Determinants of transcription factor regulatory range

Chen-Hao Chen\textsuperscript{1,2,3}, Rongbin Zheng\textsuperscript{4}, Collin Tokheim\textsuperscript{1,3}, Xin Dong\textsuperscript{4}, Jingyu Fan\textsuperscript{4}, Changxin Wan\textsuperscript{4}, Qin Tang\textsuperscript{3,5}, Myles Brown\textsuperscript{3,5}, Jun S. Liu\textsuperscript{6}, Clifford A. Meyer\textsuperscript{1,3} & X. Shirley Liu\textsuperscript{1,3,6}\textsuperscript{✉}

Characterization of the genomic distances over which transcription factor (TF) binding influences gene expression is important for inferring target genes from TF chromatin immunoprecipitation followed by sequencing (ChIP-seq) data. Here we systematically examine the relationship between thousands of TF and histone modification ChIP-seq data sets with thousands of gene expression profiles. We develop a model for integrating these data, which reveals two classes of TFs with distinct ranges of regulatory influence, chromatin-binding preferences, and auto-regulatory properties. We find that the regulatory range of the same TF bound within different topologically associating domains (TADs) depend on intrinsic TAD properties such as local gene density and G/C content, but also on the TAD chromatin states. Our results suggest that considering TF type, binding distance to gene locus, as well as chromatin context is important in identifying implicated TFs from GWAS SNPs.
ChIP-seq is broadly used for identifying the genome-wide binding sites of specific TFs. Thousands of these binding-site profiles, or cistromes, have been produced in cells and tissues. One important use of cistromes is to identify the TFs that regulate a given gene. However, it is difficult to assign most TF-binding events to a gene (or vice versa) because relatively few TF-binding sites occur very near or within a gene promoter. In most studies, target genes are designated using ad hoc methods, such as assigning TF-binding sites to the nearest gene or to genes within arbitrary genomic distance thresholds. More accurate approaches use soft thresholds and consider the effect of multiple binding sites, but still use arbitrarily defined parameters and model all genes with the same parameters. More recently, genome-wide chromatin conformation Hi-C maps have revealed that the genome is organized as a hierarchically nested structure. Topologically associating domains (TADs) represent one level of this hierarchy that influence enhancer activity by facilitating enhancer-promoter interactions within TADs. However, Hi-C maps detect fewer enhancer-promoter interactions than expected based on known regulatory interactions. How to use Hi-C observations in quantitative models of cis-regulatory interactions remains an open question. In principle, systematic analysis of Hi-C, TF cistromes, histone mark ChIP-seq, and gene expression data could reveal insights into gene regulatory mechanisms that lead to better predictions of TF target genes and, in turn, more accurate interpretation of non-coding GWAS hits.

In this study, we systematically model the genomic distance over which TFs regulate genes, and evaluate how these regulatory ranges depend on TFs and on genomic and chromatin contexts (measured by H3K27ac ChIP-seq). Our integrative analyses of large compendia of ChIP-seq, gene expression, and eQTL data reveal a previously undescribed relationship between TF regulatory ranges and local genomic and chromatin contexts. These results suggest the existence of two distinct ‘short-range’ and ‘long-range’ classes of TF with contrasting characteristics, including genome-wide binding, auto-regulatory, tissue-restricted expression, and pioneer-factor like properties. In addition, the regulatory ranges of long-range TFs are longer in repressed genomic regions than in more active ones, suggesting the widespread existence of enhancers that regulate genes up to a hundred kb away in gene-sparse regions. Finally, we apply our findings to identify TFs implicated with disease-associated SNPs. Unlike previous studies, which mostly focus on individual TFs, our findings illustrate commonalities between TFs and provided a succinct framework for interpreting complex gene regulation.

**Results**

**Distance-based model of TF binding and gene expression.** TF cistromes produced by ChIP-seq represent the genome-wide locations of TF-binding sites and potential cis-regulatory elements. To infer the gene regulatory characteristics of different TFs, we used the regulatory potential (RP) model to describe the relationships between TF cistromes and gene expression (Fig. 1a). The RP model considers contributions of all binding sites of TF i near the transcription start site (TSS) of gene j by summing the regulatory effect of individual binding sites weighted by binding-site-to-TSS distances. This model is based on two assumptions consistent with experimental observations. First, a TF-binding site’s regulatory effects on a gene’s expression level typically diminishes monotonically with the genomic distance separating the TF-binding site and the gene TSS. This assumption is motivated by the genomic distance-dependent monotonic decay observed in chromatin interaction experiments, eQTL studies showing that SNPs associated with the gene expression often fall close to the gene, and CRISPR enhancer screens showing perturbations affecting expression are enriched near the TSS. The second assumption is that the contribution of each enhancer is independent of the others so that the influence of multiple enhancers on a gene is represented by the sum of their individual contributions. With these assumptions, we used an exponential decay function, parameterized by a “decay distance” constant Δ, within which distance the TF regulatory effects are halved. We calculated the RP score, R_i(Δ), using the sum of TF i binding sites weighted by their distance to the TSS of gene j with an exponential decay distance constant (Δ), to estimate the regulatory effect of TF i on gene j. (Fig. 1a; See methods for statistical details).

Although the RP model does not encompass the full complexity of gene regulation, it does allow us to quantify the genomic distances over which TFs influence genes. When the decay distance (Δ) is small, only TF-binding sites near the TSS contribute to the regulation of the gene; when the decay distance (Δ) is large, distant binding sites contribute. In a gene expression experiment involving the perturbation of the expression of a single TF, comparison of mRNAs before and shortly after the perturbation reveals the differentially expressed (DE) genes that are likely to be directly regulated by that TF. If ChIP-seq data for a TF was available under the same experimental conditions, the RP model can be applied to predict the likely TF target genes from the ChIP-seq peaks. It is important to note that R_i(Δ) depends on the decay distance, Δ, and changing the decay distance will affect the ranking of modeled TF i regulatory effects on gene j. Systematically varying Δ across a broad range we can determine the Δ that best separates the DE gene set from the other genes by calculating the Kolmogorov–Smirnov statistic for each of the tested Δ value. We refer to the optimal decay distance as Δ_i*, the “regulatory decay distance” of TF i. Using this approach, we estimated Δ_i* for the sets of genes that are downregulated on knockdown of MYC, FOXM1, GABPA, and ERα17, or upregulated on dexamethasone stimulation of AR18 and on dexamethasone stimulation of NR3C119 (Fig. 1b, Supplementary Fig. 1a). In all cases the associations between the differentially expressed gene sets and the peak RPs were highly significant. We observed different regulatory decay distances (Δ_i*) for different TFs. The Δ_i* for MYC, GABPA, and FOXM1 were shorter (Δ_i* < 1 kb) than for ERα, and NR3C1 (Δ_i* > 10 kb). Besides, a comparison of the RP model with a model that predicts target genes from TF ChIP-seq on the basis of the distance from the TSS to the nearest peak shows that the RP model better predicts target genes (Supplementary Fig. 1b).

**Different regulatory distances define two classes of TF.** TF perturbation data sets with DE genes and TF ChIP-seq data generated in the same cell type are scarce. To infer the regulatory decay distances (Δ_i*) for more TFs, we used gene expression to identify likely TF target genes. We used correlations between gene expression levels of TFs and other genes across cell lines or tissues, reasoning that the TF expression would better correlate with expression of genes regulated by that TF than those not regulated by the TF. We estimated Δ_i* to be the Δ value that maximizes the agreement between target genes calculated from TF i binding sites and target genes calculated from TF i gene expression correlations (Fig. 1c; methods). With Δ_i*, the association between the RP scores of TF i on gene j (R_{ij}(Δ_i*)) and the correlations of TF i with gene j (p_{ij}^{corr}) is highly significant (p value < 10−8) for all TFs included in this study (Supplementary Fig. 2a).
We first inferred the $\Delta_i^*$ using the Cistrome DB TF ChIP-seq collection\(^3\) and gene expression data across ~1000 cell lines from the Cancer Cell Line Encyclopedia (CCLE). $\Delta_i^*$ is <1 kb for some TFs such as YY1, CREB1, FOXM1, ATF1, and TFDP1 (Fig. 1d left, Supplementary Fig. 2b left), but is >10 kb for TFs such as PPARG, FOXA1, GRHL2, FOSL2, and TEAD1 (Fig. 1d right, Supplementary Fig. 2b right). We repeated the analysis above using another five gene expression cohorts\(^20\)–\(^24\), and further used Genotype-Tissue Expression (GTEx)\(^9\) to infer tissue-specific $\Delta_i^*$ (methods). These analysis yielded similar estimates of $\Delta_i^*$ across different gene expression cohorts using both linear and nonlinear approaches (Methods, Supplementary Fig. 2d–e). Moreover, for the same TF, the inferred $\Delta_i^*$ is similar in different tissues (Supplementary Fig. 2f). Of 108 TFs with at least three ChIP-seq samples containing at least 20,000 peaks in Cistrome DB, we identified 60 TFs with at least three ChIP-seq samples that yielded significant $\Delta_i^*$ estimates in at least three gene expression cohorts (Supplementary Data 1). We observed a distribution of $\Delta_i^*$.
Long- and short-range TFs have distinct properties. We investigated other properties of short- and long-range TFs. First, we wondered whether different groups of TFs preferentially bind to genomic regions with different levels of H3K27ac, a marker of active promoters and enhancers. To define genomic intervals likely to be of relevance to transcriptional regulation, we selected Hi-C-derived TADs as genomic units. We characterized TADs using mean H3K27ac ChIP-seq read density from 1545 human H3K27ac ChIP-seq samples across diverse cell types in the Cistrome DB and examined their signals across 3051 previously defined TADs (Supplementary Data 2). We clustered TADs across samples using the mean H3K27ac signal, revealing coordinated TAD usage in different cell lineages (Fig. 2a, top). Genes within TAD clusters were enriched in gene ontology categories relevant to the tissue clusters. By mapping the TAD clusters to A/B compartments annotated in 21 Hi-C samples, we found the H3K27ac level corresponds with the percentage of compartment A (Fig. 2a, bottom). Therefore, we defined A-type TADs (316 TADs) and B-type TADs (649 TADs), corresponding to high and low H3K27ac signal across most samples, respectively (Fig. 2a, top). B-type TADs have more A/T-rich DNA, lower gene density, longer gene transcripts, and more tissue-restricted gene expression (Supplementary Fig. 3a–b).

Having characterized the genomic regions with differential H3K27ac marks, we went on to examine the binding density of 1396 TF ChIP-seq samples. In general, TFs prefer to bind to A-type TADs (Supplementary Data 5, Fig. 2b, top). To investigate whether regulation within a given TAD could be dominated by a given TF, we calculated a z score for each TF-TAD pair, reflecting the binding density of this TF in the TAD relative to the binding densities of other TFs in the same TAD. Certain TFs, such as YY1, favor A-type TADs, while others, such as TEADI, favor B-type TADs (Fig. 2b, bottom). In general, the short-range TFs tend to have higher z scores in the A-type TADs, whereas a large proportion of long-range TFs tend to prefer the B-type TADs (Supplementary Fig. 3c). This finding is consistent with well-known principles in which lineage-specifying genes are subject to more long-range regulation, whereas ubiquitously expressed genes are regulated at closer range.

Previous studies suggested that TADs partition the genome into functional domains that help coordinate gene expression, and genes with related functions are occasionally clustered together and expressed in the same tissues. We reasoned that if a set of TADs were heavily bound by one long-range TF, its long-range regulatory properties would enable that TF to regulate a significant proportion of the genes located within those TADs. We defined a TF’s “target” TADs as those TADs that are densely bound by the given TF relative to other TFs in the Cistrome DB (z score > 1, Fig. 2c). An example of a TEADI target TAD (Fig. 2d) demonstrates the nearly perfect colocalization of TEADI and H3K27ac ChIP-seq peaks, suggesting that TEADI dominates the activity of its target TAD. Indeed, the GO term enrichment of genes within the target TADs correspond to the known functions of the dominating TFs (Supplementary Data 6).

To investigate whether long-range TFs may influence the chromatin state of target TADs as predicted, we examined whether the expression of long-range TFs are robustly associated with H3K27ac levels in the target TADs. For each TF, we selected CCLE cell lines with available H3K27ac ChIP-seq data and sorted these cell lines by the TF expression level. We observed positive associations between the TF expression and the mean H3K27ac levels of its target TADs for most long-range TFs such as TEADI (Fig. 3a–b), but weaker or negative associations for short-range TFs such as YY1 (Fig. 3a–b). Similarly, such significant positive associations hold between the expression of long-range TFs, such as TEADI, and genes located in their target TADs (Fig. 3c–d, Supplementary Fig. 3d) and weaker associations for short-range TFs. These contrasting results between long- and short-range TFs suggest that long-range TFs might influence the chromatin state and gene expression across their target TADs than short-range TFs.

As lineage-specifying genes are subject to more long-range regulation, we assessed whether the long-range TFs are expressed in a lineage-specific manner. As predicted, their expression across 48 GTEx tissues shows that long-range TFs are generally expressed in a tissue-restricted fashion, reflected by their higher coefficients of variation across GTEx tissues (Fig. 3e). Self-activating TF comprised positive feedback loops have been proposed as a mechanism for establishing stable gene expression programs during lineage specification. Indeed, compared with short-range TFs, TADs that harbor long-range TF genes are heavily bound by these TFs themselves ($p$ value = 0.008), suggesting that long-range TFs tend to have auto-regulatory properties (Fig. 3f, Supplementary Fig. 3e). This indicates that multiple binding sites of the same TF within the TAD harboring the TF gene itself may serve as a robust auto-regulatory mechanism for maintaining lineage restricted long-range TF expression.
Having shown that a large proportion of long-range TFs bind to B-type TADs, we reasoned the abundant long-range TF binding within these TADs might require pioneer-like properties, characterized by the capability to facilitate the binding of other TFs at their binding sites. In the CCLE cell lines, we evaluated whether a TF’s DNA motif becomes more enriched in the ChIP-seq peaks of other factors as the TF’s expression increases. For instance, in cell lines with increased TEAD1 expression, the TEAD1 motif becomes more enriched in the ChIP-seq peaks of other factors (Fig. 3g). TFs with such “pioneer-factor-like” properties are generally long-range TFs (Fig. 3h, Methods, Supplementary Data 7), including known pioneer factors (FOXA1, GRHL2) and previously unknown ones (FOSL2) (Supplementary Fig. 3f).

Regulatory distances differ between A- and B-type TADs. Having shown that the regulatory distances are TF specific, we investigated other determinants. In the RP model, we have assumed that for a given TF the same regulatory distance \( \Delta_i \) applies to all target genes and found the modal regulatory distance of long-range TFs to be 47 kb (Fig. 1e). However, several genes, including BCL11A, GATA3, PAI6, SHH, and SOX9, are regulated by enhancers over much longer genomic distances. Notably, these genes occur in genomic regions that have low gene densities and low gene expression levels. We hypothesized that there exist systematic differences in TF regulatory decay distances between active and repressed genomic regions.

Having defined A-type and B-type TADs with high and low H3K27ac marks, respectively (Fig. 2a), we examined whether regulatory distances differ between these TAD types. Using the RP model while restricting the analysis to genes located within A-type and B-type TADs, respectively, we calculated TAD type-specific regulatory decay distances, \( \Delta_i^A \) (in A-type TADs) and \( \Delta_i^B \) (in B-type TADs) for each TF \( i \). For long-range TFs, their regulatory decay distances are longer in the B-type TADs (i.e. \( \Delta_i^B > \Delta_i^A \)) (Fig. 4a), suggesting the few documented long-range enhancer activity in B-type TADs being the norm rather than the exception. We obtained consistent results using CAGE gene expression data and tissue-matched GTEx data (Fig. 4b; Supplementary Fig. 4a). Although TF-binding sites and TSSs tend to be sparser in B-type TADs, the model is not merely reflecting peak to gene distance distributions. An analysis of regulatory decay distances inferred using the most significant 10,000 and 20,000 peaks produces almost the same results (Supplementary Fig. 4b), whereas the peak-to-TSS distances are inversely related to peak numbers (Supplementary Fig. 4c).

We went on to verify whether enhancers regulate genes over longer genomic distances in B-type TADs than in A-type ones using complementary data. CAGE-seq, which captures the TSSs of both mRNAs and eRNAs, allows us to compute the transcription correlation between coding genes and eRNAs as a function of their separation distances. We found higher enhancer-promoter correlations over longer genomic distances in B-type TADs (Supplementary Fig. 4d). Analysis of GTEx eQTL data revealed the distances between eQTLs and the corresponding gene TSSs to be significantly longer in B-type TADs in most
tissues (Fig. 4c), with neural tissues having the largest distance differences. This phenomenon is not due to linkage disequilibrium (LD), as the LD block sizes are similar in B-type and A-type TADs (Supplementary Fig. 4e).

We next examined whether the difference in the regulatory behavior in A-type and B-type TADs is owing to differences in chromatin interaction frequencies within these TAD types. Using Hi-C chromatin interaction data, we compared contact frequencies within A-type TADs to those within B-type TADs as a function of genomic distance. Decay rates in Hi-C data are typically quantified using power-law models. Therefore, for a fair comparison, we modified the RP model using the power-law

---

**ARTICLE NATURE COMMUNICATIONS | https://doi.org/10.1038/s41467-020-16106-x**

NATURE COMMUNICATIONS | (2020) 11:2472 | https://doi.org/10.1038/s41467-020-16106-x | www.nature.com/naturecommunications
equation (methods) and determined the decay constants of long-range TFs to lie in a range between −0.35 and −0.47. The model determined decay rates to be steeper in A-type TADs (−0.4 ~ −0.5) than in B-type TADs (−0.3 ~ −0.45) (Fig. 4d), consistent with what we described above. These decay rates are smaller (in terms of absolute magnitude) than those observed in Hi-C data, which are well beyond −0.5 (Fig. 4e). Although it is interesting to note the differences in A-type and B-type TAD Hi-C decay rates in the 250 kb to 1 Mb range, this range is beyond the regulatory range of TFs observed in our model-based analysis.

Regulatory distances change with chromatin states. Our analyses revealed that the binding of long-range TFs in A-type and B-type TADs have different regulatory ranges. However, several genetic properties of TADs contribute to the establishment of their chromatin states and influence their designation as A-type or B-type. To tease out the true chromatin effect from the genetic properties, we next assessed whether different chromatin states in the same TADs in two different cell types, have different impacts on regulatory distances (Fig. 5a). We identified the nuclear factor NF-kappa-B p65 subunit, RELA, for which gene expression data in GTEx and multiple ChIP-seq data sets are available in the lymphoblastoid cell line GM12878 and lung carcinoma cell line A549. We defined TADs with statistically significant differences in H3K27ac ChIP-seq signals between the two cell lines as GM12878-predominant (more H3K27ac in GM12878 than in A549) or A549-predominant TADs. We then used GTEx lymphoblastoid and lung tissue expression data to find the regulatory decay distances of RELA in the GM12878-predominant TADs (∆GM12878\textsuperscript{a}) and A549-predominant TADs (∆A549\textsuperscript{a}), respectively. We observed that the regulatory decay distance estimated using lymphoblastoid gene expression data are shorter in GM12878-predominant TADs (∆GM12878\textsuperscript{a} > ∆A549\textsuperscript{a}, lymphoblastoid expression). Conversely, the RELA decay distance estimated using lung gene expression data is shorter in A549-predominant TADs (∆GM12878 > ∆A549, lung expression) (Fig. 5b). This suggests that the regulatory decay distance for a given TF becomes shorter in a given TAD as that TAD becomes more active.

To explore this phenomenon more broadly, we examined the chromatin effects on regulatory decay distance using eQTL data from various tissues and cell types, including brain cortex, stomach, lymphoblastoid, and whole blood. We then identified the tissue-restricted TADs and determined whether the eQTLs within these TADs are closer to their associated genes in the tissues where these TADs are more active. For example, the eQTLs derived from brain cortex tissue tend to be closer to the TSSs of their associated genes in the TADs that are more active in brain. Likewise, lymphoblastoid eQTLs are closer to the TSSs of the associated genes in the TADs that are more active in lymphoblastoid cells (Fig. 5c). This is not a result of ascertainment bias because such bias would lead to the opposite result: there is greater statistical power to detect weak, distant, eQTL associations for highly expressed genes. In fact, this phenomenon holds true for most pairs of GTEx tissues with available H3K27ac ChIP-seq data we examined (Supplementary Fig. 5). These results support our hypothesis that chromatin states of the TADs, in addition to their genetic properties, influence regulatory decay distances. More specifically, ∆i becomes shorter in the same TADs when the TADs become active. Further investigation is needed to understand the mechanisms underlying this phenomenon.

A TAD-wise TF-regulation model helps interpret GWAS. Genome-wide association studies (GWAS) have determined many trait-associated single-nucleotide polymorphisms (SNPs)\textsuperscript{41} in the non-coding regions, but the functional interpretation of these non-coding variants remains challenging. GWAS SNPs are enriched in enhancers and DNase hypersensitive regions\textsuperscript{42}, suggesting roles in gene regulation, presumably by modulating TF-binding affinities\textsuperscript{43}. The implicated TFs were inferred as those with enriched motifs or ChIP-seq peaks significantly overlapping causal SNPs, which were fine mapped using statistical evidence or through functional annotation\textsuperscript{44}. Such approaches often cannot produce satisfactory results for the following reasons: (1) limited numbers of trait-associated SNPs; (2) sparse TF ChIP-seq peaks across the genome; and (3) differential TF-binding densities in different genomic regions. To illustrate the last point, consider two TFs, TF X and TF Y, each with a total of 10 ChIP-seq peaks genome-wide. A B-type TAD harbors one TF X and three TF Y ChIP-seq peaks, and an A-type TAD harbors nine TF X and seven TF Y peaks. Although the total numbers of peaks are the same for TF X and TF Y, TF Y would play a dominant role in the B-type TAD, whereas playing a similar role as TF X in the A-type TAD. In other words, it is important to consider the density of a particular TF’s binding relative to other TFs. Analysis of sparse trait-associated SNPs and sparse TF ChIP-seq peaks can be facilitated by relaxing the requirement for ChIP-seq peaks to overlap SNPs. The trait-associated SNPs in GWAS studies are after all only proxies for causal SNPs and the TADs harboring these causal variants can be well estimated. The same principle applies to TF ChIP-seq: although not all TF-binding sites are necessarily identified in ChIP-seq, the peaks that are observed may well approximate the relative abundances of TF binding within TADs. Therefore, we can infer implicated TFs by
Fig. 4 Long-range TFs have longer regulatory decay distances in B-type TADs. **a-b** Long-range TFs have longer regulatory decay distances in B-type TADs than in A-type TADs. **c** In 47 (out of 48) GTEx tissues, eQTL-TSS distances are longer in B-type (blue) TADs than those in A-type (red) TADs. Tissues were sorted using differential eQTL-TSS distances between A-type and B-type TADs. The box plot extends from the lower to the upper quartile values of the data, with a line at the median. **d** The power-law decay rates of 49 long-range TFs measured in A-type TADs (−0.4 ~ −0.5) are significantly larger (in terms of absolute magnitude) than those measured in B-type TADs (−0.3 ~ −0.45), (two-sided Student’s t test p value = 0.0013). The box plot extends from the lower to the upper quartile values of the data, with a line at the median. The whiskers extend from the box to show the range of the data. **e** In human embryonic stem cells (hESC), mesendoderm (ME), neural progenitor (NP), and trophoblast-like (TB) Hi-C data, the average contact frequency between two loci decreases as the distances between the loci increases, following a power-law relationship. In the range of TAD sizes (median size 680 kb), the decay rates are more negative than −0.5. Source data are provided as a Source Data file.
overlapping SNPs with TADs, characterized by different relative TF abundances (methods). We inferred the implicated TFs using fine mapped SNPs\textsuperscript{44}, and found our approach to outperform SNP-TF ChIP-seq peaks overlaps in ranking known diseases-associated TFs\textsuperscript{45} (Fig. 6a).

Our approach can also be understood as associating the non-coding SNPs that are likely to affect genes with the TAD-dominant long-range TFs that are likely to regulate these same genes (Fig. 2c–h). Applying this method to infer the possible role or long-range TFs in diseases with >20 annotated associated SNPs in the GWAS catalog\textsuperscript{41}, we identified clusters of TFs associated with distinct groups of diseases, such as the previously known association of \textit{RUNX1} with autoimmune\textsuperscript{46} diseases, \textit{KLF5} with breast cancer\textsuperscript{47}, and \textit{CEBPA} as well as \textit{HNF4A} with lipid metabolism\textsuperscript{48} (Fig. 6b). This analysis also revealed co-occurring diseases that have been reported in epidemiological studies, such as autoimmune diseases and chronic lymphocytic leukemia\textsuperscript{49}, as well as type-2 diabetes and breast cancer\textsuperscript{50}. These findings demonstrate how a TAD-wise analysis of TF relative binding enrichment could help prioritize long-range TFs germane to traits of interest in GWAS studies.

**Discussion**

Despite intensive scientific investigation into the role of TFs in regulating metazoan gene expression, the mechanisms by which TFs regulate specific genes are still not well understood. In this study, we quantitatively modeled the ranges of genomic distance over which TFs regulate genes in different genomic and
chromatin contexts. We systematically examined 3406 TF ChIP-seq, 1545 H3K27ac histone modification ChIP-seq, gene expression from 1037 CCLE cancer cell lines and 48 GTEx tissues, and further analyzed eQTL and GWAS catalog data.

Our analyses revealed two distinct classes of TF with different ranges of regulatory influence, chromatin-binding preferences, pioneer-like properties, and auto-regulatory behaviors. Although the precise mechanisms that underlie these regulatory classes remain unclear, it is clear that different TFs have different regulatory genomic ranges of influence. Regarding the TF properties explored in this study, classifying TFs into two groups is sufficient to contrast their differences. Our results suggest that the binding sites of short-range TFs that are located far away from gene TSSs, which usually account for the majority of binding sites, are unlikely to influence gene expression. Transcription is known to involve several distinct stages including recruitment of RNA polymerase II (PolII) to the promoter and release of the paused transcript, which usually account for the majority of binding sites, are sites of short-range TFs that are located far away from gene TSSs, and the y axis are inferred long-range TFs. Each entry is the color-coded cosine similarity between relative TF occupancies in TADs and numbers of trait-associated SNPs in TADs. Red squares or blue squares are trait-TF pairs with significantly positive or negative cosine similarities, respectively. TFs and traits were clustered using hierarchical clustering. Source data are provided as a Data Source file.

Fig. 6 TAD-wise analysis of non-coding GWAS hits prioritizes relevant TFs. a Evaluation of the inference of GWAS-SNP relevant TFs on the basis of rankings of known disease-associated TFs. Blue boxes are known disease-associated TFs ranked by overlapping SNPs with TF ChIP-seq peaks, whereas orange boxes are known disease-associated TFs ranked by overlapping SNPs with TADs characterized by different relative TAD-wise TF abundances. The later approach can be approximated as cosine similarities between relative TAD-wise TF occupancies and numbers of disease-associated SNPs in TADs. The two-sided Wilcoxon signed-rank test p value of the paired median TF rankings is 0.005 (n = 14). The box plot extends from the lower to the upper quartile values of the data, with a line at the median. The whiskers extend from the box to show the range of the data. b The heatmap shows the long-range TFs implicated from GWAS SNPs. The x axis represents traits in the GWAS Catalog, especially autoimmune disease, cancers, and metabolic syndromes, and the y axis are inferred long-range TFs. Each entry is the color-coded cosine similarity between relative TF occupancies in TADs and numbers of trait-associated SNPs in TADs. Red squares or blue squares are trait-TF pairs with significantly (p value < 0.001 using one-sided permutation considering multiple hypothesis test) positive or negative cosine similarities, respectively. TFs and traits were clustered using hierarchical clustering. Source data are provided as a Data Source file.
of SOX57 and SHH58, which are regulated by certain distal enhancers without direct physical enhancer–promoter interaction. In addition, a recent study in Drosophila showed that disrupting long-range inter-TAD loops does not alter expression for the majority of genes58.

Non-looping regulatory mechanisms may involve thermodynamically induced phase separation phenomena69. The compartments that have been observed in Hi-C data suggest that heterochromatin and euchromatin regions of the genome are spatially segregated5 and may form different nuclear structures such as lamina-associated domains61, nucleoli, nuclear speckles, PML bodies, and Cajal bodies62. In addition, altering the nuclear position of a locus has been observed to modify the expression of nearby genes63. Our analysis raises the possibility that TF-induced disruption of heterochromatin, by disruption of transcriptionally repressive zones or by nucleation of transcriptionally active ones, can lead to alterations in the regulatory micro-environment of genes and account for the chromatin state-dependent regulatory decay distances. Recent experimental work using a CRISPR-Cas9 based technology to induce phase-separated chromatin droplet formation at targeted genomic loci has shown that droplet formation in heterochromatin regions can cause large disruptions of heterochromatin domains64. The coordinated binding of pioneer factors to heterochromatin rich TADs could seed droplet nucleation points64 and the enhancer activity of these binding sites might be related to the release of a gene from a repressive environment. An alternative mechanism is facilitated tracking, in which enhancer-bound protein complexes move toward the promoter in a progressive, unidirectional fashion, whereas possibly remaining bound to the enhancer65. In principle, the presence of enhancers or promoters between the tracking enhancer and its target promoter may impede the tracking progress and would explain the shorter regulatory distances in A-type domains.

Our modeling of regulatory decay distances is based on large-scale TF cistrome data and gene expression cohorts of various cell lineages. Although the Cistrome DB contains most publicly available TF cistromes, and the CCLE and GTEx expression cohorts include many different cell and tissue types, we are far from covering all the TFs in all cellular contexts. Nonetheless, the high degree of consistency between analyses using large amounts of orthogonal data provides solid support that the trends we observed are likely to reflect the general behavior of most TFs. These findings may help the interpretation of GWAS SNPs: significant SNPs in B-type TADs can impact the expression of distant genes in the same TAD; significant SNPs in A-type TADs, on the other hand, are less likely to be associated with (>100 kb) distant genes. Some TADs are densely bound by certain long-range TFs, and the chromatin states of these TADs are dominated by these TFs. These TADs, which are “targets” of TFs, are predominantly of the B-type. Genes located in these target TADs usually have tissue-specific functions and the target TADs often harbor the TF gene itself, indicating TAD-wise autoregulatory control of cell lineage specification. Integration of the heterogeneous distribution of long-range TFs and GWAS SNPs can better prioritize the TFs implicated in the manifested phenotype than the conventional SNP-TF ChiP-seq overlap approach. A hierarchical probabilistic framework in which relative TAD-wise TF abundance combined with local TF information may further improve the inference. Further work are needed to understand the mechanistic basis of these TF-specific and context-dependent gene regulatory effects.

Methods

TAD annotation. The TAD annotations were downloaded from http://chromosome.soscope.edu/mouse/hi-c/download.html, and their coordinates were converted from hg18 to hg38 using liftOver software from UCSC: http://hgdownload.cse.ucsc.edu/goldenPath/hg18/liftOver/. TF ChIP-seq data processing. The raw sequence data of TF ChIP-seq were downloaded from Gene Expression Omnibus and processed through standard workflow of ChLin66, consisting of quality control and peak calling using MACS. For fair comparisons of TF occupancy distributions between samples, those TF ChIP-seq samples with <20,000 peaks were discarded, and for other TF ChIP-seq samples only the top 20,000 peaks based on peak intensities were included for downstream analysis.

Identification of TF enriched TADs. For each TF ChIP-seq sample, we calculated the TF-binding density in each TAD (number of peaks/kb). Noticing that for all TFs this density is higher in the A-type TADs than the B-type TADs, we examined whether some TADs are more densely bound by certain TFs relative to others. To achieve this, for each TAD we calculated the mean and standard deviation of the TF-binding density in the TAD across all TFs. We then applied a z score transformation to all TF ChIP-seq samples, and calculated the z score of a given TF to get the TAD-specific relative occupancies of the given TF compared with other TFs. For example, a TF i with a high z score in a given TAD j indicates that TAD j is more densely bound by TF i in comparison with other TFs. For each TF, we define TF target TADs as those TADs with z scores higher than 1.

RP model. In our model of transcription regulation by a given TF i, we assume that multiple TF i binding sites contribute additively to the regulation of a gene j. In this model, we modeled the effect of a single ChiP-seq peak at TF i on gene j with an influence function that decays exponentially with the genomic distance between the TSS of gene j and peak k. The decay distance (Δ) defines the half-life of the exponential decay function. The RP, Ri(Δ), defines the total regulatory effect of TF i on gene j by summing all binding sites of TF i within TAD j of the gene j (Eq. 1).

\[
R_i(\Delta) = \sum_{peak k of TF i} 2^{-\Delta_{ik}/\Delta}
\]

This RP model can be also built using power-law functions:

\[
R_i(\lambda) = \sum_{peak k of TF i} x_{ik}^{\lambda}
\]

Where λ is power-law decay rate.

Inferring TF regulatory ranges using TF perturbations. Assuming that the DE genes upon TF perturbation would comprise TF direct target genes, we hypothesize that the RP, Ri(Δ), with correct regulatory decay distance can best separate DE genes from other genes. In TF perturbation related gene expression data matched to TF ChIP-seq data, we calculated the Ri(Δ) for each gene using Eq. (1) and derived DE genes (absolute log2 fold change > 1 and p value < 0.01). We then used Kolmogorov-Smirnov test to measure which Δ separates DE genes from other genes. The regulatory decay distance (Δ′) was then defined as the Δ that gave rise to the most significant KS test value.

Inferring TF regulatory ranges using gene expression cohorts. The method for inferring the regulatory decay distance of a TF i is based on the relationship between TF i gene expression, the gene expression of other genes, and the ChiP-seq peaks of TF i. In step 1, we calculate a statistic (the Pearson correlation coefficient) that summarizes the association between the expression of TF i and the expression of the other genes. In step 2, the summary statistic from step 1 is combined with TF ChIP-seq data in a model of cis-regulatory effect that uses RPs to estimate the regulatory decay distance. In step 2, a range of values for the regulatory decay distance parameter Δ, for TF i is evaluated to determine which value of Δ maximizes the association between the TF i ChIP-seq data and the gene expression correlation of TF i gene with other genes.

Step 1: Association between the gene expression of a TF i and other genes. We make the following two assumptions about the relationship between a TF i gene expression and the expression of the genes it regulates: (1) TFs are regulators of gene expression, and (2) TFs do not influence gene expression unless they are expressed. We approximate the relationship between the gene expression of TF i and another gene j in sample k using the linear relationship:

\[
E_{ijk} = a + y_{ij} E_{ik} + c_j
\]

where E_{ijk} and E_{ik} are the standardized expression levels of TF i and gene j in sample k (i.e., E_{ijk} ∼ N(0,1) and E_{ik} ∼ N(0,1)), y_{ij} summarizes the transcriptional regulatory effect TF i might have on gene j, and c_j ∼ N(0,σ^2). Below we show that, under these assumptions, the Pearson correlation coefficient (ρ_{ij}) between the gene expression of TF i (E_i) and another gene j (E_j) is the maximum likelihood estimator of the regulatory effect parameter, y_{ij}.
Given expression data \((E_{i1}, E_{i2}) \ldots (E_{in}, E_{in})\), the conditional probability distribution of the data under Eq. (2) is:

\[
\prod_{i=1}^{n} p(E_{i|E_{i1}; a, y_{ij}, \sigma^{2}}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^{2}}} e^{-\frac{(y_{ij} - \gamma_{ij})^2}{2\sigma^2}}
\]

The log-likelihood of the parameters based on the observed expression data is:

\[
\ell(a, y_{ij}, \sigma^{2}) = \sum_{i=1}^{n} \log p(E_{i|E_{i1}; a, y_{ij}, \sigma^{2}}) = \frac{n}{2}\log 2\pi - \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (E_{i} - (a + y_{ij}E_{i}))^2
\]

\(\hat{\gamma}_{ij}\), the maximum likelihood estimator of \(\gamma_{ij}\), can be derived using

\[
\hat{\gamma}_{ij} = \frac{\sum_{i=1}^{n} (E_{i} - E_{i1}) \cdot (E_{i} - E_{i1})}{\sum_{i=1}^{n} (E_{i} - E_{i1})^2} \cdot \frac{C_{0}}{S_{i}^2}
\]

where \(C_{0}\) is the covariance between \(E_{i}\) and \(E_{i1}\). As \(E_{i1}\) and \(E_{i}\) are standardized expression levels of TF \(i\) and gene \(j\), and thus \(S_{i}^2 = S_{j}^2 = 1\).

\[
\hat{\gamma}_{ij} = \frac{C_{0}}{S_{i}^2}
\]

\(f(\cdot)\) can be a linear or nonlinear function, and we conduct the analysis using both linear and nonlinear model.

**Linear model (for Step 2)**. \(\hat{\gamma}_{ij}\), the cis-regulatory effect of TF \(i\) on gene \(j\), can be approximated as:

\[
\hat{\gamma}_{ij} = f(R_{ij}(\Delta^c)) + e_{ij}
\]

Where \(R_{ij}(\Delta^c)\) is the RP model of cis-regulatory effect of TF \(i\) on gene \(j\) parameterized by TF-specific regulatory decay distance \(\Delta^c\), defined as (1), and \(e_{ij} \sim N(0, \sigma^2_{ij})\).

The parameters in Eq. (5) can be inferred using maximum likelihood estimations:

\[
\hat{a}, \hat{b}, \hat{\Delta}^c, \hat{\sigma}^2 = \arg\max \left(\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (\hat{y}_{ij} - a_i - b_i \cdot R_{ij}(\Delta^c))^2\right)
\]

Where \(n\) is the number of genes. To get the maximum likelihood estimator \(\hat{a}, \hat{b}, \hat{\Delta}^c, \hat{\sigma}^2\), we first fixed \(\Delta^c\), and \(\hat{a}, \hat{b}, \hat{\sigma}^2\) are:

\[
\hat{b} = \frac{\sum_{i=1}^{n} (\hat{y}_{ij} - \hat{\gamma}_{ij}) \cdot (R_{ij}(\Delta^c) - R_{ij}(\Delta^c))}{\sum_{i=1}^{n} (R_{ij}(\Delta^c) - R_{ij}(\Delta^c))^2}
\]

\[
\hat{a} = \hat{\gamma}_{ij} - \hat{b} \cdot R_{ij}(\Delta^c)
\]

\[
\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_{ij} - \hat{a} - \hat{b} \cdot R_{ij}(\Delta^c))^2
\]

Then with fixed \(\hat{a}, \hat{b}, \hat{\sigma}^2\),

\[
\hat{\Delta}^c = \arg\max \left(\frac{n}{2}\ln(\hat{\sigma}^2)\right) = \arg\min(\hat{\sigma}^2) = \arg\max(\rho^2)
\]

Where \(\rho\) is the Pearson correlation coefficient between \(\hat{y}_{ij}\) and \(R_{ij}(\Delta^c)\), is given by:

\[
\rho^2_{\text{expr.TF}}(\Delta^c) = \frac{\sum_{i=1}^{n} (\hat{y}_{ij} - \hat{\gamma}_{ij}) \cdot (R_{ij}(\Delta^c) - R_{ij}(\Delta^c))}{\sqrt{\sum_{i=1}^{n} (\hat{y}_{ij} - \hat{\gamma}_{ij})^2} \cdot \sqrt{\sum_{i=1}^{n} (R_{ij}(\Delta^c) - R_{ij}(\Delta^c))^2}}
\]

Therefore, to approximate \(\hat{\Delta}_{c}\), the maximum likelihood estimator of \(\Delta^c\), we tested a range of different \(\Delta\) values from 100 bp to 4000 kb, and determined \(\Delta^c\) as the one that maximizes Eq. (7).

**Nonlinear model (for Step 2)**. In an alternative method we measure the association between the gene expression association, \(\gamma_{ij}\), and the ChIP-seq based RP, \(R_{ij}(\Delta)\), using “distance correlation” (dCor), a measure of association that, unlike the Pearson correlation, is independent of the underlying variable distributions. Systematically varying \(\Delta\), regarding \(R_{ij}(\Delta)\) with different \(\Delta\) as different models, we perform model selection using goodness of fit tests:

\[
\Delta^c = \arg\max_{\Delta} \left(\text{dCor}(\gamma_{ij}, R_{ij}(\Delta))\right)
\]

**Empirical assessment of assumptions**. The purpose of this model is to use gene expression to quantify the possible regulatory relationships between a TF and other genes, based on an analysis of gene expression and TF Chip-seq data. In the empirical evaluation of Step 2 (Supp. Fig. 2a) we find that \(\gamma_{ij}\) and \(R_{ij}(\Delta)\) are indeed significantly correlated, suggesting that \(\Delta^c\), inferred using a linear relation in Eq. (5) captures the association between \(\gamma_{ij}\) and \(R_{ij}(\Delta)\). For all the TFs used in this study, the \(p\) values were found to lie in a range between 10^{-200} (< machine precision – 10^{-16}) and 10^{-8}. Supp. Fig. 2d, e show that the inferred regulatory distances derived on the basis of linear and distance correlation based approaches are similar.

**Procedure for inferring regulatory ranges of TFs**. Based on Eq. (7), we can estimate TF \(i\)-specific \(\Delta^c\) by integrating TF ChIP-seq from Cistrome and gene expression data from CCLE and other gene expression cohorts. The detailed steps are as follows:

1. We modeled \(\hat{\gamma}_{ij}\), the maximum likelihood estimator of transcriptional regulatory effect of TF \(i\) on gene \(j\), with \(R_{ij}(\Delta^c)\) using Eq. (5). \(R_{ij}(\Delta^c)\) was calculated based on TF ChIP-seq data with Eq. (1).

2. In Eq. (3), \(\hat{\gamma}_{ij}\) equals to the Pearson correlation coefficient, \(\rho^2_{\text{expr.TF}}(\Delta^c)\), between the expression of TF \(i\) and gene \(j\), using normalized gene expression data. The expression data can be CCLE, GTEx, or other gene expression cohorts.

   a. CCLE: we downloaded CCLE gene expression data and normalized the log-transformed RPKM values in different cell lines using quantile normalization, so that the distributions of gene expression values were the same in each cell line. Data were then normalized gene-wise by mean centering.

   b. GTEx: to infer tissue-specific \(\Delta^c\), we matched the ChIP-seq cell line to GTEx tissue type. We downloaded normalized GTEx gene expression data, and derived tissue-specific Pearson correlation \(\rho^2_{\text{expr.TF}}(\Delta^c)\) [tissue].

3. As Eq. (6), \(\Delta^c\) can be inferred as the \(\Delta\) that gives rise to the maximum \(\rho^2_{\text{expr.RP}}(\Delta)\) as Eq. (7). Therefore, we repeated the above step 1–2 with \(\Delta\) ranging from 100 bp to 4000 kb, and determined \(\Delta^c\) as:

\[
\Delta^c = \arg\max_{\Delta} \left(\text{dCor}(\hat{\gamma}_{ij}, R_{ij}(\Delta))\right)
\]

Specifically, we derived the \(\Delta^c\) using the following steps:

   a. We defined qualified TF ChIP-seq samples if there is any \(\Delta\) giving rise to \(\rho^2_{\text{expr.RP}}(\Delta)\) greater than 0.1.

   b. If there were more than two qualified TF ChIP-seq samples, we averaged the \(\rho^2_{\text{expr.RP}}(\Delta)\) for each \(\Delta\). If the maximum average correlation is larger than 0.1, we chose the \(\Delta\) that gives rise to the maximum average correlation coefficient as the \(\Delta^c\).

   c. For GTEx, we matched the ChIP-seq cell line to GTEx tissue type and derived tissue-specific regulatory decay distance, \(\Delta^c\) [tissue] as step 3a–b.

**Modeling dynamic regulatory distances of TFs**. For the investigation of the regulatory decay distance dependence on chromatin context, we proceeded as follows:

1. To infer TAD type-specific (e.g., B-type TADs or A-type TADs) regulatory decay distances, in step 3 of “Inferring regulatory decay distances of transcriptional factors using Cistrome TF ChIP-seq data and CCLE or GTEx gene expression data” we included only genes located within the designated
TAD type (e.g., B-type TADs or A-type TADs). Other steps remained unchanged in the workflow.

2. To investigate whether the regulatory decay distances depend on chromatin status in addition to genomic features, we focused on TADs that have distinct activity status levels in different lineages. Take lung and lymphoblastoid tissues, for example. We first computed the lung H3K27ac level for each TAD by averaging the TAD H3K27ac mean values over all H3K27ac ChIP-seq samples associated with the lung cluster. The lymphoblastoid H3K27ac TAD levels were calculated in a similar way. We defined the lung-specific TAD clusters as those with significantly (p value < 1e−10) higher H3K27ac levels in lung compared with H3K27ac ChIP-seq produced from lymphoblastoid. The lymphoblastoid-specific TAD clusters were defined in a similar way. For the RELA analysis we used GTEx expression data and RELA ChIP-seq samples in A549 to compute the lung-specific regulatory decay distance. RELA ChIP-seq in lymphoblastoid and GTEx lymphoblastoid expression data was used to compute the lymphoblastoid-specific regulatory decay distance.

H3K27ac clustering. The raw sequence data of H3K27ac ChIP-seq were downloaded from Gene Expression Omnibus and processed through standard workflow of Chikun with peak calling using MACS and the processed data are accessible through the Cistrome Data Browser. We calculated the average H3K27ac occupancies in each TAD using bigWigAverageOverBed, and further normalized each sample using z transformation. We then clustered the H3K27ac ChIP-seq samples and TAD as 10 clusters, both using hierarchical clustering. We defined the TAD cluster with the weakest and strongest H3K27ac signals as B-type TADs and A-type TADs, respectively.

Pioneer-like factor identification. Pioneer factors can access closed chromatin and facilitate the binding of other TFs to the accessible genomic loci. To probe whether a TF such as TEAD1, for example, possesses pioneer-like factor properties, we investigated whether there is a gain in TEAD1 motif enrichment in peaks of non-TEAD1 TF ChIP-seq samples in cells with high levels of TEAD1 expression. We proceeded as following steps:

1. Calculate the TEAD1 motif enrichment in peaks of all non-TEAD1 TF ChIP-seq samples using Homer. TEAD1 motif enrichment represents the ratio of the proportion of non-TEAD1 ChIP-seq peaks with the TEAD1 motif to the proportion of background sequences with the TEAD1 motif.
2. Average the TEAD1 motif enrichment for redundant TF cell pairs.
3. Define the median TEAD1 enrichment across all non-TEAD1 TF ChIP-seq samples in each cell line i as y̅i and define xi as the level of TEAD1 gene expression in that cell line. Defining ỹ as the mean of y̅i across all cell lines, we further derived ỹ as follows:

\[ y̅ = \frac{1}{m} \sum_{i=1}^{m} y̅_i \]

4. Perform logistic regression ỹ = x, and define TEAD1 as “pioneer-factor like” if the logistic regression p value is smaller than 0.05.

GTEx eQTL analysis. We downloaded eQTL loci across multiple tissues from the GTEx database (accession phs000424.v7.p2), remapped the coordinates from GRCh37/hg19 to GRCh38/hap3, and removed eQTL-TSS pairs with p values larger than 1e−5. To investigate whether eQTL-TSS pairs in B-type TADs have longer distances than those in A-type TADs, we selected those eQTL-TSS pairs with TSS located in B-type and A-type TADs, and then compared their log-transformed eQTL-TSS distances using Student’s t test. To further examine whether the eQTL-TSS distances change significantly, we used the same approach to examine the changing eQTL-TSS distances across cell lines for each TAD. To probe whether a TF such as TEAD1, for example, possesses pioneer-like factor properties, we investigated whether there is a gain in TEAD1 motif enrichment in peaks of non-TEAD1 TF ChIP-seq samples in cells with high levels of TEAD1 expression. We proceeded as following steps:

1. Calculate the TEAD1 motif enrichment in peaks of all non-TEAD1 TF ChIP-seq samples using Homer. TEAD1 motif enrichment represents the ratio of the proportion of non-TEAD1 ChIP-seq peaks with the TEAD1 motif to the proportion of background sequences with the TEAD1 motif.
2. Average the TEAD1 motif enrichment for redundant TF cell pairs.
3. Define the median TEAD1 enrichment across all non-TEAD1 TF ChIP-seq samples in each cell line i as y̅i and define xi as the level of TEAD1 gene expression in that cell line. Defining ỹ as the mean of y̅i across all cell lines, we further derived ỹ as follows:

\[ y̅ = \frac{1}{m} \sum_{i=1}^{m} y̅_i \]

4. Perform logistic regression ỹ = x, and define TEAD1 as “pioneer-factor like” if the logistic regression p value is smaller than 0.05.

GWAS-SNP enrichment analysis. We mapped the disease-associated SNPs to TADs and calculated the number of SNPs in each TAD. Relative TAD-wise TF abundance was calculated as following steps:

1. For each of 216 TF ChIP-seq samples, the top 20,000 peaks were included.
2. For each of 3051 TADs, numbers of TF peaks located within were calculated. The relative TAD-wise TF abundance was calculated by dividing TF peak number with the mean TF peak number.

Using the relative TAD-wise TF abundance to approximate the TF occupancies overlapping with SNPs, we can write the equation as:

\[ T\text{score} = \frac{\sum_{\text{TAD}} \#\text{SNPs} \times (\text{relative TAD} – \text{wise TF abundance})}{\#\text{SNPs}} \]

Which can be further approximated as cosine similarities:

\[ \frac{\sum_{\text{TAD}} \#\text{SNPs} \times (\text{relative TAD} – \text{wise TF abundance})}{\#\text{SNPs}^2} \]

We plotted the diseases/traits-TF heatmap with cosine similarities between relative TF occupancies and number of disease SNPs in TADs, and clustered the diseases/trails and TFs using hierarchical clustering.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data that support the findings of this study are available as follows: Cistrome [https://cistrome.org/db/#/], CCLE [https://portals.broadinstitute.org/ccle], GTEx [https://www.gtexportal.org/home/], Hic-40 [ftp://cooler.csail.mit.edu/cooler/hic40/], GWAS [https://www.ebi.ac.uk/gwas/home], Linkage disequilibrium blocks69 [http://distild.jensenlab.org/download.html], CAGE [http://enhancer.bnf.co.uk/pretsets/]. All other relevant data supporting the key findings of this study are available within the article and its Supplementary Information files or from the corresponding authors upon reasonable request. Key processed data for Figs. 1e, 2a, 2b, 2c, 3h, and Supplementary Fig. 2d are provided as Supplementary Data. Data set used for figures in the current study are available as Source Data. A reporting summary for this Article is available as a Supplementary Information file.

Code availability

Open source software is listed below: MACS2 2.1.2: https://github.com/taoliu/MACS. LiftOver: http://hgdownload.soe.ucsc.edu/admin/exe/macOSX86_64/liftOver. Homer v4.11: http://homer.ucsd.edu/homer/index.html. For inference of TF regulatory decay distance: https://bitbucket.org/jdialab/tf_regulatory_distance/src/master/. Scripts specific for each figure are available with request to corresponding authors.

Received: 11 September 2019; Accepted: 14 April 2020; Published online: 18 May 2020

References

1. Robertson, G. et al. Genome-wide profiles of STAX1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat. Methods 4, 651–657 (2007).
2. Johnson, D., Mortazavi, A., Myers, R. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. Science 316, 1497–1502 (2007).
3. Mei, S. et al. Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. Nucleic Acids Res. 45, D658–D662 (2017).
4. Ouyang, Z., Zhou, Q. & Hung, W. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. PNAS 106, 21521–21526 (2009).
5. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485, 376–380 (2012).
6. Long, H. C., Prescott, S. L. & Wysocka, J. Review ever-changing landscapes: transcriptional enhancers in development and evolution. Cell 167, 1170–1187 (2016).
7. Rada-Iglesias, A. et al. A unique chromatin signature uncovers early developmental enhancers in humans. Nature 470, 279–283 (2011).
8. Barretina, J. et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity Supp. Nature 483, 603–607 (2012).
9. Lonsdale, J. et al. The genotype-tissue expression (GTEx) project. Nat. Genet. 45, 580–585 (2013).
10. Wang, S. et al. Target analysis by integration of transcriptome and ChIP-seq data with BETA. Nat. Protoc. 8, 2502–2515 (2013).
11. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326, 289–293 (2009).
12. Pickrell, J. K. et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 464, 768–772 (2010).
13. Fulker, C. P., Muchauser, M., Anyoh, A. & Munson, G. Systematic mapping of functional enhancer-promoter interactions with CRISPR interference. Science 36056, 1–8 (2016).
14. Cappellen, D., Schlang, T., Bauer, M., Maurer, F. & Hynes, N. E. Novel c-Myc target genes mediate differential effects on cell proliferation and migration. J. Biol. Chem. 278, 78–70 (2003).
15. Bergamaschi, A. et al. The forkhead transcription factor FOXM1 promotes endocrine resistance and invasiveness in estrogen receptor-positive breast cancer by expansion of stem-like cancer cells. Breast Cancer Res. 16, 436 (2014).
16. Petryszak, R. et al. Expression Atlas update—an integrated database of gene and protein expression in humans, animals and plants. Nucleic Acids Res. 44, D846–D852 (2016).
17. Al Saleh, S., Al Mulla, F. & Luqmani, Y. A. Estrogen receptor silencing induces epithelial to mesenchymal transition in human breast cancer cells. PLoS ONE 6, e20610 (2011).
18. Wang, Q. et al. A hierarchical network of transcription factors governs androgen receptor-dependent prostate cancer growth. Mol. Cell 27, 380–392 (2007).
19. Muzikar, K. A., Nickols, N. G. & Dervan, P. B. Repression of DNA-binding dependent glucocorticoid receptor-mediated gene expression. Proc Natl Acad Sci USA 106, 16598–16603 (2009).
20. Zhu, Q. et al. Targeted exploration and analysis of large cross-platform human transcriptomic cDNA. Nat. Methods 12, 211–214 (2015).
21. Gras, I. et al. Integrative gene expression profiling of cancers by genome-wide survey of breadth of expression in normal tissues. Genomics 86, 127–141 (2005).
22. Martens, G. A. et al. Clusters of conserved beta cell marker genes for differentiation. EMBO J. 25, 7639–7652 (2006).
23. Ali, T., Renkawitz, R. & Bartkuhn, M. Insulators and domains of gene activity and chromatin structure of the human serpin gene cluster at 14q32.1. EMBO J. 26, 275–287 (2007).
24. Dekker, J. GC- and AT-rich chromatin domains differ in conformation and spatially active regions in the human genome. Cell Rep. 20, 2349–2360 (2016).
25. Schmitt, A. D. et al. A compendium of chromatin contact maps reveals spatially active regions in the human genome. Cell 177, 2042–2059 (2019).
26. Dikker, L., Geyt, F., Lefebvre, P. & Scholl, M. Different activation domains prime cis-regulatory elements required for macrophage and B cell differentiation. Dev. Cell 21, 860–873 (2011).
27. Dekker, J. The authors declare no competing interests. Nucleic Acids Res. 45, 352–359 (2017).
28. Bagheri-Fam, S. et al. Long-range upstream and downstream enhancers for spatiotemporal activation by displacing nucleosomes. Mol. Cell 50, 1011–1020 (2012).
29. Canver, M. C. et al. BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. Nature 527, 192–197 (2015).
30. Ohmura, S. et al. Lineage-affiliated transcription factors bind the Gata3 Tce1 enhancer to mediate lineage-specific programs. J. Clin. Invest. 126, 865–878 (2016).
31. Kleinjanz, D. A. et al. Long-range downstream enhancers are essential for Pax6 expression. Dev. Biol. 299, 563–581 (2006).
32. Symmons, O. et al. The Shh topological domain facilitates the action of remote enhancers by reducing the effects of genomic distances. Dev. Cell 39, 529–543 (2016).
33. Bagheri-Fam, S. et al. Long-range upstream and downstream enhancers control distinct subsets of the complex spatiotemporal Sox9 expression pattern. Dev. Biol. 291, 382–397 (2006).
34. Andersson, R., Sandelin, A. & Danko, C. G. A uniﬁed architecture of transcriptional regulatory elements. Trends Genet. 31, 426–433 (2015).
35. Alexander, J. M., Guan, J., Huang, B., Lomvardas, S. & Weiner, O. D. Live-cell imaging reveals enhancer-dependent sox2 transcription in the absence of enhancer proximity. bioRxiv 409672 (2018).
36. Benabdallah, N. S. et al. PARP-mediated chromatin unfolding is coupled to long-range enhancer activation. bioRxiv 155325 (2017).
37. Ghiavi-Helm, Y. et al. Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. Nature 541, 1272–1282 (2019).
38. Larson, A. G. et al. Liquid droplet formation by HP1α associates with nucleosomes. J. Cell Biol. 2059, 1–13 (2016).
39. van Steensel, B. & Belmont, A. S. Lamina-associated domains: links with chromosomal architecture, heterochromatin, and gene repression. Cell 169, 780–791 (2017).
40. Andrulis, I. E., Neamn, A. M., Zappulla, D. C. & Sterniglitz, R. Perinuclear localization of chromatin facilitates transcriptional silencing. Nature 394, 592–595 (1998).
41. Shin, Y. et al. Liquid nuclear condensates mechanically sense and restructure the genome. Cell 175, 1481–1491.e13 (2018).
42. Blackwood, E. M. & Kadonaga, J. T. Going the distance: a current view of enhancer action. Science 281, 60–63 (1998).
43. Qin, Q. et al. Chilin: a comprehensive ChIP-seq and DNase-seq quality control and analysis pipeline. BMC Bioinformatics 17, 404 (2016).
44. Székely, G. J., Rizzo, M. L. & Bakirov, N. K. Measuring and testing dependence by correlation of distances. Ann. Stat. 35, 2799–2799 (2007).
45. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell 38, 576–589 (2010).
46. Gibbs, R. A. et al. The International HapMap Project. Nature 426, 789–796 (2003).

Acknowledgements
We thank Drs. Leonard Mirny, Bing Ren, Fraui Jin, and Mike Pazzi for the helpful discussions during manuscript preparation. We also acknowledge the following funding sources for supporting our work: National Key Research and Development Program of China 2017YFC0908500, NIH grant U24 HG009446, and U24 CA237617 to X.S.L. and C.M.

Author contributions
C-H.C. and C.M. conceived the project. C-H.C performed the analyses. R.Z., J.F., C.W., X.D. helped data collection. C.T., Q.T., M.B.R., and J.S.L provided suggestions. C.M. and X.S.L. supervised the whole study and wrote the manuscript with C-H.C. with the help of all other authors.

Competing interests
The authors declare no conflict of interest.
