequences of TF binding sites. Prior studies also showed that alterations in the spacing between TF binding sites can influence promoter and enhancer activity. However, the relative importance of altered TF spacing has not been systematically analyzed in the context of natural genetic variation. Here, we exploit millions of InDels provided by five diverse strains of mice to globally investigate the effects of altered spacing on TF binding and local histone acetylation in macrophages. We find that spacing alterations resulting from InDels are generally well tolerated in comparison to genetic variants that directly alter TF binding sites. These findings have implications for interpretation of non-coding genetic variation and comparative analysis of regulatory elements across species.

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). Many of these variants fall into regulatory regions of the genome, implicating their effects on gene regulation (GTEx Consortium, 2015, Farh *et al.*, 2015). Gene expression is regulated in a cell-type-specific manner by transcription factors (TFs) that bind to short, degenerate sequences in promoters and enhancers 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. Genetic variation that creates or disrupts TF binding motifs is a well-established mechanism for altering gene expression and biological function (Grossman *et al.*, 2017, Deplancke *et al.*, 2016, Heinz *et al.*, 2015). Collaborative binding of TFs required for enhancer or promoter selection can interact over a relatively broad range of spacing (e.g., 100-200 bp; Slattery *et al.*, 2014, Jiang and Singh, 2014, Heinz *et al.*, 2010). Consistent with this, flexibility in motif spacing relationships has been demonstrated using reporter assays in Drosophila (Menoret *et al.*, 2013) and HepG2 cells (Smith *et al.*, 2013).

On the contrary, substantial evidence also showed that specific spacing relationships between motifs can be important for TF binding and function (Boeva, 2016). A special category is provided by TFs that form ternary complexes recognizing composite binding sites, exemplified by CAP-SELEX studies of 9,400 TF pairs (Jolma *et al.*, 2015), MyoD and other muscle-specific factors in muscle cells (Nandi *et al.*, 2013), Sox2 and Oct4 in embryonic stem cells (Rodda *et al.*, 2005), Ets and E-box in haematopoietic cells (Ng *et al.*, 2014), etc. Similar constrained spacing between independent motifs are required for the optimal binding

and function of interacting TFs at the interferon-β enhanceosome (Panne, 2008). In addition, reporter assays examining synthetic alterations of motif spacing between collaborative factors revealed examples for high sensitivity of gene expression on spacing in Ciona (Farley *et al.*, 2015). However, these studies did not distinguish the impact of altered spacing on transcription factor binding or subsequent recruitment of co-activators required for gene activation. Moreover, it remains unknown the extent to which these findings are relevant to altered spacing resulting from natural genetic variation in human population or between animal species.

Bone marrow-derived macrophages (BMDMs) from genetically diverse strains of mice provide a powerful system for studying the genome-wide impact of natural genetic variation on gene regulation. Single nucleotide polymorphisms (SNPs) and short insertions and deletions (InDels) represent common forms of genetic variation in the genomes of different mouse strains (Keane *et al.*, 2011) and are associated with strain-specific variation in gene expression. SNPs and InDels could affect motif sequence and mutate a motif, while InDels could additionally change spacing between motifs. Initial studies in the BMDMs from two strains of mice used naturally occurring motif mutations to support a collaborative binding model between LDTFs (e.g., PU.1 and C/EBPβ) and a hierarchical binding model between LDTFs and signal dependent transcription factors (SDTFs) (e.g., PU.1 and p65) (Heinz *et al.*, 2013). Subsequent studies leveraging more than 50 million SNPs and 5 million InDels from five mouse strains linked ~60% of strain-specific TF binding sites to mutated motifs (Link *et al.*, 2018a), suggesting a possibility for the remaining strain-specific sites to be impacted by InDels that alter motif spacing.

To investigate the effects of altered spacing on TF binding and function, we first characterized the genome-wide binding patterns of macrophage LDTFs and SDTFs based on their binding sites determined by chromatin immuno-precipitation sequencing (ChIP-seq). By leveraging the local genetic variation at the TF binding sites from the five strains of mice, we linked the alteration of motif spacing to the change of TF binding activity and 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 find that InDels altering spacing between specific pairs of LDTFs and SDTFs can be associated with significant changes in their respective binding, but this relationship can largely be explained by effects of these InDels on the binding motifs of other collaborative factors, suggesting a general tolerance of spacing alterations resulting from natural genetic variation. These findings have implications for understanding mechanisms underlying enhancer selection, interpretation of non-coding variants associated with phenotypic variation, and comparisons of regulatory elements between species.

Results

Characterization of the spacing between macrophage LDTFs

As a starting point, we characterized the spacing relationship between the macrophage LDTFs, PU.1 and C/EBPβ (**Figure 1A**), which have been found to bind in a collaborative manner at regulatory regions of macrophage-specific genes (Heinz *et al.*, 2010). We first determined reproducible PU.1 and C/EBPβ binding sites from the replicate ChIP-seq data of C57BL/6J (C57) mice (Link *et al.*, 2018a) and then categorized them into three groups: co-bound by both factors, bound by PU.1 only, and bound by C/EBPβ only (**Figure 1B**). For every binding site, we identified the DNA sequence best matching the motifs of PU.1 and C/EBPβ as determined by position weight matrices (PWMs) (Stormo, 2000; **Materials and Methods**). We then computed the spacing (i.e., distance) between the centers of best-

matching sequences and plotted its distribution for sites within the same group. Co-binding sites showed a preference, but not strictness, for PU.1 and C/EBPβ motifs to occur within ±75 bp of each other (**Figure 1C**; **Figure 1—figure supplement 1**), in agreement with prior studies (Heinz *et al.*, 2010). Noticeably, a discontinuity occurred at where the two motifs overlap (spacing < 12 bp), potentially due to a steric inhibition of co-binding in these instances. For the sites bound by PU.1 or C/EBPβ alone, the spacing relationship between PU.1 and C/EBPβ motifs was statistically similar to the background distribution, consistent with few collaborative interactions between PU.1 and C/EBPβ at these sites.

After observing the overall proximity between PU.1 and C/EBPβ motifs at their co-binding sites, we investigated whether this spacing preference had an impact on TF binding. The binding activities of PU.1 and C/EBPβ were quantified by ChIP-seq reads at the co-binding sites. We correlated the number of reads with either motif spacing or motif score, which represents the similarity of a sequence in comparison to PWMs. Both PU.1 and C/EBPβ binding activities were positively correlated with the motif scores of their respective motifs but showed a much weaker correlation with spacing (**Figure 1D**). Interestingly, PU.1 binding activity is negatively correlated with C/EBPβ motif score, implicating a synergistic binding model between these two TFs, which would allow the recognition of more degenerate motif sequence when they bind together to DNA.

Effect of altered spacing on transcription factor binding based on natural genetic variation across mouse strains

Table 1. P-values and effect sizes for the effects of different genetic variation between C57 and PWK on PU.1 binding, C/EBPβ binding, and H3K27ac.

	Mutated PU.1 motif		Mutated C/EBPβ motif		Altered spacing (unfiltered)		Altered spacing (filtered by collabor. factors)		Altered spacing (filtered by unrelated factors)	
	p	*d*	*p*	*d*	*p*	*d*	*p*	*d*	*p*	*d*
PU.1 binding	<1e-4	[1.07, 1.12]	<1e-4	[0.71, 0.78]	<1e-4	[0.38, 0.58]	0.002	[0.16, 0.52]	<1e-4	[0.37, 0.66]
C/EBPβ binding	<1e-4	[0.76, 0.84]	<1e-4	[1.03, 1.10]	<1e-4	[0.27, 0.50]	0.16	[0.04, 0.37]	6e-4	[0.20, 0.55]
H3K27ac	<1e-4	[0.56, 0.65]	<1e-4	[0.55, 0.65]	<1e-4	[0.24, 0.45]	0.17	[0.04, 0.29]	9e-4	[0.16, 0.45]

P-values are calculated based on 10,000 iterations of permutation tests by comparing the log fold changes of ChIP-seq reads against the variant-free category. P-value below 1e-4 is beyond the specified testing power. Effect sizes are represented by Cohen's *d*, displayed by its 90% confidence interval, which is based on 10,000 iterations of sampling the variant-free regions. For the comparison of different filters for InDels altering spacing, we computed the statistics for the complete set (i.e., unfiltered) and the filtered sets after excluding InDels simultaneously mutating motifs of collaborative factors or non-collaborative factors.

To investigate the requirement and tolerance of spacing proximity observed for collaborative factors, we leveraged the natural genetic variation across genetically diverse strains of mice as a mutagenesis screen. We selected five strains from which the ChIP-seq data of macrophage LDTFs and SDTFs were previously published (Link *et al.*, 2018a): C57BL/6J (C57), BALB/cJ (BALB), NOD/ShiLtJ (NOD), PWK/PhJ (PWK), and SPRET/EiJ (SPRET). Independent comparisons were conducted between C57 and one of the other four strains, which provide 4-40 million SNPs and 1-4 million InDels with respect to C57 (Keane *et al.*, 2011). We first identified the co-binding sites of PU.1 and C/EBPβ for every strain based on ChIP-seq data. For each pairwise analysis, the co-binding sites from C57 and the compared strain were pooled and distributed into four categories based on the impacts of local variants: altered spacing, mutated PU.1 motif, mutated C/EBPβ motif, and variant-free (**Figure 2A**; **Materials and Methods**). We quantified the effects of genetic variation on TF binding based on log2 fold changes of ChIP-seq reads between the compared strains. PU.1 binding is

143 significantly affected by mutated PU.1 motif and mutated C/EBPβ motif, which have the
144 largest and second largest effect sizes represented by Cohen's *d* (**Table 1**; **Figure 2B**; **Figure**
145 **2—figure supplement 1**). Similarly, C/EBPβ binding is most significantly affected by
146 mutated C/EBPβ motif, followed by mutated PU.1 motif (**Table 1**; **Figure 2C**; **Figure 2—**
147 **figure supplement 1**). Altered spacing resulting from InDels showed a much smaller, but
148 still significant effect on both PU.1 and C/EBPβ binding (**Table 1** "unfiltered"). In many
149 cases, an alteration of several nucleotides between PU.1 and C/EBPβ motifs has no
150 observable effect on TF binding, while one SNP that alters the core sequence of a motif can
151 disrupt TF binding (**Figure 2—figure supplement 2**).
152
153 After observing a significant effect of altered spacing on the binding of PU.1 and C/EBPβ,
154 we investigated whether the effect size is influenced by the scale or direction of spacing
155 alterations. By correlating the change of binding activity with the size of InDels (positive for
156 insertions and negative for deletions), spacing alteration demonstrated an effect independent
157 from the scale or direction of InDels (**Figure 2D; Figure 2—figure supplement 3**). On the
158 contrary, changes of motif score are strongly correlated with changes of ChIP-seq reads,
159 consistent with the important role of motif for TF binding. The invariable effects of InDels
160 were unexpected because, based on the spacing relationship of PU.1 and C/EBPβ, we
161 expected a preference for closer spacing and a larger effect from longer InDels. However,
162 InDels altering 1 or 2 bp between motifs can often have an effect as large as relatively long
163 InDels, and such effect is not affected by the initial spacing between motifs (**Figure 2—**
164 **figure supplement 4**), suggesting that the significant effect of InDels might not be directly
165 resulted from the alteration of spacing but from other reasons.
166
167 **InDels that alter spacing may simultaneously mutate motifs of other collaborative**
168 **factors**
169 One possibility for seeing a significant effect of InDels that reside between PU.1 and C/EBPβ
170 motifs could be alterations of motifs recognized by other collaborative factors. To test this
171 hypothesis, we developed a computational framework to confidently identify collaborative
172 factors. Considering that it would be a vast undertaking to perform ChIP-seq on all expressed
173 TFs, our framework leverages TF binding sites identified from single ChIP-seq data and
174 predicts collaborative factors based on high-score and closely spaced motifs (**Figure 3A**).
175 This design is supported by our observations on PU.1 and C/EBPβ binding sites where an
176 increasing threshold on the motif score of collaborative factors recovered a larger proportion
177 of co-binding sites (**Figure 3—figure supplement 1**) and also recovered the spacing
178 relationships previously identified from ChIP-seq data (**Figure 3—figure supplement 2**). To
179 compare the spacing distribution predicted by our framework to the distribution identified
180 from co-binding sites, we tested on PU.1 binding sites measured by ChIP-seq and searched
181 for high-score motif of cJun (i.e., FOS::JUN or TGAG/CTCA), which is a known LDTF of
182 macrophages and a collaborative factor of PU.1. The predicted distribution of cJun around
183 PU.1 aligned well with experimentally determined distribution based on cJun ChIP-seq (Link
184 *et al.*, 2018a), showing the utility of identifying collaborative factors based on closely spaced
185 high-score motifs (**Figure 3B**). Therefore, we applied this approach to uncover the
186 collaborative factors of PU.1 and C/EBPβ from over five hundred TFs whose motifs are
187 available in the JASPAR database (Fornes *et al.*, 2020). To facilitate the comparison across
188 motifs, we used the top 4,000 regions ranked by the motif score of every computed motif to
189 obtain spacing distribution and compared each distribution against the background
190 distribution using KS tests. P-values from KS tests were given signs to distinguish positive or
191 negative associations with proximal spacing. Most TFs indicate no spacing relationship with
192 either PU.1 or C/EBPβ (**Figure 3C; Figure 3—figure supplement 3**). Motifs with proximal

193 spacing relationships tend to have relatively high expression based on RNA-seq data (Link *et*
194 *al.*, 2018a).
195
196 Based on our computational framework, we selected twelve predicted collaborative factors,
197 which are closely spaced with PU.1 or C/EBPβ (KS test p-value < 1e-5) and also highly
198 expressed in mouse macrophages (TPM value > 16). We refined the testing of co-binding
199 sites by filtering out those with motif mutations of any collaborative factors on at least one
200 core position (roughly equivalent to a change of motif score greater than 1). The remaining
201 sites with InDels between PU.1 and C/EBPβ motifs, which should represent a clean set of
202 spacing alterations, showed a diminished effect on TF binding (**Table 1** "filtered by collabor.
203 factors"; **Figure 3D**). When we filtered on unrelated factors identified as non-collaborative
204 by our framework, the effect sizes were not affected (**Table 1** "filtered by unrelated factors";
205 **Figure 3—figure supplement 4**).
206
207 To investigate whether the effects of altered spacing on PU.1 and C/EBPβ binding can be
208 generalized to hierarchical interactions with signal-dependent transcription factors, we
209 repeated our analyses on another pair of TFs, PU.1 and the NFκB subunit p65. Upon
210 macrophage activation with the TLR4-specific ligand Kdo2 lipid A (KLA), p65 enters the
211 nucleus and primarily binds to poised enhancer elements that are selected by pioneering
212 factors including PU.1. By leveraging the ChIP-seq data of the two TFs from C57 and PWK
213 macrophages treated for one hour with KLA (Link *et al.*, 2018a), we observed a preference
214 for proximal spacing between PU.1 and p65 motifs at their co-binding sites (**Figure 3—**
215 **figure supplement 5**) and a diminished effect of altered spacing after excluding InDels that
216 affect motifs of the predicted collaborative factors of PU.1 (**Figure 3E**; **Figure 3—figure**
217 **supplement 6**), consistent with our finding from PU.1 and C/EBPβ that spacing alterations
218 are well tolerated by TF binding.
219
220 **Effect of altered spacing on promoter and enhancer function**
221 Although alterations in motif spacing were generally well tolerated at the level of DNA
222 binding, it remained possible that changes in motif spacing could influence subsequent steps
223 in enhancer and/or promoter activation. To examine this, we extended our analysis to local
224 histone acetylation as a surrogate of promoter and enhancer function. We leveraged the
225 H3K27ac ChIP-seq data for the five strains of mice (Link *et al.*, 2018a) and calculated the log
226 fold changes of H3K27ac level within the extended 1000-bp regions of the PU.1 and C/EBPβ
227 co-binding sites. Testing all the co-binding sites demonstrated significant effects of both
228 spacing alteration and motif mutation (**Table 1**; **Figure 4A; Figure 4—figure supplement**
229 **1**). However, the significance for altered spacing disappeared after filtering out sites
230 potentially having motif mutations for the previous twelve collaborative factors (**Table 1**;
231 **Figure 4B**; **Figure 4—figure supplement 2**). Again, filtering for unrelated factors did not
232 influence the effect size (**Table 1**; **Figure 4—figure supplement 3**). The tolerance of
233 spacing alteration was further reflected by a weak correlation between the change of
234 acetylation level and the size of InDels, in comparison to a much stronger correlation with
235 motif scores of both PU.1 and C/EBPβ (**Figure 4C**). Similar to what was observed for TF
236 binding, altered spacing demonstrated trivial effects on histone acetylation, which is
237 supported by the high consistency between change of TF binding and change of acetylation
238 (**Figure 4—figure supplement 4**). Noticeably, the acetylation level has an overall smaller
239 scale of change compared to TF binding activity, reflecting its more complex dependency on
240 TF binding (Reiter *et al.*, 2017).
241
242 **Consideration of gap penalties in cross species sequence alignments**

The insignificant effect of InDels between TF binding sites on TF binding and local histone acetylation at a genome-wide scale suggested that evolutionary pressure on enhancer selection and function would be relatively tolerant of these forms of genetic variation in comparison to InDels that directly affect the sequences of TF binding motifs. This is in contrast to effects of InDels in protein coding regions of the genome, in which insertions or deletions of bases other than multiples of three would result in frame-shift mutations. To explore this possibility, we performed sequence alignments using BLAST (Boratyn *et al.*, 2013) for well-established regulatory elements of macrophage-specific genes in the mouse with the human genome using i) standard parameters that impose significant penalties for gaps or ii) lenient parameters in which gap penalties were diminished. These comparisons frequently resulted in relatively short sequence alignments when standard gap penalties were applied but much more extended alignments that contained multiple relevant TF binding motifs using lenient gap penalties (**Figure 5A; Figure 5—figure supplements 1 and 2**). Examples are provided for putative regulatory elements of genes with known functions in macrophages (**Figure 5B**; **Figure 5—figure supplement 3**), including *Anxa7* (Li *et al.*, 2013), *Fos* (Hop *et al.*, 2018), *Vmp1* (Dziuba *et al.*, 2012), *Max* (Ayer *et al.*, 1993), and *Sema4d* (Li *et al.*, 2018). These regions are bound by PU.1 and C/EBPβ in mouse BMDMs and are acetylated at H3K27 in both mouse BMDMs and human monocytes. A standard alignment of the human genome using 300-bp sequences from the mouse genome resulted in homologies that contain neither PU.1 nor C/EBPβ motif (**Figure 5C; Figure 5—figure supplement 4**). In contrast, lenient gap penalties captured much more extended regions, containing high-score motifs of both PU.1 and C/EBPβ. The motif sequences of PU.1 and C/EBPβ are well preserved between human and mouse, but the motif spacing is altered by 1-6 bp, further supporting the general tolerance of spacing alterations.

Discussion

We investigated the global dependencies of collaborative TFs on spacing, using LDTFs and SDTFs of macrophages as the study model. PU.1 and C/EBPβ demonstrated a preference for proximal motif spacing at their co-binding sites, but this preference for proximal spacing is not a strong modifier of TF binding in comparison to the high correlation between motif scores and TF binding activities. By leveraging natural genetic variation across genetically diverse strains of mice, we revealed the effects of spacing alterations and motif mutations on TF binding and function. InDels that alter spacing between PU.1 and C/EBPβ motifs were associated with a smaller, but significant, change of TF binding and histone acetylation compared to motif mutations. However, by excluding InDels that potentially affect motifs of other collaborative factors identified by our newly developed framework, we observed an insignificant effect of the remaining sites. This finding suggests that the significant effects observed for InDels at some sites are very likely due to the motif mutations of other collaborative factors instead of spacing alterations between PU.1 and C/EBPβ motifs. This result is consistent with the slope seen in the spacing distributions of PU.1 and C/EBPβ at their co-binding sites (**Figure 1C**). For example, an InDel resulting in a change in spacing from 20 bp to 30 bp would still place the motifs well within the range of collaborative interactions. Similar relationships were observed for PU.1 and cJun, and for PU.1 and p65. Although these relationships are likely to be general, studies of additional LDTFs and SDTFs in other cell types will be required to establish this point.

These findings provide evidence that a subset of transcriptional regulatory elements does not require strict spacing relationships between transcription factors, in contrast to the examples provided by functional and structural studies of the interferon-β enhanceosome (Panne, 2008) and demonstrated in vivo in the case of synthetically modified enhancer elements in Ciona

293 (Farley *et al.*, 2015). However, these two examples represent regulatory elements in which
294 key TF motifs are tightly spaced in their native contexts (i.e., 6-13 bp between motif centers).
295 Direct protein-protein interactions are observed between bound TFs at the interferon-β
296 enhanceosome, analogous to interactions defined for cooperative TFs that form ternary
297 complexes (Morgunova and Taipale, 2017, Reményi *et al.*, 2003). Insertions or deletions
298 between these tightly spaced motifs may result in sequence alterations as well as the potential
299 for steric inhibition of DNA binding. Consistent with this, spacing distributions for most
300 collaborative TFs exhibit a discontinuity at spacings of less than 12 bp between motif centers
301 due to overlap of their sequences (**Figure 1C**). The present studies were thus not able to
302 distinguish effects of spacing from effects of motif mutations below this motif distance
303 threshold.
304
305 Another question raised by the discrepancy between the spacing dependencies discovered by
306 previous studies and the spacing tolerance concluded by the present studies is the relative
307 proportion of regulatory elements overall in which strict spacing relationships have functional
308 importance. The current studies are limited by the ~5 million InDels provided by five strains
309 of mice. Of the approximately 14,000 genomic locations co-bound by PU.1 and C/EBPβ and
310 associated with local histone acetylation, informative InDels to test for impact of spacing
311 (i.e., between PU.1 and C/EBPβ motifs, not affecting other collaborative TF motifs, and not
312 complicated by other variants) were present at ~300 sites, representing ~2% of these regions.
313 While this set of genomic locations enabled clear conclusions based on comparisons to ~4000
314 variant free sites, the extent to which this set of binding sites is representative of all
315 regulatory elements is unclear. In particular, the interferon-β enhancer is among many
316 regulatory elements that have no InDels across the five mouse strains examined. It thus
317 remains possible that a subset of enhancers is dependent on strict spacing relationships.
318
319 Regardless of the extent of potential spacing-dependent regulatory elements, the present
320 studies provide strong evidence that naturally occurring alterations in spacing between TF
321 binding sites within putative regulatory elements are generally well tolerated. The
322 conclusions are likely transferrable to explain the effects of InDels observed in human
323 genomes, considering the similar number and size of InDels observed in human population
324 (Mills *et al.*, 2011). To leverage an additional source of genetic variation, we compared the
325 regulatory elements of mouse macrophages lacking InDels to human genomic sequences.
326 Standard gap penalties generally resulted in short sequence fragments, whereas more lenient
327 penalties recovered extended regions of homology containing corresponding LDTF motifs.
328 These findings support that InDels are tolerated by a large fraction of regulatory elements and
329 provide a basis for decreasing gap penalties for sequence comparisons of putative regulatory
330 elements across species. Nevertheless, these studies rely on natural genetic variation, which is
331 subject to natural selection. It will therefore be of interest to systematically introduce variable
332 sizes of InDels between LDTFs in representative variant free enhancers to obtain an unbiased
333 answer to the generality of the tolerance of spacing alterations.
334

335 Materials and Methods

336 Sequencing data processing

337 The mouse sequencing data used in this study were downloaded from the GEO database with
338 accession number GSE109965 (Link *et al.*, 2018a). We mapped the ChIP-seq reads using
339 Bowtie2 v2.3.5.1 (Langmead and Salzberg, 2012) and mapped the RNA-seq reads using
340 STAR v2.5.3a (Dobin *et al*., 2013) all with default parameters. Data from C57BL/6J mice
341 were mapped to mm10 genome. Reads from BALB/cJ, NOD/ShiLtJ, PWK/PhJ, and
342 SPRET/EiJ were mapped to their respective genomes built by MMARGE v1.0 with default

343 variant filters and were then shifted to mm10 genome using MMARGE v1.0 "shift" function
344 (Link *et al.*, 2018b) to facilitate comparison at homologous regions. The reproducible TF
345 binding sites were identified from mapped ChIP-seq data by first using HOMER v4.9.1
346 (Heinz *et al.*, 2010) to call unfiltered 300-bp peaks (command "findPeaks -style factor -L 0 -
347 C 0 -fdr 0.9 -size 200") and then running IDR v2.0.3 (Li *et al*., 2011) on replicates with
348 default parameters. Gene expression was quantified by TPM to represent normalized RNA-
349 seq reads mapped to exons using HOMER v4.9.1 (command "analyzeRepeats.pl rna mm10 -
350 count exons -condenseGenes -tpm"). Activity of TF binding was quantified by the number of
351 TF ChIP-seq reads within 300-bp TF binding sites normalized by library size using HOMER
352 v4.9.1 (command "annotatePeaks.pl mm10 -norm 1e7"). Activity of promoter and enhancer
353 was quantified by normalized H3K27ac ChIP-seq reads within extended 1000-bp regions
354 around TF binding sites.
355
356 **Motif score and motif spacing calculation**
357 We extracted the DNA sequences of TF binding sites from the genomes of different mouse
358 strains using the MMARGE v1.0 "extract_sequences" function (Link *et al.*, 2018b). Based on
359 DNA sequences, we computed motif scores and identified TF binding motifs as previously
360 described (Shen *et al.*, 2020). Generally, we first calculated dot products between position
361 weight matrices (PWMs) and sequence vectors using Biopython package (Cock *et al.*, 2009).
362 PWMs for PU.1, C/EBPβ, and over 500 other TFs were obtained from the JASPAR
363 vertebrate core database (Fornes *et al.*, 2020). Then the highest score for each PWM and its
364 position across 300 bp were recorded to represent the entire sequence. Changes of motif
365 scores were computed between the highest motif scores in two compared strains at the same
366 regions. To obtain the confident binding positions of the measured TFs, we excluded TF
367 binding sites whose corresponding motifs are larger than 40 bp away from the peak centers or
368 have a score lower than zero (i.e., less likely to occur than random chance). Approximately
369 70% of total peaks passed these criteria for both PU.1 and C/EBPβ. Motif spacing was
370 calculated from center of one motif to another, but only for sites whose highest motif scores
371 are greater than zero.
372
373 **Background sequence generation**
374 We generated background sequences by shuffling the sequences of TF binding sites in a unit
375 of dimers, which can well preserve the GC content. We then manually replaced the central
376 part of each background sequence with a TF binding motif by sampling nucleotides based on
377 the probabilities in its PWM. Motif score and motif spacing were calculated in the same way
378 for these shuffled sequences as for the TF binding sites.
379
380 **Categorization of regions based on genetic variation**
381 To investigate the effects of genetic variation, we separated the PU.1 and C/EBPβ co-binding
382 sites into four categories. "Mutated PU.1" and "Mutated CEBPB" include sites with variants
383 that change the motif scores of PU.1 and C/EBPβ motifs, respectively. "Altered spacing"
384 category includes sites where InDels exist between PU.1 and CEBPB motifs, which are not
385 altered by any other variant. Co-binding sites classified into these three categories all
386 experience a single impact from their local variants (either altered spacing or mutated motif,
387 not both) so that the effect size can be clearly traced. "Variant free" is the control category,
388 which contains sites with no genetic variation. The information about genetic variation across
389 mouse strains were extracted using MMARGE v1.0 "mutation_info" function (Link *et al.*,
390 2018b).
391

392 **Statistical testing of effect size**
393 Effect size of genetic variation was computed by the ratio of ChIP-seq read counts between
394 two compared strains followed by log2 transformation. We conducted permutation tests with
395 10,000 iterations to compare the absolute log ratios between "Variant free" and other
396 categories. During every iteration, we randomly selected a comparable size of regions from
397 the "Variant free" category and computed the mean of the selected set. Based on 10,000
398 mean values, we generated the null distribution and computed the percentile of the mean
399 from the testing category on the null distribution as p-value. We also obtained the Cohen's *d*
400 between the sampled variant-free set and the testing category as the effect size (Sullivan and
401 Feinn, 2012) and summarized the 90% confidence interval from 10,000 *d* values.
402
403 **Identification of collaborative factors based on motif score and spacing**
404 The TF binding sites identified from ChIP-seq data were first centered around the
405 corresponding motif based on the highest motif score within 300-bp regions. Again, we
406 filtered out those with motif score below zero or motif located more than 40 bp away from
407 peak center. Next, we searched for the motifs of other TFs within ±150 bp. If the motif has a
408 score greater than zero and does not overlap with the motif of the bound TF, we compute the
409 distance from motif center to region center (i.e., center of the bound TF motif) and obtain a
410 predicted spacing distribution by aggregating all the distances for each motif. The predicted
411 distribution is further smoothed by a sliding average window of 8 bp for visualization. Each
412 spacing distribution is compared to the distribution obtained from background sequences with
413 the Kolmogorov-Smirnov test (KS test) using Scipy package (Virtanen *et al.*, 2019). We
414 conducted KS test for both halves of the distribution, upstream and downstream, generating
415 two p-values for each motif. The mean p-values are used to represent the significance of
416 dissimilarity from background distribution. Additionally, we gave signs to the p-values
417 depending on whether more distances occur within or beyond 75 bp. Positive sign shows a
418 preference for close spacing while negative sign represents inhibition of close spacing.
419 Collaborative factors are predicted to have preference for close spacing. During the analyses
420 of PU.1 and C/EBPβ binding, the twelve predicted collaborative TFs used to filter for InDels
421 include IRF3, E2F6, SP1, ATF4, USF family (USF1, USF2), ETS family (ELF4, ETV6,
422 ELK4), and AP-1 family (FOS::JUN, FOSL2::JUN, JDP2), while the unrelated factors used
423 as controls include EGR1, OLIG1, NEUROD2, STAT1, KLF13, CTCF, and BARHL2.
424 During the analyses of PU.1 and p65 binding, ten out of the twelve predicted collaborative
425 TFs were used after excluding USF family, which was only predicted to be collaborators of
426 C/EBPβ.
427
428 **Sequence alignment between mouse and human**
429 Among 3,917 variant-free PU.1 and C/EBPβ co-binding sites merged from C57 and PWK,
430 we quantified the H3K27ac level within the extended 1,000-bp regions and set a cutoff of
431 H3K27ac ChIP-seq reads at 16 to obtain active regulatory elements. We extracted 300-bp
432 sequences of these co-binding sites from the mm10 genome and aligned them to the hg38
433 genome using BLASTn algorithm (Boratyn *et al.*, 2013). Except for the different gap
434 penalties ("Gap Costs" on the BLAST web interface) tested in our studies, the other
435 parameters were used as default settings.
436
437 **Acknowledgements**
438 We would like to thank Leslie Van Ael for assistance with manuscript preparation.

Competing interests

None declared

Figure 1. Spacing relationship of PU.1 and C/EBPβ. (A) Schematic of the collaborative binding model between PU.1 and C/EBPβ, which recognize their own motifs spaced in macrophage-specific enhancers. (B) Numbers of singly binding and co-binding sites of PU.1 and C/EBPβ identified from ChIP-seq data. (C) Distributions of C/EBPβ motif around PU.1 binding sites. The distributions for non-overlapping sites (spacing > 12 bp) of each category were compared against the background distribution generated from shuffled sequences using Kolmogorov–Smirnov test (KS test). P-values from KS test are displayed in brackets. The spacing distributions were smoothed by an 8-bp sliding window for visualization purpose. (D) Hexbin plots showing the correlation between TF binding activity and motif spacing or motif score for the 9849 co-binding sites. Log2 ChIP-seq reads were calculated within 300 bp to quantify the binding activity of PU.1 and C/EBPβ. The color gradients represent the density of sites. Spearman correlation coefficients together with p-values are displayed to show the level of correlation.

The following figure supplements are available for figure 1.

Figure supplement 1. Spacing relationship of PU.1 and C/EBPβ. (A) Spacing distributions of PU.1 motif around C/EBPβ motif at co-binding sites and C/EBPβ-singly-binding sites. P-values display the comparison against the background distribution using KS tests. (B) Spacing distributions regarding different orientation of the motifs. Co-binding sites and PU.1-singly-binding sites were divided into two subgroups representing same or opposite orientation of the PU.1 and C/EBPβ motifs. The overall distributions are very similar for both subgroups.

Figure 2. Effects of spacing alterations resulting from natural genetic variation across mouse strains. (A) Schematic showing impacts of genetic variation on motif sequence or motif spacing. PU.1 and C/EBPβ co-binding sites can be classified into four categories based on the impacts of local variants: “altered spacing”, “mutated PU.1”, “mutated C/EBPβ”, and “variant free”. (B, C) Absolute log2 fold changes of ChIP-seq reads between C57 and PWK for (B) PU.1 binding and (C) C/EBPβ binding. Boxplot shows the median and quartiles of every distribution with its sample size displayed on top. (*) indicates a significant effect size with $p < 0.001$ from permutation tests compared against the "variant free" category (Materials and Methods). (D) Correlations between change of C/EBPβ binding and change of motif spacing or motif score. The co-binding sites used for change of spacing, PU.1 motif score, and C/EBPβ motif score are from the previously defined categories “altered spacing”, “mutated PU.1”, “mutated C/EBPβ”, respectively.

The following figure supplements are available for figure 2.

Figure supplement 1. Change of PU.1 and C/EBPβ binding affected by genetic variation for the other three pairwise comparisons. (*) indicates significance value $p < 0.001$ based on permutation test of every category against "variant free" category. The results from C57 vs. BALB and C57 vs. NOD are similar to what we saw for C57 vs. PWK. C57-SPRET comparison did not show significant results for “altered spacing” category, likely due to much more genetic variants between these two strains than other pairs, which introduced stronger trans effects to the “variant-free” category making the baseline effects high and potentially complicating the effects from InDels altering motif spacing.

Figure supplement 2. Example sites of motif mutation and spacing alteration. (A) a 5-bp increase in spacing has little effect on TF binding. (B) an A-to-G mutation on PU.1 motif yields a loss of both PU.1 and C/EBPβ binding.

Figure supplement 3. Correlations between change of PU.1 binding and change in spacing or motif score. The co-binding sites here are from the C57-PWK comparison and are the same as those in Figure 2. Motif scores showed high correlation, while scale of spacing alteration is not associated with change of PU.1 binding.

493 **Figure supplement 4**. Effect size of genetic variation in relation with the initial spacing
494 between PU.1 and C/EBPβ motif. Co-binding sites from C57-PWK comparison are binned
495 based on the initial motif spacing and then used to calculate the absolute log2 fold change
496 between the two strains, which were aggregated to compute mean values for each bin. The
497 effect size of InDels altering spacing is overall not affected by the initial motif spacing.
498
499 **Figure 3.** Refining InDels to exclude those potentially mutating motifs of other collaborative
500 factors. (A) Overview of our newly developed framework for identifying collaborative
501 factors from single ChIP-seq data. Given the binding sites for TF of interest, our method
502 searches for other motifs and uses regions with high-score motifs to compute the spacing
503 distribution, which is further compared against the background distribution using KS test.
504 Those with significant proximal distribution are predicted collaborative factors. (B)
505 Comparison between the actual spacing relationship obtained from co-binding sites and the
506 predicted spacing distribution of cJun and PU.1. P-values from KS test by comparing to the
507 background distribution are shown in brackets. (C) Signed p-values of over five hundred
508 motifs for PU.1 and C/EBPβ binding sites. Color gradients indicate the level of gene
509 expression measured by RNA-seq and quantified by TPM. The complete list of p-values is
510 available in Figure 3—source data 1. (D) Effect size of a refined set of PU.1 and C/EBPβ co-
511 binding sites for C57-PWK comparison. About half of the original "altered spacing" sites
512 were excluded due to their impacts on at least one of the twelve predicted collaborative factor
513 motifs. (*) indicates $p < 0.001$ based on permutation tests against the "variant free" category.
514 (E) Effect size of refined PU.1 and p65 co-binding sites for C57-PWK comparison. "Altered
515 spacing" category has excluded InDels that impacts motifs of the collaborative factors of
516 PU.1 identified from our framework.
517 The following figure supplements are available for figure 3.
518 **Figure supplement 1**. Fractions of recovered co-binding sites by filtering with different
519 motif score thresholds. (A) PU.1 binding sites identified from PU.1 ChIP-seq data were
520 filtered with different thresholds on C/EBPβ motif. (B) C/EBPβ binding sites identified from
521 C/EBPβ ChIP-seq data were filtered with different thresholds on PU.1 motif. Both
522 demonstrated an increase in fraction of co-binding sites by a larger threshold.
523 **Figure supplement 2**. Predicted spacing distributions of PU.1 and C/EBPβ. Recovered from
524 (A) PU.1 binding sites with top C/EBPβ motif (CEBPB), and (B) C/EBPβ binding sites with
525 top PU.1 motif (SPI1). Both predicted distributions are similar to the spacing distribution
526 obtained from the actual co-binding sites identified from PU.1 and C/EBPβ ChIP-seq data.
527 **Figure supplement 3**. Examples of predicted spacing distributions. (A) PU.1 and GFI1 as an
528 example of no spacing relationship, and (B) PU.1 and ZEB1 as an example of distant spacing
529 relationship. P-values shown in brackets are obtained from KS tests by comparing to the
530 background distribution (shuffled sequences) without assigning signs to distinguish proximal
531 and distant spacing relationship.
532 **Figure supplement 4**. Fold changes of TF binding after filtering out mutations on non-
533 collaborative factors. The remaining sites in "Altered spacing" category still have a
534 significant effect on TF binding based on permutation tests ($p < 0.001$).
535 **Figure supplement 5**. Spacing relationship of PU.1 and p65 in mouse macrophages at pro-
536 inflammatory state induced by KLA treatment for 1 hour. Co-binding sites show clear
537 preference for PU.1 and p65 motifs to be proximal, while p65-singly-binding sites do not
538 have the same preference. The distributions exclude sites where PU.1 and p65 motifs overlap
539 with a shift of 3 or 4 bp (overlapping "GGAA"/"TTCC").
540 **Figure supplement 6**. Fold changes of TF binding for four categories of PU.1 and p65 co-
541 binding sites. "Altered spacing" includes all co-binding sites where InDels occur between
542 PU.1 and p65 motifs and alter the motif spacing without considering any impact on motifs of

543 other collaborative factors. (*) indicates significance value $p < 0.001$ based on permutation
544 test of every category against "variant free" category.
545 The following source data are available for figure 3.
546 **Source data 1**. Complete list of signed p-values indicating predicted spacing relationships.
547
548 **Figure 4.** Effects of spacing alteration on promoter and enhancer activity measured by local
549 histone acetylation. (A, B) Absolute log2 fold changes of H3K27ac ChIP-seq reads between
550 C57 and PWK for (A) unfiltered co-binding sites and (B) refined co-binding sites after
551 excluding InDels that mutate motifs of potential collaborative factors. (*) indicates $p < 0.001$
552 based on permutation tests against the "variant free" category. (C) Correlations between
553 change of H3K27ac level and change of motif score or motif spacing. The co-binding sites
554 used here are unfiltered.
555 The following figure supplements are available for figure 4.
556 **Figure supplement 1**. Results from the other three pairwise comparisons on the change of
557 H3K27ac level for four categories of PU.1 and C/EBPβ co-binding sites. (*) indicates
558 significance value $p < 0.001$ based on permutation test of every category against "variant
559 free" category. Again, C57-SPRET comparison did not show significant results for "altered
560 spacing" category, likely due to the much larger genetic diversity between these two strains,
561 which complicates the effects from InDels altering motif spacing with trans effects from the
562 variants nearby.
563 **Figure supplement 2**. Results from the other three pairwise comparisons on the change of
564 H3K27ac level after filtering out InDels in "Altered spacing" category that impact motifs of
565 predicted collaborative factors. "Altered spacing" category no longer shows a significant
566 effect on the acetylation level.
567 **Figure supplement 3**. Change of H3K27ac level affected by genetic variation after filtering
568 out InDels that mutate motifs of non-collaborative factors. The remaining sites in "Altered
569 spacing" category still have a significant effect on local acetylation of H3K27 based on
570 permutation tests ($p < 0.001$).
571 **Figure supplement 4**. Correlation between change of TF binding and change of H3K27ac
572 level at all PU.1 and C/EBPβ co-binding sites. Fold changes were calculated by dividing C57
573 by PWK. The larger fold change between PU.1 and C/EBPβ binding was used for plotting.
574 Overall, a strong correlation exists between H3K27ac level and TF binding, represented by a
575 Pearson correlation of 0.7.
576
577 **Figure 5.** Implications of reducing gap penalties in cross species sequence alignments. (A)
578 Lengths of aligned sequences for the co-binding sites of PU.1 and C/EBPβ in mouse
579 BMDMs that are enriched with local H3K27ac and have no genetic variation between C57
580 and PWK. 300-bp sequences were aligned to the human genome using BLAST with either
581 the standard or the lenient gap penalties. The complete list of regions together with alignment
582 results are available in Figure 5—source data 1. (B) Example showing a co-binding site
583 within intron of *Anxa7* that was successfully aligned to a homology region in human. The
584 aligned regions are enriched with H3K27ac in both mouse BMDMs and human monocytes.
585 (C) The alignment results of the example co-binding site using standard or lenient gap
586 penalties. Lenient gap penalties resulted in the recovery of a PU.1 motif and a C/EBPβ motif.
587 The spacing between PU.1 and C/EBPβ motif centers is 39 bp in mouse and 42 bp in human,
588 which did not impact the binding of PU.1 or the activity of this region.
589 The following figure supplements are available for figure 5.
590 **Figure supplement 1**. Alignment results of all PU.1 and C/EBPβ co-binding sites in mouse
591 macrophages compared with human genome. All the co-binding sites were identified as 300-
592 bp regions in mouse genome and compared to the human genome using BLAST. Lenient gap

593 penalties resulted in longer alignments than standard parameters. Despite that, the majority of
594 these co-binding sites have less than a third of the complete sequences aligned. By comparing
595 to Figure 5A, our results suggest that acetylated co-binding sites are much more conserved
596 between mouse and human than the rest with less acetylation and potentially less functional
597 importance.
598 **Figure supplement 2**. Alignment results of PU.1 and C/EBPβ co-binding sites that are
599 enriched with local H3K27ac and have no genetic variation between C57 and PWK using
600 other possible gap penalties. All the co-binding sites were identified as 300-bp regions in
601 mouse genome and compared to the human genome using BLAST. The lenient gap penalty
602 ({2, 2}) still produced much longer aligned sequences than other penalty options.
603 **Figure supplement 3**. Examples of PU.1 and C/EBPβ co-binding sites in BMDMs that are
604 aligned to homology regions in human using lenient gap penalties. (A) Enhancer closest to
605 *Fos*, which encodes AP-1 family transcription factor and is known to be important for
606 macrophage function. (B) Enhancer proximal to *Vmp1*, which has been found to be
607 associated with inflammatory response of macrophages. (C) Intron of *Max*, which encodes a
608 basic-helix-loop-helix-zipper protein and is found to accumulate during macrophage
609 differentiation. (D) Intron of *Sema4d*, which is found to be regulated by macrophages in
610 tumor.
611 **Figure supplement 4**. The alignment results of the example sites shown in Figure
612 supplement 3. (A) *Fos* enhancer, (B) *Vmp1* enhancer, (C) *Max* intron, and (D) *Sema4d* intron.
613 The following source data are available for figure 5.
614 **Source data 1**. List of aligned regions.
615

References

Ayer, D. E., and Eisenman, R. N. (1993). A switch from Myc: Max to Mad: Max heterocomplexes accompanies monocyte/macrophage differentiation. Genes & development, 7(11), 2110-2119.

Boeva, V. (2016). Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in Eukaryotic cells. Frontiers in Genetics, 7, 24.

Boratyn, G. M., Camacho, C., Cooper, P. S., Coulouris, G., Fong, A., Ma, N., Madden, T. L.,Matten, W. T., McGinnis, S. D., Merezhuk, Y., et al. (2013). Blast: a more efficient report with usability improvements. Nucleic acids research, 41(W1):W29–W33.

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

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., and 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.

Deplancke, B., Alpern, D., and Gardeux, V. (2016). The genetics of transcription factor DNA binding variation. Cell, 166(3):538–554.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M.,and Gingeras, T. R. (2013). Star: ultrafast universal rna-seq aligner. Bioinformatics, 29(1):15–21.

Dziuba, N., Ferguson, M. R., O'Brien, W. A., Sanchez, A., Prussia, A. J., McDonald, N. J., Friedrich, B. M., Li, G., Shaw, M. W., Sheng, J., Hodge, T. W., Rubin, D. H., Murray, J. L. (2012). Identification of cellular proteins required for replication of human immunodeficiency virus type 1. AIDS research and human retroviruses, 28(10), 1329-1339.

Farh, K. K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., Shoresh, N., Whitton, H., Ryan, R. J., Shishkin, A. A., et al. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature, 518(7539):337–343.

Farley, E. K., Olson, K. M., Zhang, W., Brandt, A. J., Rokhsar, D. S., and Levine, M. S. (2015). Suboptimization of developmental enhancers. Science, 350(6258):325–328.

Fornes, O., Castro-Mondragon, J. A., Khan, A., van der Lee, R., Zhang, X., Richmond, P. A., Modi, B. P., Correard, S., Gheorghe, M., Baranašić, D., et al. (2020). Jaspar 2020: update of the open-access database of transcription factor binding profiles. Nucleic acids research, 48(D1):D87–D92.

Grossman, S. R., Zhang, X., Wang, L., Engreitz, J., Melnikov, A., Rogov, P., Tewhey, R., Isakova, A., Deplancke, B., Bernstein, B. E., Mikkelsen, T. S., and Lander, E. S. (2017). Systematic dissection of genomic features determining transcription factor binding and enhancer function. Proceedings of the National Academy of Sciences, 114(7):E1291–E1300.

GTEx Consortium (2015). The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. Science, 348(6235):648–660.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., and 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.

Heinz, S., Romanoski, C. E., Benner, C., Allison, K. A., Kaikkonen, M. U., Orozco, L. D., and Glass, C. K. (2013). Effect of natural genetic variation on enhancer selection and function. Nature, 503:487.9

Heinz, S., Romanoski, C. E., Benner, C., and Glass, C. K. (2015). The selection and function of cell type-specific enhancers. Nature Reviews Molecular Cell Biology, 16(3):144–154.

Hop, H.T., Arayan, L.T., Huy, T.X., Reyes, A.W., Vu, S.H., Min, W., Lee, H.J., Rhee, M.H., Chang, H.H. and Kim, S., 2018. The key role of c-Fos for immune regulation and bacterial dissemination in brucella infected macrophage. Frontiers in Cellular and Infection Microbiology, 8, 287.

Jiang, P. and Singh, M. (2014). Ccat: combinatorial code analysis tool for transcriptional regulation. Nucleic acids research, 42(5):2833–2847.

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

Keane, T. M., Goodstadt, L., Danecek, 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., McIntyre, R. E., Bhomra,A., Nicod, J., Gan, X., Yuan, W., Van Der

Weyden, L., Steward, C. A., Bala, S., Stalker, J., Mott,R., Durbin, R., Jackson, I. J., Czechanski, A., Guerra-Assunçao, J., Donahue, L. R., Reinholdt,L. G., Payseur, B. A., Ponting, C. P., Birney, E., Flint, J., and Adams, D. J. (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. Nature, 477(7364):289–294.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods, 9:357.

Li, H., Huang, S., Wang, S., Zhao, J., Su, L., Zhao, B., Zhang, Y., Zhang, S., and Miao, J. (2013). Targeting annexin A7 by a small molecule suppressed the activity of phosphatidylcholine-specific phospholipase C in vascular endothelial cells and inhibited atherosclerosis in apolipoprotein E−/− mice. Cell death & disease, 4(9), e806-e806.

Li, H., Wang, J. S., Mu, L. J., Shan, K. S., Li, L. P., and Zhou, Y. B. (2018). Promotion of Sema4D expression by tumor-associated macrophages: Significance in gastric carcinoma. World journal of gastroenterology, 24(5), 593.

Li, Q., Brown, J. B., Huang, H., and Bickel, P. J. (2011). Measuring reproducibility of high-throughput experiments. Annals of Applied Statistics, 5(3):1752–1779.

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., and Glass, C. K. (2018a). Analysis of Genetically Diverse Macrophages Reveals Local and Domain-wide Mechanisms that Control Transcription Factor Binding and Function. Cell, 173(7):1796–1809.e17.

Link, V. M., Romanoski, C. E., Metzler, D., and Glass, C. K. (2018b). MMARGE: Motif mutation analysis for regulatory genomic elements. Nucleic Acids Research, 46(14):7006–7021.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon,A., Milano, A., Morales, J., et al. (2017). The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). Nucleic acids research, 45(D1):D896–D901.

Menoret, D., Santolini, M., Fernandes, I., Spokony, R., Zanet, J., Gonzalez, I., Latapie, Y., Ferrer,P., Rouault, H., White, K. P., et al. (2013). Genome-wide analyses of shaven baby target genes reveals distinct features of enhancer organization. Genome biology, 14(8):R86.

Mills, R. E., Pittard, W. S., Mullaney, J. M., Farooq, U., Creasy, T. H., Mahurkar, A. A., Kemeza, D. M., Strassler, D. S., Ponting, C. P., Webber, C., and Devine, S. E. (2011). Natural genetic variation caused by small insertions and deletions in the human genome. Genome research, 21(6), 830-839.

Morgunova, E. and Taipale, J. (2017). Structural perspective of cooperative transcription factor binding. Current Opinion in Structural Biology, 47:1–8.

Nandi, S., Blais, A., and 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.

Ng, F. S., Schütte, J., Ruau, D., Diamanti, E., Hannah, R., Kinston, S. J., and Göttgens, B. (2014). Constrained transcription factor spacing is prevalent and important for transcriptional control of mouse blood cells. Nucleic Acids Research, 42(22):13513–13524.10

Panne, D. (2008). The enhanceosome. Current Opinion in Structural Biology, 18(2):236–242.

Reiter, F., Wienerroither, S., and Stark, A. (2017). Combinatorial function of transcription factors and cofactors. Current Opinion in Genetics & Development, 43:73–81.

Reményi, A., Lins, K., Nissen, L. J., Reinbold, R., Schöler, H. R., and Wilmanns, M. (2003). Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4and Sox2 on two enhancers. Genes and Development, 17(16):2048–2059.

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

Shen, Z., Hoeksema, M., Ouyang, Z., Benner, C., and Glass, C. (2020). MAGGIE: leveraging genetic variation to identify DNA sequence motifs mediating transcription factor binding and function. bioRxiv.

Slattery, M., Zhou, T., Yang, L., Machado, A. C. D., Gordân, R., and Rohs, R. (2014). Absence of a simple code: how transcription factors read the genome. Trends in biochemical sciences,39(9):381–399.

Smith, R. P., Taher, L., Patwardhan, R. P., Kim, M. J., Inoue, F., Shendure, J., Ovcharenko, I., and Ahituv, N. (2013). Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. Nature genetics, 45(9):1021.

Stormo, G. D. (2000). DNA binding sites: representation and discovery. Bioinformatics, 16(1):16–23.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski,E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J.,Mayorov, N., Nelson, A.

R. J., Jones, E., Kern, R., Larson, E., Carey, C., Polat, , Feng, Y., Moore,E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris,C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and Contributors, S. . .(2019). SciPy 1.0–Fundamental Algorithms for Scientific Computing in Python. pages 1–22.

Sullivan, G. M., and Feinn, R. (2012). Using effect size—or why the P value is not enough. Journal of graduate medical education, 4(3), 279-282.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. (2017). 10 years of gwas discovery: biology, function, and translation. The American Journal of Human Genetics, 101(1):5–22.

Figure 1

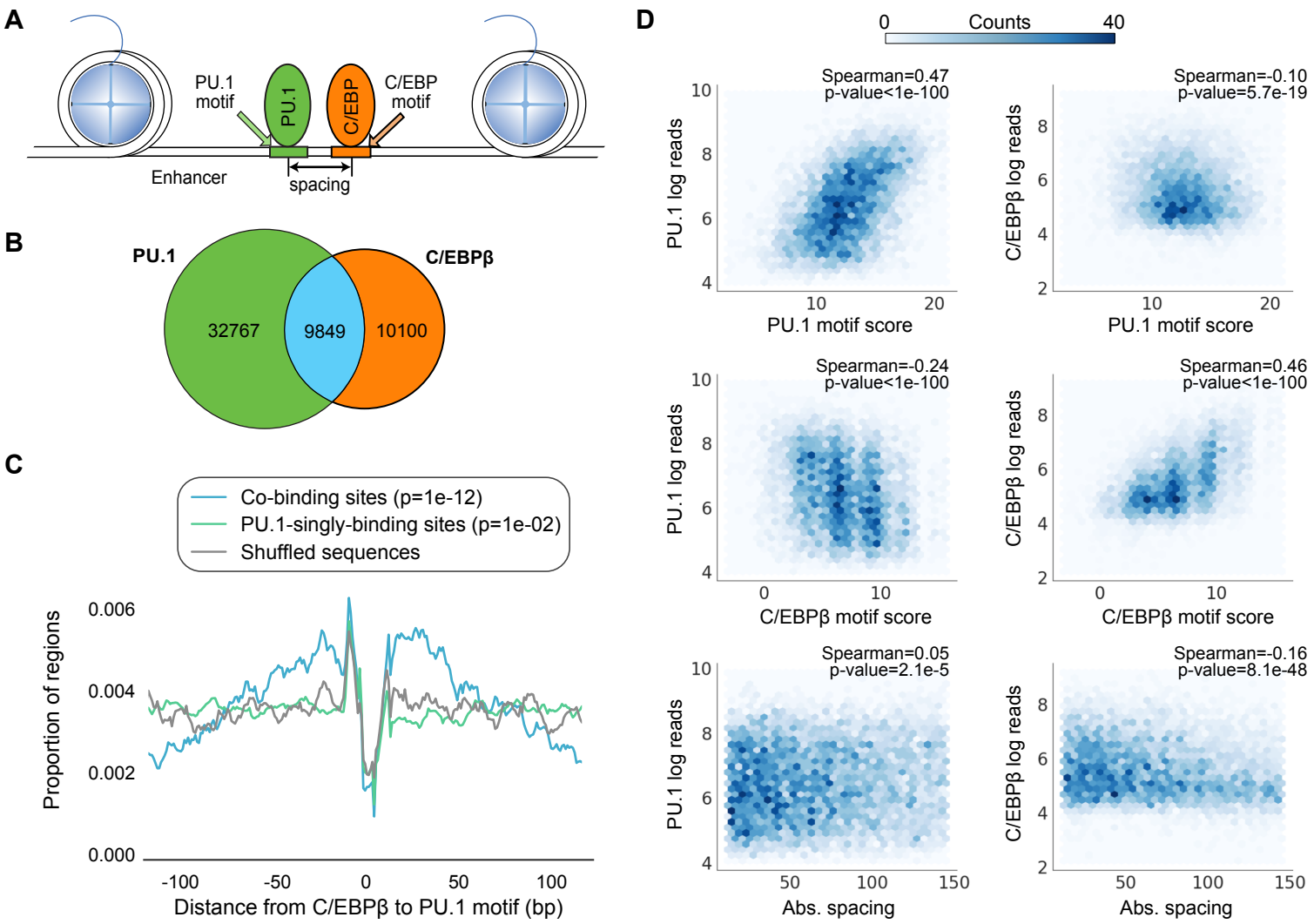
A
PU.1 motif
PU.1
C/EBP
C/EBP motif
Enhancer
spacing
B
PU.1
C/EBPβ
32767
9849
10100
C
Co-binding sites (p=1e-12)
PU.1-singly-binding sites (p=1e-02)
Shuffled sequences
Proportion of regions
0.006
0.004
0.002
0.000
-100
-50
0
50
100
Distance from C/EBPβ to PU.1 motif (bp)
D
0
Counts
40
Spearman=0.47
p-value<1e-100
PU.1 log reads
PU.1 motif score
Spearman=-0.10
p-value=5.7e-19
C/EBPβ log reads
PU.1 motif score
Spearman=-0.24
p-value<1e-100
PU.1 log reads
C/EBPβ motif score
Spearman=0.46
p-value<1e-100
C/EBPβ log reads
C/EBPβ motif score
Spearman=0.05
p-value=2.1e-5
PU.1 log reads
Abs. spacing
Spearman=-0.16
p-value=8.1e-48
C/EBPβ log reads
Abs. spacing

Figure 2

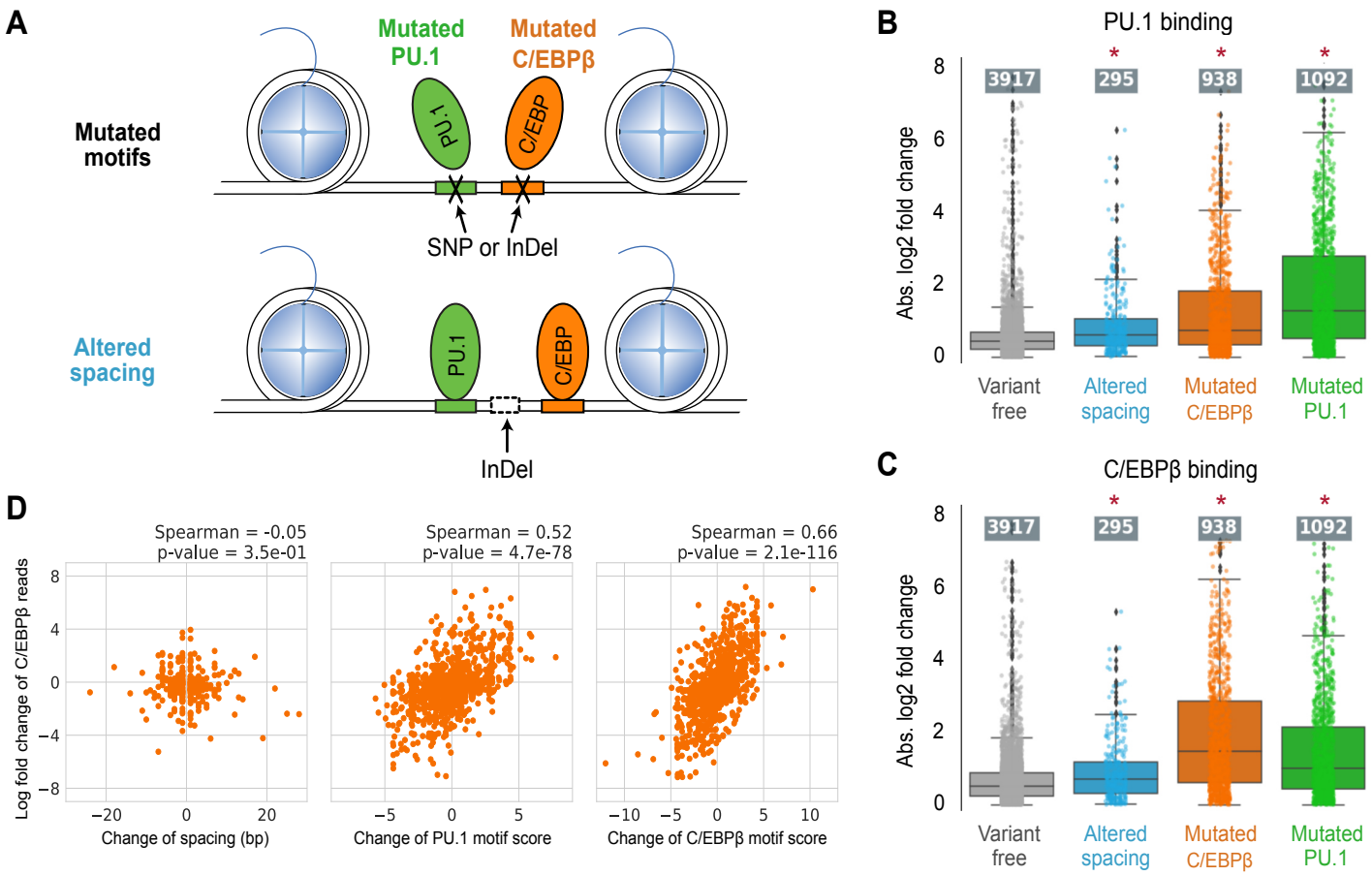
A
Mutated PU.1
Mutated C/EBPβ
Mutated motifs
PU.1
C/EBP
SNP or InDel
Altered spacing
PU.1
C/EBP
InDel
B
PU.1 binding
Abs. log2 fold change
3917
295
938
1092
Variant free
Altered spacing
Mutated C/EBPβ
Mutated PU.1
C
C/EBPβ binding
Abs. log2 fold change
3917
295
938
1092
Variant free
Altered spacing
Mutated C/EBPβ
Mutated PU.1
D
Spearman = -0.05
p-value = 3.5e-01
Spearman = 0.52
p-value = 4.7e-78
Spearman = 0.66
p-value = 2.1e-116
Log fold change of C/EBPβ reads
Change of spacing (bp)
Change of PU.1 motif score
Change of C/EBPβ motif score

Figure 3

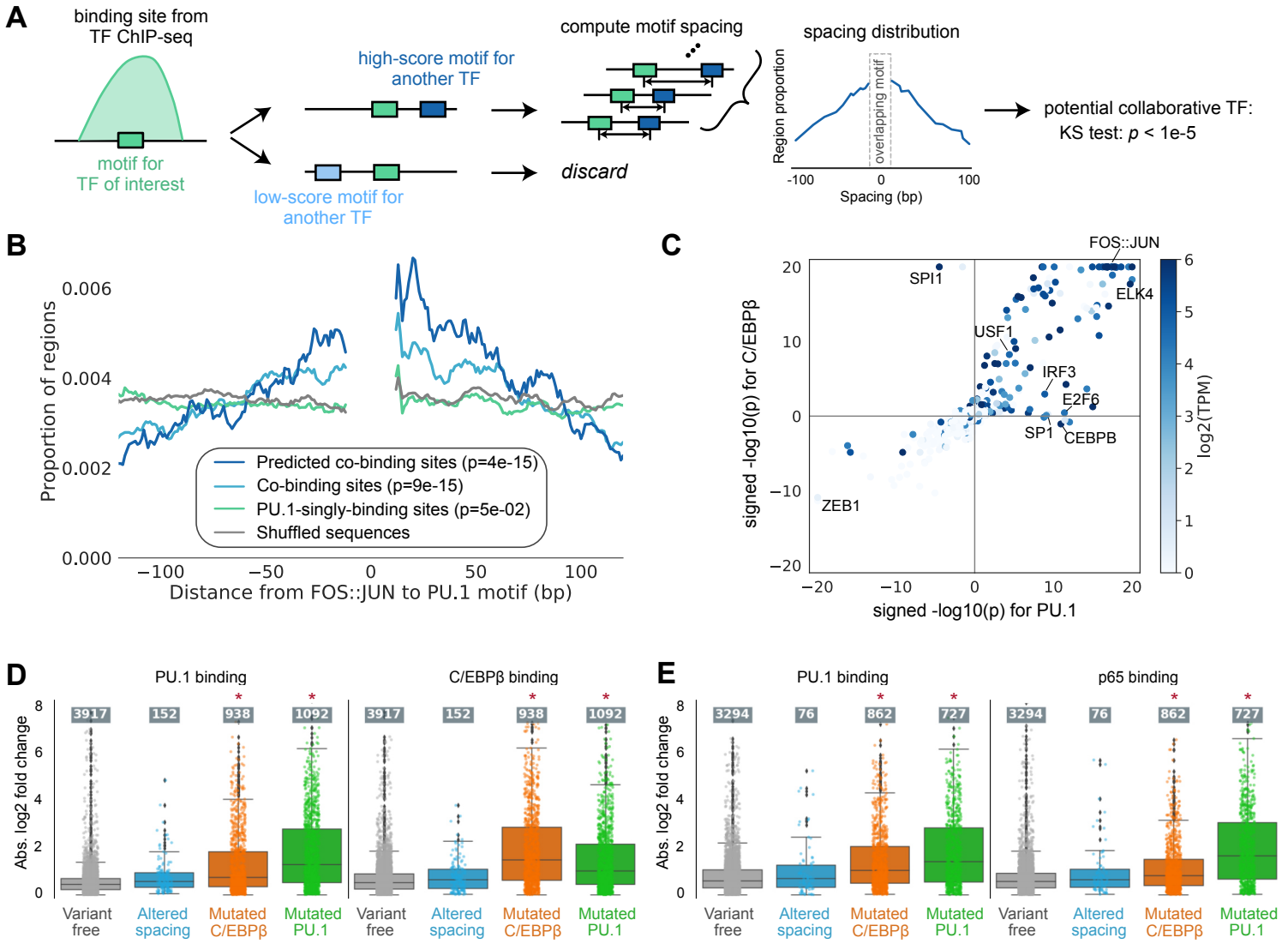
A
binding site from TF ChIP-seq
motif for TF of interest
high-score motif for another TF
low-score motif for another TF
compute motif spacing
discard
spacing distribution
Region proportion
overlapping motif
-100
0
100
Spacing (bp)
potential collaborative TF:
KS test: p < 1e-5
B
Proportion of regions
0.006
0.004
0.002
0.000
−100
−50
0
50
100
Distance from FOS::JUN to PU.1 motif (bp)
Predicted co-binding sites (p=4e-15)
Co-binding sites (p=9e-15)
PU.1-singly-binding sites (p=5e-02)
Shuffled sequences
C
FOS::JUN
SPI1
ELK4
USF1
IRF3
E2F6
SP1
CEBPB
ZEB1
signed -log10(p) for C/EBPβ
signed -log10(p) for PU.1
log2(TPM)
D
PU.1 binding
C/EBPβ binding
Abs. log2 fold change
3917
152
938
1092
Variant free
Altered spacing
Mutated C/EBPβ
Mutated PU.1
E
PU.1 binding
p65 binding
3294
76
862
727

Figure 4

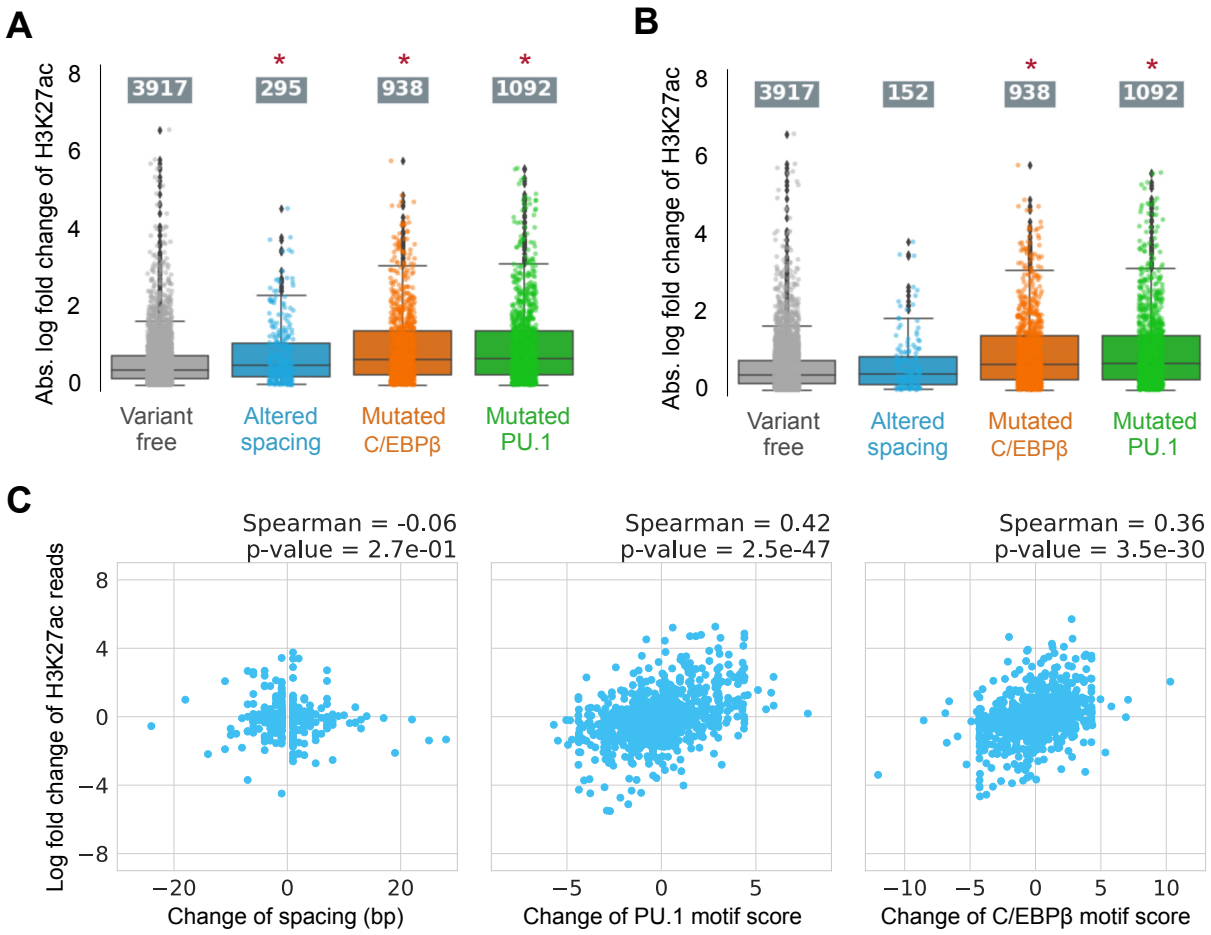
A
Abs. log fold change of H3K27ac
3917
295
938
1092
Variant free
Altered spacing
Mutated C/EBPβ
Mutated PU.1
B
Abs. log fold change of H3K27ac
3917
152
938
1092
Variant free
Altered spacing
Mutated C/EBPβ
Mutated PU.1
C
Log fold change of H3K27ac reads
Spearman = -0.06
p-value = 2.7e-01
Change of spacing (bp)
Spearman = 0.42
p-value = 2.5e-47
Change of PU.1 motif score
Spearman = 0.36
p-value = 3.5e-30
Change of C/EBPβ motif score

Figure 5

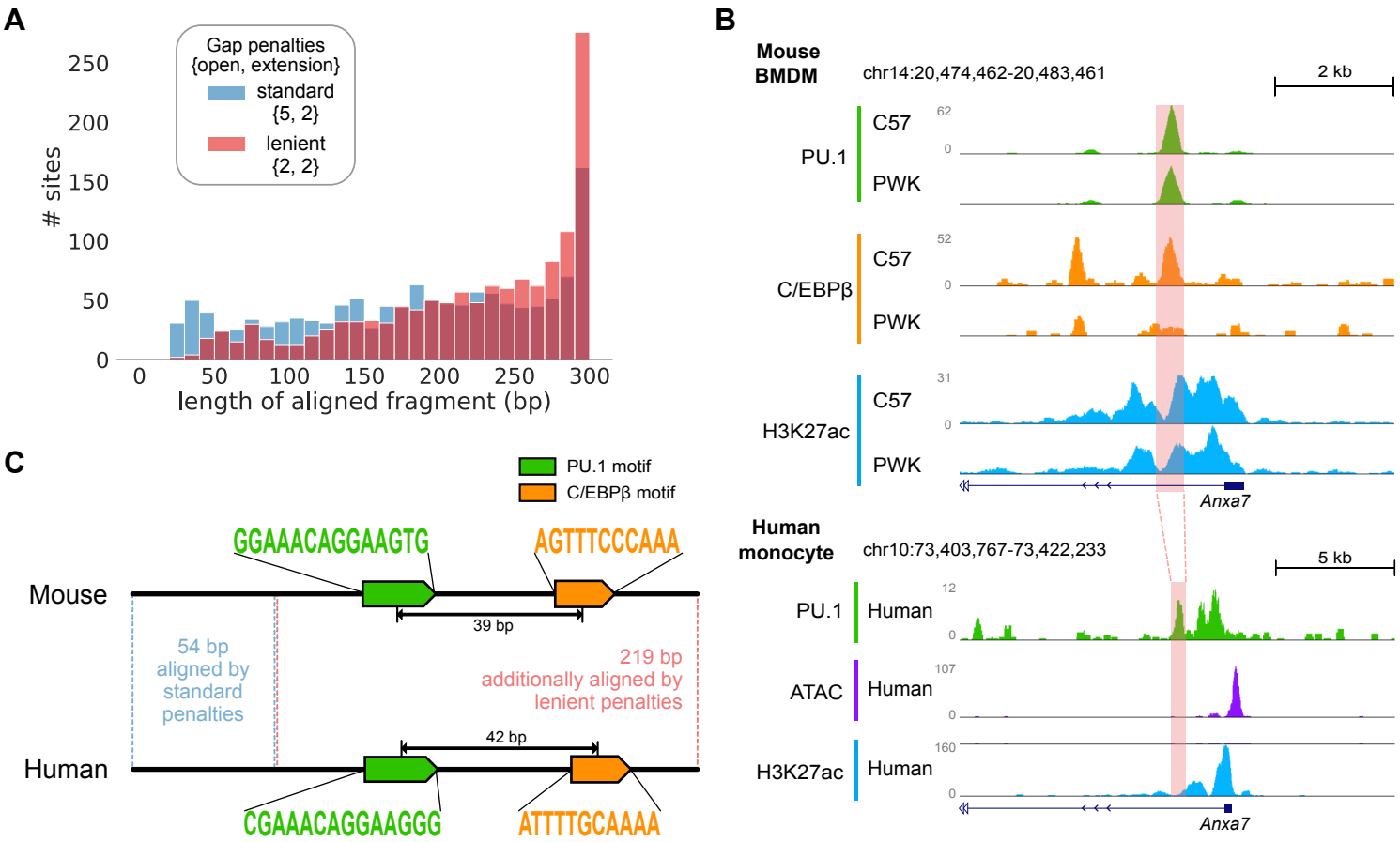
A
Gap penalties
{open, extension}
standard
{5, 2}
lenient
{2, 2}
sites
length of aligned fragment (bp)
B
Mouse
BMDM
chr14:20,474,462-20,483,461
2 kb
PU.1
C57
PWK
C/EBPβ
H3K27ac
Anxa7
Human
monocyte
chr10:73,403,767-73,422,233
5 kb
Human
ATAC
C
PU.1 motif
C/EBPβ motif
GGAAACAGGAAGTG
AGTTTCCCAAA
Mouse
39 bp
54 bp
aligned by
standard
penalties
219 bp
additionally aligned by
lenient penalties
42 bp
Human
CGAAACAGGAAGGG
ATTTTGCAAAA

Figure 1—figure supplement 1

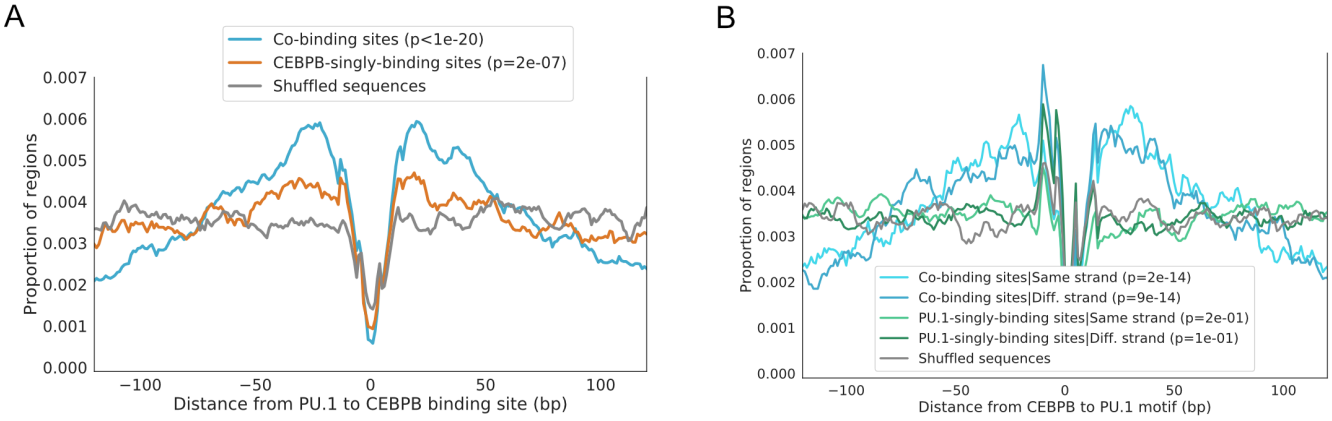

Figure 2—figure supplement 1

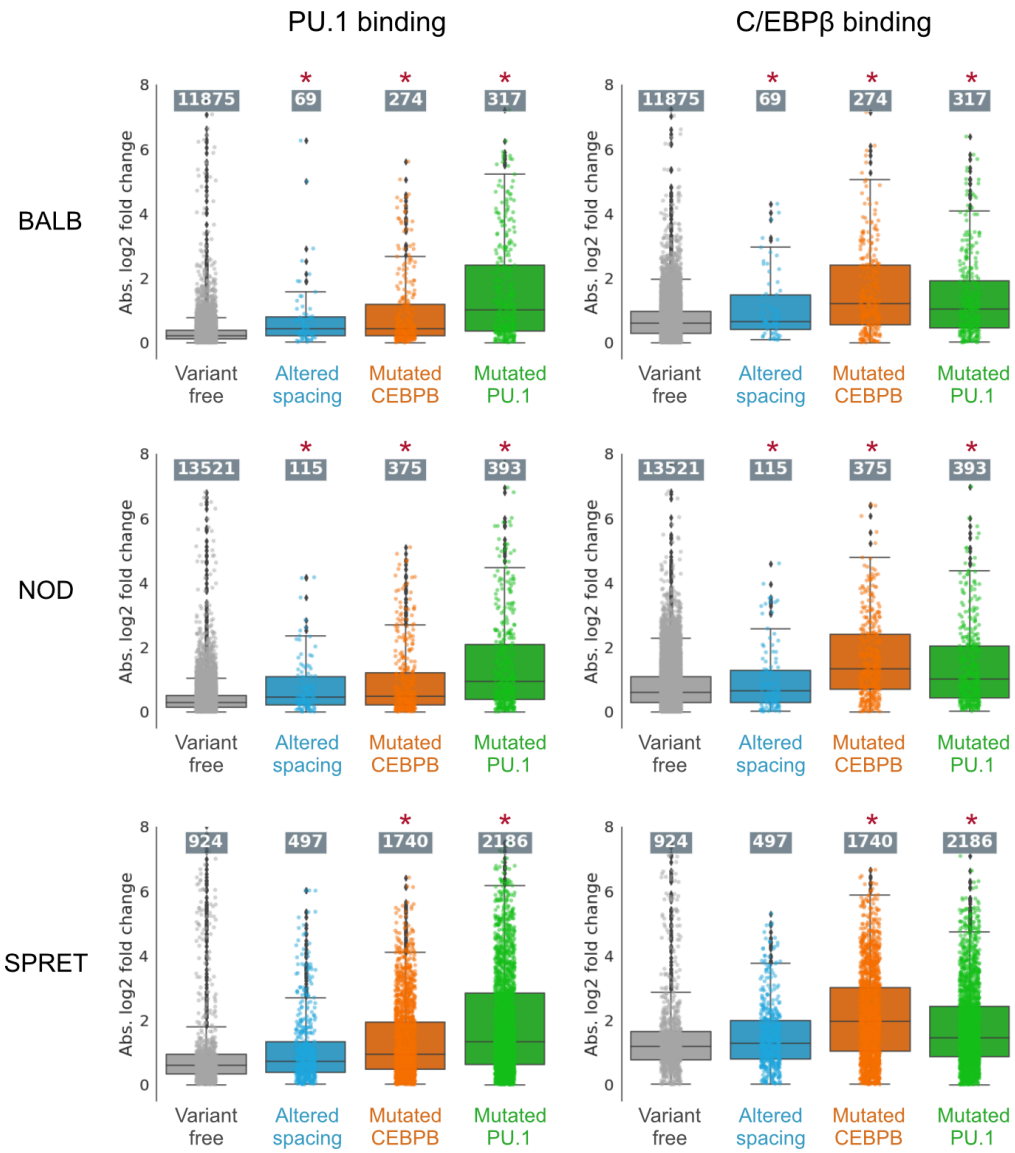

Figure 2—figure supplement 2

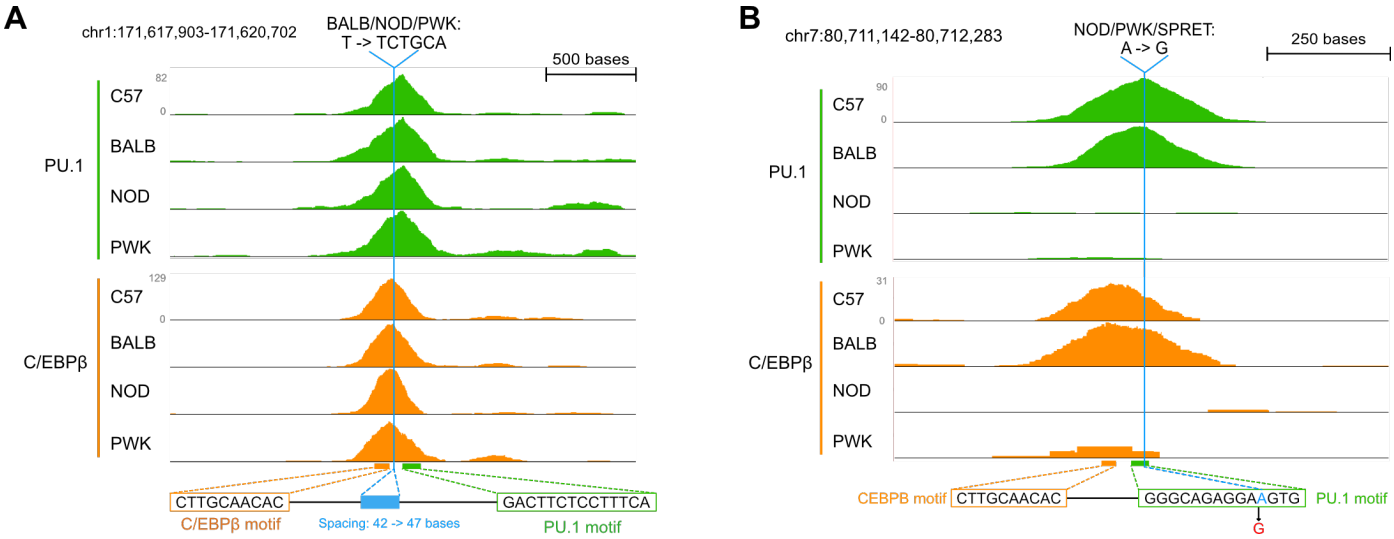

Figure 2—figure supplement 3

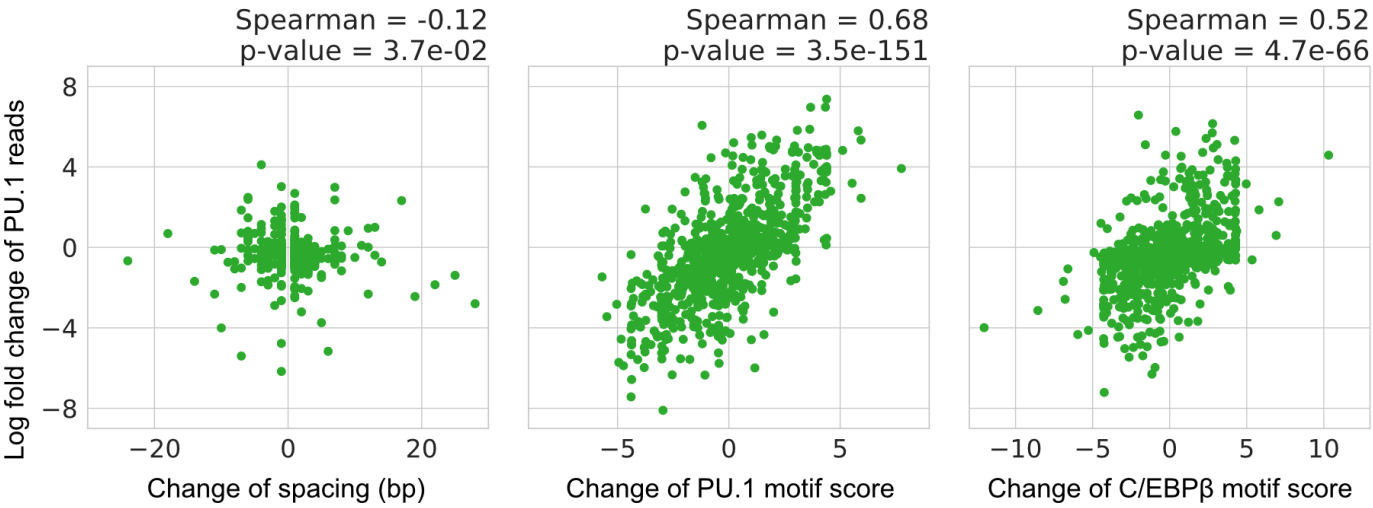

Figure 2—figure supplement 4

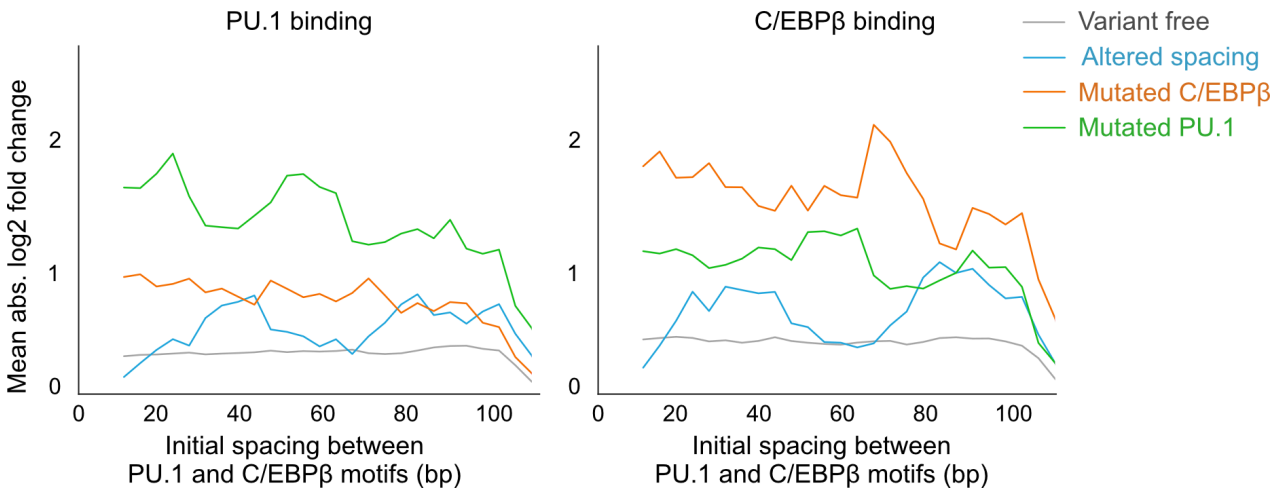

Figure 3—figure supplement 1

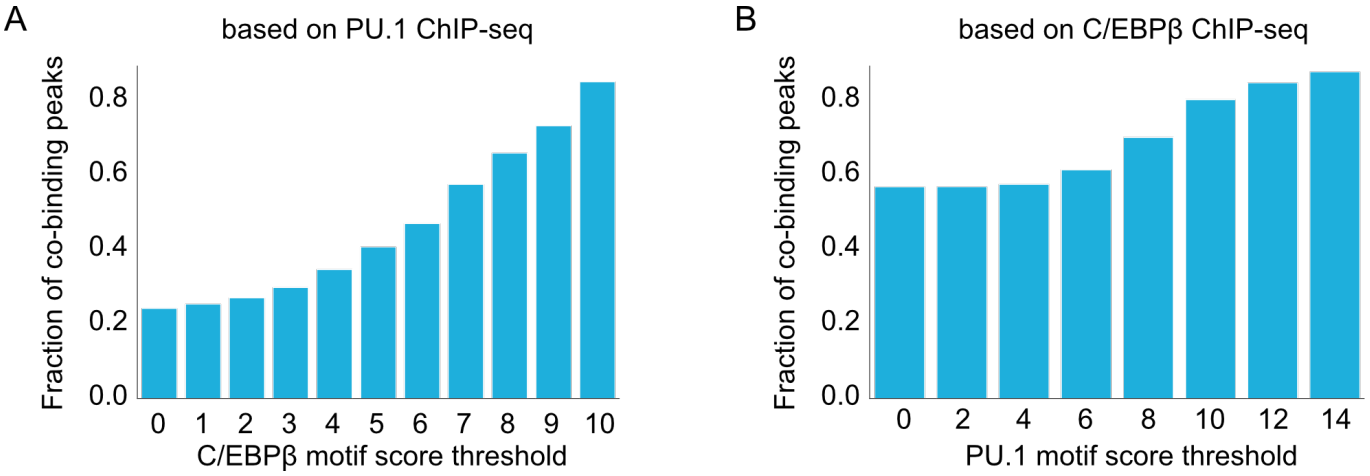

Figure 3—figure supplement 2

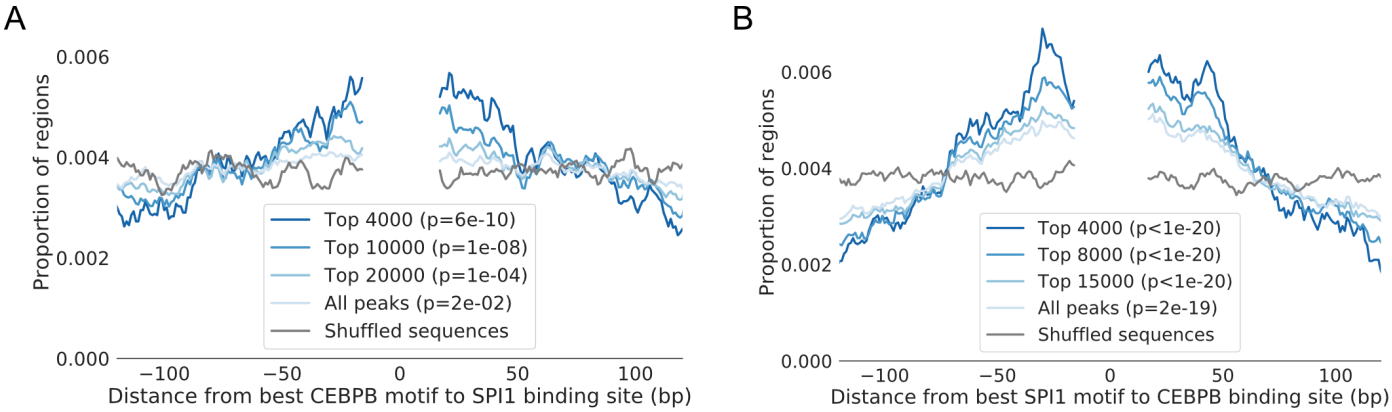

Figure 3—figure supplement 3

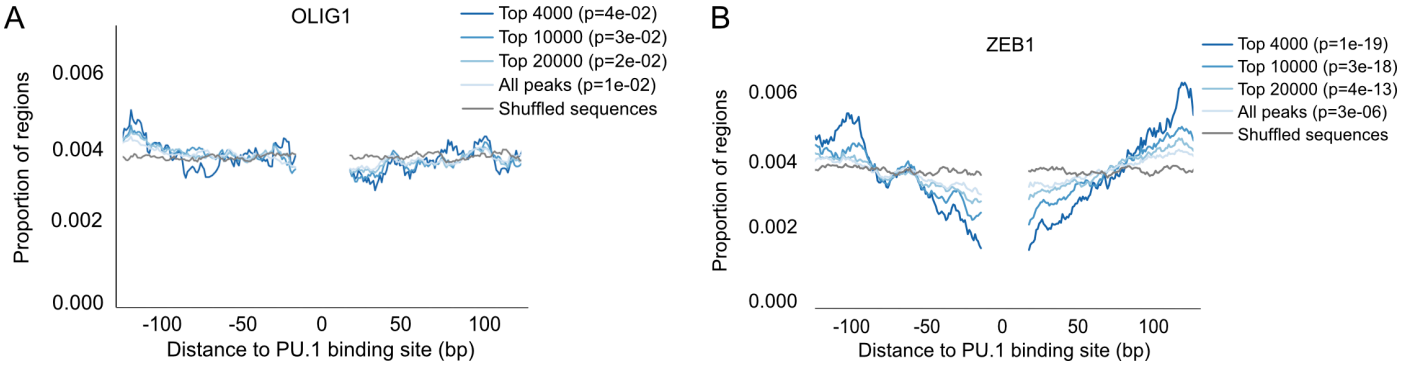

Figure 3—figure supplement 4

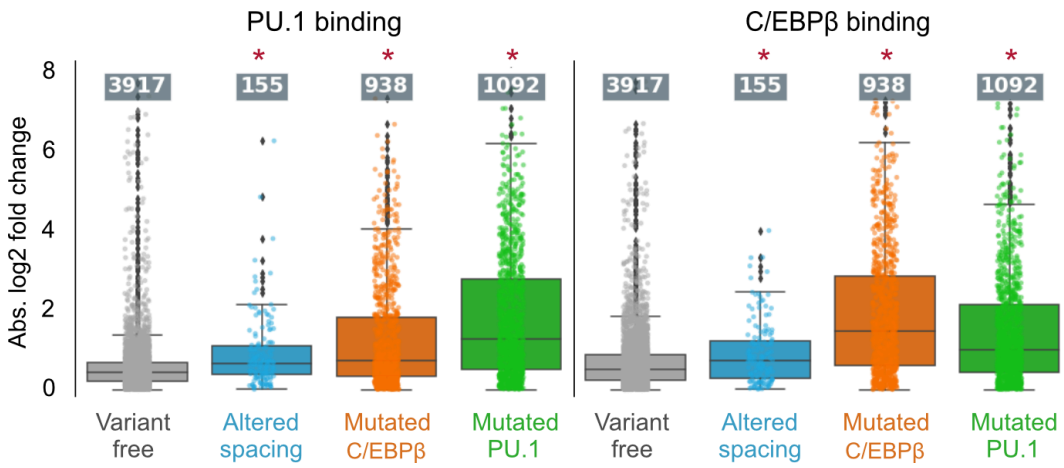

Figure 3—figure supplement 5

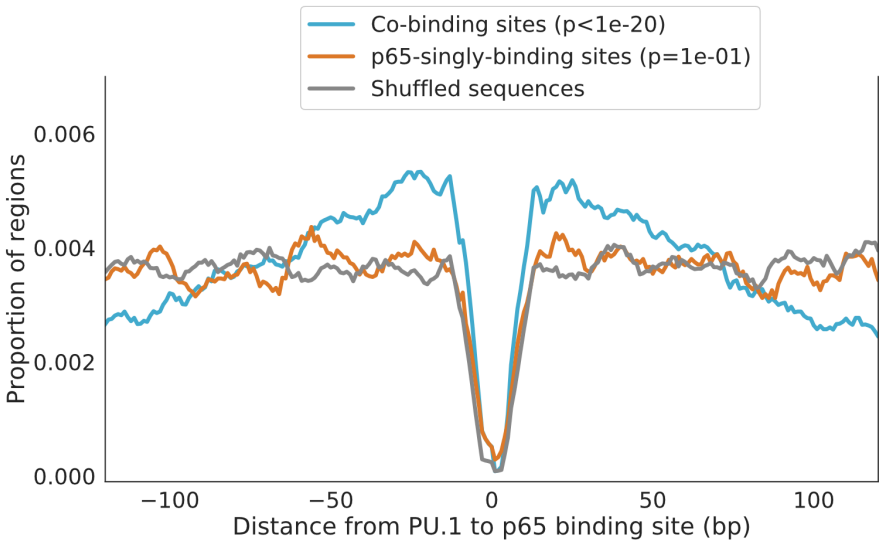

Figure 3—figure supplement 6

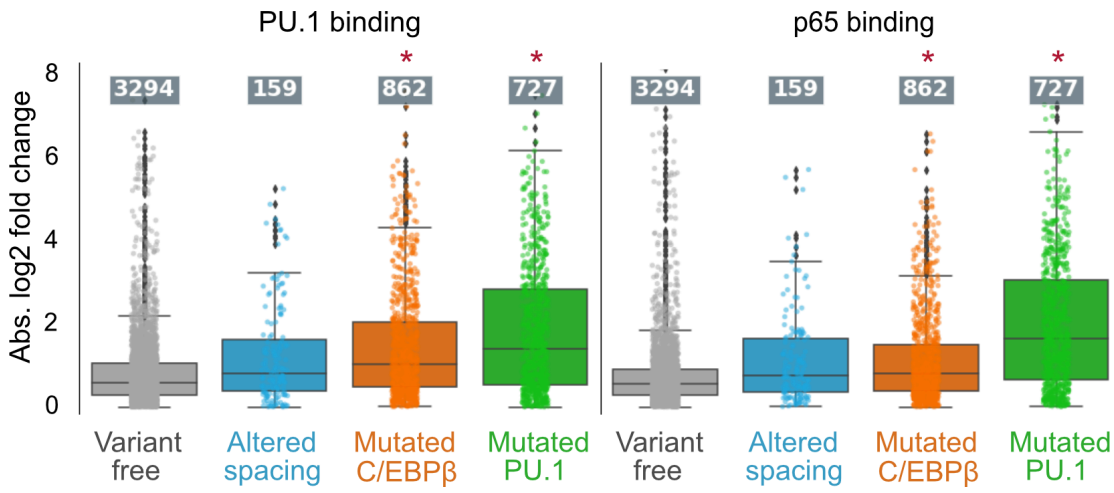

Figure 4—figure supplement 1

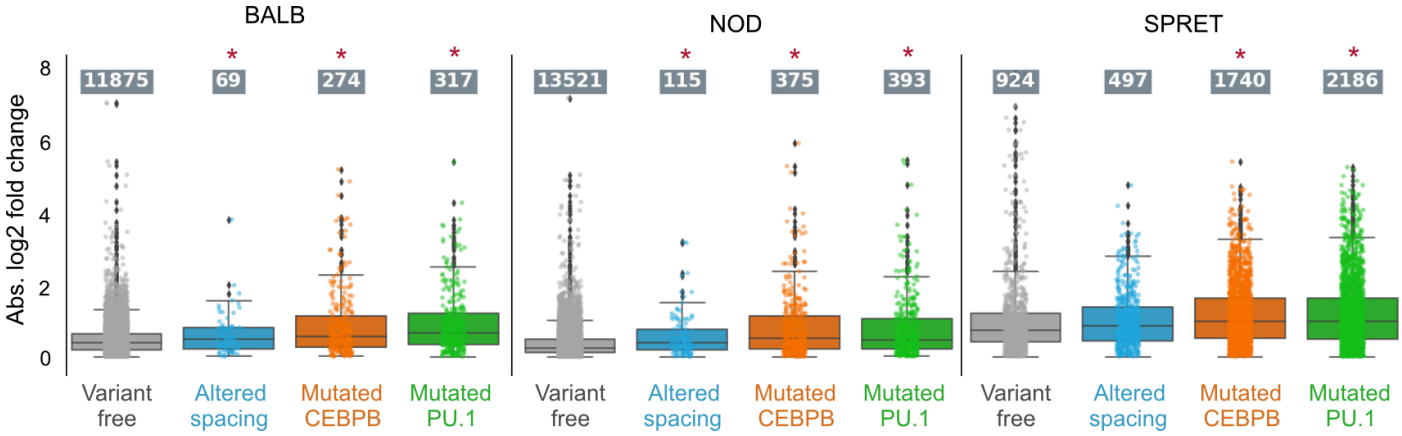

Figure 4—figure supplement 2

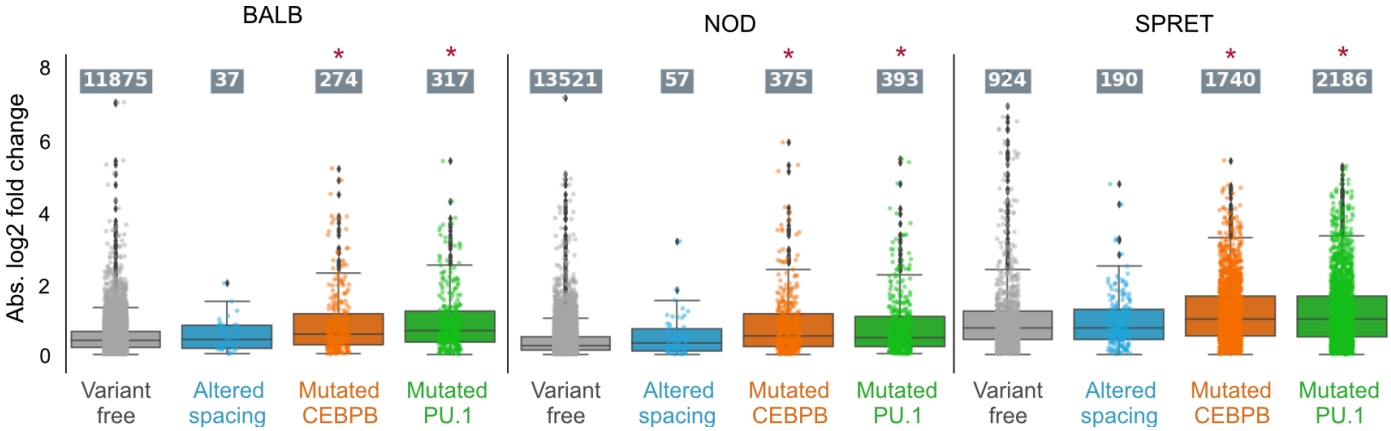

Figure 4—figure supplement 3

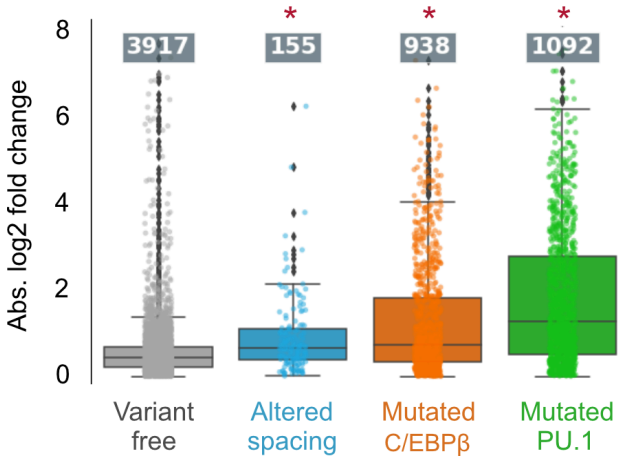

Figure 4—figure supplement 4

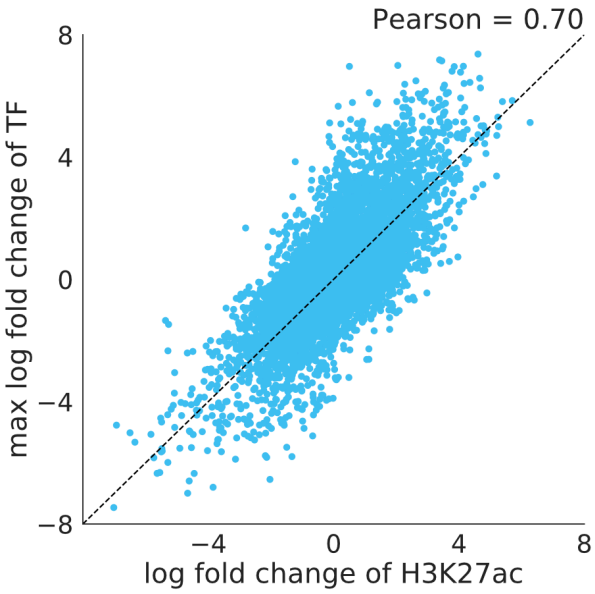

Figure 5—figure supplement 1

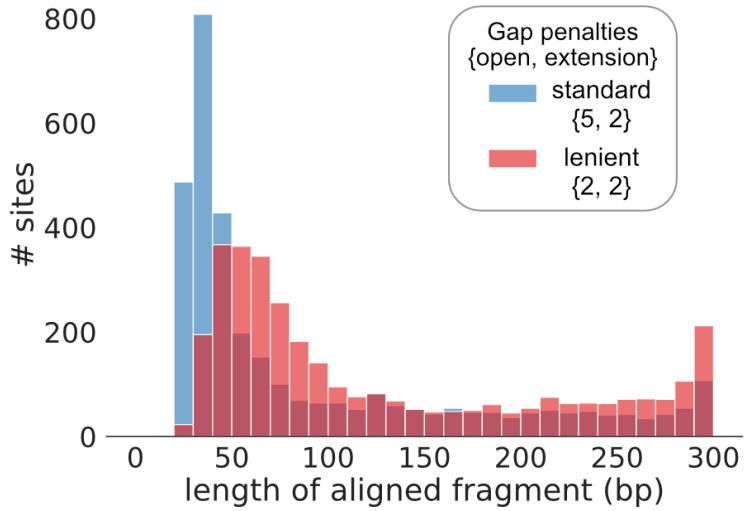

Figure 5—figure supplement 2

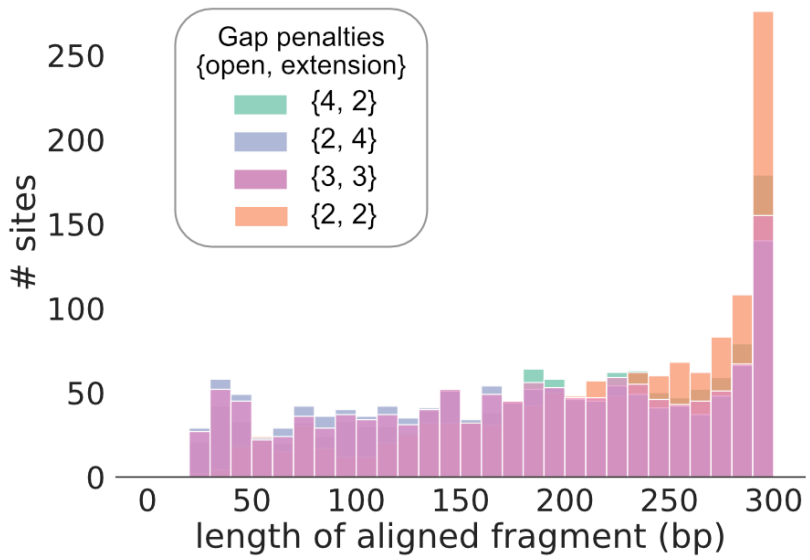

Figure 5—figure supplement 3
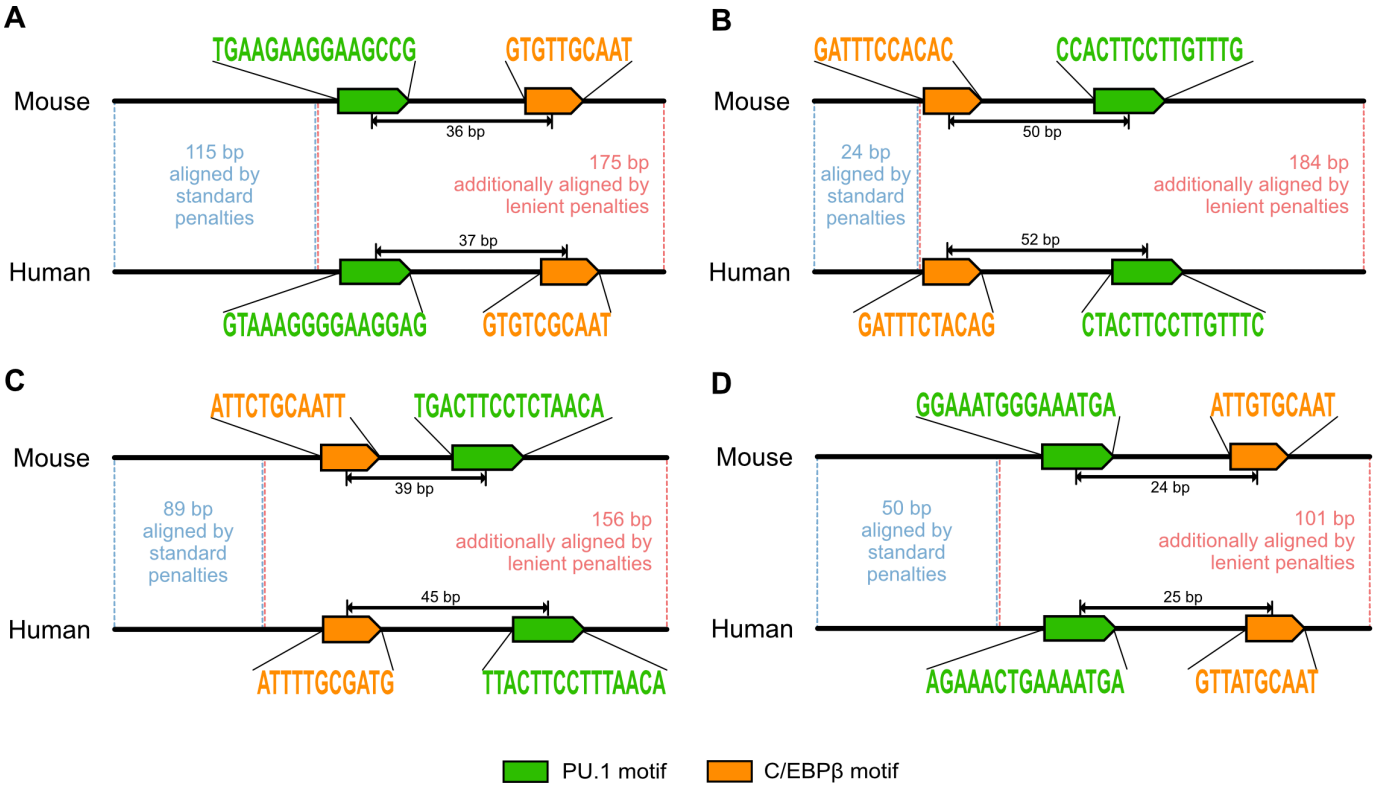
Figure 5—figure supplement 4