CT-FOCS: a novel method for inferring cell type-specific enhancer-promoter maps

Tom Aharon Hait[1,2], Ran Elkon[2,3†] and Ron Shamir[1†]

[1]The Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel. [2]Department of Human Molecular Genetics and Biochemistry, Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel. [3]Sagol School of Neuroscience, Tel Aviv University, Tel Aviv, Israel. [†]equal contribution.

Abstract

Spatiotemporal gene expression patterns are governed to a large extent by enhancer elements, typically located distally from their target genes. Identification of enhancer-promoter (EP) links that are specific and functional in individual cell types is a key challenge in understanding gene regulation. We introduce CT-FOCS, a new statistical inference method that utilizes multiple replicates per cell type to infer cell type-specific EP links. Computationally predicted EP links are usually benchmarked against experimentally determined chromatin interactions measured by ChIA-PET and promoter-capture HiC techniques. We expand this validation scheme by using also loops that overlap in their anchor sites. In analyzing 1,366 samples from ENCODE, Roadmap epigenomics and FANTOM5, CT-FOCS inferred highly cell type-specific EP links more accurately than state-of-the-art methods. We illustrate how our inferred EP links drive cell type-specific gene expression and regulation.

Keywords

DNase-seq, CAGE, enhancers, promoters, ChIA-PET, pc-HiC, gene regulation, motif-finding, FANTOM5, ENCODE, Roadmap Epigenomics

Introduction

Understanding the effect of the noncoding part of the genome on gene expression in specific cell types is a central challenge [1]. Cell identity is, to a large extent, determined by cell-type specific transcriptional programs driven by lineage-determining transcription factors (TFs; reviewed in [2]). Such TFs mostly bind to enhancer elements located distally from their target promoters [3]. To find cell type-specific enhancer-promoter links (ct-links) one needs to compare links across multiple and diverse cell types. 3D chromatin conformation data that enables deciphering ct-links, e.g., ChIA-PET [4] and HiC [5,6], are still not available for many distinct cell types and tissues [5–10]. Consequently, there is high need for computational methods that would predict ct-links based on other broadly available data. Such resources include large-scale epigenomic data available for a variety of human cell types and tissues, which enable concurrent quantification of enhancer and promoter activities.

A key challenge is to identify which of the numerous candidate enhancer-promoter (EP) links are actually (1) functional (or active) and (2) specific to a cell type of interest. We define an EP link to be specific to a certain cell type if it is active in the cell type and its activity is limited to a small fraction of cell types. Ernst et al. [11] predicted ct-links based on correlated cell type-specific enhancer and promoter activity patterns from nine chromatin marks across nine cell types. Similarly, the Ripple method [12] predicted ct-links in five cell types. The cell type specificity of the inferred EP links was measured by their occurrence in other cell types. Additional methods that predicted EP links for a low number of cell types are IM-PET [13] and

TargetFinder [14]. All these methods rely on data of multiple chromatin marks and expression data for the studied cell types. The JEME algorithm finds global and cell type-active EP links (but not necessarily cell type-specific) using only 1-5 different omics data types [15]. Each reported EP link is given a score for its tendency to be active in a given cell type. JEME reported an average of 4,095 active EP links per cell type, and most of these may be nonspecific.

Several recent studies aimed at finding ct-links experimentally. Rajarajan et al. [16] used in-situ HiC for a schizophrenia risk locus to identify 1,702 and 442 neuronal progenitor cell (NPC) specific and neuron specific 3D chromatin interactions for 386 and 385 genes, respectively. Some of the NPC and neuron-specific interactions may be EP interactions (or ct-links). Gasperini et al. [17] used CRISPR screening to perturb 5,920 human candidate enhancers that may affect gene expression at the single-cell level in combination with eQTL analysis, and identified 664 EP links covering 479 genes enriched with K562-specific genes and lineage-specific TFs. Jung et al. [18] generated long-range chromatin maps from capture Hi-C data across 27 human cell types. When analyzing this dataset using ChromHMM models derived from Roadmap Epigenomics data [19], we found that the median number of interactions between active enahncers and promoters that were unique to a specific cell type was 630 per cell type. (**Methods**). Notably, the number of cell-type specific interactions reported by these studies is far lower than the number of EP interactions reported per cel type by JEME, indicating that only a small portion of EP links that are active in a cell type are specific to it.

Here, we introduce CT-FOCS, a novel method for inferring ct-links from large-scale compendia of hundreds of cell types measured by a single omic technique (e.g., DNase Hypersensitive Sites sequencing; DHS-seq). Given the omic profile for a set of cell types, each one with replicates, CT-FOCS uses linear mixed effect models (LMMs) to infer ct-links. CT-FOCS was applied on public DHS profiles from ENCODE and Roadmap Epigenomics [19–21], and cap analysis of gene expression (CAGE) profiles from FANTOM5 [22]. Overall, CT-FOCS inferred ~230k ct-links for 651 cell types. We demonstrate that the inferred ct-links drive cell type-specific regulation and gene expression.

Results

The CT-FOCS procedure for predicting cell type-specific EP links

We developed a novel method called CT-FOCS (*C*ell *T*ype *F*DR-corrected *O*LS with *Cross*-validation and *S*hrinkage) for inferring cell type-specific EP links (ct-links). The method utilizes single omic data from large-scale datasets. We applied CT-FOCS on three public datasets: (1) ENCODE and Roadmap epigenomics DHS profiles [19–21], which contain 208 and 350 samples from 106 and 73 distinct cell lines, respectively; and (2) FANTOM5's CAGE profiles [22], which contain 808 samples from 472 cell lines, primary cells, and tissues (**Methods**).

The input to CT-FOCS is enhancer and promoter activity matrices over the same samples and a cell type label for each sample. The output is a set of ct-links for each cell type. CT-FOCS is based on FOCS [23], which discovers global EP links showing correlated enhancer and promoter activity patterns across many samples. FOCS performs linear regression to the levels of the k enhancers that are closest to the target promoter (with k=10) followed by two non-parametric statistical tests for producing initial promoter models, and regularization to

retrieve the most informative enhancers per promoter model. To find ct-links based on the global links identified by FOCS, CT-FOCS starts with the full (that is, non-regularized) promoter model. It uses a mixed effect regression-based method, utilizing groups of replicates available for each cell type to adjust a specific regression curve per cell type-group in one promoter model (**Fig. 1; Methods**).

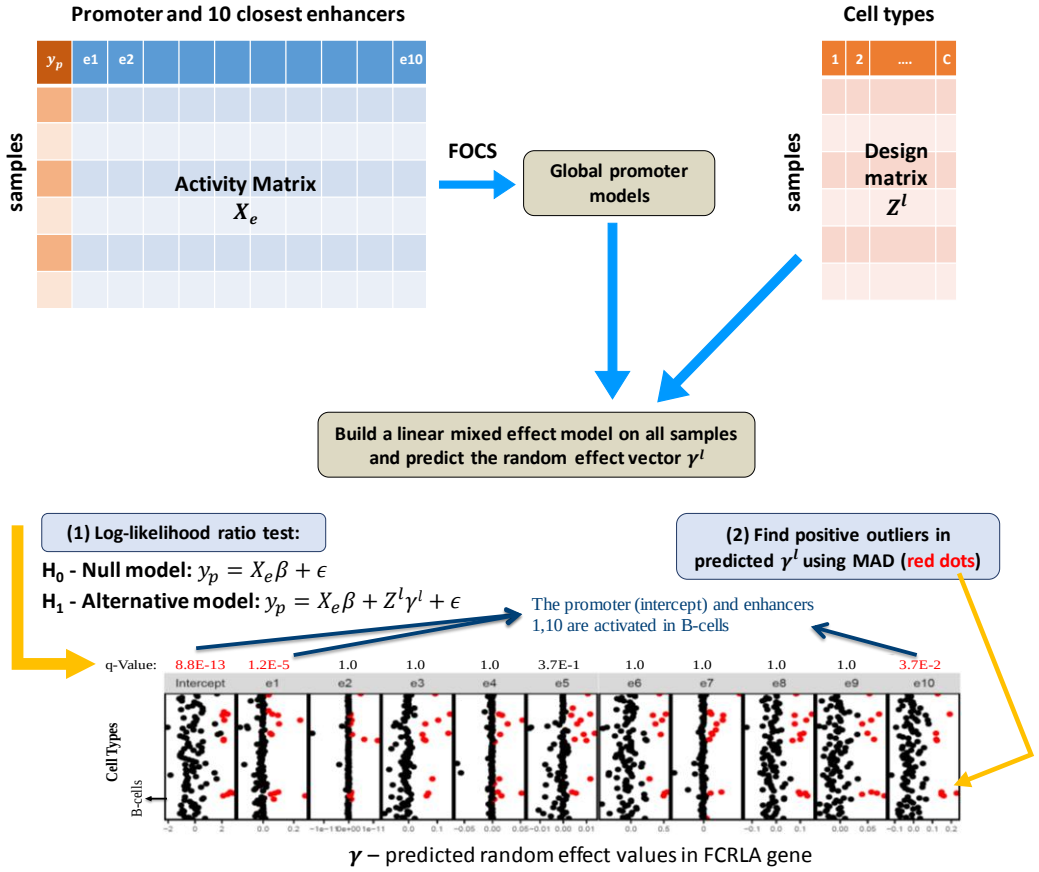

Figure 1. Outline of the CT-FOCS algorithm for promoter *p*. Let y_p denote the observed promoter activity, X_e be the activity matrix of the $k = 10$ closest enhancers to p, $l \in \{1, \dots, k+1\}$ one of the variables (enhancer or promoter, i.e. the intercept), and the matrix $Z^l[i,j]$ for variable l equals to $X_e[i,l]$ if sample i belongs to cell type j and 0 otherwise (see **Methods**). First, a robust global promoter model is inferred by applying the leave-cell-type-out cross validation step from FOCS (see [23] for details). Second, CT-FOCS builds a linear mixed effects model (LMM) on all samples using y_p, X_e, and Z^l. Then, the algorithm performs two tests for every l: (1) log-likelihood ratio test (LRT) to compare between the simple linear regression and the LMM model, which includes the component $Z^l\gamma^l$ where γ^l is a vector of the predicted random effect values for each variable (i.e., enhancer or promoter) per cell type. The tests are carried out eleven times (testing the $k = 10$ enhancers and the intercept). The p-values for these LRT tests are adjusted for multiple testing (q-values). (2) The γ^l values produced by the LMM are standardized using the Median Absolute Deviation (MAD) technique and positive outliers (red dots) are identified. A cell type-specific EP link (ct-link) is called if: (1) both enhancer and promoter (i.e., the intercept) have q-value <0.1 (marked in red), and (2) the enhancer and the promoter are found as positive outliers in the same cell type. In the FCRLA gene given as an example, the promoter p and enhancers e_1, e_{10} are significant and found as positive outliers in B-cells. Therefore, E_1p and $E_{10}p$ are called by CT-FOCS as B-cell-specific EP links.

Overall, CT-FOCS identified 17,672, 16,614 and 195,232 ct-links in ENCODE, Roadmap, and FANTOM5 datasets, respectively **(Table 1).** These included an average of 167, 234 and 414 ct-links, respectively, per cell type in (median 94, 73, and 594, respectively, **Table 1; Supplementary Fig. 1**). These numbers are in line with the low number of ct-links experimentally observed for NPC, neurons, and K562 cells [16,17], and indicate that cell-type specific EP links constitute only a small portion of the EP links that are active in a cell type. The predicted EP links are on average shared across 3, 1.65, and 2.5 cell types in ENCODE, Roadmap, and FANTOM5 datasets, respectively (**Supplementary Fig. 2**). CT-FOCS predicted both proximal and distal interactions with an average distance between the enhancer and promoter center positions of ~20kb, ~28kb, and ~160kb (median ~17kb, ~21kb, and ~110kb) in ENCODE, Roadmap and FANTOM5 datasets, respectively (**Supplementary Fig. 3A-C**). The complete set of predicted ct-links for each cell type is available at http://acgt.cs.tau.ac.il/ct-focs.

Table 1. Statistics on the number of predictions per cell type

Dataset	Avg. ct-links	Avg. enhancers	Avg. promoters	Avg. genes*	Tot. ct-links	Cell type with maximum ct-links
ENCODE	167	158	86	82	17,672	Caco-2 (1,572)
Roadmap	234	226	131	130	16,614	CD8 primary cells (2,123)
FANTOM5	414	318	146	134	195,232	Temporal lobe (13,354)

(*) Ensembl protein-coding genes

Since links are expected to function mostly within topological associated domains (TADs) [24,25], we also tested the fraction of predicted ct-links falling within 9,274 GM12878 TADs reported by Rao et al. [6] compared to randomly shuffled EP links. Random EP links were generated by randomly shifting each of the original ct-links across their chromosomes while keeping the original distances between the ceneters of the linked enhancers and promoters. We applied 1,000 such shuffles and tested if the original fraction is significantly higher than the random one. For CT-FOCS's predictions, in all datasets (ENCODE, Roadmap, and FANTOM5) the obtained empirical p-value was < 0.001 (**Supplementary Fig. 4**). These results suggest that predicted ct-links lie within TADs as expected.

ChIA-PET and pc-HiC connected loops as validation for inferred EP links

We used 3D chromatin contact loops mediated by RNAP2 or promoter capture HiC (pc-HiC) loops as a gold standard for validating inferred ct-links. We defined a loop as 3D chromatin contact generated from either ChIA-PET or pc-HiC data. The straightforward validation of an inferred ct-link is to check whether the E and P regions overlap the two anchors of the same loop. However, as loops indicate 3D proximity of their anchors, overlapping anchors of different loops indicate proximity of their other anchors as well [26,27]. In addition, ct-links that span a linear distance of < 20kb, where ChIA-PET loops may perform poorly [28], may not be supported by a single loop. Thus, we broadened the set of anchors that are considered to be proximal to connected loop sets (**CLSs**): We considered two anchors of different loops to be proximal if their loops have overlapping anchors. More formally, the CLS of a loop is defined

as the set of anchors of all loops that overlap with at least one of its anchors (**Fig. 2A**). Hence, if the enhancer and promoter regions of a ct-link overlap different anchors from the same CLS then we view this as support of the predicted link (**Fig. 2B; Methods**). As initial filtering, we used TADs inferred from HiC data and removed loops crossing TAD boundaries, since functional loops are usually confined to TADs (**Methods**).

We tested the ct-links predicted on the cell line GM12878 using reported RNAP2-mediated ChIA-PET recorded in this cell and pc-HiC loops detected on a compendium of cell types [10,18]. Out of 280 CT-FOCS inferred ct-links, 30% were validated by ChIA-PET single loops, and 66% were supported by CLSs. For pc-HiC loops, the numbers were 12% and 29%, respectively (**Fig. 2B;** see **Supplementary Fig. 5** for an additional example). To test the significance of the validation, we generated random sets of 280 EP links, residing within TADs, with the same linear distances between E and P as the ct-links predicted by CT-FOCS (**Methods**). Each random link was taken from the same chromosome as the true link in order to account for chromosome-specific epigenetic state [29]. In 1,000 random sets, CLSs supported on average 22% (61 out of 280) and at most 29% (81 out of 280) (**Supplementary Fig. 6**). Hence, according to this test, the number of predicted ct-links supported by ChIA-PET data is significant with P<0.001. The random sets overlap with pc-HiC CLSs was on average 16% (45 out of 280) and at most 23% (65 out of 280), again showing significance with P<0.001.

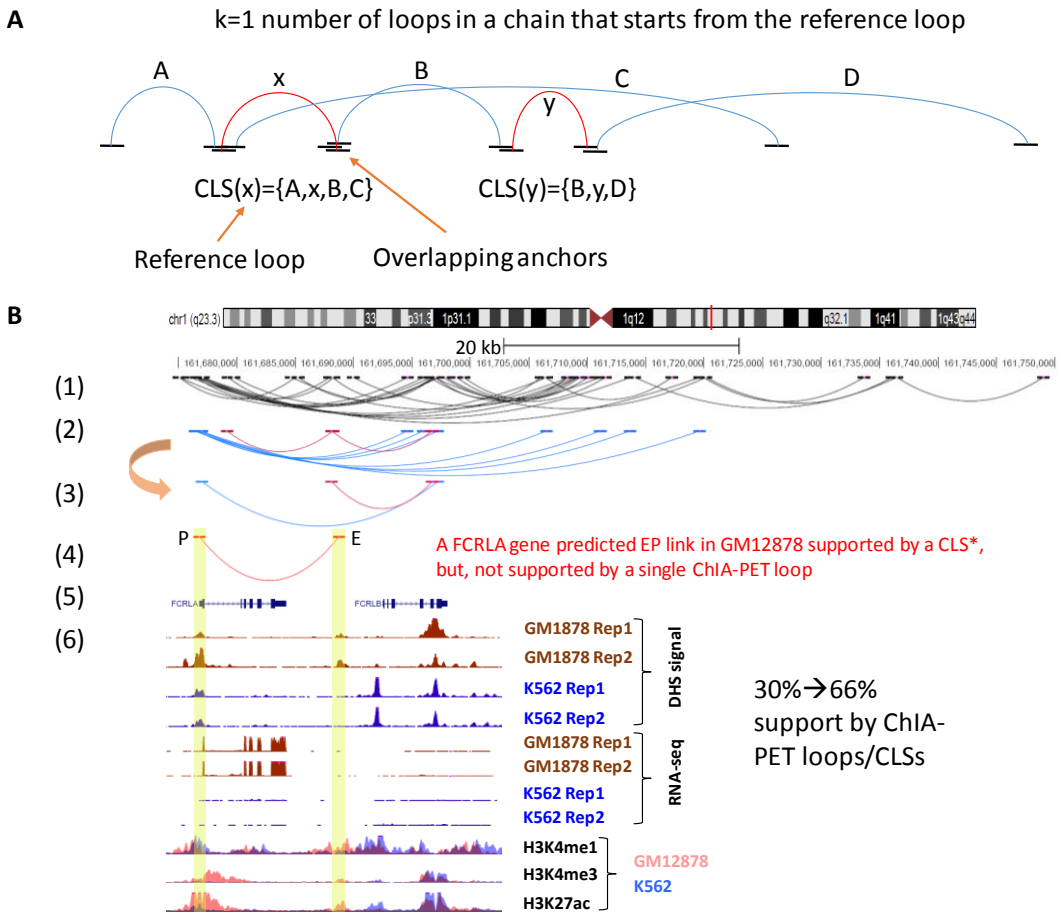

Figure 2. ChIA-PET CLSs support predicted ct-links. A CLS of reference loop x is defined as the set of all adjacent loops overlapping one of x's anchors including loop x. (A) Examples of two connected loop sets (CLSs) for reference loops x and y. Loop x's anchors overlap with at least one of the anchors of loops A, B, and C, and, therefore, the CLS of x is composed of loops A, x, B, C. Similarly, the CLS of y is composed of loops B, y, and D. (B) (1) A 70kb region of chromosome 1 showing ChIA-PET loops detected in cell type GM12878. (2) The same region showing only loops that have anchors overlapping the enhancer or promoter of the examined ct-link shown in (4). Pink loops: loops overlapping the enhancer; blue loops: loops overlapping the promoter. (3) A CLS that collectively support the predicted ct-link shown in (4). Therefore, this CT-FOCS inferred ct-link in (4) is validated by a CLS, but not by individual ChIA-PET loops. (5) A predicted ct-link of FCRLA gene in GM12878. (6) Gene expression (RNA-seq), epigenetics (DHS-seq) and gene annotations for the predicted ct-link region. Tracks are shown using UCSC genome browser for data from GM12878 and K562 cell lines.

CT-FOCS inferred ct-links correlate with cell type-specific gene expression

To evaluate the specificity of CT-FOCS predictions, we compared the activity of the set of ct-links inferred for a particular cell type with their activity in all other cell types. We defined the EP signal of a link in a cell type as the logarithm of the product of enhancer and promoter activity in that cell type, and used these signals to compute cell-type specificity as defined in [30] (**Methods**). We used this score on 280 ct-links predicted by CT-FOCS on the B-lymphocyte GM12878 cell line using the ENCODE data. Indeed, the lymphocyte group of cell types (GM12878, other B-cells, and T-cells) exhibited the highest EP signals (**Fig. 3A**), and GM12878 ranked first by specificity (**Fig. 3C;** see **Supplementary Fig. 7** for additional example).

Next, we examined the cell type specificity in gene expression (GE) of the genes involved in the ct-links (**Methods**). For this task, we analyzed expression data for 112 cell types [31] for the set of 124 genes whose promoter was included in the 280 GM12878-specific ct-links. Here too, the lymphocyte group showed the highest expression levels compared to other cell type (**Fig. 3B**), and GM12878 ranked fourth by GE specificity, after three other B-cells (**Fig. 3D**). These results show that for GM12878, ct-links predicted by CT-FOCS based on DHS data are correlated with GM12878-specific GE programs.

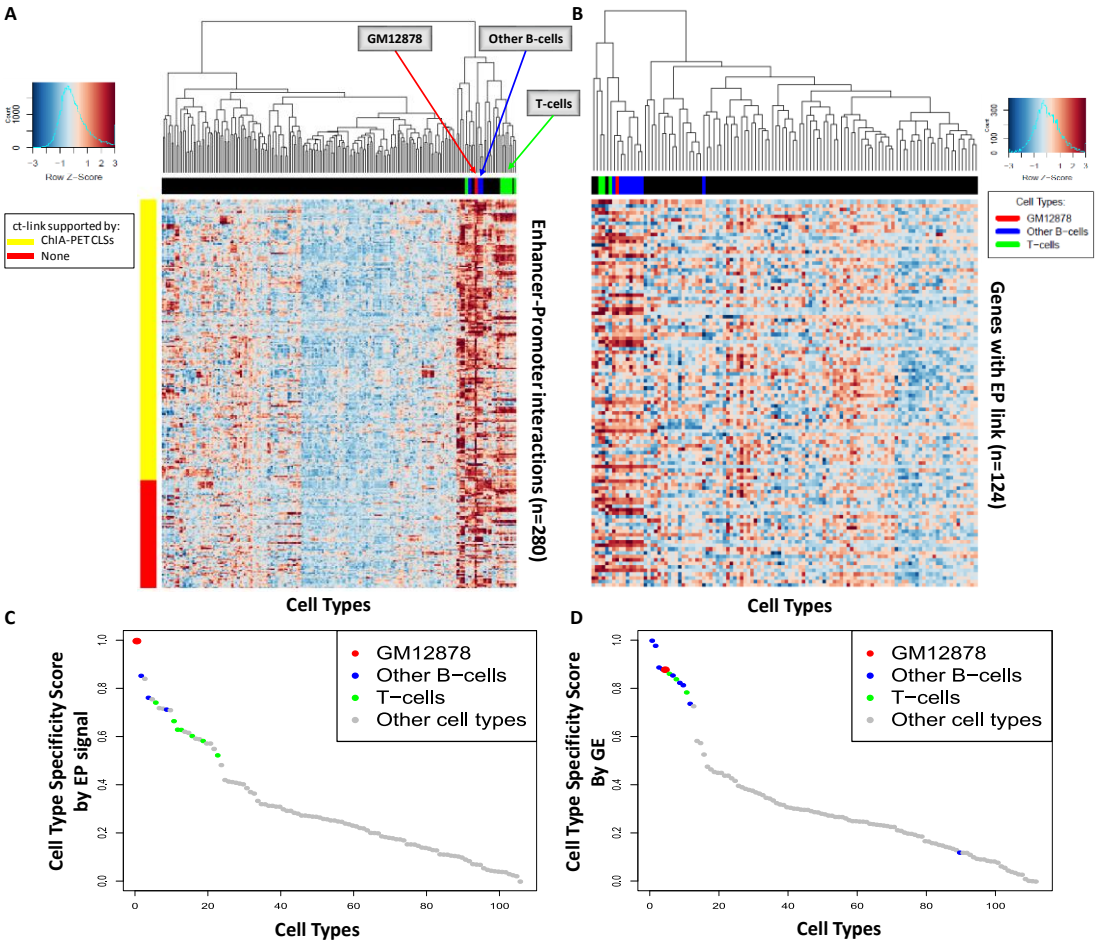

Figure 3. Specificity of ct-links predicted for GM12878. (A) Heatmap of EP signals for 280 ct-links predicted on GM12878 based on ENCODE data. Rows – EP links, columns – cell types, color – z-score of EP signal. ct-links supported by GM12878 ChIA-PET CLSs are marked in yellow. Cell types related to lymphocytes (B/T-cells) are highlighted in color. **(B)** Heatmap of gene expression (GE) for 124 genes involved in the predicted ct-links. Rows – genes, columns – cell types, color – z-score of GE. **(C)** Cell type specificity scores based on the EP signals. **(D)** Cell type specificity scores based on expression for the gene set in B (**Methods**). In A and C, 106 cell types are included in the analysis; in B and D, 112 ENCODE cell types with ENCODE gene expression are included [31].

Comparison of CT-FOCS to other methods

We compared CT-FOCS predictions with those of two other methods. (1) JEME [15], which predicts EP links that are active in a particular cell type but are not necessarily cell type-specific. (2) A naive method termed 'MAD-FOCS', which takes the shrunken promoter models from FOCS and predicts ct-links using the median absolute deviation (MAD) technique (**Methods**; see **Supplementary Fig. 1-3** for the properties of the solutions provided by the three methods). To compare the specificity of the three methods we used their results on FANTOM5 data, where all three used profiles of the single omic CAGE for prediction (JEME predictions on ENCODE and Roadmap data were based on multiple omics data types). We also compared CT-FOCS and MAD-FOCS performance on the ENCODE and Roadmap datasets.

We computed the EP specificity ranking of the results of the three methods on 276 FANTOM5 cell types that had at least 50 CT-FOCS predicted ct-links. MAD-FOCS ranks were significantly highest (**Supplementary Fig. 8A**). On the ENCODE data, CT-FOCS ranks were significantly higher than those of MAD-FOCS (**Supplementary Fig. 8B-C**). For both ENCODE and Roadmap predictions, CT-FOCS had significantly higher ranks of GE specificity than MAD-FOCS (**Supplementary Fig. 8B-C**). These results suggest that CT-FOCS predicts EP-links regulating genes that are more cell type specific compared to MAD-FOCS and JEME.

Next, we tested to what extent the links inferred for GM12878 by each method were supported by the GM12878 RNAP2 ChIA-PET assay compared to the support to links inferred on other cell types. We expected ct-links inferred for GM12878 to show higher support by GM12878 ChIA-PET data compared to ct-links predicted for other cell types. For each examined cell type, we computed the ratio between the percentage of ct-links predicted in GM12878 that were supported by GM12878 ChIA-PET CLSs to the percentage of predicted links in that cell types that were supported. Indeed, on FANTOM5 data, CT-FOCS ct-links predicted for GM12878 showed significantly higher support (median $\log_2$(ratio) ~4.8; **Fig. 4A**). Most of the cell types that had a ratio<1 were biologically related to GM12878 (e.g., B cell line and Burkitt's lymphoma cell line). MAD-FOCS had a smaller support (median $\log_2$(ratio) ~3.6). In contrast, JEME's predicted links for GM12878 had similar support rate by GM12878 ChIA-PET CLS as the EP linked predicted for the other cell types (median $\log_2$(ratio) ~0.1; P<1.7E-45; one sided Wilcoxon paired test between CT-FOCS's and JEME's predicitions). Similar results were achieved when validating against ChIA-PET single loops (**Fig. 4B**). Similar advantage of CT-FOCS over MAD-FOCS was observed for ENCODE and Roadmap (**Supplementary Fig. 9**). We confirmed these results using pc-HiC assays for 14 cell types (**Supplementary Fig. S10;** see the results for six other FANTOM5 cell types in **Fig. 4C**). These results indicate that the links identified by JEME are active across many different cell types, and that the particularity of the links of CT-FOCS is higher than those of MAD-FOCS.

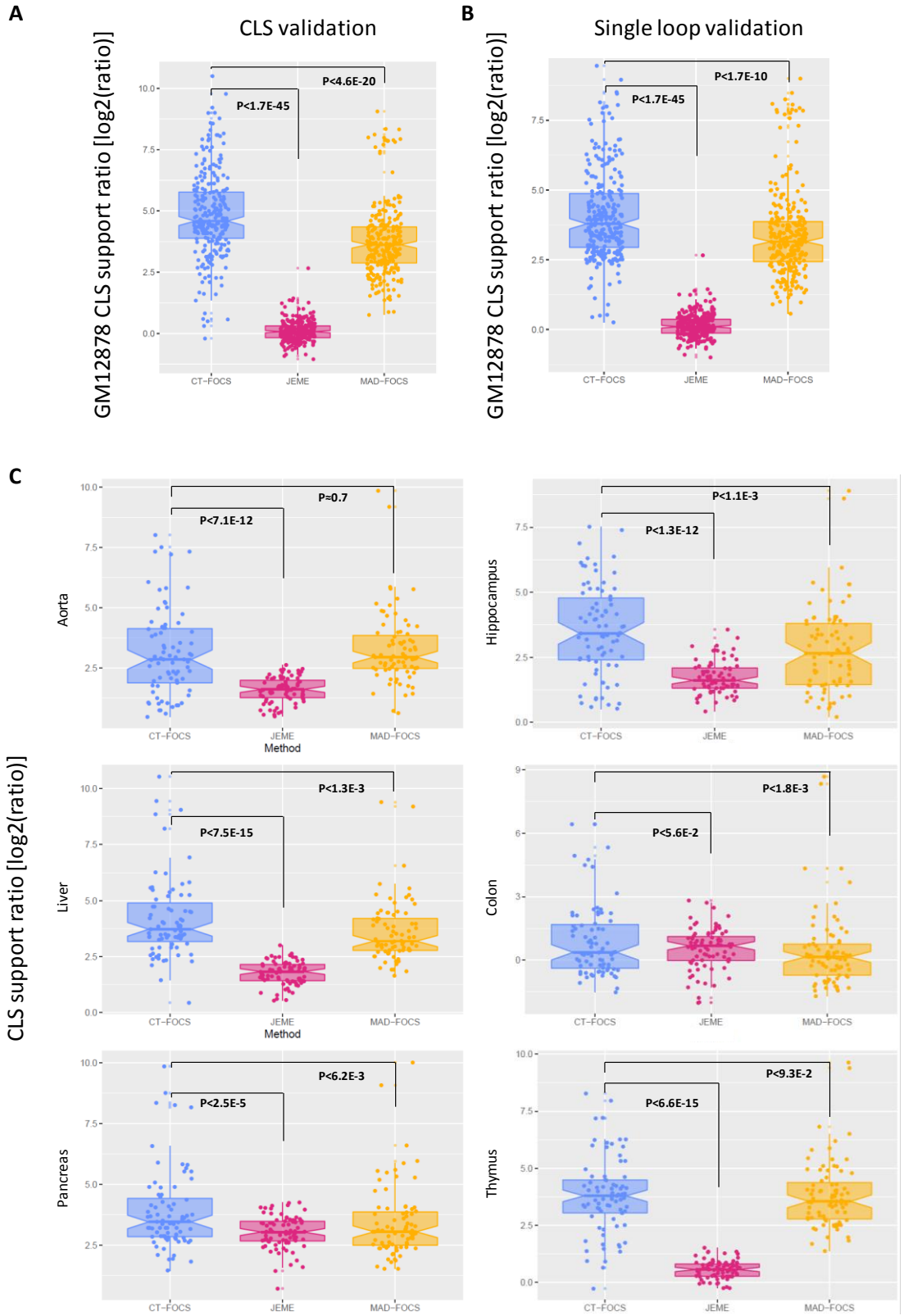

Figure 4. The particularity of each algorithm's predictions as measured by ChIA-PET and pc-HiC assays. (A-B) Each algorithm was applied to each cell type on FANTOM5 dataset, and the predicted links were compared to GM12878 ChIA-PET loops and CLSs. Comparison included 327 cell types that had at least 50 predicted EP links in CT-FOCS, MAD-FOCS, and JEME. The plots show the for each cell type the ratio between the percentage of predicted EP links on GM12878 that had GM12878 ChIA-PET support to the percentage of predicted links in that cell type that had

GM12878 ChIA-PET support (**Methods**). **A.** CLS support**. B.** Single loop support. **(C)** The same analysis as in (A) for 6 additional cell types compared to EP links derived from pc-HiC assays. p-values are based on one sided Wilcoxon paired test.

Predicted ct-links drive cell type-specific gene regulation

We asked whether the enhancers and promoters in the inferred ct-links demonstrate signals of cell type-specific gene regulation, as shown previously for lineage-determining TFs [32] and in K562 [17]. To this end, we searched for occurrence of 402 known TF motifs (position weight matrices; PWMs) within the enhancers and promoters of the inferred links. To lessen false discoveries, we restricted our search to digital genomic footprints (DGFs; **Methods**), which are short genomic regions (~20 bp on average) identified by DHS that tend to be stably bound by TFs [33]. We used ~8.4M reported DGFs covering 41 diverse cell and tissue types derived from ENCODE DHS data [34]. For each TF and cell type, we calculated the overrepresentation factor of the TF motif in the target set (enhancers or promoters of the inferred ct-links) compared to a matched control set harboring a similar nucleotide distribution (**Methods**).

We applied this test on the ct-links predicted on GM12878 using the ENCODE dataset. A set of 13 overrepresented TFs was identified in promoters, and a different set of 13 overrepresented TFs was identified in enhancers. These TFs showed on average higher overrepresentation in both enhancers and promoters compared to their occurrence in the ct-links inferred for other cell types (**Fig. 5A-B**). In terms of the specificity score of the TF overrepresentation factors, GM12878 ranked first in both enhancers and promoters (**Fig. 5C-D**). Unlike the EP signal and GE specificity-based results (**Fig. 3C-D**), here other cell types from the lymphocyte group were not highly ranked, suggesting that the regulatory TF modules detected by our analysis are strictly GM12878-specific. While individual TFs from that group are enriched across many cell types, as a group the signal for GM12878 stands out.

Among the TFs that were discovered in analysis of GM12878, a B-lympoblastoid cell line, the early B-cell factor 1 (EBF1) had the 3rd highest overrepresentation factor in promoters, and the paired box gene 5 (PAX5), which drives B-cell lineage commitment [35], ranked 7th in enhancers. EBF1, SPI1, BATF, RUNX3, IRF4, and PAX5, detected by our analysis, were shown to cooperate with STAT5A-CEBPB-PML complex, predicted to be involved in chromatin looping. Since these cofactors exhibit GM12878-specific expression (**Supplementary Fig. S11)**, they define the cell-type-specific binding of STAT5A-CEBPB-PML complex in GM12878 compared to K562 [36]. Note that Zhang et al. [36] used ChIP-seq data from multiple TFs as well as HiC data to identify TF complexes involved in chromatin looping in GM12878 and K562 cell lines. Our method requires only a single omic data to find possible TF complexes mediating chromatin looping for hundreds of cell types.

We applied the same TF analysis and specificity ranking on the ct-links inferred from ENCODE for 68 cell types that had at least 50 predicted EP links. The analysis identified 12 TFs on average in enhancers and 19 in promoters per cell type (**Supplementary Table S1**). In enhancers, 57 out of the 68 cell types ranked first in specificity, while in promoters, 58 out of 68 ranked first. Overall, the ct-links inferred on the ENCODE dataset appear to drive cell type-specific gene regulation.

We applied the same analyses on 328 FANTOM5 cell types that had at least 50 predicted EP links by CT-FOCS, MAD-FOCS, and JEME. CT-FOCS analysis identified an average 15 TFs in enhancers and 25 in promoters per cell type. JEME identified 33 and 34, and MAD-FOCS identified 6 and 14, respectively (**Supplementary Tables S2-4;** See also CT-FOCS results on Roadmap in **Supplementary Table S5**). CT-FOCS ranked ~57% of the cell types first in enhancers and promoters, while the other methods ranked ~36% in enhancers and 31-48% in promoters. Overall, CT-FOCS tends to find gene regulation modules that are more cell type-specific.

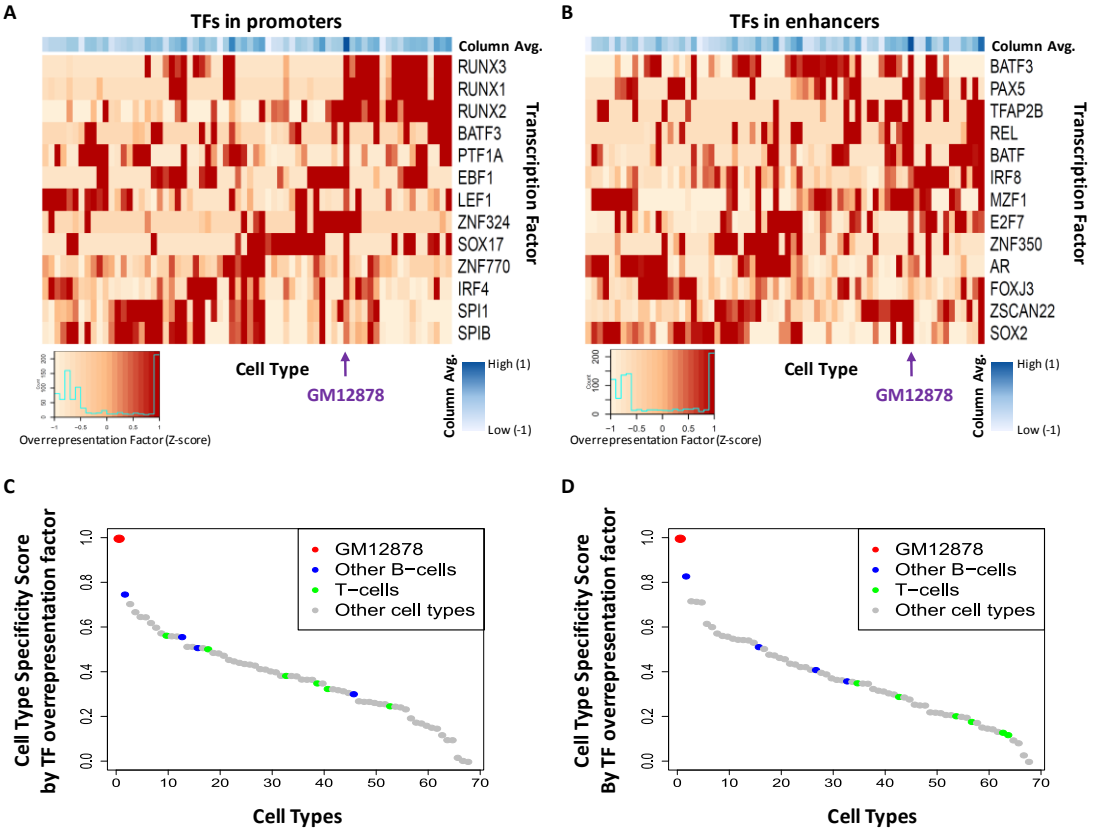

Figure 5. Overrepresented transcription factor motifs in enancers and promoters of ct-links. **(A,B)** Heatmaps of TF motif overrepresentation factor (after Z-score transformation) in promoters (A) and enhancers (B) of GM12878-specific EP links identified by CT-FOCS on ENCODE data. TFs shown had q-value < 0.1 (Hyper Geometric test). **(C-D)** Cell type specificity score ranks based on GM12878-specific TF overrepresentation factors in promoters (C) and enhancers (D) compared to other cell types (**Methods**).

Discussion

We introduced CT-FOCS, a novel method for inferring cell type-specific EP links (ct-links) based on activity patterns derived from large-scale single omic. We applied CT-FOCS on three datasets from two different omics, DHS profiles in ENCODE and Roadmap [19,20], and CAGE profiles from FANTOM5 [22], and derived a rich resource of statistically validated ct-link maps for hundreds of cell types.

To validate predicted ct-links, we proposed a simple scheme based on experimental data on ChIA-PET and pc-HiC using the notion of connected loop sets (CLSs, **Methods**; **Fig. 2** and **Supplementary Fig. 5**). On both data types, the fraction of predicted ct-links that had experimental support more than doubled when using CLSs vs. single loops. Single loop validation does not take into account the possible interaction of multiple promoters and enhancers affecting a gene's expression. By using CLSs, one can support EP link where, for example, the enhancer indirectly links to the promoter via an intermediate element. It can also overcome the "blind spot" of chromosome conformation capture methods in discovering interactions shorter than 20kb [28]. We believe that evaluation using CLS support may help to improve future methods for EP inference.

We used specificity scores [30] to assess the cell type-specificity of multiple features of the inferred ct-links: EP signal, expression of linked genes, and overrepresentation of TFs (**Fig. 3A-B** and **Fig. 5**; **Methods**). It allowed us to create global summaries of the cell type-specificity across many cell types, and we used it to compare with different EP link inference methods. Overall, across the three datasets, CT-FOCS predictions linked genes that are more cell type-specific than JEME and MAD-FOCS (**Supplementary Fig. S8**). The predicted ct-links also revealed overrepresented TFs in their enhancers and promoters. Consistent high specificity of the predictions was observed in EP DHS signals, GE, and TF overrepresentation factors (**Fig. 3A-B** and **Fig. 5A-B**).

Several comments are in order regarding our inferred ct-links. First, a common naïve practice was to map enhancers to their nearest gene. Among the CT-FOCS predicted EP links, on average per cell type, only ~10% contained enhancers that map to the nearest gene. While this low proportion is lower than previous reports (~26% in FOCS and ~40% in FANTOM5 [22]), it may have been affected by two confounders: (1) The lower limit set on the linear genomic distance between enhancer and promoter in a link (e.g., in ENCODE we set the limit to 10kb, in order to prevent activity sharing between the promoter and its candidate enhancers; **Methods**). This may lead to missing shorter links. (2) The low number of FANTOM5 reported enhancers (~43k). FANTOM5 enhancers tend to be located within intergenic regions, possibly reducing the correlation of the enhancers with the nearest gene. As a result, fewer EP links are identified using correlation-based techniques (e.g., linear regression). In contrast, the nearest gene links had poor validation results in ChIA-PET and HiC 3D loops and eQTL data [15]. Second, an average of ~60% of the predicted EP links involve intronic enhancers, similar to the report by FOCS (70%). Third, the average number of predicted ct-links per cell type was rather modest: 167 in ENCODE, 234 in Roadmap, and 414 in FANTOM5 (**Table 1**). These numbers are very low, considering that ENCODE reported a total of ~3M regulatory elements [20], but they are in line with the small number of ct-links reported previously in experiments on NPC, neuron, and K562 cells [16,17]. Fourth, each promoter involved in a ct-link was linked to ~2 enhancers on average (and a maximum of 9) in each cell type.

CT-FOCS uses linear mixed models for modeling two effects. The first is the joint contribution of multiple enhancers to the promoter activity, which was previously shown to predict gene expression more accurately compared to pairwise enhancer-gene correlations [15]. The second is the contribution of the disjoint cell type groups of samples to the promoter activity (**Methods**). By taking into account the cell type of each sample we can ask whether we should

treat the promoter activity prediction separately for each cell type group. This means that an estimated regression coefficient will not be the same for all samples but rather adjusted according to their cell type. Therefore, intuitively, using the difference in the regression coefficients between cell type groups, one can infer ct-links.

A limitation of CT-FOCS is that it considers only the ten closest enhancers to each promoter when building the models. A recent study suggested that 60% of the causal interactions between two ATAC-seq peaks occur within <20Kbp distance [28]. In agreement with this study, CT-FOCS predicts ct-links with a median linear span of ~17kb and ~21kb in ENCODE and Roadmap Epigenomics datasets, respectively (**Supplementary Fig. 3**). In addition, there is growing evidence that not all long-range enhancers require proximity to their target genes in order to activate them [37]. This may suggest that considering only the ten closest enhancers in these datasets is not a substantial limitation. In contrast, in FANTOM5, the median span of our ct-links is ~110 kb, possibly due to the small number of FANTOM5 predicted enhancers (~43k). A possible future improvement to CT-FOCS is to include all enhancers within a window of 1Mb around each promoter, e.g., by using Bayesian hierarchical models, considering possible confounders and a-priori information such as ChIA-PET and pc-HiC loops and eQTLs.

We used CT-FOCS to construct a large compendium of ct-links for 651 cell types. It is publicly available and can be useful for multiple genomic inquiries. For example, it can improve identification of known and novel cell type-specific TFs and enhance our understanding of key transcriptional cascades that determine cell fate decisions. Furthermore, the integration of protein-protein interactions (PPIs) with TF identification in predicted ct-links may help identify cell type-specific PPI modules [38]. These modules may contain additional new proteins (e.g., co-factors and proteins that are part of the mediator complex) that shape the 3D chromatin in a cell type-specific manner. Overall, the new method and compendium may advance our understanding of cell type-specific genome regulation.

Conclusions

- CT-FOCS identified cell type-specific enhancer-promoter links (ct-links) for 651 cell types inferred from ENCODE, Roadmap Epigenomics, and FANTOM5 data. On average, ~354 ct-links were discovered per cell type. The inferred ct-links showed substantially higher cell type-specificity compared to previous methods.
- The inferred ct-links correlate with cell type-specific gene expression and regulation.
- We validate predicted links with ChIA-PET and pc-HiC experimental data by using the notion of connected loops.

Methods

ENCODE, Roadmap, and FANTOM5 data preprocessing

Please refer to the Supplementary Methods for additional information.

CT-FOCS model Implementation

Our model for promoter p (**Fig. 1**) includes its k closest enhancers. The activity of the promoter across the n samples is denoted by the n-long vector $y_{p,}$ and the activity level of the enhancers

across the samples is summarized in the matrix X_e of dimensions $n \times (k+1)$, with the first column of ones for the intercept and the last k columns corresponding to the candidate enhancers. There are $C < n$ cell types and each sample is labeled with a cell type. $k = 10$ was used.

The application of an appropriate mixed effects model to the data depends on the distribution of the promoter and enhancer activities. We observed that FANTOM5 and Roadmap data have normal-like distribution and ENCODE data have zero-inflated negative binomial (ZINB) distribution (**Supplementary Fig. 12A-C**). For Roadmap and FANTOM5, we applied regular linear mixed effect regression. For ENCODE, we applied generalized linear mixed effect regression (GLMM).

For each promoter, we defined a null model and $k+1$ alternative models, each corresponding to a single random effect (i.e., random slope for enhancer or random intercept for the promoter). We defined the null model as the simple linear regression $y_p = X_e\beta + \epsilon$, and each of the alternative models as the LMM model $y_p = X_e\beta + Z\gamma^l + \epsilon$, where $X_e\beta$ is the fixed effect, $Z\gamma^l$ is the random effect, and ϵ is a random error. $l \in \{1, \dots, k+1\}$ is one of the variables (enhancer or the intercept). γ^l is a C-long vector of random effects to be predicted. Z is a nxC design matrix that groups the samples by their cell types, namely:

$$Z[i,j] = \begin{cases} X_e[i,l] & sample\ i\ belongs\ to\ cell\ type\ j \\ 0 & otherwise \end{cases}$$

We applied a likelihood ratio test between the residuals of the $k+1$ alternative models and the null model, and got $k+1$ p-values. Such p-values were calculated for each of the $|P|$ promoters, and corrected together for multiple testing using FDR [39], with the number of tests performed $|P| \cdot (k+1)$.

Each predicted random effect vector $\gamma^l = (\gamma_1^l, \dots, \gamma_C^l)$ of the alternative models was normalized using the median absolute deviation (MAD), i.e., $\gamma'^l_i = |(\gamma_i^l - median(\gamma^l)|/mad(\gamma^l)$, where $mad(\gamma^l) = median(|\gamma^l - median(\gamma^l)|)$ is calculated over all cell types together. If $\gamma'^l_i > 2.5$ then enhancer l (or the promoter, if $l = 1$) was identified as outlier in cell type i. We chose to use the MAD statistic since the mean and the standard deviation are known to be sensitive to outliers [40].

Finally, we defined cell type-specific EP links (abbreviated *ct-links)* as those that had: (1) significant random effect intercept of the promoter (P) (2) significant random effect slope of the enhancer (E), both with q-value < 0.1, and (3) E and P were identified as outliers in the same cell type according to the MAD criterion.

MAD-FOCS model implementation

Please refer to the Supplementary Methods for additional information.

External validation of predicted EP links using ChIA-PET and pc-HiC loops

We used ChIA–PET interactions to evaluate the performance of CT-FOCS and of other methods for EP linking. We downloaded ChIA-PET data of GM12878 cell line (GEO accession: GSE72816)

assayed with RNAP2 [10], and pc-HiC data across 27 tissues (GEO accession: GSE86189) [18]. Each loop identifies an interaction between two genomic intervals called its *anchors*. In ChIA-PET data, to focus on high confidence interactions, we filtered out loops with anchors' width >5kb or overlapping anchors. Loop anchors were resized to 1kbp (5kbp in pc-HiC) intervals around the anchor's center position. We filtered out loops crossing topologically associated domain (TAD) boundaries, as functional links are usually confined to TADs [7,41–43]. For this task, we downloaded 3,019 GM12878 TADs [44], which are largely conserved across cell types [6], and used them for filtering ChIA-PET and pc-HiC loops from all cell types.

To overcome the sparseness of the ChIA-PET loops, and the 8kb minimum distance between loop anchors[9,10], we combined loops into connected components of loop sets (**CLSs**) as follows: for every reference loop, x, its CLS is defined as the set of anchors of all loops that overlap with at least one of x's anchors two loops by at least 250 bp (**Fig. 2A**). We used the igraph R package [45] for this analysis.

To evaluate if a ct-link is confirmed by the ChIA-PET data, we checked if both the enhancer and the promoter fall into the same CLS. Specifically, we defined 1Kbp genomic intervals (±500 bp upstream/downstream; 5Kbp genomic intervals: ±2.5Kbp upstream/downstream in pc-HiC) for the promoters (relative to the center position; relative to the TSS in FANTOM5 dataset) and the enhancers (relative to the enhancer's center position) as their genomic positions. An EP link was considered supported by a CLS if the genomic intervals of both its promoter and enhancer overlapped different anchors from the same CLS **(Fig. 2B** and **Supplementary Fig. 5)**.

We used randomization in order to test the significance of the total number of supported EP links by ChIA-PET single loops. We deonted that number by N_t. We performed the test as follows: (1) For each predicted EP link, we randomly matched a control EP link, taken from the set of all possible EP pairs that lie within 9,274 GM12878 TADs from Rao et al. [6], with similar linear distance between E and P center positions. We restricted the matching to the same chromosome in order to account for chromosome-specific epigenetic state [29]. The matching was done using MatchIt R package (method='nearest', distance='logit', replace='FALSE') [46]. This way, the final set of matched control EP links had the same set of linear interaction distances as the original EP links. (2) We counted N_r, the number of control EP links that were supported by ChIA-PET single loops. We repeated this procedure for 1,000 times. The empirical P-value was $P = \frac{\#(N_r \geq N_t)}{1000}$, or P<0.001 if the numerator was zero A similar empirical p-value was computed for the CLSs.

We used the following formula to calculate the GM12878 ChIA-PET CLS support ratio (**Fig. 4**):

$$ratio\left(\frac{GM12878}{CellType}\right) = \frac{\%GM12878\ specific\ EPs\ in\ GM12878\ CLS}{\%CellType\ specific\ EPs\ in\ GM12878\ CLS}$$

Calling cell-type specific active EP links reported in a capture Hi-C study

We wished to count how many of the EP links reported in capture Hi-C data [18] were unique to a single cell-type. We downloaded 906,721 promoter-other (PO) capture Hi-C loops generated across 27 tissues (GEO accession: GSE86189) [18]**.** These loops involve a known

promoter of a specific gene and a non-promoter region, which may be an enhancer. We retained PO loops that appeared in exactly one cell type. We set the PO anchors to 1kb intervals around their center positions.

To call promoter and enhancer regions, we downloaded 474,004 enhancer and 33,086 promoter regions predicted by a 15-state ChromHMM model on Roadmap epigenetic data across 127 tissues (**see URLs**) [19]. We kept the enhancers of state Enh or EnhG (genic enhancers) in any of 127 Roadmap tissues. Similarly, we kept the promoters of state TssA (active TSS) or TssAFlnk (Flanking Active TSS). Then, we resized each region to 1kb interval around its center position. We called the resulting sets active promoters and enhancers.

A retained PO loop whose P and O anchors had at least 250 bp overlap with active ChromeHMM promoter and enhancer, respectively, was considered as cell type-specific active EP loop.

Cell type specificity score

We quantified the intensity of an EP link in a given sample by $\log_2 a + \log_2 b$ where a and b are the enhancer and promoter activities in that sample. The *EP signal* of the link for a particular cell type is the average of the signal across the samples from that cell type. Define $x_c = (x_{c1}, \ldots, x_{cn})$ as the vector of signals in cell type c, where n is the total number of EP links discovered in cell type c, and define $d_{c,i}$ as the Euclidean distance between the vectors of cell types c and i, both with the same EP links from cell type c. Following the definition of [30], the *specificity score* of EP links predicted in cell type c is:

$$S_c = \frac{1}{\sum_{i \neq c} d_{c,i}} \sum_{i \neq c} d_{c,i} \sum_{k=1}^{n} \left(x_{c,k} - x_{i,k}\right)$$

Similarly, cell-type specificity can be computed for the expression values of the genes annotated with EP links, or on the overrepresentation factors of TFs found at enhancers and promoters.

Motif finding on ct-links

We wished to check occurrences of transcription factor (TF) binding site motifs in our ct-links. Finding all TF motif occurrences (called hits) in a large set of promoter and enhancer sequences, each hundreds of bases long is prone to high false positive rate. We therefore limited the search for hits to digital genomic footprint regions (DGFs), very short segments that are more likely to contain genuine TF binding sites. We downloaded ~8.4M DGF sequences inferred from DNase-seq in ENCODE [34]. The mean DGF length was $L \approx 20$ bp, with a maximum length of 68 bp.

We looked for hits of 402 HOCOMOCO V11 [47] TF core motifs (taken from MEME suite database [48]; **see URLs**) in DGFs within enhancer and promoter regions of predicted ct-links. We call the resulting set of sequences the *target set*. Hits were found using FIMO [49] with 0-order Markov model as background created using fasta-get-markov command line from MEME suite [48]. Matches with FIMO q-value<0.1 were considered hits for each TF. To evaluate the significance of the findings we repeated the search on a control set from matched

regions (one per target region) having similar distribution of single nucleotides and dinucleotides. Matching was done using MatchIt R package [46] (method='nearest', distance='mahalanobis'). For each TF we used a one sided Hyper-Geometric (HG) test to compare the prevalence of TF hits in the target set with that in the background (target+control) sets. Motifs having *q*-value < 0.1 were selected.

If a k-long TF motif had l_t hits in a target set containing m_t possible k-mers in total (in both strands) and the same motif had l_b hits in the background set containing m_b possible k-mers, then the *overrepresentation factor* of the TF is defined as $(l_t/m_t)/(l_b/m_b)$. To avoid division by zero we used the Laplace correction (adding +1 to all four terms). If l_t is zero then we set the overrepresentation factor as 1.

Statistical methods, visualization and tools

All computational analyses and visualizations were done using the R statistical language environment [50]. To correct for multiple testing we used the p.adjust() function (method='BY'). We used 'GenomicRanges' package [51] for finding overlaps between genomic intervals. We used 'rtracklayer' [52] and 'GenomicInteractions' [53] packages to import/export genomic positions. Linear mixed effect regression models were created using lme R function from nlme package [54]. Generalized linear mixed effect with zero inflated negative binomial models were created using glmmTMB R function from glmmTMB package [55]. Counting reads in genomic intervals was done using BEDTools [56]. Graphs were created using graphics [57], ggplot2 [58], gplots [59], and the UCSC genome browser (**see URLs**).

URLs

474,004 Roadmap putative enhancers predicted by a 15-state ChromHMM model, https://personal.broadinstitute.org/meuleman/reg2map/HoneyBadger2-intersect_release/DNase/p10/enh/15/state_calls.RData ; 33,086 Roadmap putative enhancers predicted by a 15-state ChromHMM model, https://personal.broadinstitute.org/meuleman/reg2map/HoneyBadger2-intersect_release/DNase/p10/prom/15/state_calls.RData ; GencodeV10 TSS annotations, ftp://genome.crg.es/pub/Encode/data_analysis/TSS/Gencodev10_TSS_May2012.gff.gz ; JEME FANTOM5 promoter/enhancer processed data, https://www.dropbox.com/sh/wjyqyog3p5d33kh/AACx5qgwRPIij44ImnzvpFxUa/Input%20files/FANTOM5/1_first_step_modeling?dl=0&subfolder_nav_tracking=1 ; FANTOM5 sample annotation biomart, http://biomart.gsc.riken.jp/ ; FANTOM5 DB, http://fantom.gsc.riken.jp/ ; UCSC genome browser, https://genome.ucsc.edu/ ; MEME HOCOMOCO v11 402 core mono TF motifs, http://meme-suite.org/meme-software/Databases/motifs/motif_databases.12.18.tgz

Data access

- Materials (code and data) are available at http://acgt.cs.tau.ac.il/ct-focs
- The code for reproducing CT-FOCS output and figures is available at https://github.com/Shamir-Lab/CT-FOCS (under BSD 3-Clause "New" or "Revised" license).

- The database of ct-links predicted by CT-FOCS is available at http://acgt.cs.tau.ac.il/ct-focs/download.html.
- ENCODE DNase-seq samples (106 cell types) were downloaded from GEO dataset GSE29692 [21,60,61].
- Roadmap Epigenomics DNase-seq samples (73 cell types) were downloaded from GEO dataset GSE29692 [34,60–64].
- FANTOM5 CAGE data were downloaded from http://fantom.gsc.riken.jp/ [22].

Acknowledgements

The study is supported in part by the German-Israeli Project DFG RE 4193/1-1 (to R.S. and R.E.), Israel Science Foundation (grant No. 1339/18 to R.S.), ISF grant No. 3165/19, within the Israel Precision Medicine Partnership program (to R.S.), the Koret-UC Berkeley-Tel Aviv University Initiative in Computational Biology and Bioinformatics (to R.E.) and Len Blavatnik and the Blavatnik Family foundation (to R.S). T.A.H. is supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University. R.E. is a Faculty Fellow of the Edmond J. Safra Center for Bioinformatics at Tel Aviv University. This work was carried out in partial fulfillment of the requirements for the Ph.D. degree at The Blavatnik School of Computer Science at Tel Aviv University of T.A.H.

Author contributions

T.A.H., R.E., and R.S. designed the research. T.A.H. developed the computational methods and performed the analyses. R.E. and R.S. supervised the study. All authors analyzed the data and wrote the manuscript.

Disclosure declaration

The authors declare no competing financial interests.

List of abbreviations

EP – Enhancer-Promoter, CAGE – Cap Analysis of Gene Expression, CLS – connected loop set, ct-link - Cell-Type specific enhancer-promoter link, DHS – DNase-I Hypersensitive Site, TF – transcription factor.

Supplementary material

Supplementary Table S1: TF overrespresentation q-values and overrespresentation factors in promoters and enhancers involved in ct-links identified by CT-FOCS on ENCODE data

Supplementary Table S2: TF overrespresentation q-values and overrespresentation factors in promoters and enhancers involved in ct-links identified by CT-FOCS on FANTOM5 data

Supplementary Table S3: TF overrespresentation q-values and overrespresentation factors in promoters and enhancers involved in ct-links identified by MAD-FOCS on FANTOM5 data

Supplementary Table S4: TF overrespresentation q-values and overrespresentation factors in promoters and enhancers involved in ct-links identified by JEME on FANTOM5 data

Supplementary Table S5: TF overrespresentation q-values and overrespresentation factors in promoters and enhancers involved in ct-links identified by CT-FOCS on Roadmap data

Supplementary Table S6: FANTOM5 sample annotation. The annotation maps between 808 sample IDs, 472 cell types, and three cell type categories (cell line, primary cell, or tissue).

References

1. Gloss BS, Dinger ME. Realizing the significance of noncoding functionality in clinical genomics. Exp. Mol. Med. 2018;50:97.

2. Heinz S, Romanoski CE, Benner C, Glass CK. The selection and function of cell type-specific enhancers. Nat. Rev. Mol. Cell Biol. 2015;16:144–54.

3. Bulger M, Groudine M. Enhancers: the abundance and function of regulatory sequences beyond promoters. Dev. Biol. 2010;339:250–7.

4. Fullwood MJ, Ruan Y. ChIP-based methods for the identification of long-range chromatin interactions. J. Cell. Biochem. 2009;107:30–9.

5. Lieberman-aiden E, Berkum NL Van, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science. 2009;326:289–93.

6. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell. 2014;159:1665–80.

7. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012;485:376–80.

8. Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature. 2013;503:290–4.

9. Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell. 2012;148:84–98.

10. Tang Z, Luo OJ, Li X, Zheng M, Zhu JJ, Szalaj P, et al. CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. Cell. 2015;163:1611–27.

11. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473:43–9.

12. Roy S, Siahpirani AF, Chasman D, Knaack S, Ay F, Stewart R, et al. A predictive modeling approach for cell line-specific long-range regulatory interactions. Nucleic Acids Res. 2015;43:8694–712.

13. He B, Chen C, Teng L, Tan K. Global view of enhancer-promoter interactome in human cells. Proc. Natl. Acad. Sci. 2014;111:E2191–9.

14. Whalen S, Truty RM, Pollard KS. Enhancer – promoter interactions are encoded by complex genomic signatures on looping chromatin. Nat. Genet. 2016;48:488.

15. Cao Q, Anyansi C, Hu X, Xu L, Xiong L, Tang W, et al. Reconstruction of enhancer-target

networks in 935 samples of human primary cells, tissues and cell lines. Nat. Genet. 2017;201:7.

16. Rajarajan P, Borrman T, Liao W, Schrode N, Flaherty E, Casiño C, et al. Neuron-specific signatures in the chromosomal connectome associated with schizophrenia risk. Science. 2018;362:eaat4311.

17. Gasperini M, Hill AJ, McFaline-Figueroa JL, Martin B, Kim S, Zhang MD, et al. A genome-wide framework for mapping gene regulation via cellular genetic screens. Cell. 2019;176:377–90.

18. Jung I, Schmitt A, Diao Y, Lee AJ, Liu T, Yang D, et al. A compendium of promoter-centered long-range chromatin interactions in the human genome. Nat. Genet. 2019;51:1442–9.

19. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Kheradpour P, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317.

20. Consortium EP, others. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.

21. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. Nature. 2012;489:75–82.

22. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014;507:455–61.

23. Hait TA, Amar D, Shamir R, Elkon R. FOCS : a novel method for analyzing enhancer and gene activity patterns infers an extensive enhancer-promoter map. Genome Biol. 2018;19:59.

24. Krijger PHL, de Laat W. Regulation of disease-associated gene expression in the 3D genome. Nat. Rev. Mol. Cell Biol. 2016;17:771–82.

25. Pombo A, Dillon N. Three-dimensional genome architecture: players and mechanisms. Nat. Rev. Mol. Cell Biol. 2015;16:245–57.

26. Olivares-Chauvet P, Mukamel Z, Lifshitz A, Schwartzman O, Elkayam NO, Lubling Y, et al. Capturing pairwise and multi-way chromosomal conformations using chromosomal walks. Nature. 2016;540:296.

27. Song W, Sharan R, Ovcharenko I. The first enhancer in an enhancer chain safeguards subsequent enhancer-promoter contacts from a distance. Genome Biol. 2019;20:197.

28. Kumasaka N, Knights AJ, Gaffney DJ. High-resolution genetic mapping of putative causal interactions between regions of open chromatin. Nat. Genet. 2019;51:128.

29. Xi W, Beer MA. Local epigenomic state cannot discriminate interacting and non-interacting enhancer--promoter pairs with high accuracy. PLoS Comput. Biol. 2018;14:e1006625.

30. Javierre BM, Burren OS, Wilder SP, Kreuzhuber R, Hill SM, Sewitz S, et al. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. Cell. 2016;167:1369–84.

31. Sheffield NC, Thurman RE, Song L, Safi A, Stamatoyannopoulos JA, Lenhard B, et al. Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. Genome Res. 2013;23:777–88.

32. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of

lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell. 2010;38:576–89.

33. Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, et al. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nat. Methods. 2009;6:283.

34. Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, et al. An expansive human regulatory lexicon encoded in transcription factor footprints. Nature. 2012;489:83–90.

35. Nechanitzky R, Akbas D, Scherer S, György I, Hoyler T, Ramamoorthy S, et al. Transcription factor EBF1 is essential for the maintenance of B cell identity and prevention of alternative fates in committed cells. Nat. Immunol. 2013;14:867.

36. Zhang K, Li N, Ainsworth RI, Wang W. Systematic identification of protein combinations mediating chromatin looping. Nat. Commun. 2016;7:12249.

37. Alexander JM, Guan J, Li B, Maliskova L, Song M, Shen Y, et al. Live-cell imaging reveals enhancer-dependent Sox2 transcription in the absence of enhancer proximity. Singer RH, Struhl K, Liu Z, editors. Elife. 2019;8:e41769.

38. Duren Z, Chen X, Jiang R, Wang Y, Wong WH. Modeling gene regulation from paired expression and chromatin accessibility data. Proc. Natl. Acad. Sci. 2017;114:E4914–23.

39. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann. Stat. 2001;1165–88.

40. Leys C, Ley C, Klein O, Bernard P, Licata L. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. J. Exp. Soc. Psychol. 2013;49:764–6.

41. Hou C, Li L, Qin ZS, Corces VG. Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. Mol. Cell. 2012;48:471–84.

42. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature. 2012;485:381.

43. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, et al. Three-dimensional folding and functional organization principles of the Drosophila genome. Cell. 2012;148:458–72.

44. Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat. Genet. 2015;47:598.

45. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal. 2006;Complex Sy:1695.

46. Ho DE, Imai K, King G, Stuart EA. MatchIt: nonparametric preprocessing for parametric causal inference. J. Stat. Softw. 2011;42:1–28.

47. Kulakovskiy I V, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI, et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res. 2017;46:D252–9.

48. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009;37:W202–8.

49. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011;27:1017–8.

50. R Core Team. R: A Language and Environment for Statistical Computing. R Found. Stat. Comput. Vienna, Austria; 2018;

51. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for Computing and Annotating Genomic Ranges. PLOS Comput. Biol. 2013;9:e1003118.

52. Lawrence M, Gentleman R, Carey V. rtracklayer: an R package for interfacing with genome browsers. Bioinformatics. 2009;25:1841–2.

53. Harmston, N., Ing-Simmons, E., Perry, M., et al. GenomicInteractions: An R/Bioconductor package for manipulating and investigating chromatin interaction data. BMC Genomics. 2015;16:963.

54. Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team. nlme: linear and nonlinear mixed effects models. 2018.

55. Brooks ME, Kristensen K, van Benthem KJ, Magnusson A, Berg CW, Nielsen A, et al. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. R J. 2017;9:378–400.

56. Quinlan AR, Hall IM. BEDTools : a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.

57. R Core Team. A Language and Environment for Statistical Computing. R Found. Stat. Comput. 2018;

58. Wickham H. ggplot2: elegant graphics for data analysis. Springer; 2009.

59. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw WHA, Lumley T, et al. gplots: various R programming tools for plotting data. 2016.

60. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337:1190–5.

61. Polak P, Lawrence MS, Haugen E, Stoletzki N, Stojanov P, Thurman RE, et al. Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. Nat. Biotechnol. 2014;32:71.

62. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, et al. The NIH roadmap epigenomics mapping consortium. Nat. Biotechnol. 2010;28:1045–8.

63. Polak P, Karlić R, Koren A, Thurman R, Sandstrom R, Lawrence MS, et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature. 2015;518:360.

64. Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, et al. Human body epigenome maps reveal noncanonical DNA methylation variation. Nature. 2015;523:212.