SUPPLEMENTARY METHODS
Discovering epistatic feature interactions from neural network models of regulatory DNA sequences

Peyton Greenside, Tyler Shimko, Polly Fordyce, Anshul Kundaje
Stanford University
Contact: pgreens@stanford.edu, akundaje@stanford.edu

Simulated sequences and embedding motifs

We simulated sequences using the simdna package (https://github.com/kundajelab/simdna) with bases randomly sampled from fixed GC frequency of 0.46. We simulated 60,000 sequences – 20,000 for each Set in Figure 2A – each of length 200bp. Embedding locations were uniformly sampled across the entire sequence. If a location was sampled that already had a motif embedded, the location would be re-sampled until there was sufficient room for the new motif to be embedded. We embedded the best hit motif from the position weight matrix (PWM) and thus every embedded motif for the same factor had the exact same sequence. The specific motifs used from Kheradpour et al.[10] were "ELF1_known2", "SIX5_known1", "TAL1_known1", and "AP1_disc3."

The number of embeddings for all motifs was determined by sampling from a Poisson distribution. For ELF1 and SIX5 motifs, this distribution had a mean of 2 and was truncated to allow only values of 1 or 2. For AP1 and TAL1 motifs, this distribution had a mean of 1 and was truncated to allow only values of 0 or 1. Truncation means that if a value outside the allowed range was sampled, the number was re-sampled until it fell in the allowed range.

We trained a Convolutional Neural Network (CNN) with one convolutional layer with 40 filters of size 19 and relu activation, followed by a max pooling layer of pool length 10 and then a fully connected layer of size 200 with dropout (p=0.5) and relu activation. The final layer had a sigmoid activation and the model was trained with Adam and binary cross-entropy loss until no improvement was seen for 3 epochs. In our training procedure we used cross validation by holding out 10,000 sequences each for testing and validation and training on the other 40,000 sequences.

We compute importance scores using DeepLIFT with a fixed GC reference, which can be thought of as the average over many sequences with similar GC content. A fixed GC reference is a sequence where that $X_0[(A,T),p] = 0.27$ and $X_0[(G,C),p] = 0.23$ for each position $p = [1, L]$, which we felt was the best importance scoring method given that we constructed the reference background sequence given this exact frequency.

TAL1-GATA1-GATA2 model

We trained a multi-task neural network to model in vivo binding of TAL1, GATA1 and GATA2 in the K562 cell line. We trained this 3-task CNN model to classify 1kb sequences underlying GATA1, GATA2 and TAL1 ENCODE ChIP-seq peaks (positive class) in K562 from all other chromatin accessible DNase-seq peaks (negative class) in K562. Thus for each task all positive examples overlapped the factor's ChIP-seq peaks and DNase-seq peaks and all negative examples overlapped DNase-seq peaks but not ChIP-seq peaks.

Our model was a series of 5 convolutional layers each with 25 convolutions of size 10 followed by a max pooling layer with stride 25 and a sigmoid activation. We held out chromosomes 8 and 9 for our testing set and used the rest of the data for training. Our model achieved mean auROC of 0.953 and mean auPRC of 0.459 on the held-out test set across all three tasks.

For determining motif sites to compare FIS between putative TAL1 and GATA1 binding sites, we found all exact matches in the input sequences to the pattern 'GATA' for GATA1 and to 'CA**TG' for TAL1 where * can be any base {A,C,G,T}. While we do not expect every such location to be bound by its corresponding TF,

these motifs greatly enrich for bound sites of these factors relative to the rest of the sequence. We computed DFIM by mutating these discovered GATA1 locations and assessing the FIS of all TAL1 motif locations. We also performed the reverse procedure of mutating TAL1 and found a similar effect on GATA1 motifs. When multiple TAL1 and/or GATA1 motifs appeared in the sequence, we found all combinations of a single TAL1 location and a single GATA1 location and performed the analysis for this pair of locations while holding the rest of the sequence fixed.

bQTL model

We trained an 18-task CNN to model chromatin accessibility (measured by either ATAC-seq or DNase-seq) profiles in 16 primary hematopoietic cell types from Corces et al. [17] and two ENCODE cell lines [18]. For each task, reproducible peaks (Irreproducible Discovery Rate < 5%) across replicates were used as positive examples. The negative set for each task consisted of all other regions overlapping the union of peaks from all 18 cell types, excluding the positive set regions. The bQTLs were not included in any way in the input to the model as the labels were constructed only from accessibility data.

We define binary labels on 1kb sequences based on the presence/absence of overlapping ATAC-seq peaks in each of 16 primary hematopoietic cell types - 13 healthy and 3 leukemic - as well as 2 DNase-seq samples of cell lines K562 and GM12878 from the ENCODE project[18]. Our convolutional architecture had 3 convolutional layers with 300 filters of length 19, 200 filters of length 11 and 200 filters of length 7, each with batch normalization and max pooling with pool lengths 3, 4 and 4, respectively. This was followed by two fully connected layers of size 1000 with batch normalization and dropout (p=0.3) after each layer. We trained the model with the Adam optimizer and binary cross-entropy loss. We initialized our model with weights learned from a multi-task model trained on 900 DNase-seq samples in the Roadmap and ENCODE projects. The final performance of our model reached means of 0.69 auPRC, 0.91 auROC, unbalanced accuracy of 0.90 and balanced accuracy of 0.74. The 18 classes had an average class imbalance of 13.9% positives with a minimum of 4.7% (erythroid task) and a maximum of 21.7% (LSC task).

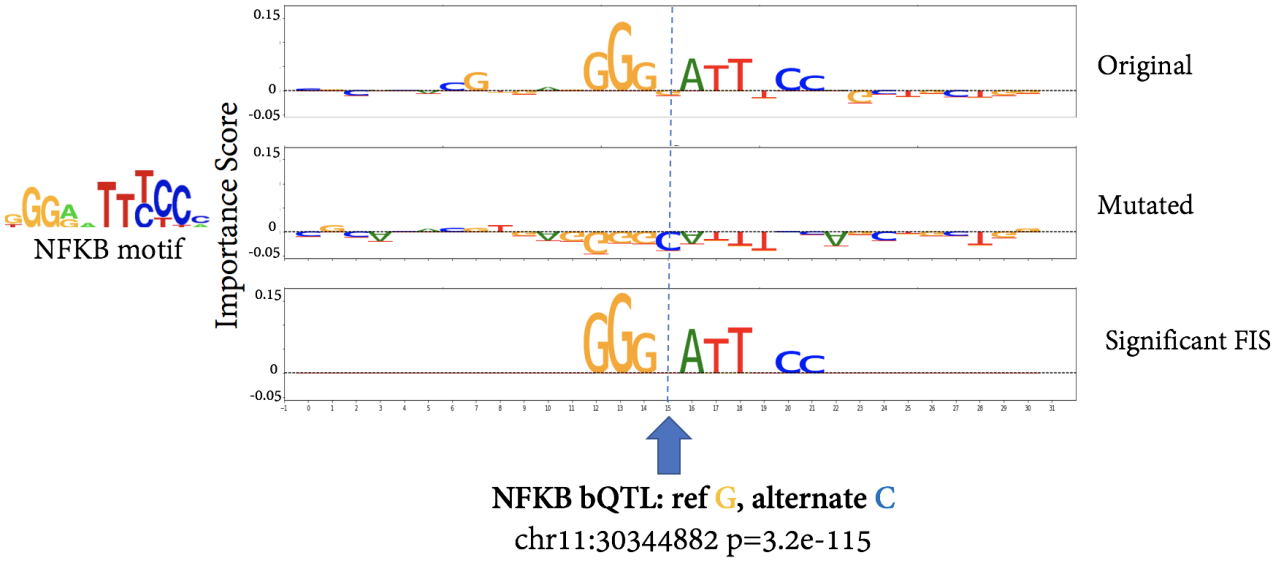
After training the model, we only further analyzed the significant bQTLs and other background SNVs with no significant allelic effects that overlapped the TF ChIP-seq peaks and the DNase-seq peaks in any of the 18 cell types. We performed feature interaction analyses using the GM12878 (lymphoblastoid cell-line) task of our model since the allelic effects of the bQTLs were estimated from ChIP-seq data in pooled lymphoblastoid lines.

In Silico Mutagenesis (ISM)

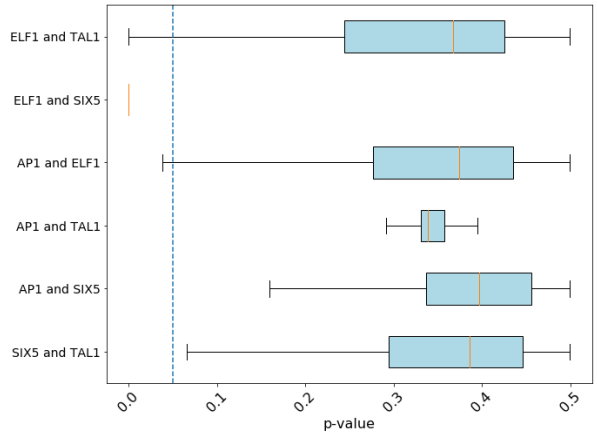
Motif-level ISM was performed by mutating the motif to a specific GC fraction (for the TAL1 and GATA1 analysis presented, this was set to 0.46) and computing the difference in model prediction between the original sequence and the mutated sequence. The pairwise ISM interaction score between two motifs is computed by first computing ISM for each motif separately, while leaving the other motif intact, and then jointly mutating both motifs in the same sequence, and calculating the difference between the joint ISM score and the sum of the two individual ISM scores. This score captures the synergy between the two motifs that is not captured by their additive effect.

To compare pairwise ISM interaction scores for sequences with both TAL1 and GATA1 motifs, we used the same motif matches described above in section "**TAL1-GATA1-GATA2 model**" and computed scores for all pairs of these TAL1 and GATA1 locations. We color each point by the absolute difference in base pairs between the start of the two motifs. The FIS here represents the average over the entire target motif's response to the source mutation. We plot the absolute value on both axes.

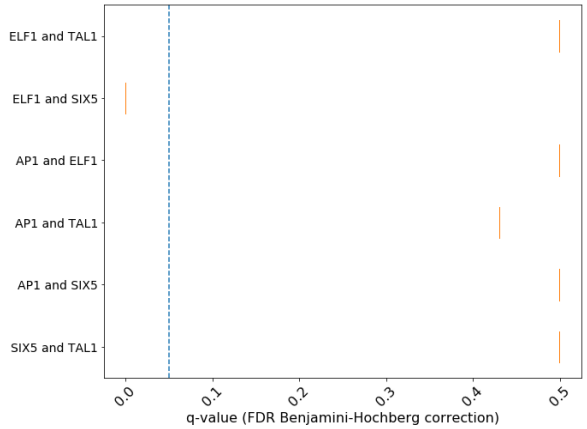
Supplementary Figures



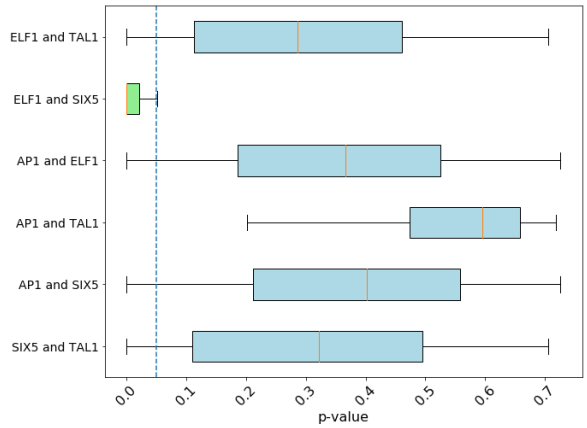
SFig. 1 The NFKB QTL at chr11:30344883 interrupts a known NFKB binding site. The original importance scores are in the top row, the mutated scores after converting the reference allele "G" to a "C." The delta profile is pictured in the bottom row where all non-significant bases (p>0.05) have been omitted leaving just the responding motif.



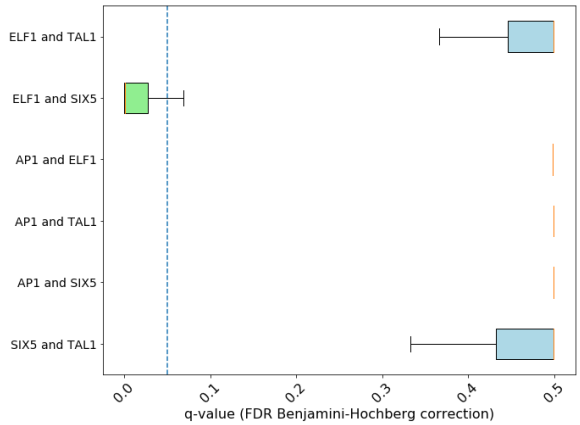
SFig. 2A P-values determined from fitting a NULL distribution to dinucleotide shuffled sequences using DeepLIFT with a fixed GC reference for computing importance scores.



SFig. 2B Q-values – corrected with the Benjamini-Hochberg FDR procedure – of **SFig. 2A.**

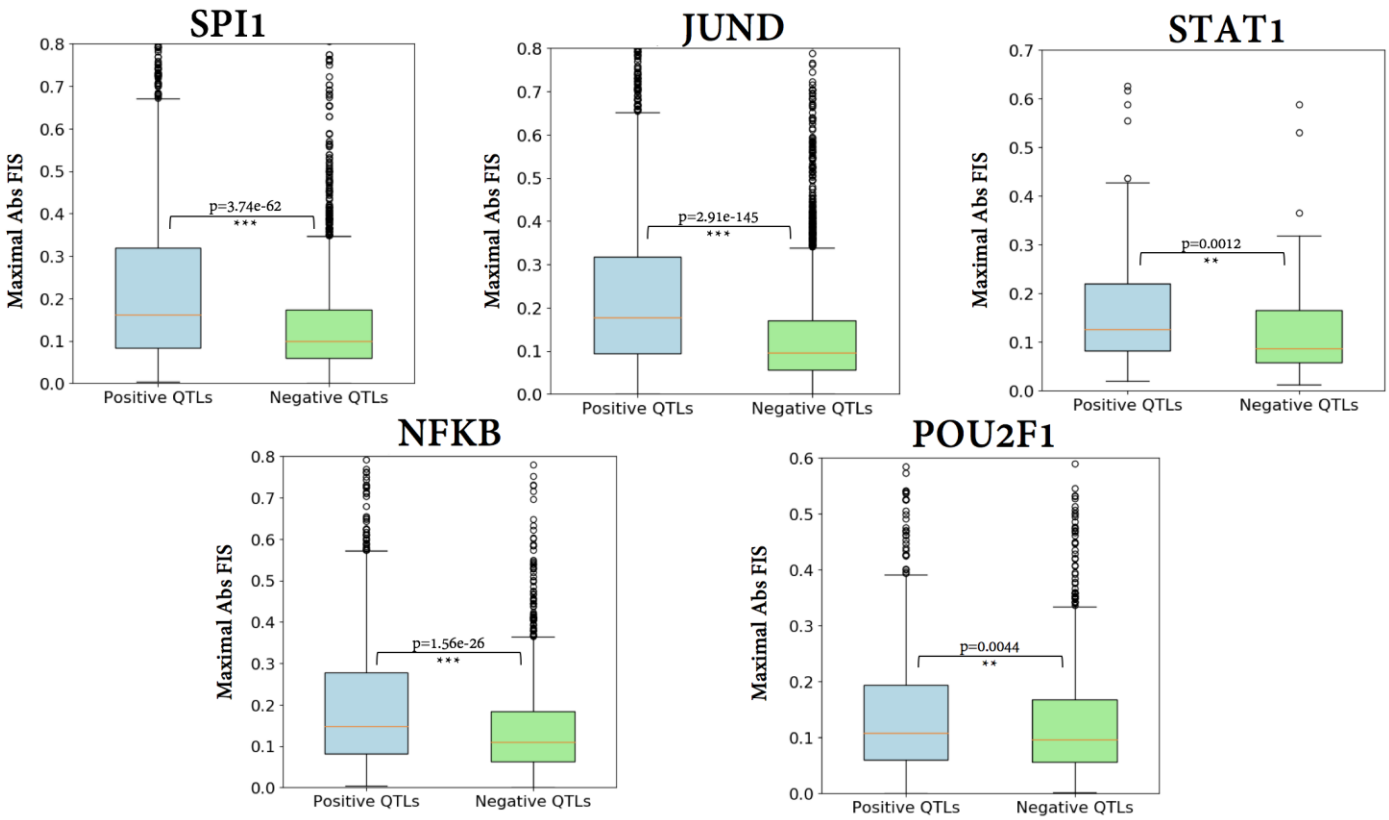


SFig. 2C P-values determined from fitting a NULL distribution to dinucleotide shuffled sequences using saliency

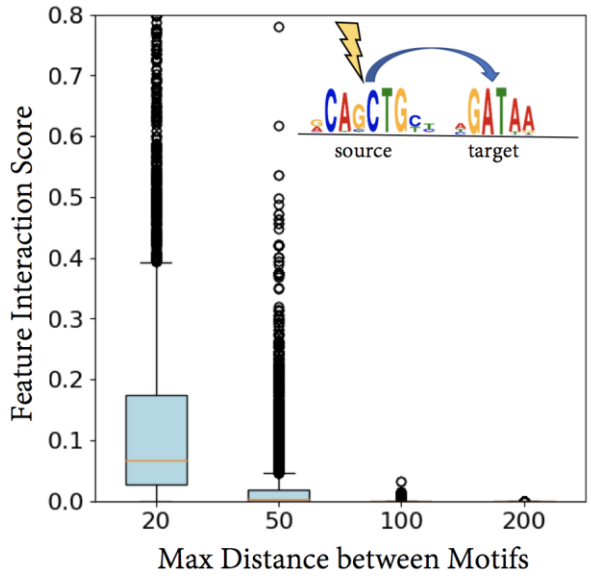


SFig. 2D Q-values – corrected with the Benjamini-Hochberg FDR procedure – of **SFig. 2C.**

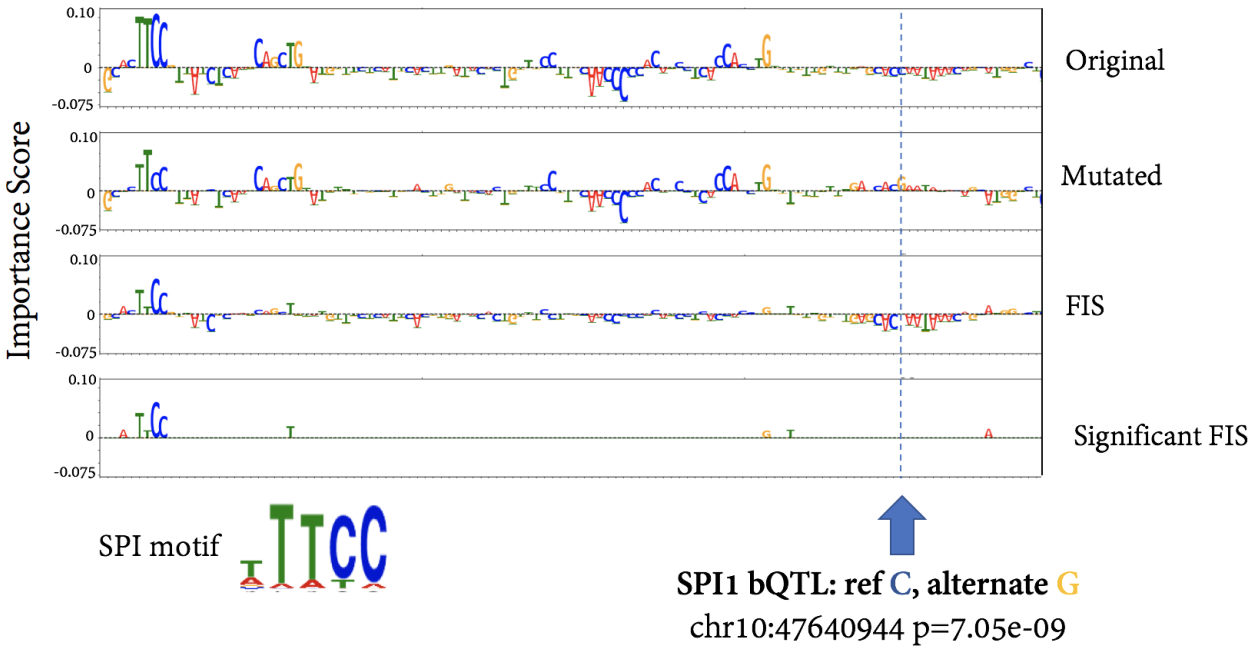
maps for computing importance scores.



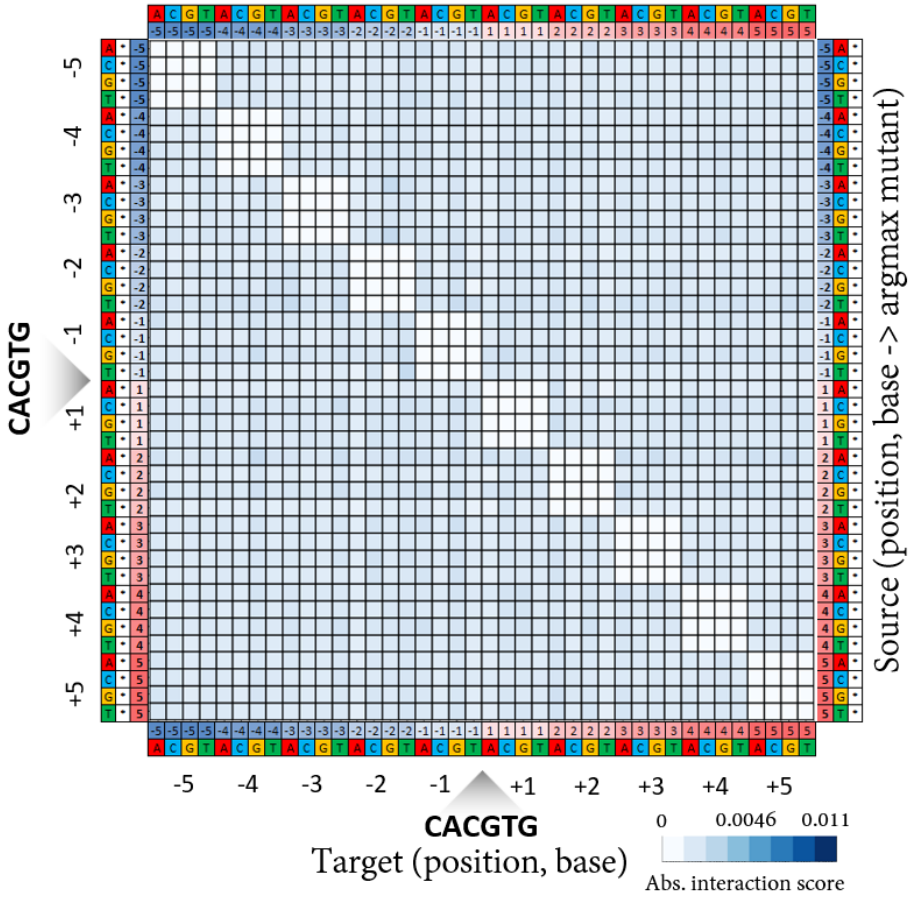
SFig. 3 Significant differences between positive (significant p<5e-5) and insignificant (p=1) bQTLs are recapitulated using importance scores computed with saliency maps, showing robustness of the method across multiple importance score methods.



SFig. 4 Mutating TAL1 also has an effect on GATA1 motifs within 20bp in comparison to those that are greater than 20bp away, showing generally symmetric results to those in **Fig. 3B.**



SFig. 5 This SPI1 QTL appears to modulate the strength of an SPI1 binding site 100 base pairs away from the actual variant site.



SFig. 6 We observe weak pairwise interactions between positions in the marginalized aggregate DFIM for Cbf1 across the 5K lowest binding affinity sequences. The rows correspond to (source position, source base, argmax mutant base). The columns correspond to (target position, target base).