bioRxiv preprint doi: https://doi.org/10.1101/707158; this version posted July 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 interaction measured by ChIA-PET. We expand this validation scheme by introducing the concept of connected loop set, which combines loops that overlap in their anchor sites. Analzying 1,366 samples from ENCODE, Roadmap epigenomics and FANTOM5, CT-FOCS inferred highly cell type-specific EP links more accurately than a state-of-the-art method. We illustrate how our inferred EP links drive cell type-specific gene expression and regulation.

Keywords

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

Background

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). Such TFs mostly bind to enhancer elements located distally from their target promoters [2]. To find cell type-specific enhancer-promoter links (ct-links) one needs to compare EP links across multiple and diverse cell types. Deciphering ct-links using 3D chromatin conformation sequencing data (e.g., ChIA-PET [3] and HiC [4,5]) is still not available for many distinct cell types and tissues [4–9]. Consequently, there is high demand 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 enabe concurrent quantification of enhancer and promoter elements.

A key challenge is to identify which of the numerous candidate 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. [10] 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 [11] 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 [12] and

bioRxiv preprint doi: https://doi.org/10.1101/707158; this version posted July 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

TargetFinder [13]. 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 [14]. Each reported EP link is given a score denoting tendency to be active in a given cell type. JEME reports 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. [15] used in-situ HiC and 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 enhancer-promoter interactions (or ct-links). Gasperini et al. [16] 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 transcription factors (TFs; reviewed in [17]). Remarkably, both studies reported far fewer links than JEME, indicating that only a small portion of EP links that are active in a cell type are specific for it.

Here, we introduce CT-FOCS (*C*ell *T*ype *F*DR-corrected *O*LS with *Cross*-validation and *S*hrinkage), 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). It is built upon our previously published method, FOCS [18], which infers global EP links that show high correlation between the enhancer and the promoter activity patterns across many samples. 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 DNase Hypersensitive Sites (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. The ct-links inferred using CT-FOCS are available at http://acgt.cs.tau.ac.il/ct-focs.

Results

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

We developed a novel method called CT-FOCS (Cell Type FOCS) 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 Rodamap epigenomics DHS profiles [19–21], which contain 208 and 350 samples covering 106 and 73 distinct cell lines, respectively; and (2) FANTOM5's CAGE profiles [22], which contain 808 samples covering 472 cell lines, primary cells, and tissues (**Methods**).

CT-FOCS is based on FOCS [18], which discovers global EP links showing correlated enhancer and promoter activity patterns across many samples. FOCS uses linear regression followed by two non-parametric statistical tests for producing initial promoter models, and applies a model shrinkage regularization to retrieve the most informative enhancers per promoter

bioRxiv preprint doi: https://doi.org/10.1101/707158; this version posted July 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

model (out of the k enhancers that are closest to the target promoter; we used k=10). To find ct-links based on the global links identified by FOCS, CT-FOCS starts with the full (that is, non-regularized) promoter models. It uses a mixed effect regression-based method, which utilizes 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**). The input to CT-FOCS is enhancer and promoter activity matrices over the same samples and a unique cell type labeling for each sample. The output is a set of ct-links for each cell type (**Methods**).

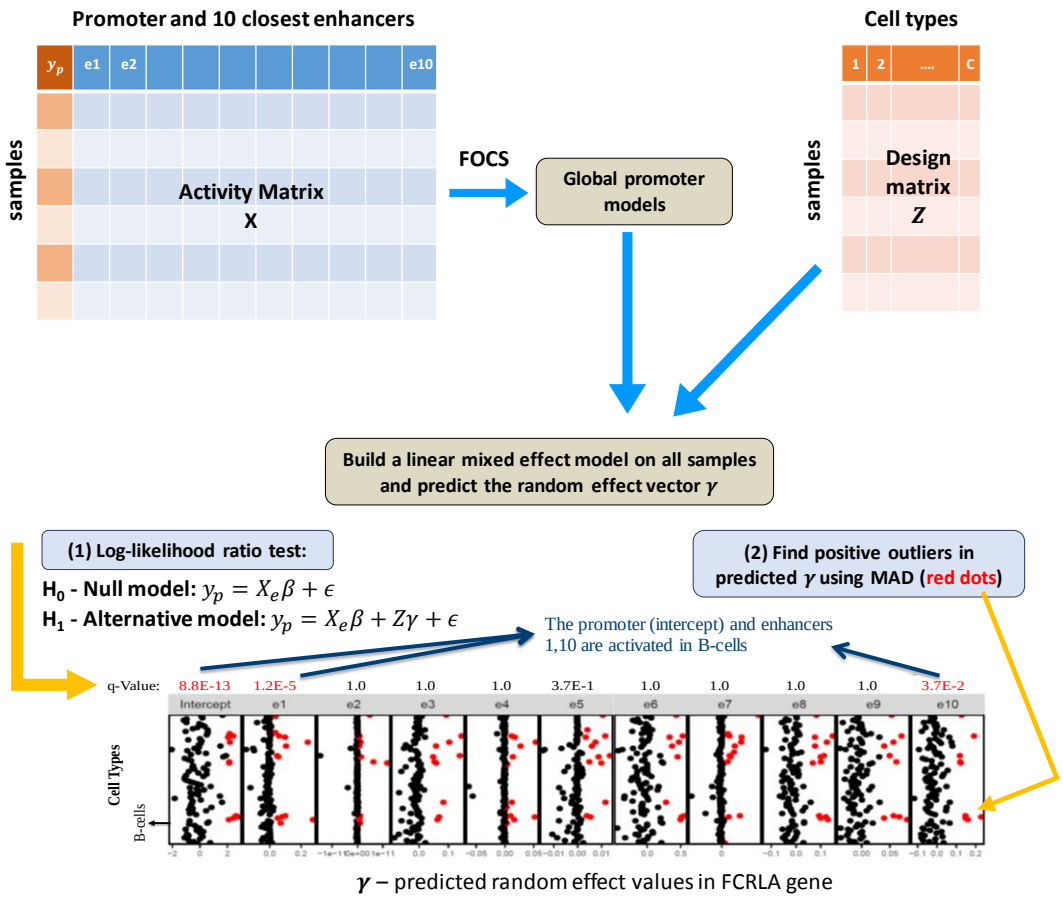

Figure 1. Outline of the CT-FOCS algorithm. Let y_p denote the observed promoter p activity, X be the activity matrix of the 10-closest enhancers to the promoter p, and Z a design matrix classifying each sample to its unique cell type (**Methods**). First, CT-FOCS infers robust global promoter models by applying the leave-cell-type-out cross validation step from FOCS on each promoter model p (**see FOCS** [18] **for further description**). Second, CT-FOCS builds a linear mixed effects model (LMM) on all samples using y_p, X, and Z. Then, CT-FOCS performs two tests: (1) log-likelihood ratio test (LRT) to compare between the Null model, which is the simple linear regression without $Z\gamma$, and the LMM model. For each promoter model, such tests are carried out eleven (k+1) times (testing the k enhancers and the intercept), such that in each test Z is a design matrix of either the intercept or one of the ten enhancer coefficients (**Methods**). Then, p-values for these LRT tests are adjusted for multiple testing (q-values). (2) Each LMM predicts a γ vector of size C (number of cell types) for either the intercept or one of the enhancers. CT-FOCS standardizes γ values using the Median Absolute Deviation (MAD) technique and finds positive outliers (red dots) (**Methods**). Finally, CT-FOCS calls cell type-specific EP links (ct-links) defined as: (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

bioRxiv preprint doi: https://doi.org/10.1101/707158; this version posted July 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

gene given as an example, the promoter 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 (median) of 167 (94), 234 (73) and 414 (594) ct-links per cell type in (**Table 1; Supplementary Fig. 1A-C**). These numbers are in line with the low number of reported ct-links experimentally observed for NPC, neurons, and K562 cells [15,16], and indicates that cell-type specific EP links constitute only a small portion of the EP links that are active in any particular cell type. CT-FOCS predicted EP links are on average shared across 3, 1.65, and 2.5 cell types in ENCODE, Roadmap, and FANTOM5 datasets (**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 Epigenomics, and FANTOM5 datasets, respectively (**Supplementary Fig. 3A-C**). Subsequent analyses are performed on ENCODE and FANTOM5 datasets. The complete set of predicted ct-links for each cell type is available at http://acgt.cs.tau.ac.il/ct-focs.

Table 1. Average number of predictions per cell type

Dataset	Avg. ct-links	Avg. enhancers	Avg. promoters	Avg. genes*	Tot. ct-links	Cell type with max. 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

ChIA-PET connected loops as validation for inferred EP links

A gold standard for validating inferred ct-links is 3D chromatin contact loops mediated by RNAP2. The straightforward validation of an inferred ct-link is to check whether the E and P regions overlap different anchors of a certain ChIA-PET loop. However, as loops indicate 3D proximity of their anchors, overlapping anchors of different loops indicate proximity of their other anchors as well. Thus, we expand the procedure of validating computationally inferred EP links by introducing the concept of connected loop sets (**CLS**). We consider two ChIA-PET loops as connected if they have at least one overlapping anchor. More generally, a CLS is defined as the set of anchors of all loops that have a chain of connected loops between them (**Supplementary Fig. S4A**). Hence, if the enhancer and promoter regions of an EP link overlap different anchors from the same ChIA-PET CLS then we can view this as support of the predicted link as well (**Methods**) [23]. In addition, EP links that span a linear distance of < 20kb, where ChIA-PET loops may perform poorly [24], may not be supported by a single ChIA-PET loop, but a CLS may support such short EP links.

Using ChIA-PET RNAP2-mediated loops measured in GM12878 cell line [9] (see **Supplementary Fig. 5** for the summaries of the CLSs found), while ~30% of CT-FOCS inferred

bioRxiv preprint doi: https://doi.org/10.1101/707158; this version posted July 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

GM12878-specific EP links were validated by GM12878 ChIA-PET single loops, 72% were supported by CLSs (**Fig. 2;** see **Supplementary Fig. 6-7** for additional examples). To test the significance of the presicion obtained (72%, 202/280 GM12878-specific EP links) we selected random sets of 280 EP links with the same linear distances between E and P as the true EP links (**Methods**). Each random link was taken from the same chromosome as the true link in order to account for chromosome-specific epigenetic state [25]. In 1,000 random sets, CLSs supported on average 24% (68 out of 280) and at most 31% (87 out of 280) (P<0.001; **Supplementary Fig. 4B**).

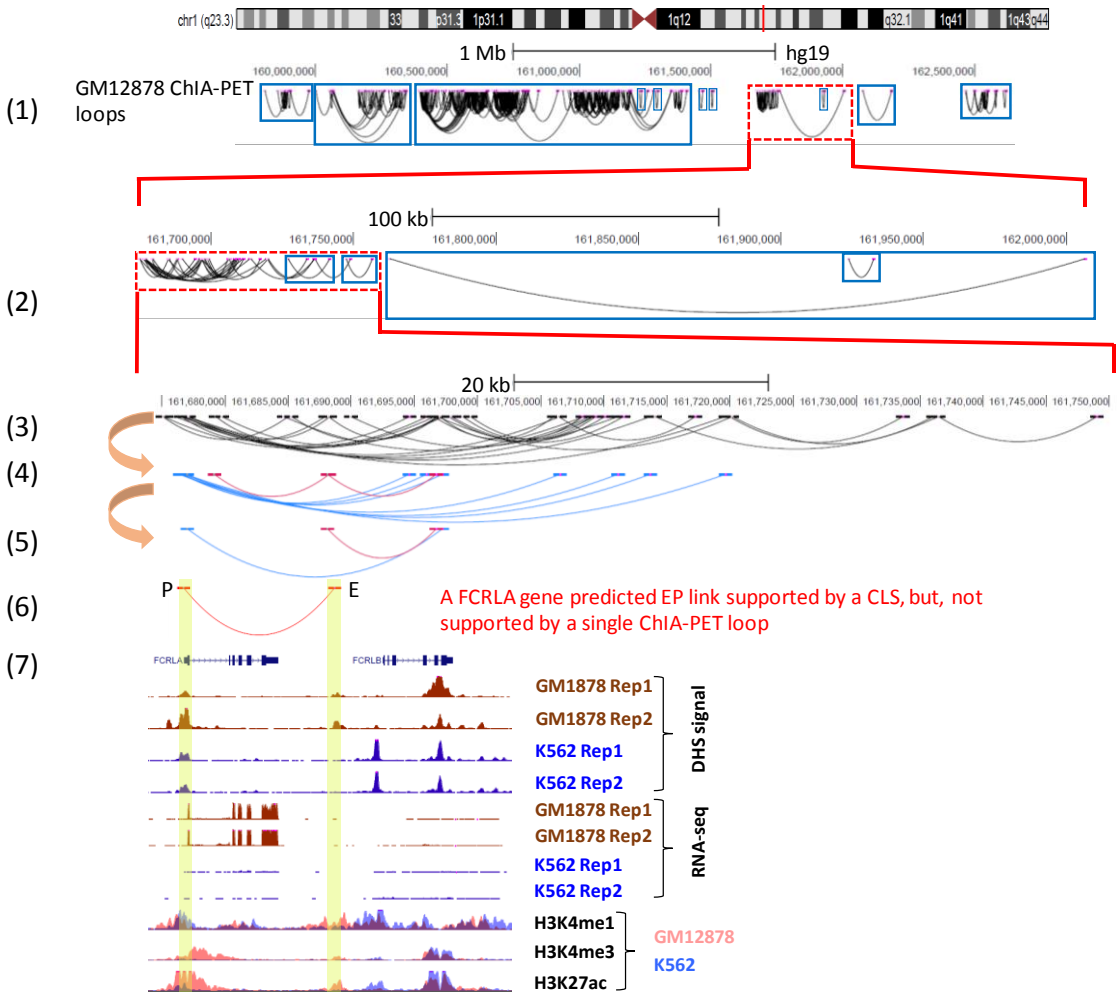

Figure 2. ChIA-PET CLSs support predicted ct-links. (1) A 2Mb region of chromosome 1 showing all ChIA-PET loops detected for cell type GM12878. Rectangles indicate connected loop sets (CLSs). (2) Zoom-in on 300kb region. Note the nested CLSs on the right. (3) Another zoom-in on a 70kb region, showing a single ChIA-PET CLS. (4) The same region showing only loops that have anchors overlapping the enhancer or promoter of the examined ct-link shown in (6). Pink loops: loops overlapping the enhancer; blue loops: loops overlapping the promoter. (5) Two connected loops in a CLS that collectively support the predicted ct-link shown in (6). Therefore, this CT-FOCS inferred ct-link in (6) is validated by a CLS, but not by individual ChIA-PET loops. (6) A predicted ct-link of FCRLA gene. (7) 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.

bioRxiv preprint doi: https://doi.org/10.1101/707158; this version posted July 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

CT-FOCS aims to pinpoint EP links that are highly cell type-specific. To evaluate the specificity of the predictions we used a cell type specificity formula, which compares activity of ct-links inferred for a particular cell type with their activity in all other cell types [26] (**Methods**). On ENCODE data, CT-FOCS predicted 280 GM12878-specific EP links (with 72% support by ChIA-PET connected loops). The lymphocyte group of cell types (GM12878, other B-cells, and T-cells) exhibited the highest EP signals, while GM12878 ranked first by EP signal specificity (**Fig. 3A-B**).

Next, we also examined the cell type specificity in gene expression of the genes annotated with CT-FOCS inferred ct-links (**Methods**). For this task, we analyzed gene expression (GE) data for 112 cell types [27] and as a test case focused on the expression of the set of 124 genes whose promoter was included in the 280 GM12878-specific ct-links. Here too, the lymphocyte group showed the highest GE levels compared to other cell type (**Fig. 3C**), and GM12878 ranked fourth by GE specificity (**Fig. 3D**). These results indicate that the GM12878-specific EP links predicted by CT-FOCS based on DHS data are correlated with GM12878-specific GE programs.

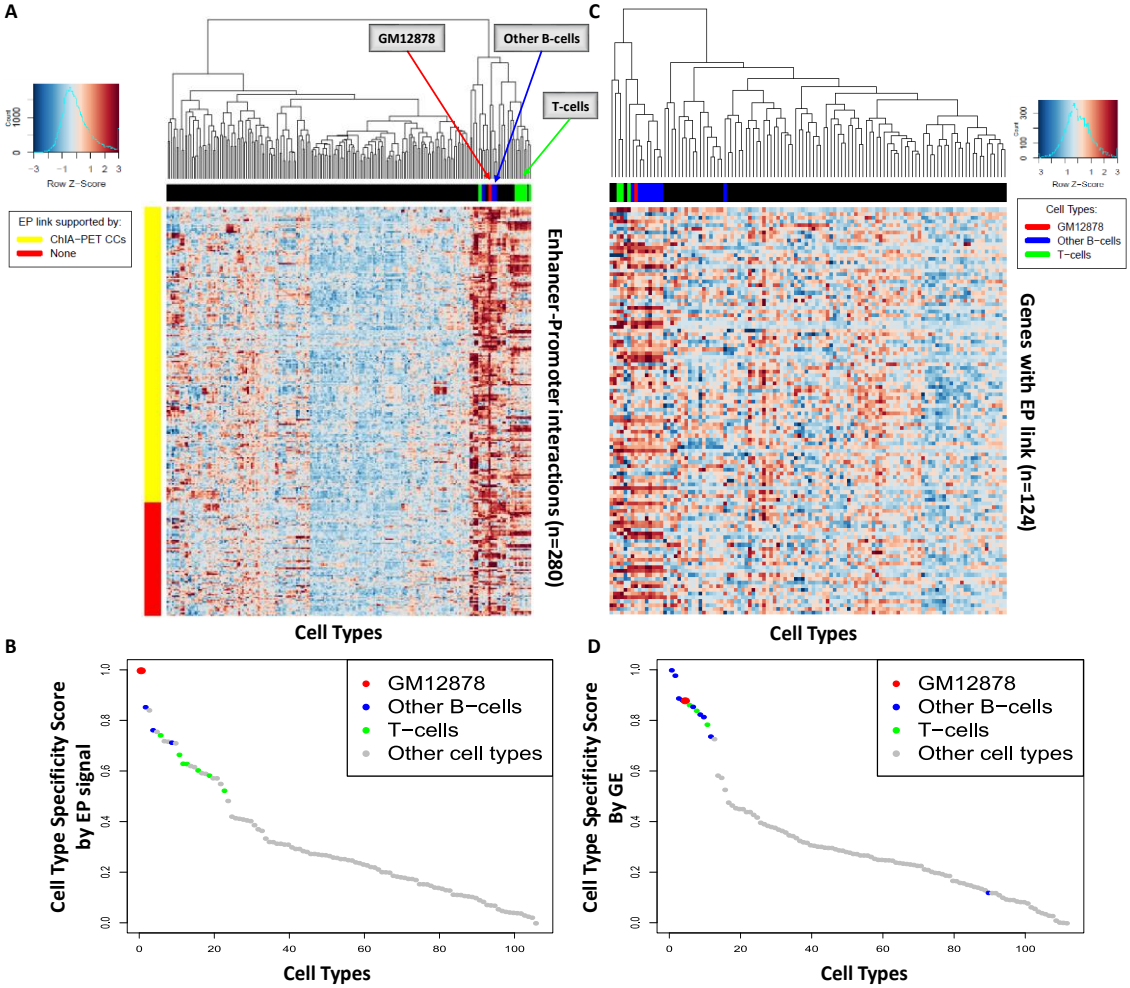

Figure 3. Specific interaction activity and target gene expression of ct-links predicted for GM12878. (A) Heatmaps of EP signals for 280 ct-links predicted on GM12878 based on ENCODE data spanning 106 cell types. Rows – EP links, columns – cell types, color – z-score of interaction

bioRxiv preprint doi: https://doi.org/10.1101/707158; this version posted July 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

signal. ct-links supported by CLSs are marked in yellow. Cell types related to lymphocytes (B/T-cells) are highlighted in color. **(B)** Cell type specificity score ranks based on the signal of GM12878-specific EP links for all 106 cell types. **(C)** Heatmap of gene expression (GE) for 124 genes involved in ct-links predicted on GM12878. The ENCODE GE data covers 112 cell types [27]. Rows – genes, columns – cell types, color – z-score of GE. **(D)** Cell type specificity score ranks based on GE of the 124 genes annotated with a GM12878 ct-link for the 112 cell types included in the GE dataset (**Methods**).

Comparison of CT-FOCS and JEME

We compared CT-FOCS predictions with those of JEME [14], which infers EP links that are active in a particular cell type but are not neccesarily cell type-specific (see **Supplementary Fig. 1D, 2D, and 3D** for JEME inferred EP links, how many of them are shared across cell types, and their distances per cell type). Predictions were made on the FANTOM5 dataset, and subjected to specificity analysis. We used FANTOM5 data for this comparios as JEME prediction based on ENCODE/Roadmap data were based on multiple omic data types measured by this project, while FANTOM5 is a single omic dataset. For cell type GM12878, CT-FOCS identified 210 genes (340 ct-links, 252 enhancers and 228 promoters). These genes showed high GE in the lymphocyte group compared to other cell types (**Fig. 4A**). JEME reported 9,065 GM12878-active EP links covering 4,268 genes (5,216 promoters and 2,338 enhancers), with a GE profile that is only weakly specific to the lymphocyte group (**Fig. 4A**). In terms of cell type specificity of their GE profiles, CT-FOCS ranked GM12878 2nd (out of 328 cell types having at least 50 CT-FOCS predicted ct-links, with GM18507 ranked first), while JEME ranked GM12878 17th (**Fig. 4B**).

We also calculated GE specificity for cell types HepG2, K562, and MCF-7 based on the links identified by both algorithms. In CT-FOCS, all three cell types ranked first in specificity, while in JEME, HepG2, K562, and MCF-7 ranked 2, 3, and 14, respectively. In addition, we computed the distribution of EP specificity ranking of all 328 cell types analyzed. CT-FOCS cell type ranks were significantly higher than those of JEME (average 34 and median 21 for CT-FOCS, vs. average 83 and median 69 for JEME; P<9.6E-39; one sided Wilcoxon paired test; **Fig. 4B**).

Next, we tested to what extent the EP links inferred for GM12878 by both methods were supported by the ChIA-PET assay of that cell type compared to the links inferred on other cell types. Biologically, ct-links inferred for other cell type are expected to show lower support rate by GM12878-CLSs compared to GM12878-specific EP links. For GM12878 and each of the other examined cell types, we compared the fraction of ct-links supported by GM12878 ChIA-PET CLSs. Indeed, CT-FOCS ct-links predicted for GM12878 showed significantly higher support rate (median $\log_2$(ratio) ~2.5; **Fig. 4C**). Most of the cell types that had $\log_2$(ratio) <0 were related to B-cells as GM12878 (e.g., B cell line, burkitt's lymphoma cell line). 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 $\log_2$FC ~0; P<2.5E-33; one sided Wilcoxon paired test; **Fig. 4C**). These results indicate again that a large portion of EP links are active across many different cell types, while a minority of them are highly cell type specific.

bioRxiv preprint doi: https://doi.org/10.1101/707158; this version posted July 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

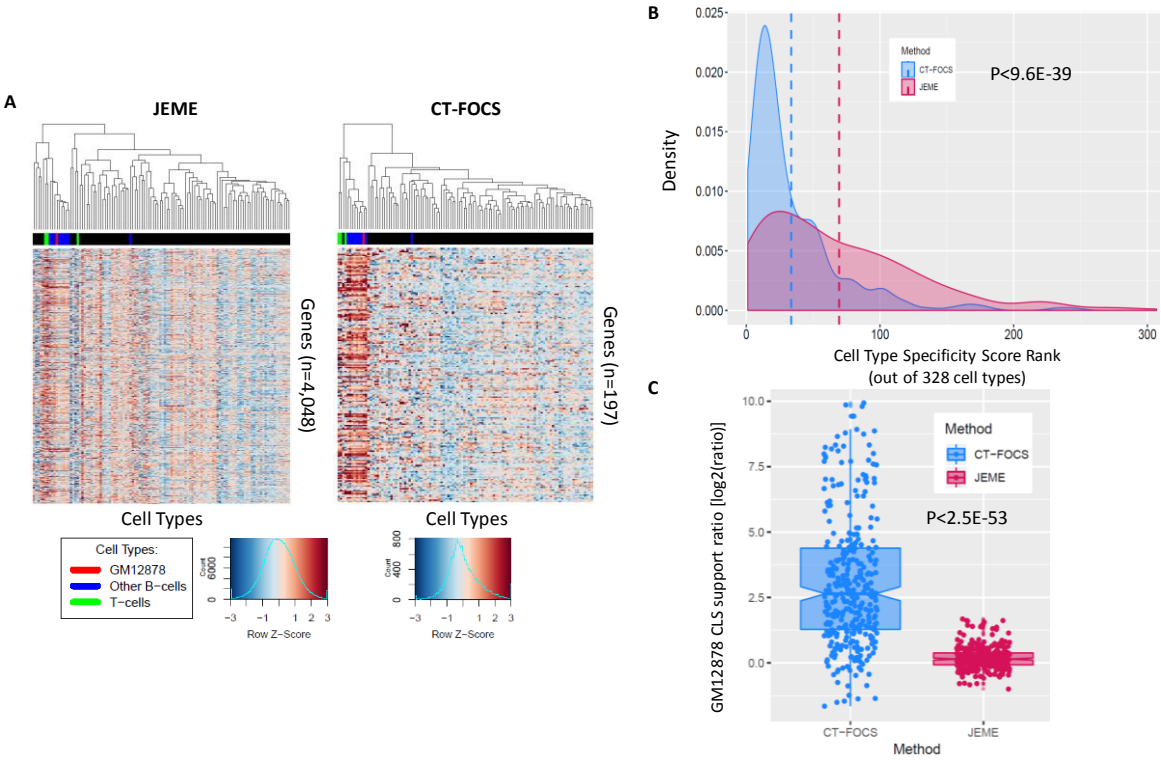

Figure 4. Cell-type specifity of EP links inferred by CT-FOCS and JEME. Active EP links were predcited by JEME and ct-links were predicted by CT-FOCS on 472 cell types from FANTOM5. **(A)** Heatmaps of gene expression (GE) from ENCODE microarray data covering 112 cell types [27]. For each method, the analysis includes the genes whose promoters were contained in the set of EP-links predicted for GM12878. Cell types (columns) related to lymphocytes (B/T-cells) are highlighted. **(B)** Distribution of specificity ranks of EP signal across 328 cell types. The plots show the density of cell type specificity score ranks based on EP signals of the EP links predicted by CT-FOCS (blue) and JEME (red). Dashed lines denote the average rank. The scores were calculated on 328 cell types that had at least 50 predicted EP links in both CT-FOCS and JEME. **(C)** The ratio between the fraction of predicted EP links on GM12878 that had a CLS support to the fraction in 327 other cell types (**Methods**). P-values computed using 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-specfic gene regulation, as shown previously for lineage-determining TFs [17] and in K562 [16]. To this end, we searched for occurrences of 402 known TF motifs (position weight matrices; PWMs) within the enhancers and promoters of inferred EP 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, which tend to be stably bound by TFs [28]. We used publicly available ~8.4M DGFs covering 41 diverse cell and tissue types derived from ENCODE DHS data [29]. For each TF PWM and cell type, we calculated the TF overrepresentation factor 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. 13 TFs were identified in promoters and ehnahcers. The identified TFs show as a regulatory group

bioRxiv preprint doi: https://doi.org/10.1101/707158; this version posted July 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

higher overrepresentation factor in both enhancers and promoters compared to the other cell types (**Fig. 5A-B**). In terms of the specificity score of the TF overrepresentation factors, GM12878 is 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 discovered TFs are strictly GM12878-specific regulation group.

Among the TFs that were discovered in analysis of the ct-links 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) ranked 7th in enhancers known to drive B-cell lineage commitment [30]. EBF1 was predicted to modify 3D organization of chromatin by cooperating with PAX5, and it was shown that discovered TFs BATF, RUNX3, IRF4, and PAX5 are enriched in GM12878 [31].

We applied the same TF analysis and specificity ranking on the ct-links inferred from ENCODE for 68 cell types (out of 106) 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 by overrepresentation factor 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 above analyses on the FANTOM5 dataset for 328 cell types (out of 472) that had at least 50 predicted EP links by both CT-FOCS and JEME. The analysis identified 15 TFs on average in enhancers and 25 in promoters per cell type in CT-FOCS EP links compared to 33 and 69 in JEME (**Supplementary Tables S2-3**). In enhancers, 186 out of the 328 cell types ranked first by overrepresentation factor specificity, while in promoters, 190 out 328 ranked first, in CT-FOCS. JEME had 112 cell types in enhancers and 159 in promoters ranking fisrt. Our results show that CT-FOCS tend to find gene regulation that is more cell type-specific compared to JEME, possibly suggesting that many of the TFs found in JEME act in multiple cell types' gene regulation.

bioRxiv preprint doi: https://doi.org/10.1101/707158; this version posted July 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

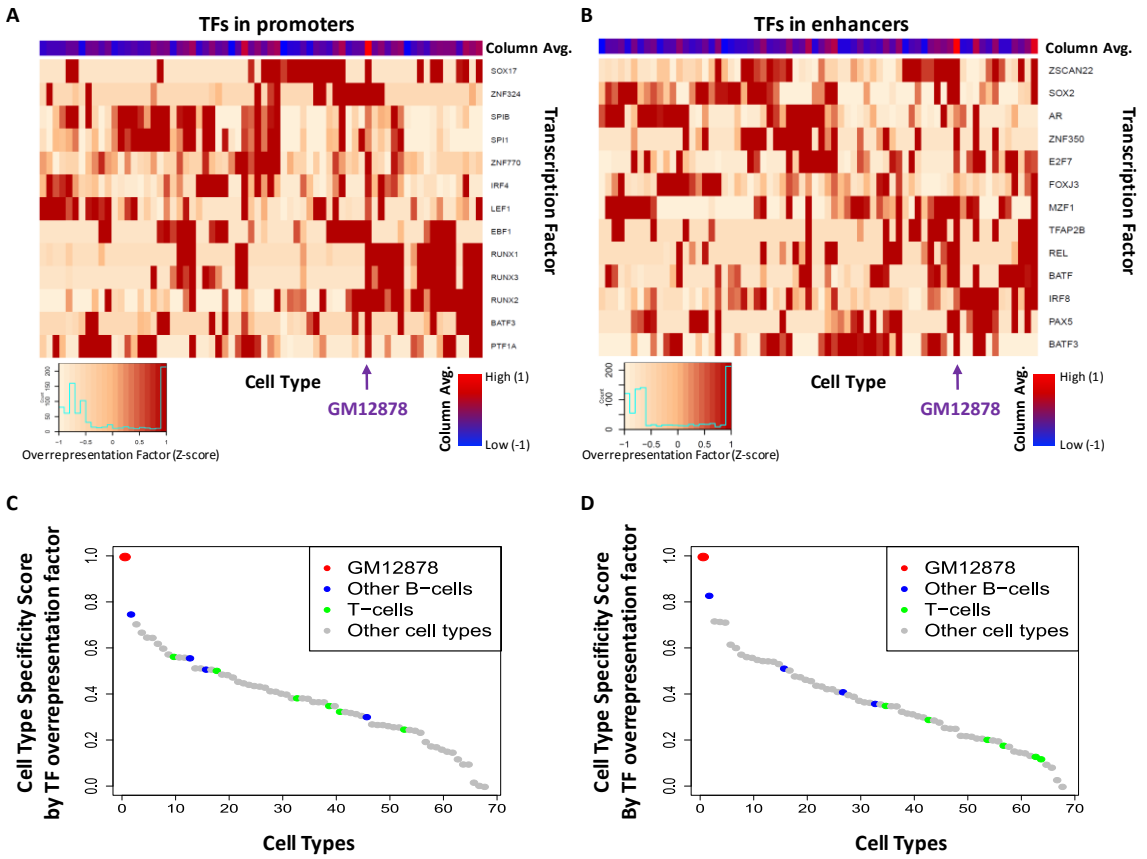

Figure 5. Overrepresented TF PWMs in ct-links. (A-B) Heatmaps TF PWM overrepresentation in promoters (A) and enhancers (B) of GM12878-specific EP links from CT-FOCS. TFs shown had P-value < 0.05 (Hyper Geometric test; **Methods**). **(C-D)** Cell type specificity score ranks based on GM12878-specific 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 two different datasets, DHS profiles in ENCODE and Roadmap [19,21], and CAGE profiles from FANTOM5 [22], and derived a rich resource of statistically validated ct-link maps.

To support the predicted ct-links, we developed a simple novel scheme based on ChIA-PET connected loop sets (CLSs, **Methods**; **Fig. 2** and **Supplementary Fig. 6-7**). We showed, using GM12878 ChIA-PET dataset [9], that ~72% of the ct-links predicted by CT-FOCS for GM12878 on the ENCODE data were supported by ChIA-PET CLSs whereas only 30% of these ct-links were supported by 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 other enhancers or promoters, and has a functional effect on the indirectly linked promoter. It can also overcome the "blind spot" of ChIA-PET in discovering interactions shorter than 8kb. We believe that evaluation using CLS support may help to improve future methods for EP inference.

bioRxiv preprint doi: https://doi.org/10.1101/707158; this version posted July 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The specificity score suggested in [26] allowed us 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**). Also, it allowed us to present a global summary of the cell type-specificity of the inferred features across many cell types, and use it to compare with different EP inference method. Based on FANTOM5 dataset, we showed that CT-FOCS ct-link predictions were more specific than JEME's (**Fig. 4C**). We further showed that the predicted ct-links drove cell type-specific GE and revealed overrepresented TFs in the link's enhancers and promoters. Taken together, we saw high correlation between cell type-specific EP DHS signals, GEs, and TF overrepresentation factors (**Fig. 3A-B** and **Fig. 5A-B**).

Several comments are in order regarding our CT-FOCS inferred EP maps. 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 [18] and ~40% in FANTOM5 [22]), it may result from two confounders: (1) The lower limit was set on the linear genomic distance between putative enhancer and promoter links (e.g., in ENCODE we set the distance to at least 10kb 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, less EP links are identified using correlation-based techniques (e.g., linear regression). In contrast, the naïve nearest gene mapping to each enhancer has been previously shown to result with a poor performance when validating with ChIA-PET and HiC 3D loops and eQTL data [14]. Second, ~60% of the predicted EP links, on average per cell type, involve intronic enhancers, similar to the report by FOCS (70%). Third, the average number of CT-links per cell type was 167 in ENCDOE, 234 in Roadmap, and 414 in FANTOM5, respectively (**Table 1**). While these numbers are very low, considering that ENCODE reported a total of ~3M regulatory elements [19], they are in line with the small number of cell type-specific EP links reported previously for NPC, neuron, and K562 cells [15,16]. Fourth, each promoter was linked to ~2 enhancers on average (and a maximum of 9) in each cell type.

CT-FOCS uses LMM 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 [14]. The second effect is the contribution of the disjoint cell type groups of samples to the promoter activity. By taking into account the cell type of each sample we can ask whether should we 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 (**Methods**).

CT-FOCS is limited by considering only the ten closest enhancers to each promoter when building the LMM models. A possible future improvement is to include all enhancers within a window of say 1Mb around each promoter, e.g., by using Bayesian hierarchical models, considering possible confounders and a-priori information such as ChIA-PET loops and eQTLs.

bioRxiv preprint doi: https://doi.org/10.1101/707158; this version posted July 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We used CT-FOCS to construct a broad, publicly available compendium of ct-links for 651 cell types, which can be useful for multiple genomic inquiries. For example, it can improve identification of known and novel cell type-specific TFs and enhance our understading of key transcriptional cascades that determine cell fate decisions. Furthermore, the integration of protein-protein intractions (PPIs) with TF identification in predicted ct-links may help identifiy cell type-specific PPI modules [32]. 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 idenified 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 for FANTOM5 data showed substantially higher cell type-specificity scores compared to a previous state-of-the-art method.
- The inferred ct-links correlate with cell type-specific gene expression and regulation.
- We provided a novel way to support predicted EP links that uses ChIA-PET connected loops insead of single loops. On GM12878 cell line, 72% of the predicted GM12878-specific EP links were supported by connected loops compared to 30% by single loops.

Methods

ENCODE DHS data preprocessing

ENCODE DHS peaks of enhancers and promoters [19] were processed as in FOCS [18] with the following changes: (1) we analyzed only promoters of annotated protein-coding genes according to GencodeV10 TSS annotations (**see URLs**). (2) We applied a relative-log-expression (RLE) normalization [33], as implemented in edgeR [34,35]. (3) We retained promoters and enhancers that showed robust activity in at least one cell type: signal ≥ 5 RPKM in all samples of at least one cell type. Overall, we analyzed 208 sampels from 106 cell types. Our preprocessing resulted with 36,056 promoters (mapped to 13,105, 13,464, and 13,197 protein-coding genes according to HGNC_symbols, Ensembl, and Entrez, respectively) and 658,231 putative enhancers.

Enhancers closer than 10kb to the nearest promoter were discarded since we wanted to reduce false positive links due to the high signal correlation at short distances, and to predict distal interactions as suggested by [13]. The candidate enhancers for each promoter were defined as the 10 closest enhancers located within a window of 1Mb (±500Kb upstream/downstream) from the promoter's center position.

We first applied the FOCS pipeline, including leave-cell-type-out cross validation (LCTO CV) on the promoters and their candidate enhancers, and accepted promoter models with q-value≤ 0.1 in the activity level test (see [18] for details). Unlike FOCS, we did not apply here regularization on the predicted EP links. Overall, the procedure resulted with 17,832 promoter

bioRxiv preprint doi: https://doi.org/10.1101/707158; this version posted July 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

models (mapped to 9,090, 9,320, and 9,160 HGNC_symbols, Ensembl, and Entrez protein-coding genes, respectively).

Roadmap Epigenomics DHS data preprocessing

Roadmap DHS peaks of enhancers and promoters [21] were processed as described above for the ENCODE data with the following change: we retained promoters and enhancers with signal ≥1RPKM in all samples of at least one cell type. Overall, we analyzed 350 sampels from 73 cell types. Our preprocessing resulted with 12,493 promoters (mapped to 10,022, 10,358 and 10,101 protein-coding genes according to HGNC_symbols, Ensembl and Entrez, respectively) and 448,356 putative enhancers.

For each promoter, the candidate enhancers for each promoter were defined as the 10 closest enhancers located within a window of 1Mb (±500Kb upstream/downstream) from the promoter's center position. Unlike ENCODE, we did not enforce a lower bound on the distance here since ~91% of the enhancers are located within intergenic regions, thus, reducing the probability of FP links due to the high signal correlation at short distances. We applied the same FOCS pipeline on the promoters and their candidate enhancers as described above for the ENCODE data. This resulted with 8,505 promoter models for further analysis.

FANTOM5 CAGE data preprocessing

We downloaded the FANTOM5 CAGE data from JEME [14] repository (**see URLs**). Overall, the data contained 24,048 promoters and and 42,656 enhancers, covering 808 samples. Enhancer and promoter expression matrices were RLE normalized. We manually annotated the 808 samples with 472 cell types (**Supplementary Table S5**) using **Table S1** from FANTOM5 [22].

For each promoter, the candidate enhancers were defined as the 10 closest enhancers located within ±1Mb from the promoter's TSS as performed in JEME [14]. Unlike ENCODE, we did not enforce a lower bound on the distance here. We applied the same pipeline on the promoters and their candidate enhancers as described above for the ENCODE data. This resulted with 21,468 promoter models for further analysis.

CT-FOCS model Implementation

Consider a model for promoter p and its k closest enhancers that passed the preprocessing step. The activity vector 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 same n samples is summarized in the matrix X_e of dimensions $n \cdot (k+1)$, with the first k columns corresponding to the candidate enhancers and a column of ones for the intercept. There are C cell types and each sample is labeled with a cell type.

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. 8A-C**). For Roadmap and FANTOM5, we applied the regular linear mixed effect regression using lmm R function from nlme R package [36]. For ENCODE, we applied the generalized linear mixed effect regression (GLMM) using glmmTMB R function

bioRxiv preprint doi: https://doi.org/10.1101/707158; this version posted July 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

from glmmTMB package [37]. We defined the null model as the simple linear regression $y_p = X_e\beta + \epsilon$, and each of the *k+1* alternative models as the LMM model $y_p = X_e\beta + Z\gamma + \epsilon$, where $X_e\beta$ is the fixed effect, $Z\gamma$ is the random effect, and ϵ is a random error. Z is a nxC design matrix that groups the samples by their cell types, namely:

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

Where $l \in \{1, \dots, k+1\}$ is one of the variables (enhancer or the intercept). γ is a C-long vector of random effects to be predicted. In LMM, we estimate the coefficients $\beta_1, \dots, \beta_{k+1}$ of the fixed effect, the random effect variance $Var(\gamma)$ assuming that $\gamma \sim N\big(0, Var(\gamma)\big)$ under the normal-like distribution (or, ZINB in GLMM), and the residual variance σ_ϵ^2. The estimated $\widehat{Var(\gamma)}$ and $\hat{\beta}_1, \dots, \hat{\beta}_{k+1}$ are used to predict γ, which is then utilized for calling cell-type specific EP links, as described below.

For each promoter model, we defined $k+1$ alternative models, each corresponding to a single random effect (i.e., random slope for enhancer or random intercept for the promoter). We applied the likelihood ratio test between the residuals of the alternative and the null models $k+1$ times (one for each random effect), and got $k+1$ p-values. Such p-values were calculated for each of the $|P|$ promoters, and corrected for multiple testing using the Benjamini-Hochberg FDR [38], where the number of tests performed is $|P| \cdot (k+1)$.

The predicted random effect vectors, $\gamma_1, \dots, \gamma_{k+1}$, of the alternative models were normalized using the median absolute deviation (MAD), i.e., $\gamma'_i = |(\gamma_i - median(\gamma_i)|/mad(\gamma_i)$. Positive outliers with $\gamma'_{ij} > 2.5$ were identified as active enhancers or promoters. We chose to use the MAD statistic since the mean and the standard deviation are known to be sensitive to outliers [39].

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 positive outliers in the same cell type according to the MAD criterion (that is, both E and P have normalized coefficient > 2.5).

External validation of predicted EP links using ChIA-PET 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 [9]. Each ChIA-PET loop identifies an interaction between two genomic intervals called its *anchors*. To focus on high confidence interactions, we filtered out loops with anchors' width >5kb or overlapping anchors. Retained loop anchors were resized to 1kbp intervals around to the anchor's center position. We filtered out loops crossing topologically associated domain (TAD) boundaries, as functional links are usually confined to TADs [6,40–42]. For this task, we downloaded 3,019 GM12878 TADs [43], which are largely conserved across cell types [5], and used them for filtering ChIA-PET loops from all cell types.

To overcome the sparseness of the ChIA-PET loops, and the 8kb minimum distance between loop anchors[8,9], we combined loops into connected components of loop sets (**CLSs**) as

bioRxiv preprint doi: https://doi.org/10.1101/707158; this version posted July 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

follows: When two loops had anchors that overlapped by at least 250 bp, we put them in the same CLS (**Supplementary Fig. S4A**). We used the igarph R package [44] 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 fell into the same CLS. Specifically, we defined 1Kbp genomic intervals (±500 bp upstream/downstream) for the promoters (relative to the center position) and the enhancers (±500 bp from the enhancer center) 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 **(Supplementary Fig. S4A)**.

We used randomization in order to test the significance of the total number of supported EP links by ChIA-PET single loops defined as 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, 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 [25]. The matching was done using MatchIt R package (method='nearest', distance='logit', replace='FALSE') [45]. 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 $P = \frac{\#(N_r \geq N_t)}{1000}$. 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. 4C**):

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

Cell type specificity score

We quantify 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 signal of the link a particular cell type is the average of the signal across the samples from that cell type. Define x_i as the vector of signals in cell type i of length equal to the total number n of EP links discovered in the study, and define $d_{c,i}$ as the Euclidean distance between the vectors x_c and x_i of two cell types. Following the definition of cell-type specificity from [26], cell type specificity in cell type c of a set of EP links is:

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

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. Given a short TF motif, finding all its occurrences (called hits) in a large set of promoters may

bioRxiv preprint doi: https://doi.org/10.1101/707158; this version posted July 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

require checking millions of possibilities and is prone to high false positive rate. We therefore limited the serath for PWM hits to digital genomic footprint regions (DGFs) that are more likely to represent genuine TF binding sites. We downloaded ~8.4M DGFs sequences inferred from DNase-seq data from ENCODE [29]. The mean DGF length was $L \approx 20$ bp, with a maximum length of 68 bp.

We looked for hits of 402 HOCOMOCO V11 [46] TF core motifs (taken from MEME suite database [47]; **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 [48] with 0-order Markov model as background created using fasta-get-markov command line from MEME suite [47]. Hits with FIMO *q*-value <0.1 were considered 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 [45]; method='nearest', distance='mahalanobis'. For each TF motif 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 motifs in the background set containing m_b possible k-mers, then the *overrepresentation factor* of the TF is defined as $(l_t/m_t)/(l_b/m_b)$

Statistical tests, visualization and tools used

All computational analyses and visualizations were done using the R statistical language environment [49]. To correct for multiple testing we used the p.adjust() function (method='BY'). We used 'GenomicRanges' package [50] for finding overlaps between genomic intervals. We used 'rtracklayer' [51] and 'GenomicInteractions' [52] packages to import/export genomic positions. Linear mixed effect regression models were created using lme R function from nlme package [36]. Generalized linear mixed effect with zero inflated negative binomial models were created using glmmTMB R function from glmmTMB package [37]. Counting reads in genomic intervals was done using BEDTools [53]. Graphs were created using graphics [49], ggplot2 [54], gplots [55], and the UCSC genome browser (**see URLs**).

URLs

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

bioRxiv preprint doi: https://doi.org/10.1101/707158; this version posted July 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

List of abbreviations

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

Funding

The study is supported in part by the DIP German-Israeli Project cooperation (to R.S. and R.E.), Israel Science Foundation (grant No. 1339/18 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.

Authors' contributions

T.A.H., R.E., and R.S. designed the research. T.A.H. developed the computational methods. T.A.H. performed the analyses, and parsed the ENCODE, Roadmap, and FANTOM5 data. R.E. and R.S. supervised the study. All authors analyzed the data and wrote the manuscript.

Acknowledgements

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.

Competing interests

The authors declare no competing financial interests.

Ethical approval

Not applicable.

Availability of data and material

- 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-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 [20,56,57].

bioRxiv preprint doi: https://doi.org/10.1101/707158; this version posted July 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

- Roadmap Epigenomics DNase-seq samples (73 cell types) were downloaded from GEO dataset GSE29692 [29,56–60].
- FANTOM5 CAGE data were downloaded from http://fantom.gsc.riken.jp/ [22].

References

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

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

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

4. 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. 2009;33292:289–93.

5. 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 [Internet]. Elsevier Inc.; 2014 [cited 2014 Dec 11];159:1665–80. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0092867414014974

6. 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. Nature Publishing Group; 2012;485:376.

7. 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.

8. 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. Elsevier; 2012;148:84–98.

9. 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 [Internet]. Elsevier Inc.; 2015;163:1611–27. Available from: http://dx.doi.org/10.1016/j.cell.2015.11.024

10. 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.

11. 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. Oxford University Press; 2015;43:8694–712.

12. He B, Chen C, Teng L, Tan K. Global view of enhancer--promoter interactome in human cells. Proc. Natl. Acad. Sci. National Acad Sciences; 2014;111:E2191--E2199.

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

bioRxiv preprint doi: https://doi.org/10.1101/707158; this version posted July 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

14. 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. Nature. 2017;201:7.

15. 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 (80-.). American Association for the Advancement of Science; 2018;362:eaat4311.

16. 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. Elsevier; 2019;176:377–90.

17. Heinz S, Romanoski CE, Benner C, Glass CK. The selection and function of cell type-specific enhancers. Nat. Rev. Mol. Cell Biol. [Internet]. Nature Publishing Group; 2015;16:144–54. Available from: http://dx.doi.org/10.1038/nrm3949

18. 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 Biology; 2018;1–14.

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

20. 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.

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

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. 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. Nature Publishing Group; 2016;540:296.

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

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

26. 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. Elsevier; 2016;167:1369–84.

27. 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.

28. 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. Nature Publishing Group; 2009;6:283.

bioRxiv preprint doi: https://doi.org/10.1101/707158; this version posted July 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

29. 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.

30. Nechanitzky R, Akbas D, Scherer S, Györy 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. Nature Publishing Group; 2013;14:867.

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

32. Duren Z, Chen X, Jiang R, Wang Y, Wong WH. Modeling gene regulation from paired expression and chromatin accessibility data. Proc. Natl. Acad. Sci. [Internet]. 2017;114:E4914–23. Available from: http://www.pnas.org/lookup/doi/10.1073/pnas.1704553114

33. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. BioMed Central; 2010;11:R106.

34. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. Oxford University Press; 2010;26:139–40.

35. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. Oxford University Press; 2012;40:4288–97.

36. Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team. {nlme}: Linear and Nonlinear Mixed Effects Models [Internet]. 2018. Available from: https://cran.r-project.org/package=nlme

37. 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. [Internet]. 2017;9:378–400. Available from: https://journal.r-project.org/archive/2017/RJ-2017-066/index.html

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

39. 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. Elsevier; 2013;49:764–6.

40. 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. Elsevier; 2012;48:471–84.

41. 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. Nature Publishing Group; 2012;485:381.

42. 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. Elsevier; 2012;148:458–72.

43. 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. Nature Publishing Group; 2015;47:598.

bioRxiv preprint doi: https://doi.org/10.1101/707158; this version posted July 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

44. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal [Internet]. 2006;Complex Sy:1695. Available from: http://igraph.org

45. Ho DE, Imai K, King G, Stuart EA. {MatchIt}: Nonparametric Preprocessing for Parametric Causal Inference. J. Stat. Softw. [Internet]. 2011;42:1–28. Available from: http://www.jstatsoft.org/v42/i08/

46. 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. Oxford University Press; 2017;46:D252--D259.

47. 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. Oxford University Press; 2009;37:W202--W208.

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

49. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria; 2017. Available from: https://www.r-project.org/

50. Aboyoun P, Carlson M, Lawrence M, Huber W, Gentleman R, Morgan MT, et al. Software for Computing and Annotating Genomic Ranges. 2013;9:1–10.

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

52. Harmston, N., Ing-Simmons, E., Perry, M., et al. GenomicInteractions: R package for handling genomic interaction data [Internet]. 2015. Available from: https://github.com/ComputationalRegulatoryGenomicsICL/GenomicInteractions/

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

54. Wickham H. ggplot2: Elegant Graphics for Data Analysis [Internet]. Springer-Verlag New York; 2009. Available from: http://ggplot2.org

55. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw WHA, Lumley T, et al. gplots: Various R Programming Tools for Plotting Data [Internet]. 2016. Available from: https://cran.r-project.org/package=gplots

56. 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 (80-.). American Association for the Advancement of Science; 2012;337:1190–5.

57. 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. Nature Publishing Group; 2014;32:71.

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

59. Polak P, Karlić R, Koren A, Thurman R, Sandstrom R, Lawrence MS, et al. Cell-of-origin

bioRxiv preprint doi: https://doi.org/10.1101/707158; this version posted July 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

chromatin organization shapes the mutational landscape of cancer. Nature. Nature Publishing Group; 2015;518:360.

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