Nucleic Acids Research, 2022 **1**
https://doi.org/10.1093/nar/gkac048

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 69978, Israel, [2]Department of Human Molecular Genetics and Biochemistry, Sackler School of Medicine, Tel Aviv University, Tel Aviv 69978, Israel and [3]Sagol School of Neuroscience, Tel Aviv University, Tel Aviv 69978, Israel

Received September 17, 2021; Revised January 09, 2022; Editorial Decision January 10, 2022; Accepted January 15, 2022

ABSTRACT

Spatiotemporal gene expression patterns are governed to a large extent by the activity of enhancer elements, which engage in physical contacts with their target genes. Identification of enhancer–promoter (EP) links that are functional only in a specific subset of cell types is a key challenge in understanding gene regulation. We introduce CT-FOCS (cell type FOCS), a statistical inference method that uses linear mixed effect models to infer EP links that show marked activity only in a single or a small subset of cell types out of a large panel of probed cell types. Analyzing 808 samples from FANTOM5, covering 472 cell lines, primary cells and tissues, CT-FOCS inferred such EP links more accurately than recent state-of-the-art methods. Furthermore, we show that strictly cell type-specific EP links are very uncommon in the human genome.

INTRODUCTION

Understanding the effect of the noncoding part of the genome on gene expression (GE) in specific cell types is a central genomic challenge (1). Cell identity is, to a large extent, determined by transcriptional programs driven by lineage-determining transcription factors [TFs; reviewed in (2)]. TFs mostly bind to enhancer elements located distally from their target promoters (3). Furthermore, the expression of a gene can be regulated by different enhancers in different cell types. For example, TAL1 transcription is regulated by three enhancers, two of which are active in different cell types (HUVEC and K562) (2). To find enhancer–promoter (EP) links that are active in only very few cell types, one has to compare links across multiple and diverse cell types. We term these links as cell type-specific links (ct-links). Data based on chromatin conformation capture (3C) genomic assays, which can identify ct-links, e.g. ChIA-PET (4), HiChIP (5) and Hi-C (6,7), are still not available for many cell types and tissues (6–11). Consequently, there is a high need for computational methods that predict ct-links based on other data. A key resource for such prediction is large-scale epigenomic data, which are available for a variety of human cell types and tissues, and enable quantification of both enhancer and promoter activities.

A key challenge is to identify which of the numerous candidate EP links (i) are actually functional (or active) and (ii) show their activity only in a specific small subset of cell types of interest. Ernst *et al*. (12) 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 (13) predicted ct-links in five cell types. The cell type specificity of the inferred EP links was quantified by comparison of their occurrence in other cell types. Additional methods that predicted EP links that are specifically active in a low number of cell types are IM-PET (14), EpiTensor (15), TargetFinder (16) and DeepTACT (17). All these methods used data of sequences, chromatin accessibility, multiple chromatin marks and GE data for the studied cell types. The JEME algorithm finds global and cell type-active EP links (but not necessarily cell type specific), using one to five different types of omics data (18). Each EP link reported by JEME is given a score for its tendency to be active in a given cell type. JEME reported an average of 4183 active EP links per cell type, and many of these may show a broad activity profile. Fulco, Nasser and coworkers (19,20) recently introduced the activity-by-contact (ABC) score for inferring cell type-specific functional EP links in 131 human biosamples with an average of 48 441 EP links per biosample. The ABC score was calculated using read counts of DNase hypersensitive site (DHS) and H3K27ac chromatin immunoprecipitation sequencing (ChIP-seq) at enhancer elements, and Hi-C contact frequency between enhancers and promoters.

Evidence of several sources suggests that while each cell type manifests tens of thousands of EP links, most of them are not unique and are shared across cell types. In a recently

*To whom correspondence should be addressed. Tel: +972 3 640 5384; Email: rshamir@tau.ac.il
Correspondence may also be addressed to Ran Elkon. Tel: +972 3 640 9865; Email: ranel@tauex.tau.ac.il
†These authors contributed equally to the paper.

published compendium of EP chromatin interactions across 27 human cell types (21), the number of EP loops that were unique to a specific cell type was rather low (a median of 630 unique EP links, compared to a median of 31 250 total EP links per cell type) (see the 'Materials and Methods' section). In line with these numbers, comparing 3D genome architecture between neuronal progenitor cells (NPCs) and mature neurons, Rajarajan *et al*. (22) identified 1702 and 442 NPC- and neuron-specific chromatin loops linked to 386 and 385 genes, respectively.

Here, we develop a novel statistical method for inferring ct-links from large-scale compendia of cell types measured by a single omics technique. We take advantage of linear mixed effect models (LMMs) to estimate cell type activity coefficients based on replicates available for each cell type. We compared the results to those of extant methods in terms of concordance with experimentally derived chromatin interactions and cell specificity of GE.

MATERIALS AND METHODS

FANTOM5 and ENCODE data preprocessing

Details on data preprocessing are provided in the 'FANTOM5 CAGE data preprocessing' and 'ENCODE DHS data preprocessing' sections in the Supplementary Methods.

CT-FOCS model implementation

Our model for promoter p (Figure 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 next 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.

To find ct-links based on the global links identified by FOCS, CT-FOCS (cell type FOCS) starts with the full (i.e. nonregularized) promoter model. We use the nonregularized promoter model as regularization reduces the overall model variance needed for making inferences. In principle, one could apply ordinary least squares regression with the cell types as additional coefficients to estimate cell type specificity. However, such models will perform poorly when the sample size is not much larger than the number of coefficients (e.g. in FANTOM5 we have 808 samples and a total of 483 coefficients: 472 cell types + $k = 10$ enhancers + intercept). By using an LMM, we can treat the cell type group level as a random effect coefficient, splitting the samples (replicates) based on their cell type of origin, at the cost of assuming a random effect distribution.

The application of an appropriate mixed effect model to the data depends on the distribution of the promoter and enhancer activities. We observed that FANTOM5 data have normal-like distribution and ENCODE data have zero-inflated negative binomial distribution (Supplementary Figure S1). For FANTOM5, we applied regular linear mixed effect regression. For ENCODE, we applied generalized linear mixed effect regression.

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 $y_p = X_e\beta + Z^l\gamma^l + \epsilon$, where $X_e\beta$ is the fixed effect, $Z^l\gamma^l$ is the random effect and ϵ is a random error. $l \in \{1, \ldots, k+1\}$ is one of the variables (enhancer or the intercept). γ^l is a C-long vector of random effects to be predicted. Z^l is an $n \times C$ design matrix that groups the samples by their cell types, namely

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

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 false discovery rate (23), with the number of tests performed $|P| \cdot (k+1)$.

Each predicted random effect vector $\gamma^l = (\gamma^l_1, \ldots, \gamma^l_C)$ of the alternative models was normalized using the MAD, i.e. $\gamma'^l_i = \left|\gamma^l_i - \text{median}(\gamma^l)\right| / \text{mad}(\gamma^l)$, where $\text{mad}(\gamma^l) = \text{median}(|\gamma^l - \text{median}(\gamma^l)|)$ is calculated over all cell types together. If $\gamma'^l_i > 2.5$, then enhancer l (or the promoter, if $l = 1$) was regarded as having an outlier activity in cell type i. We chose a moderately conservative MAD threshold, 2.5, as suggested in (24). We chose to use the MAD statistic since the mean and the standard deviation are known to be sensitive to outliers (24).

Finally, we defined ct-links as those that had (i) significant random effect intercept of the promoter and (ii) significant random effect slope of the enhancer, both with q-value <0.1, and (iii) enhancer and promoter random effect values were identified as outliers in the same cell type according to the MAD criterion.

MAD-FOCS model

MAD-FOCS takes the global EP links predicted by FOCS (25). Then, for every global EP link, MAD-FOCS calculates the enhancer and promoter median activity values across the multiple replicates per cell type. Last, it normalizes the median activities across cell types using the MAD method. EP links are identified as ct-links in a certain cell type if both enhancer and promoter activities are positive outliers in that cell type using MAD cutoff >2.5.

Filtered EP link sets

To validate the cell type specificity of predicted EP links, we use experimental 3D loops as a benchmark (see the next section). The very small number of cell types assayed does not allow us to identify true cell type-specific loops and exclude those common to many cell types. Therefore, the benchmark does not provide a gold standard of positive and negative ct-links (validations against all experimentally detected loops without considering the cell type specificity of predicted EP links are available in the Supplementary Results and Supplementary Figures S2 and S3). To allow a fair

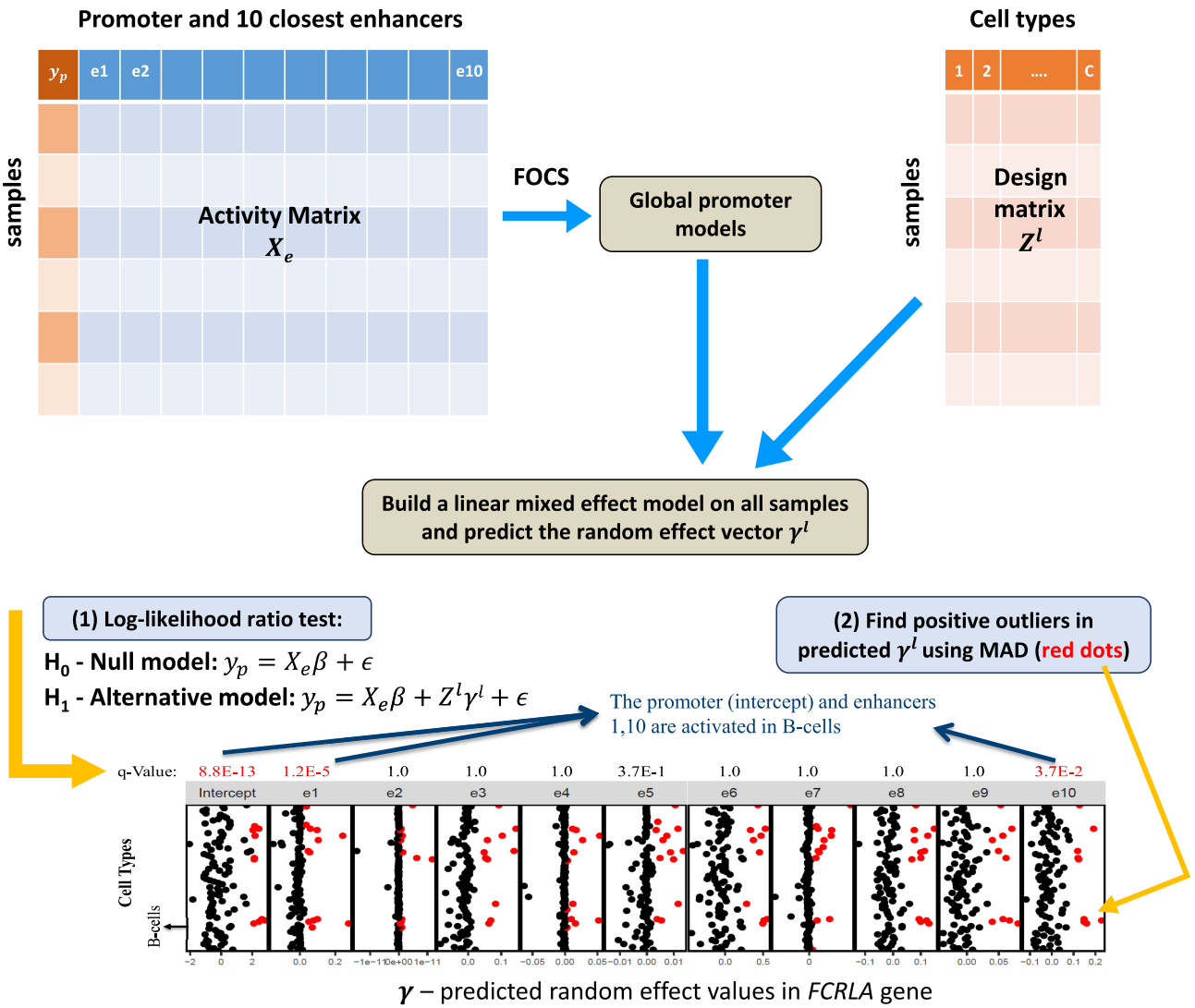

Figure 1. Outline of the CT-FOCS algorithm. Let y_p denote the observed activity of promoter p, and X_e be the activity matrix of the $k = 10$ closest enhancers to p. If $l \in \{1, \ldots, k+1\}$ is one of the variables (enhancer or promoter, i.e. the intercept), then $Z^l[i, j]$ is equal to $X_e[i, l]$ if sample i belongs to cell type j and 0 otherwise (see the 'Materials and Methods' section). First, a robust global promoter model is inferred by applying the leave-cell-type-out cross-validation step in FOCS [see (25) for details]. Second, an LMM is built on all samples using y_p, X_e and Z^l. The LMM 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. Then, the algorithm performs two tests for every l: (1) log-likelihood ratio test (LRT) to compare the simple linear regression and the LMM model. The test is carried out 11 times (testing the 10 enhancers and the intercept). The P-values for these LRTs 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 ct-link is called if (i) both enhancer and promoter (i.e. the intercept) have q-value <0.1 (marked in red), and (ii) 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 and e_{10} are significant and are commonly found as positive outliers in B cells. Therefore, $e_1 p$ and $e_{10} p$ are called by CT-FOCS as B-cell-specific EP links.

comparison between the performance of prediction methods that produce very different numbers of links, for each method and cell type, if CT-FOCS gave n links, then we took the subset of n top scored links predicted by that method. We call these subsets CT-X, where X is the method's name.

CT-JEME. JEME reports a classification score (between 0.3 and 1) for every EP link representing how active the EP link is in each cell type. We created a subset of the original JEME EP links called CT-JEME. For cell type j in FANTOM5 with n CT-FOCS ct-links, we chose the top n scoring EP links of JEME as the CT-JEME subset for that cell type. For cell types in which JEME had a lower number of EP links than CT-FOCS, we included all JEME EP links for that cell type in CT-JEME. Supplementary Figure S4A shows that the number of EP links per cell type is similar between CT-FOCS and CT-JEME. In addition, the average number of cell types sharing an EP link is 2.9 in CT-JEME compared to 11 in JEME (Supplementary Figure S4B).

CT-MAD-FOCS. To allow a fair comparison between the predictions of CT-FOCS and MAD-FOCS, we created a

subset of MAD-FOCS EP links called CT-MAD-FOCS, as described for CT-JEME earlier. We sorted the EP links by their log EP signal.

CT-TargetFinder and CT-ABC. Data for the ABC model were taken from ftp://ftp.broadinstitute.org/outgoing/lincRNA/ABC/AllPredictions.AvgHiC.ABC0.015.minus150.ForABCPaperV3.txt.gz. Among the 131 biosamples analyzed in ABC, 75 were taken from ENCODE and Roadmap epigenomic consortia (26,27) and 8 of them were also present in the CT-FOCS database and used for comparison (GM12878, HeLa-S3, K562, HCT-116, HepG2, A549 and H1-hESC). As for TargetFinder, we applied the program (https://github.com/shwhalen/targetfinder) on five cell types from ENCODE (GM12878, HeLa-S3, HUVEC, NHEK and K562) for which preprocessed multi-omics data were available on the TargetFinder website, using as input candidate DHS sites representing enhancers and promoters from ENCODE DHS data. For each cell type in ENCODE with n CT-FOCS ct-links, we chose the top n scoring EP links of TargetFinder (by classification score) and of the ABC model (by ABC score) as the predicted ct-links for that cell type for the two models, and called these subsets CT-TargetFinder and CT-ABC, respectively. Statistics on the analyzed data are summarized in Supplementary Table S1A.

External validation of predicted EP links using ChIA-PET, HiChIP and PCHi-C loops

We used 3C loops 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; ∼100 bp resolution) assayed with *POLR2A* (11), HiChIP data of Jurkat, HCT-116 and K562 cell lines (GEO accession: GSE99519; 5 kb resolution) assayed with *YY1* (28), and promoter-capture Hi-C (PCHi-C) data across 27 tissues (GEO accession: GSE86189; 5 kb resolution) (21). 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 >5 kb or overlapping anchors. Loop anchors were resized to 1 kb (5 kb in HiChIP and PCHi-C) 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 (8,29–31). For this task, we downloaded 3019 GM12878 TADs (32), which are largely conserved across cell types (7), and used them for filtering ChIA-PET and PCHi-C loops from all cell types.

To overcome the sparseness of the ChIA-PET loops, and the 8 kb minimum distance between loop anchors (10,11), we combined loops into two-step loop sets (TLSs) as follows: for every reference loop, x, its TLS is defined as the set of anchors of all loops that overlap with at least one of x's anchors by at least 250 bp (Figure 2A). We used the igraph R package (33) for this analysis.

To evaluate whether a ct-link is confirmed by the ChIA-PET data, we checked whether both the enhancer and the promoter fall in the same TLS. Specifically, we defined 1 kb genomic intervals (±500 bp upstream/downstream; 5 kb genomic intervals: ±2.5 kb upstream/downstream in HiChIP and PCHi-C) for the promoters (relative to the center position; relative to the TSS in the FANTOM5 dataset) and the enhancers (relative to the enhancer's center position) as their genomic positions. Both inter- and intra-TAD predicted EP links were included in the validation. An EP link was considered supported by a TLS if the genomic intervals of both its promoter and enhancer overlapped different anchors from the same TLS (Figure 2B and Supplementary Figure S5).

We used randomization in order to test the significance of the total number of EP links supported by ChIA-PET single loops. We denoted that number by N_t. We performed the test as follows: (i) For each predicted EP link, we randomly matched a control EP link, taken from the set of all possible EP pairs that lie within 9274 GM12878 TADs from (7), with similar linear distance between enhancer and promoter center positions. We restricted the matching to the same chromosome in order to account for chromosome-specific epigenetic state (34). The matching was done using MatchIt R package (method = 'nearest', distance = 'logit', replace = 'FALSE') (35). This way, the final set of matched control EP links had the same set of linear interaction distances as the original EP links. (ii) We counted N_r, the number of control EP links that were supported by ChIA-PET single loops. We repeated this procedure for 1000 times. The empirical P-value was $P = \#(N_r \geq N_t)/1000$, or $P < 0.001$ if the numerator was zero. A similar empirical P-value was computed for the validation rate obtained by using single loops and TLSs.

We used the following formula to calculate the GM12878 ChIA-PET TLS support ratio:

$$\text{ratio}\left(\frac{\text{GM12878}}{\text{CellType}}\right) = \frac{\%\text{GM12878-specific EPs in GM12878 TLS}}{\%\text{CellType-specific EPs in GM12878 TLS}}.$$

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

We wished to identify ct-links reported in capture Hi-C data (21). We downloaded 906 721 promoter–other (PO) capture Hi-C loops generated across 27 tissues (GEO accession: GSE86189) (21). These loops involve a known gene's promoter and a nonpromoter region, which may be an enhancer. To define a set of strictly ct-specific loops, we retained PO loops that were detected in exactly one cell type. We set the PO anchors to 1 kb intervals around their center positions. This analysis detected a median of 630 EP loops that were unique to a specific cell type.

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 (https://personal.broadinstitute.org/meuleman/reg2map/HoneyBadger2-intersect_release/DNase/p10/enh/15/state_calls.RData; https://personal.broadinstitute.org/meuleman/reg2map/HoneyBadger2-intersect_release/DNase/p10/prom/15/state_calls.RData) (27). 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 a 1 kb interval around its center position. We called the

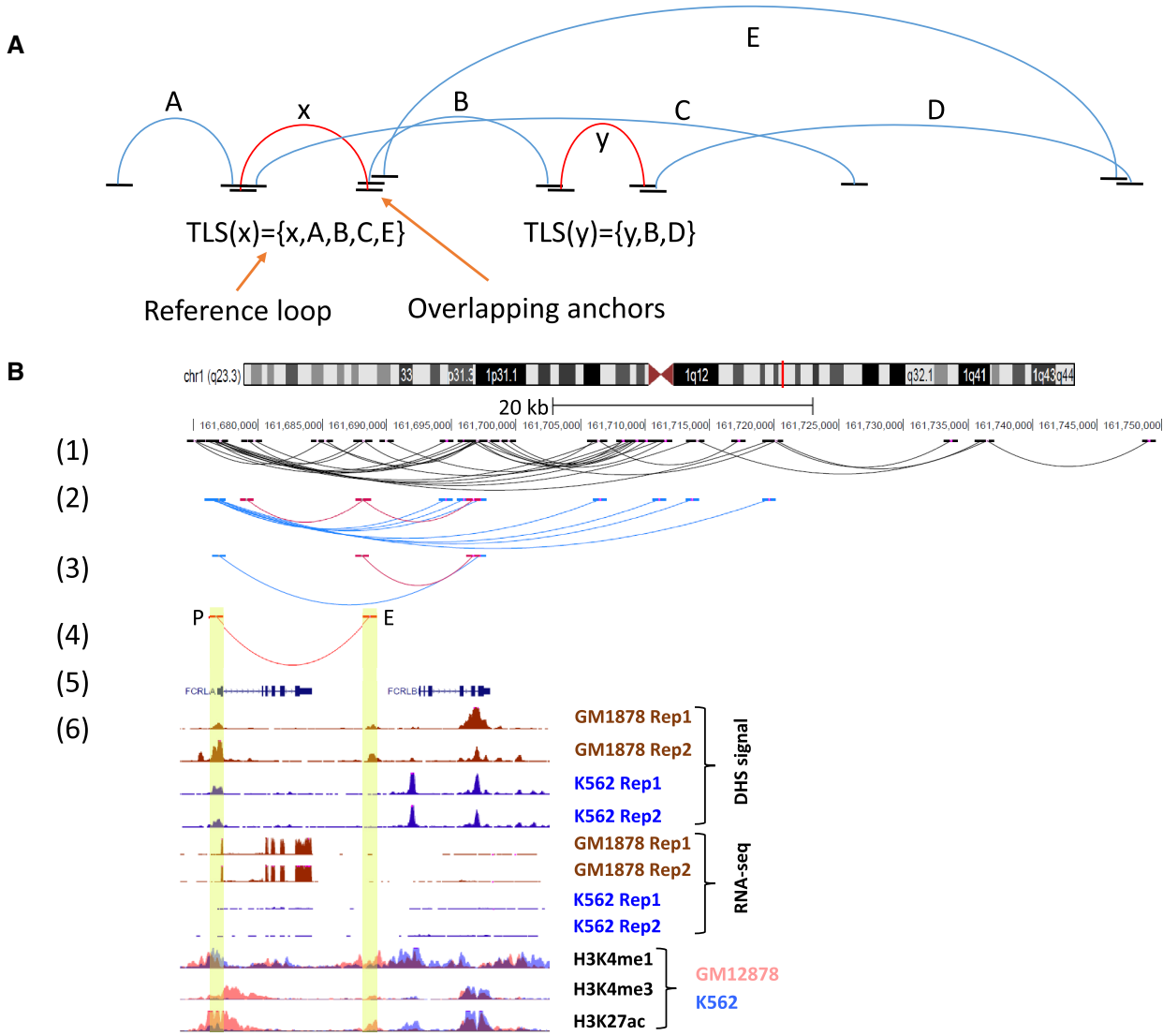

Figure 2. ChIA-PET TLSs support predicted ct-links. The TLS of a reference loop x is defined as the set of all loops that have an anchor overlapping one of x's anchors including loop x. **(A)** Examples of TLSs. Loop x's anchors overlap with at least one of the anchors of loops A, B, C and E; therefore, the TLS of x is composed of loops x, A, B, C and E. Similarly, the TLS of y is composed of loops B, y and D. Loop E overlaps anchors of both B and D, but is not part of TLS(y) as it does not overlap y's anchors. **(B)** (1) A 70 kb region of chromosome 1 showing ChIA-PET loops detected in GM12878. (4) A ct-link predicted by CT-FOCS. (2) The same region showing only loops that have anchors overlapping the anchors of the ct-link. Pink: loops overlapping the enhancer; blue: loops overlapping the promoter. (3) A TLS that supports the predicted ct-link. The ct-link in (4) is validated by the TLS, but not by any single ChIA-PET loop. (5) Gene annotations. (6) GE (RNA-seq) and epigenetic signals (DHS-seq and selected histone modifications) for the region. Tracks are shown using UCSC Genome Browser for data from GM12878 and K562 cell lines. The data indicate that this link is active in GM12878 but not in K562.

resulting sets active promoters and enhancers. A retained PO loop whose promoter and other anchors had at least 250 bp overlap with active ChromHMM 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 (36), the *specificity score* of EP links predicted in cell type c is

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

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 examined the occurrence of TF binding site motifs in sequences of ct-links' promoters and enhancers. Finding all TF motif occurrences (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 (DGF) regions, very short segments that are more likely to contain genuine TF binding sites. We downloaded ∼8.4 million DGF sequences inferred from DNase-seq in ENCODE (37). The mean DGF length was $L \approx 20$ bp, with a maximum length of 68 bp.

We intersected the DGFs with enhancer and promoter regions of predicted ct-links. We call the resulting set of sequences the *target set*. We looked for hits of 402 HOCOMOCO V11 (38) TF core motifs [taken from MEME suite database (39); http://meme-suite.org/meme-software/Databases/motifs/motif_databases.12.18.tgz] in the target sets. Hits were found using FIMO (40) with a zero-order Markov model as background created using fasta-get-markov command from MEME suite (39). For each TF, matches with FIMO q-value <0.1 were considered hits. To evaluate the statistical 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 (35) (method = 'nearest', distance = 'mahalanobis'). For each TF, we used a one-sided hypergeometric test to compare between the prevalence of its hits in the target and background (target + control) sets. Motifs having q-value <0.1 were selected.

If a k-long TF motif had l_t hits on 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 was 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 (41). To correct for multiple testing, we used the p.adjust() function (method = 'BY'). We used 'GenomicRanges' package (42) for finding overlaps between genomic intervals. We used 'rtracklayer' (43) and 'GenomicInteractions' (44) packages to import/export genomic positions. Linear mixed effect regression models were created using lme R function from nlme package (45). Generalized linear mixed effect with zero-inflated negative binomial models were created using glmmTMB R function from glmmTMB package (46). Counting of reads in genomic intervals was done using BEDTools (47). Graphs were created using graphics (41), ggplot2 (48), gplots (49), ComplexHeatmap (50) and the UCSC Genome Browser (https://genome.ucsc.edu/).

RESULTS

The CT-FOCS algorithm

We developed a novel method called CT-FOCS for inferring ct-links. The method utilizes a single type of omics data [e.g. cap analysis of gene expression (CAGE) or DHS] from large-scale datasets.

The input to CT-FOCS is enhancer and promoter activity profiles for a set of cell types. The output is the set of ct-links called for each cell type. Note that the enhancers or promoters involved in ct-links can be broadly active separately. In contrast to methods that seek global correlations between the activity profiles of enhancers and promoters, the aspect emphasized and detected by CT-FOCS is the specificity of the link between the two elements; that is, links reported by CT-FOCS highlight the few cell types in which the enhancer and promoter are predicted to functionally interact.

CT-FOCS builds on FOCS (25), which discovers global EP links showing correlated enhancer and promoter activity patterns across many samples. FOCS performs linear regression on the levels of the 10 enhancers that are closest to the target promoter, followed by two nonparametric statistical tests for producing initial promoter models, and regularization to retrieve the most informative enhancers per promoter model. CT-FOCS starts with the full (nonregularized) FOCS promoter model (see the 'Materials and Methods' section), and uses an LMM, utilizing groups of replicates available for each cell type to adjust a distinct regression curve per cell type group in one promoter model (Figure 1; see the 'Materials and Methods' section). We call a ct-link in a certain cell type if it meets the following criteria: (i) both the enhancer and the promoter show markedly positive activity levels in that cell type compared to other cell types, and (ii) both promoter and enhancer have significantly high random effect coefficients, reflecting an advantage of the LMM over the global FOCS model (see the 'Materials and Methods' section). The second criterion increases our confidence that the high activity detected by the first is specific to this cell type.

To demonstrate the difference between the linear and LMM predictions, Supplementary Figure S6 shows, for the same promoter (P), two links involving distinct enhancers (E1 and E2), one predicted by CT-FOCS (E1P) and the other by FOCS (E2P). The link between E1 and P is active only in neurons, while the link between E2 and P is active over a wider range of cell types of distinct lineages (amniotic membrane cells, whole blood cells, fibroblasts, endothelial cells and preadipocytes).

Note that choosing links by setting a threshold only on the log EP value would produce many false-positive calls, as the signals in promoters tend to be higher than those in enhancers (51) (see the examples in Supplementary Figure S6A and B).

We applied CT-FOCS on FANTOM5 CAGE profiles, which include 808 samples from 225 cell lines, 157 primary cells and 90 tissues (51) (see the 'Materials and Methods' section). CAGE quantifies the activity of both enhancers and promoters, and overall this dataset covers 42 656 enhancers and 24 048 promoters (mapped to 20 597 Ensembl protein-coding genes). For some analyses, we also applied CT-FOCS to ENCODE's DHS profiles (26,52),

which cover 106 cell types, each with typically two replicates. This dataset includes measurements for 36 056 promoters (mapped to 13 464 Ensembl protein-coding genes) and 658 231 putative enhancers (see the 'Materials and Methods' section). Unlike the FANTOM5 dataset, which builds on the expression of enhancer RNAs as a robust readout for enhancer activity, open genomic regions identified by DHS do not necessarily mark functionally active enhancers and promoters. Thus, EP maps inferred using the ENCODE dataset may be less reliable, and we focus our analyses mainly on the FANTOM5 dataset.

Overall, CT-FOCS identified 195 232 ct-links in the FANTOM5 dataset (Table 1), with an average of 414 ct-links per cell type (median 94; Table 1 and Supplementary Figure S4A). These results are in line with the low number of ct-links observed experimentally by the above-mentioned studies, including for NPCs and neurons (22,53), and further indicate that the EP links specific to a cell type constitute only a small portion of the EP links that are active in it. The EP links called by CT-FOCS were on average shared across 2.5 cell types (Supplementary Figure S4B). CT-FOCS predicted both proximal and distal interactions, with an average EP distance of ∼160 kb (median ∼110 kb; Supplementary Figure S4C). The complete set of predicted ct-links for each cell type is available at http://acgt.cs.tau.ac.il/ct-focs.

Since EP links are expected to function mostly within TADs (54,55), we next tested whether ct-links detected by CT-FOCS are enriched for intra-TAD genomic intervals. As TADS are largely cell type invariant (7), we used for these tests the 9274 TADs reported by Rao *et al*. in GM12878 (7). Indeed, comparison with randomly matched EP links demonstrated that predicted ct-links tend to lie within TADs (Supplementary Figure S7).

Inferred ct-links correlate with cell type-specific GE

To evaluate the specificity of the 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 activity of an EP link in a cell type as the logarithm of the product of the enhancer and promoter activities in that cell type. We used these measures to compute the cell type specificity for the set of ct-links detected in each cell type, using a score akin to (36) (see the 'Materials and Methods' section). As an example, CT-FOCS called 340 ct-links on the GM12878 lymphoblastoid cell line. We scored the cell type specificity of these 340 ct-links for each cell type. Reassuringly, GM12878 was the top scoring cell type, and other high scoring cell types were enriched for related lymphocyte cells (other B cells and T cells; Figure 3A and C). GM12878 was also ranked first in cell type specificity scores calculated separately for the promoters and enhancers of these 340 ct-links (Supplementary Figure S8).

Next, we examined how the effect of ct-links is reflected by cell type-specific expression of the linked genes (see the 'Materials and Methods' section). The 340 ct-links called by CT-FOCS in GM12878 involve 197 genes. We examined their expression profiles over 112 cell types using an independent GE dataset (56). In this analysis, we now scored each of the 112 cell types for the specificity in the expression of these 197 genes. Notably, here too, the lymphocyte group (B and T cells) showed the highest expression levels (Figure 3B) with GM12878 ranking first by GE specificity (Figure 3D). Overall, these results show that for GM12878, the ct-links predicted by CT-FOCS based on CAGE data are correlated with lymphocyte-specific GE programs. Supplementary Figure S9 shows similar results for neuron cells.

Comparison of CT-FOCS to other methods

We compared CT-FOCS predictions on the FANTOM5 dataset with those made by four alternative methods: (i) JEME (18), which predicts EP links that are active in a particular cell type but are not necessarily cell type specific. (ii) A naïve variant of FOCS, which takes the shrunken promoter models from FOCS, and predicts ct-links by detecting cell types in which the promoter and any of the model's enhancers show exceptionally high activity, based on the MAD index. We call this variant MAD-FOCS (see the 'Materials and Methods' section). (iii, iv) To overcome large differences among methods in the numbers of predicted links, we created subsets of the solutions of JEME and MAD-FOCS by filtering of their reported links to produce sets of links of the same size as the ones detected by CT-FOCS (see the 'Materials and Methods' section). We call these subsets cell-type-JEME (CT-JEME) and cell-type-MAD-FOCS (CT-MAD-FOCS), respectively.

Supplementary Figure S4 shows basic properties of the solutions provided by the five methods. EP links predicted by JEME and MAD-FOCS were, on average, shared across 11 and 12 cell types (median = 3 and 13, respectively; Supplementary Figure S4B). In contrast, the CT-FOCS, CT-MAD-FOCS and CT-JEME EP links were, on average, shared across <4 cell types (median = 2, 2 and 1, respectively), demonstrating that they identified EP links that are more specific. The same number of predicted links allows fair comparison between CT-FOCS, CT-MAD-FOCS and CT-JEME.

Next, we calculated cell type specificity scores for the EP links called by CT-FOCS, CT-MAD-FOCS and CT-JEME on the 276 FANTOM5 cell types. For each cell type, we used the ct-links called on it to calculate its specificity score on all cell types, and ranked the cell types by their scores. We expect the given cell type to score the top. In this analysis, CT-MAD-FOCS and CT-FOCS performed similarly, and significantly better than CT-JEME (Supplementary Figure S10A). In terms of GE of the genes associated with the EP links, examining the four cell types (GM12878, K562, HepG2 and MCF-7) that were present in both FANTOM5 and the independent GE dataset of Sheffield *et al*. (56), CT-FOCS was the only method that ranked first all the four cell types (Supplementary Figure S10B). Overall, these three methods seem to capture ct-links with highly specific EP and GE signals.

Then, we ranked the cell types according to cell type specificity scores obtained when considering separately the signals of the linked enhancers and promoters. Using ct-link enhancer signals, the median rank of the 'root' cell type (the cell type in which the link was found) was first by all methods, possibly because enhancers tend to be cell type specific.

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

Dataset	Average ct-links	Average enhancers	Average promoters	Average genes[a]	Total ct-links	Cell type with maximum ct-links
FANTOM5	414	318	146	134	195 232	Temporal lobe (13 354)
ENCODE	167	158	86	82	17 672	Caco-2 (1572)

[a]Ensembl protein-coding genes.

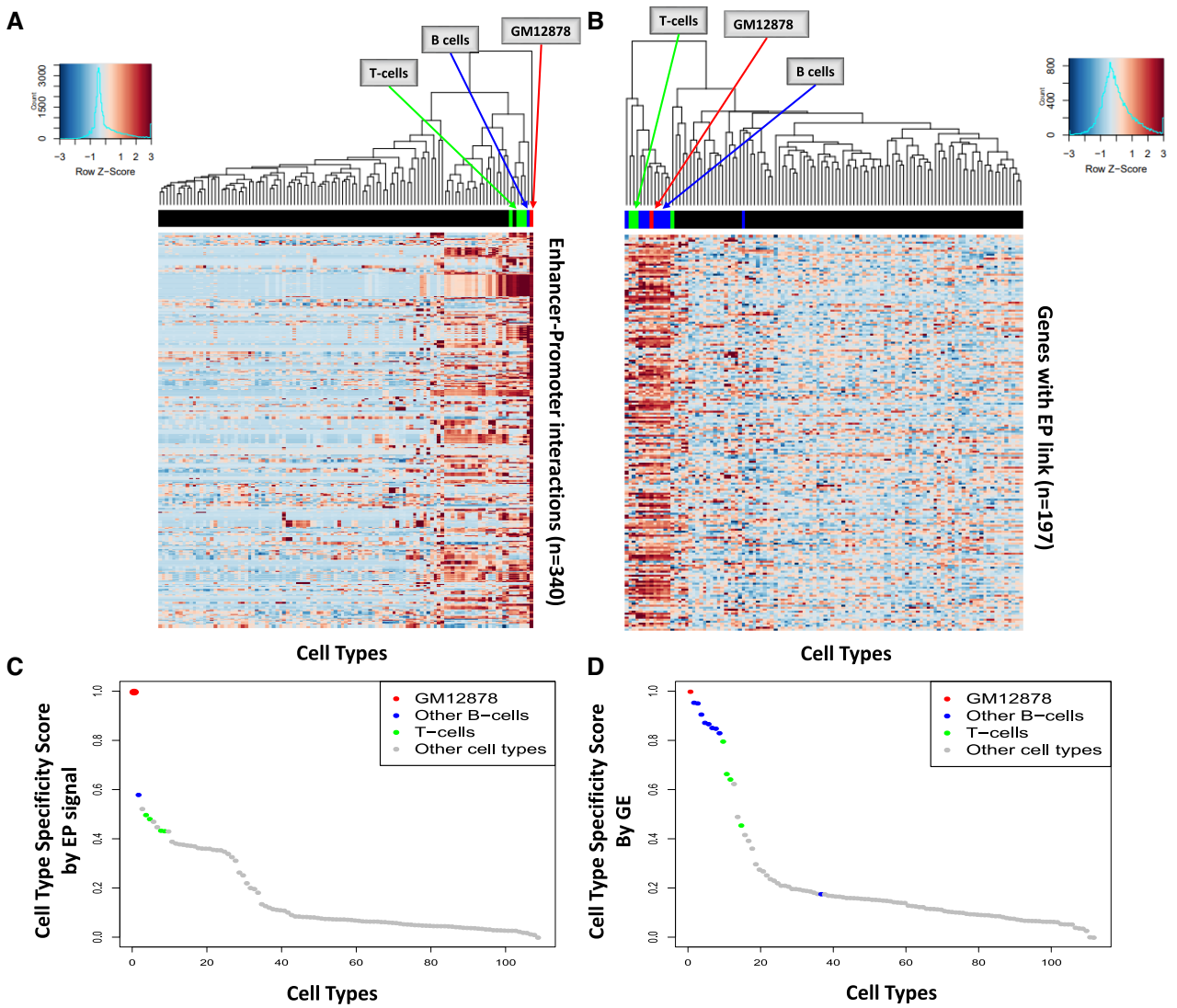

Figure 3. Specificity of ct-links predicted for GM12878 cell line. **(A)** Heatmap of EP signals for 340 ct-links predicted on GM12878 cells. Rows, EP links; columns, cell types; color, *z*-score of EP signal. Cell types related to lymphocytes (B/T cells) are highlighted in color. **(B)** Heatmap of GE for 197 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 panel (B) (see the 'Materials and Methods' section). In panels (A) and (C), 109 cell types with at least three replicates are included in the analysis; in panels (B) and (D), 112 cell types with ENCODE GE data are included (56).

However, when using ct-link promoter signals, the median rank of the root cell type obtained by CT-JEME was only 23rd, while reassuringly it was 1st for CT-FOCS and CT-MAD-FOCS. The low ranks of CT-JEME's linked promoters can explain why its predicted ct-links ranked lower compared to CT-FOCS and CT-MAD-FOCS.

Last, we compared the CT-FOCS predictions on ENCODE's DHS dataset with those obtained by six other methods: (i, ii) CT-MAD-FOCS and MAD-FOCS; (iii) TargetFinder (16), which predicts EP links based on features in enhancer, promoter and the window between them using GradientBoosting trees; (iv) ABC score model (19,20), which inferred cell type-specific functional EP links in 131 human biosamples; and (v, vi) subsets of TargetFinder and ABC model solutions having, for each cell type, a similar number of predictions to CT-FOCS (see

the 'Materials and Methods' section). We call these subsets CT-TargetFinder and CT-ABC, respectively. Note that while our evaluation of the different methods using the FANTOM5 data was done on 276 cell types (that had at least 50 predicted EP links in all methods), the evaluation using the ENCODE dataset is done only on 5–10 cell types (see the 'Materials and Methods' section). Overall, considering the specificity scores of the ct-links calculated based on DHS signals, CT-FOCS, CT-MAD-FOCS and ABC ranked the root cell type first for most cell types, better than the other three methods. On the basis of GE specificity, CT-FOCS, ABC and CT-ABC ranked the root cell type first for most cell types, performing better than the other three methods (Supplementary Table S1B).

Introducing 'two-step connected loop sets' in 3C assays to improve the evaluation of ct-links

We validated the ct-links predicted on GM12878 using empirical loops that were detected in this cell type by both *POLR2A* ChIA-PET and PCHi-C (11,21). The direct way to validate a predicted ct-link is to check whether the enhancer and promoter 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 (57,58). Furthermore, predicted ct-links that span a linear distance of <20 kb, a range where ChIA-PET loops perform poorly (59), may not be directly supported by that assay. Thus, for the validation of ct-links, we broadened the set of anchors that are considered to be proximal as follows: We define the 'two-step connected loop set' of a loop as the set of anchors of all loops that overlap with at least one of its anchors (Figure 2A). We consider a predicted ct-link as validated if its enhancer and promoter regions overlap different anchors from the same TLS (Figure 2B; see Supplementary Figure S5 for an additional example; see also the 'Materials and Methods' section). To increase our confidence that TLSs indeed represent genuine chromatin interactions, we checked for each TLS if there is a loop from the same assay that is not part of the TLS but has both anchors overlapping TLS anchors (e.g. in Figure 2A, loop E and the TLS of loop *y*). In the *POLR2A* ChIA-PET (from GM12878) and YY1 HiChIP (from K562), 54% and 64% of the TLSs were supported by such loops, respectively.

Out of the 340 ct-links inferred by CT-FOCS in GM12878, 10% were supported by ChIA-PET single loops, and 33% were supported by TLSs. Using loops from PCHi-C in GM12878, validation rates were 7.6% and 15%, respectively. (Although these rates might seem low, in the next section we show that most methods predicting EP links have a low support from 3D conformation data.) To test the significance of the observed validation rate, we generated random sets of 340 intra-TAD links having the same linear distances between enhancer and promoter regions as the ct-links predicted by CT-FOCS (see the 'Materials and Methods' section). In 1000 random sets, TLSs supported, on average, 9.4% (32 out of 340) and at most 14% (46 out of 340) (Supplementary Figure S11A), and the number of predicted ct-links supported by ChIA-PET data was significant with $P < 0.001$. Similar significance was achieved when validating the predicted ct-links directly against single loops (Supplementary Figure S11C). The same tests for PCHi-C loops gave an average overlap of matched random loops with PCHi-C TLSs of 8.5% (29 out of 340) and at most 12.4% (42 out of 340), with $P = 0.003$ for TLS (Supplementary Figure S11B) and $P = 0.048$ for single loops (Supplementary Figure S11D).

Validating predicted links by 3D conformation data

We compared the links predicted by CT-FOCS, CT-JEME and CT-MAD-FOCS to experimentally measured 3D chromatin loops, defined as the positive set. We chose the CT versions of these algorithms, which make the same number of calls, in order to allow fair comparison. In GM12878, using POLR2A ChIA-PET, CT-JEME achieved the best precision (21%) followed by CT-MAD-FOCS (19%) and CT-FOCS (10%). In K562, using YY1 HiChIP, CT-FOCS achieved the best precision (17.5%) followed by CT-MAD-FOCS (14%) and CT-JEME (3.45%). The low precision shows that single loops do not support the majority of the links predicted by any method.

Repeating the comparison using TLSs instead of single loops resulted in 2–3-fold increase in precision compared to single-loop validation in all methods. On GM12878 loops, precision was 54%, 50% and 30% in CT-JEME, CT-MAD-FOCS and CT-FOCS, respectively. On K562 loops, precision was 33%, 28% and 22% in CT-FOCS, CT-MAD-FOCS and CT-JEME, respectively. Again, the precision obtained by TLS validation for all methods was still low.

We repeated the same analysis on the ENCODE DHS dataset, comparing CT-FOCS to CT-TargetFinder and CT-ABC. Here, CT-FOCS performed markedly better in validation based on both single loops and TLSs. For example, on GM12878 with single-loop validation, CT-FOCS achieved 31% precision, while CT-TargetFinder and CT-ABC model achieved 10% and 13%, respectively. With TLS validation, CT-FOCS had 66% precision, while CT-TargetFinder and CT-ABC model achieved 30% and 47%, respectively. Similarly, on K562 with single-loop validation, CT-FOCS had 54% precision, CT-ABC 30% and CT-TargetFinder 1.4%. With TLS validation, CT-FOCS had 74% precision, CT-ABC 43% and CT-TargetFinder 3.7%.

Overall, ct-links predicted by all methods had relatively low support from 3D chromatin loops. CT-FOCS tended to achieve higher precision than the other tested methods.

Assessing cell type specificity via 3D experimental loops

As an additional test, we checked to what extent ct-links called on different cell types are supported by TLS loops that are called from GM12878's POLR2A ChIA-PET data. If ct-links called by a certain prediction method on GM12878 are indeed highly specific, we expect GM12878 to show the highest support rate in this analysis. To quantify this, we defined for each cell type the logarithm of the ratio between the validation rate observed in GM12878 and the validation rate observed for that cell type. For most cell types, we expect to obtain values >0. Indeed, CT-FOCS ct-links predicted for GM12878 showed significantly higher support rate compared to the ct-links that were predicted

in most other cell types (median $\log_2$(ratio) $\sim$ 1.7; Figure 4A). Moreover, the six cell types that showed higher validation rate than GM12878 (i.e. had $\log_2$(ratio) $<$ 0; Figure 4A, CT-FOCS boxplot) were all biologically related to GM12878 (e.g. B cell line and Burkitt's lymphoma cell line). CT-MAD-FOCS and MAD-FOCS performance was significantly lower (median $\log_2$(ratio) $\sim$ 1.1), followed by CT-JEME ($\sim$0.7) and JEME ($\sim$0.6). Note that in this analysis too, the comparisons between CT-FOCS, CT-MAD-FOCS and CT-JEME are more proper, since these methods have a similar number of predictions per cell type (and thus comparable recall). The results for MAD-FOCS and JEME are added only for reference. The results were more significant in favor of CT-FOCS when considering only TLS anchors overlapping GM12878 H3K27ac peaks downloaded from ENCODE (Supplementary Table S2A). We obtained similar results when validating against ChIA-PET single loops (Figure 4B), and when using HiChIP from K562 (Figure 4C). When using PCHi-C, HiChIP and ChIA-PET for eight individual tissues, CT-FOCS performed best overall (Figure 4D and Supplementary Table S2A).

We repeated the analysis of CT-FOCS, CT-MAD-FOCS, CT-TargetFinder and CT-ABC, now using ct-link predictions derived from the ENCODE dataset (Supplementary Table S2B). Interestingly, CT-MAD-FOCS obtained the highest precision and TLS support on GM12878. On K562, all methods had rather low performance ($\log_2$(ratio) $\simeq$ 0). Note, however, that the number of cell types compared was very low (5–10 cell types, compared to 276 for FANTOM5), so these results are anecdotal.

Overall, for the FANTOM5 dataset, the particularity of the links of CT-FOCS was higher than those of CT-MAD-FOCS and CT-JEME.

Predicted ct-links drive cell type-specific gene regulation

We next asked whether the enhancers and promoters in the ct-links inferred by CT-FOCS demonstrate signals of cell type-specific transcriptional regulation, as shown previously for lineage-determining TFs (60) and in K562 (53). To this end, we searched for occurrence of 402 known TF motifs (position weight matrices) within the enhancers and promoters of the inferred links. To lessen false discoveries, we restricted our search to DGFs (see the 'Materials and Methods' section), which are short genomic regions ($\sim$20 bp on average) identified by DHS that tend to be stably bound by TFs (61). We used $\sim$8.4 million reported DGFs in the human genome, covering 41 diverse cell and tissue types derived from ENCODE DHS data (37). 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 (see the 'Materials and Methods' section).

We first applied this test to the ct-links predicted on GM12878 using the ENCODE DHS dataset. Thirteen overrepresented TFs were identified in promoters, and a different set of 13 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 (Figure 5A and B). In terms of the specificity score of the TF overrepresentation factors, GM12878 ranked first in both enhancers and promoters (Figure 5C and D).

Many of the TFs whose motifs were detected as overrepresented on GM12878 ct-links have known roles in regulation of B-cell lineage commitment (62,63). Among them are the EBF TF 1 (*EBF1*) and the interferon regulatory factor 4 (*IRF4*) (which had, respectively, the 2nd and 8th highest overrepresentation factors in GTM12878 ct-link promoters), and the paired box 5 (*PAX5*) and the interferon regulatory factor 8 (*IRF8*) (ranked 7th and 11th in enhancers, respectively). Furthermore, *EBF1*, *SPI1*, *BATF*, *RUNX3*, *IRF4* and *PAX5*, detected by our analysis, were shown to cooperate with the *STAT5A–CEBPB–PML* complex, predicted to be involved in chromatin looping. Since these cofactors exhibit GM12878-specific expression (Supplementary Figure S12), they define highly specific chromatin binding profile for the *STAT5A–CEBPB–PML* complex in GM12878, which does not appear in the related K562 cell line (64). Note that while Zhang *et al*. (64) used ChIP-seq data from multiple TFs as well as Hi-C data to identify TF complexes involved in chromatin looping in GM12878 and K562 cell lines, our method requires data generated by only a single omics technique to pinpoint putative TF complexes that mediate EP chromatin looping for hundreds of cell types.

Next, we applied this TF motif overrepresentation analysis and specificity ranking on the ct-links inferred from ENCODE DHS data for 68 cell types that had at least 50 predicted EP links. The analysis identified an average of 12 overrepresented TF motifs in enhancers and 19 in promoters, per cell type (Supplementary Table S3). Calculating cell type specificity scores based on the set of overrepresented TFs detected on the ct-link's enhancers in each cell type ranked the studied cell type as the top one in 57 out of the 68 cell types. Similarly, using the set of overrepresented TFs detected on the ct-link's promoters ranked the studied cell type as the top one in 58 out of 68 cell types.

Last, we applied this analysis on 276 FANTOM5 cell types that had at least 50 predicted EP links in all methods. CT-FOCS analysis identified an average of 16 TFs in enhancers and 25 in promoters per cell type (Supplementary Table S4). JEME identified 33 and 69 TFs, CT-JEME identified 17 and 35, MAD-FOCS identified 9 and 20, and CT-MAD-FOCS identified 9 and 5, respectively. CT-FOCS ranked the studied cell types first in $\sim$57% and $\sim$61% of the cases for enhancers and promoters, respectively, while the other methods ranked first $\sim$1–37% in enhancers and 2–53% in promoters, with CT-MAD-FOCS showing the lowest numbers. Overall, CT-FOCS tended to find TFs that are more cell type specific.

DISCUSSION

In this study, we investigated the cell type specificity of predicted EP links by state-of-the-art methods and introduced CT-FOCS, a novel method for inferring ct-links based on activity patterns derived from a single type of omics data. The novelty of CT-FOCS is in detection of ct-links that are active in only very few cell types among hundreds of cell types, by utilizing an LMM. We applied CT-FOCS on

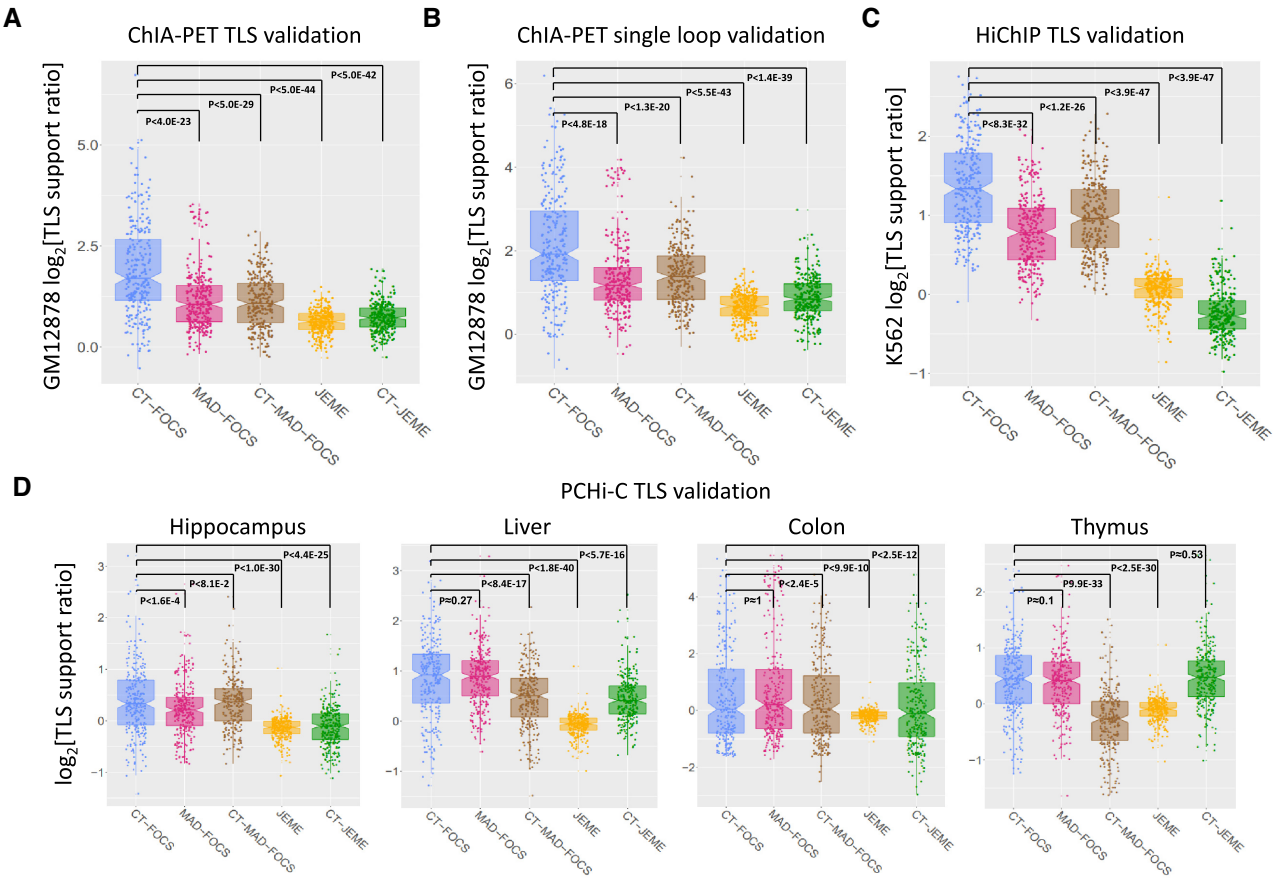

Figure 4. The particularity of each algorithm's predictions as measured by ChIA-PET, HiChIP and PCHi-C assays. **(A, B)** Each algorithm was applied to each cell type, and the predicted links were benchmarked against GM12878 ChIA-PET loops and TLSs. Comparison included 276 FANTOM5 cell types that had at least 50 predicted EP links in CT-FOCS, MAD-FOCS, CT-MAD-FOCS, JEME and CT-JEME. The plots show, for the indicated cell type, the distribution of the ratios between the percentage of predicted EP links on GM12878 that had GM12878 ChIA-PET support and the percentage of predicted links in that cell type that had GM12878 ChIA-PET support (see the 'Materials and Methods' section). **(A)** ChIA-PET TLS support. **(B)** ChIA-PET single loop support. **(C)** The same analysis as in panel (A) for K562 cell line compared to TLSs derived from K562 HiChIP assay. **(D)** The same analysis as in panel (A) but here using TLSs derived from PCHi-C in four additional cell types and tissues. All comparisons are summarized in Supplementary Table S2. *P*-values are based on one-sided Wilcoxon paired test.

CAGE profiles from FANTOM5 (51). The resulting compendium of 195 232 ct-links for 472 cell types and the program are available for use at http://acgt.cs.tau.ac.il/ct-focs and enable further inquiry on gene regulation.

We compared the cell type specificity of links predicted by each method on FANTOM5 data. We computed cell type specificity scores by using either EP signals or target GE (Supplementary Figure S10 and Supplementary Table S1A and B; see the 'Materials and Methods' section). We found that CT-FOCS and CT-MAD-FOCS achieve similar and slightly better cell type specificity ranks compared to CT-JEME on EP signal and target GE (Supplementary Figure S10). Additionally, we introduced the TLS support ratio for benchmarking predicted ct-links against chromatin interaction datasets (Figure 4 and Supplementary Table S2; see the 'Materials and Methods' section). Using this criterion, we showed that the cell type particularity of ct-links predicted by CT-FOCS was significantly higher than those of CT-JEME and CT-MAD-FOCS in five to six out of eight examined cell types with available 3D conformation data (Figure 4 and Supplementary Table S2A).

Several comments are in place regarding our inferred ct-links. First, a common naïve practice is to map enhancers to their nearest gene. Among the CT-FOCS predicted EP links, on average per cell type, only ∼10% of the enhancers mapped to the nearest gene. While this proportion is lower than that observed in previous reports [∼74% in FOCS and ∼40% in FANTOM5 (51)], it may have been affected by the relatively low number of enhancers reported by FANTOM5 (∼43 000) due to lower sensitivity of detecting enhancers using CAGE data (65). FANTOM5 enhancers tend to be located in intergenic regions, possibly reducing the correlation of the enhancers with the nearest gene, which is more apparent for intragenic enhancers located within introns of the target genes. As a result, fewer EP links are identified using correlation-based techniques (e.g. linear regression). On the other hand, low-distance links were reported to have poor validation results in ChIA-PET and Hi-C 3D loops and eQTL data (18). Second, an average of ∼60% of the predicted ct-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: 414 in FAN-

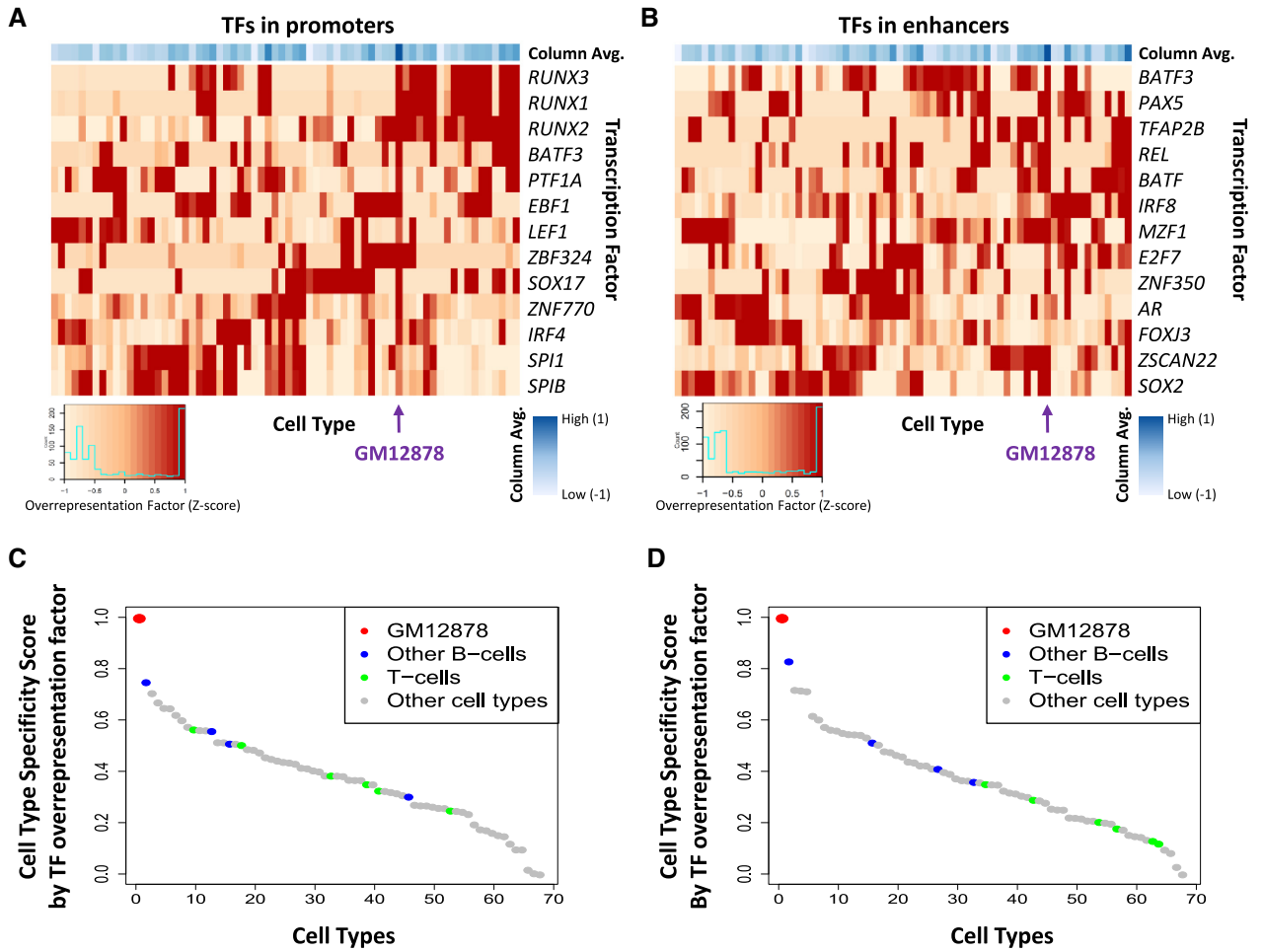

Figure 5. Overrepresented TF motifs in enhancers and promoters of GM12878 ct-links. Heatmaps of TF motif overrepresentation factor (after *z*-score transformation) in promoters **(A)** and enhancers **(B)** of GM12878-specific EP links identified by CT-FOCS on ENCODE DHS data. TFs shown had *q*-value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

TOM5 (Table 1). This relatively low number is in line with the small number of ct-links reported previously in experiments on NPCs, neurons and K562 cells (22,53), suggesting that only a small portion of the EP links that are active in a cell type are specific for it. Fourth, on average, per cell type, promoters were linked by ct-links to ∼2 (and a maximum of 9) enhancers.

In terms of methodology, CT-FOCS uses LMMs to account for two effects. The first is the joint contribution of multiple enhancers to the promoter activity, which was shown to predict GE more accurately than to pairwise enhancer–gene correlations (18). The second is the contribution of distinct cell type groups to promoter activity. By considering the cell type effect, prediction of promoter activity can be done separately for each cell type group. Thus, the estimated regression coefficient will not be the same for all samples but rather adjusted according to their cell type. In this way, ct-links are inferred based on the difference in the regression coefficients estimated for different cell type groups.

FOCS predictions are based on leave-cell-type-out cross-validation. As such, by design, it cannot infer models that are strictly cell type specific (25) (i.e. EP pairs that are active in only one specific cell type and have completely null activity in all the rest). As CT-FOCS is built upon FOCS predictions, this limitation is true for CT-FOCS predictions as well. However, we confirmed in the broad epigenomic datasets that we analyzed that cases in which an enhancer is active in only one cell type are very rare (Supplementary Results—'Loops involving enhancers active in a single cell type' and Supplementary Table S5). Nevertheless, CT-FOCS EP links show very high cell type specificity: they were shared, on average, by not more than three cell types (Supplementary Figure S4B), and >44% of them were called in a single cell type. The links identified by CT-FOCS correspond to much more prevalent (and therefore biologically more relevant) cases, in which an enhancer shows activity in several (typically, highly related) cell types, but its impact on the activity of the target promoter is markedly more prominent in one or very few of them.

A limitation of CT-FOCS is that it considers only the 10 closest enhancers to each promoter when building the models. A possible future improvement to CT-FOCS is to include all enhancers within a window of 1 Mb around each promoter, e.g. by using Bayesian hierarchical models, considering possible confounders and *a priori* information such as ChIA-PET and PCHi-C loops and eQTLs.

Another limitation of CT-FOCS is the need for cell type replicates. Cell types with at least two replicates provide variance estimate for the random effects. Cell types with a single replicate are also included in our LMM model as they can contribute to estimating the fixed effect coefficients. In FANTOM5, 179 out of 472 cell types had at least two replicates. When CT-FOCS was applied only on these 179 cell types, performance in terms of TLS support ratios improved (Supplementary Results; properties of the CT-FOCS solutions on the 179 cell types are summarized in Supplementary Figure S13). We therefore provide these predictions as well, and recommend to use them when available.

CT-FOCS can be useful for multiple genomic inquiries. 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 (Figure 5). Furthermore, integration of protein–protein interactions (PPIs) with TF identification in predicted ct-links may help identify cell type-specific PPI modules (66). These modules may contain additional new proteins (e.g. cofactors and proteins that are part of the mediator complex) that shape the 3D chromatin in a cell type-specific manner. Overall, the new method we introduced and the compendium of ct-links can advance our understanding of cell type-specific genome regulation.

DATA AVAILABILITY

CT-FOCS predicted ct-links and data are freely available at http://acgt.cs.tau.ac.il/ct-focs. Our database also contains ct-link predictions for 73 Roadmap epigenomic cell types (27). CT-FOCS source code is freely available at http://acgt.cs.tau.ac.il/ct-focs and GitHub (https://github.com/Shamir-Lab/CT-FOCS). A manual annotation of 808 FANTOM5 samples with 472 cell types can be found in Supplementary Table S6.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

R.E. is a Faculty Fellow of the Edmond J. Safra Center for Bioinformatics, Tel Aviv University. This work was carried out in partial fulfillment of the requirements for the Ph.D. degree of T.A.H. at the Blavatnik School of Computer Science, Tel Aviv University.

FUNDING

German–Israeli Project [DFG RE 4193/1-1 to R.S. and R.E.]; Israel Science Foundation [3165/19, within the Israel Precision Medicine Partnership program, and 1339/18 to R.S.]; Koret–UC Berkeley–Tel Aviv University Initiative in Computational Biology and Bioinformatics [to R.E. and R.S.]; Blavatnik Family Foundation [to R.S.]; Raymond and Beverly Sackler Chair in Bioinformatics, Tel Aviv University [to R.S.]; Edmond J. Safra Center for Bioinformatics, Tel Aviv University [to T.A.H., in part].
Conflict of interest statement. None declared.

REFERENCES

1. Gloss,B.S. and Dinger,M.E. (2018) Realizing the significance of noncoding functionality in clinical genomics. *Exp. Mol. Med.*, **50**, 97.
2. Heinz,S., Romanoski,C.E., Benner,C. and Glass,C.K. (2015) The selection and function of cell type-specific enhancers. *Nat. Rev. Mol. Cell Biol.*, **16**, 144–154.
3. Bulger,M. and Groudine,M. (2010) Enhancers: the abundance and function of regulatory sequences beyond promoters. *Dev. Biol.*, **339**, 250–257.
4. Fullwood,M.J. and Ruan,Y. (2009) ChIP-based methods for the identification of long-range chromatin interactions. *J. Cell. Biochem.*, **107**, 30–39.
5. Mumbach,M.R., Rubin,A.J., Flynn,R.A., Dai,C., Khavari,P.A., Greenleaf,W.J. and Chang,H.Y. (2016) HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods*, **13**, 919.
6. Lieberman-Aiden,E., Berkum,N.L. Van, Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
7. Rao,S.S.P., Huntley,M.H., Durand,N.C., Stamenova,E.K., Bochkov,I.D., Robinson,J.T., Sanborn,A.L., Machol,I., Omer,A.D., Lander,E.S. *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.
8. Dixon,J.R., Selvaraj,S., Yue,F., Kim,A., Li,Y., Shen,Y., Hu,M., Liu,J.S. and Ren,B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.
9. Jin,F., Li,Y., Dixon,J.R., Selvaraj,S., Ye,Z., Lee,A.Y., Yen,C.-A., Schmitt,A.D., Espinoza,C.A. and Ren,B. (2013) A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, **503**, 290–294.
10. Li,G., Ruan,X., Auerbach,R.K., Sandhu,K.S., Zheng,M., Wang,P., Poh,H.M., Goh,Y., Lim,J., Zhang,J. *et al.* (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, **148**, 84–98.
11. Tang,Z., Luo,O.J., Li,X., Zheng,M., Zhu,J.J., Szalaj,P., Trzaskoma,P., Magalska,A., Wlodarczyk,J., Ruszczycki,B. *et al.* (2015) CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell*, **163**, 1611–1627.
12. Ernst,J., Kheradpour,P., Mikkelsen,T.S., Shoresh,N., Ward,L.D., Epstein,C.B., Zhang,X., Wang,L., Issner,R., Coyne,M. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.
13. Roy,S., Siahpirani,A.F., Chasman,D., Knaack,S., Ay,F., Stewart,R., Wilson,M. and Sridharan,R. (2015) A predictive modeling approach for cell line-specific long-range regulatory interactions. *Nucleic Acids Res.*, **43**, 8694–8712.
14. He,B., Chen,C., Teng,L. and Tan,K. (2014) Global view of enhancer–promoter interactome in human cells. *Proc. Natl Acad. Sci. U.S.A.*, **111**, E2191–E2199.
15. Zhu,Y., Chen,Z., Zhang,K., Wang,M., Medovoy,D., Whitaker,J.W., Ding,B., Li,N., Zheng,L. and Wang,W. (2016) Constructing 3D interaction maps from 1D epigenomes. *Nat. Commun.*, **7**, 10812.
16. Whalen,S., Truty,R.M. and Pollard,K.S. (2016) Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat. Genet.*, **48**, 488.
17. Li,W., Wong,W.H. and Jiang,R. (2019) DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning. *Nucleic Acids Res.*, **47**, e60.

18. Cao,Q., Anyansi,C., Hu,X., Xu,L., Xiong,L., Tang,W., Mok,M.T.S., Cheng,C., Fan,X., Gerstein,M. *et al.* (2017) Reconstruction of enhancer–target networks in 935 samples of human primary cells, tissues and cell lines. *Nat. Genet.*, **201**, 7.
19. Fulco,C.P., Nasser,J., Jones,T.R., Munson,G., Bergman,D.T., Subramanian,V., Grossman,S.R., Anyoha,R., Doughty,B.R., Patwardhan,T.A. *et al.* (2019) Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.*, **51**, 1664–1669.
20. Nasser,J., Bergman,D.T., Fulco,C.P., Guckelberger,P., Doughty,B.R., Patwardhan,T.A., Jones,T.R., Nguyen,T.H., Ulirsch,J.C., Lekschas,F. *et al.* (2021) Genome-wide enhancer maps link risk variants to disease genes. *Nature*, **593**, 238–243.
21. Jung,I., Schmitt,A., Diao,Y., Lee,A.J., Liu,T., Yang,D., Tan,C., Eom,J., Chan,M., Chee,S. *et al.* (2019) A compendium of promoter-centered long-range chromatin interactions in the human genome. *Nat. Genet.*, **51**, 1442–1449.
22. Rajarajan,P., Borrman,T., Liao,W., Schrode,N., Flaherty,E., Casiño,C., Powell,S., Yashaswini,C., LaMarca,E.A., Kassim,B. *et al.* (2018) Neuron-specific signatures in the chromosomal connectome associated with schizophrenia risk. *Science*, **362**, eaat4311.
23. Benjamini,Y. and Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, **29**, 1165–1188.
24. Leys,C., Ley,C., Klein,O., Bernard,P. and Licata,L. (2013) Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. *J. Exp. Soc. Psychol.*, **49**, 764–766.
25. Hait,T.A., Amar,D., Shamir,R. and Elkon,R. (2018) FOCS: a novel method for analyzing enhancer and gene activity patterns infers an extensive enhancer–promoter map. *Genome Biol.*, **19**, 59.
26. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature, **489**, 57–74.
27. Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Kheradpour,P., Zhang,Z., Heravi-Moussavi,A., Liu,Y., Amin,V. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317.
28. Weintraub,A.S., Li,C.H., Zamudio,A. V., Sigova,A.A., Hannett,N.M., Day,D.S., Abraham,B.J., Cohen,M.A., Nabet,B., Buckley,D.L. *et al.* (2017) YY1 is a structural regulator of enhancer–promoter loops. *Cell*, **171**, 1573–1579.
29. Hou,C., Li,L., Qin,Z.S. and Corces,V.G. (2012) Gene density, transcription, and insulators contribute to the partition of the *Drosophila* genome into physical domains. *Mol. Cell*, **48**, 471–484.
30. Nora,E.P., Lajoie,B.R., Schulz,E.G., Giorgetti,L., Okamoto,I., Servant,N., Piolot,T., van Berkum,N.L., Meisig,J., Sedat,J. *et al.* (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, **485**, 381.
31. Sexton,T., Yaffe,E., Kenigsberg,E., Bantignies,F., Leblanc,B., Hoichman,M., Parrinello,H., Tanay,A. and Cavalli,G. (2012) Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell*, **148**, 458–472.
32. Mifsud,B., Tavares-Cadete,F., Young,A.N., Sugar,R., Schoenfelder,S., Ferreira,L., Wingett,S.W., Andrews,S., Grey,W., Ewels,P.A. *et al.* (2015) Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.*, **47**, 598.
33. Csárdi,G. and Nepusz,T. (2006) The igraph software package for complex network research. *InterJ. Complex Syst.*, **1695**, 1–9.
34. Xi,W. and Beer,M.A. (2018) Local epigenomic state cannot discriminate interacting and non-interacting enhancer–promoter pairs with high accuracy. *PLoS Comput. Biol.*, **14**, e1006625.
35. Ho,D.E., Imai,K., King,G. and Stuart,E.A. (2011) MatchIt: nonparametric preprocessing for parametric causal inference. *J. Stat. Softw.*, **42**, 1–28.
36. Javierre,B.M., Burren,O.S., Wilder,S.P., Kreuzhuber,R., Hill,S.M., Sewitz,S., Cairns,J., Wingett,S.W., Várnai,C., Thiecke,M.J. *et al.* (2016) Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell*, **167**, 1369–1384.
37. Neph,S., Vierstra,J., Stergachis,A.B., Reynolds,A.P., Haugen,E., Vernot,B., Thurman,R.E., John,S., Sandstrom,R., Johnson,A.K. *et al.* (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, **489**, 83–90.
38. Kulakovskiy,I. V, Vorontsov,I.E., Yevshin,I.S., Sharipov,R.N., Fedorova,A.D., Rumynskiy,E.I., Medvedeva,Y.A., Magana-Mora,A., Bajic,V.B., Papatsenko,D.A. *et al.* (2017) HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-seq analysis. *Nucleic Acids Res.*, **46**, D252–D259.
39. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.
40. Grant,C.E., Bailey,T.L. and Noble,W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.
41. R Core Team (2020) R: a language and environment for statistical computing. *R Foundation for Statistical Computing*. https://www.R-project.org, (01 March 2019, date last accessed).
42. Lawrence,M., Huber,W., Pagès,H., Aboyoun,P., Carlson,M., Gentleman,R., Morgan,M.T. and Carey,V.J. (2013) Software for computing and annotating genomic ranges. *PLoS Comput. Biol.*, **9**, e1003118.
43. Lawrence,M., Gentleman,R. and Carey,V. (2009) rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics*, **25**, 1841–1842.
44. Harmston,N., Ing-Simmons,E., Perry,M., Baresic,A. and Lenhard,B. (2015) GenomicInteractions: an R/Bioconductor package for manipulating and investigating chromatin interaction data. *BMC Genomics*, **16**, 963.
45. Pinheiro,J., Bates,D., DebRoy,S., Sarkar,D. and R Core Team (2021) nlme: linear and nonlinear mixed effects models description. https://cran.r-project.org/package=nlme(07 September 2021, date last accessed).
46. Brooks,M.E., Kristensen,K., van Benthem,K.J., Magnusson,A., Berg,C.W., Nielsen,A., Skaug,H.J., Maechler,M. and Bolker,B.M. (2017) glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *R J.*, **9**, 378–400.
47. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
48. Wickham,H. (2009) ggplot2: Elegant Graphics for Data Analysis. Springer, NY, https://ggplot2.tidyverse.org, (25 June 2021, date last accessed).
49. Warnes,G.R., Bolker,B., Bonebakker,L., Gentleman,R., Liaw,W.H.A., Lumley,T., Maechler,M., Magnusson,A., Moeller,S., Schwartz,M. *et al.* (2016) gplots: various R programming tools for plotting data. https://cran.r-project.org/package=gplots, (28 November 2020, date last accessed).
50. Gu,Z., Eils,R. and Schlesner,M. (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, **32**, 2847–2849.
51. Andersson,R., Gebhard,C., Miguel-Escalada,I., Hoof,I., Bornholdt,J., Boyd,M., Chen,Y., Zhao,X., Schmidl,C., Suzuki,T. *et al.* (2014) An atlas of active enhancers across human cell types and tissues. *Nature*, **507**, 455–461.
52. Thurman,R.E., Rynes,E., Humbert,R., Vierstra,J., Maurano,M.T., Haugen,E., Sheffield,N.C., Stergachis,A.B., Wang,H., Vernot,B. *et al.* (2012) The accessible chromatin landscape of the human genome. *Nature*, **489**, 75–82.
53. Gasperini,M., Hill,A.J., McFaline-Figueroa,J.L., Martin,B., Kim,S., Zhang,M.D., Jackson,D., Leith,A., Schreiber,J., Noble,W.S. *et al.* (2019) A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell*, **176**, 377–390.
54. Krijger,P.H.L. and de Laat,W. (2016) Regulation of disease-associated gene expression in the 3D genome. *Nat. Rev. Mol. Cell Biol.*, **17**, 771–782.
55. Pombo,A. and Dillon,N. (2015) Three-dimensional genome architecture: players and mechanisms. *Nat. Rev. Mol. Cell Biol.*, **16**, 245–257.
56. Sheffield,N.C., Thurman,R.E., Song,L., Safi,A., Stamatoyannopoulos,J.A., Lenhard,B., Crawford,G.E. and Furey,T.S. (2013) Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome Res.*, **23**, 777–788.
57. Olivares-Chauvet,P., Mukamel,Z., Lifshitz,A., Schwartzman,O., Elkayam,N.O., Lubling,Y., Deikus,G., Sebra,R.P. and Tanay,A. (2016) Capturing pairwise and multi-way chromosomal conformations using chromosomal walks. *Nature*, **540**, 296.

58. Song,W., Sharan,R. and Ovcharenko,I. (2019) The first enhancer in an enhancer chain safeguards subsequent enhancer–promoter contacts from a distance. *Genome Biol.*, **20**, 197.
59. Kumasaka,N., Knights,A.J. and Gaffney,D.J. (2019) High-resolution genetic mapping of putative causal interactions between regions of open chromatin. *Nat. Genet.*, **51**, 128.
60. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
61. Hesselberth,J.R., Chen,X., Zhang,Z., Sabo,P.J., Sandstrom,R., Reynolds,A.P., Thurman,R.E., Neph,S., Kuehn,M.S., Noble,W.S. *et al.* (2009) Global mapping of protein–DNA interactions *in vivo* by digital genomic footprinting. *Nat. Methods*, **6**, 283.
62. Nechanitzky,R., Akbas,D., Scherer,S., György,I., Hoyler,T., Ramamoorthy,S., Diefenbach,A. and Grosschedl,R. (2013) Transcription factor EBF1 is essential for the maintenance of B cell identity and prevention of alternative fates in committed cells. *Nat. Immunol.*, **14**, 867.
63. Wang,H., Lee,C.H., Qi,C., Tailor,P., Feng,J., Abbasi,S., Atsumi,T. and Morse,H.C., III (2008) IRF8 regulates B-cell lineage specification, commitment, and differentiation. *Blood*, **112**, 4028–4038.
64. Zhang,K., Li,N., Ainsworth,R.I. and Wang,W. (2016) Systematic identification of protein combinations mediating chromatin looping. *Nat. Commun.*, **7**, 12249.
65. Core,L.J., Martins,A.L., Danko,C.G., Waters,C.T., Siepel,A. and Lis,J.T. (2014) Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.*, **46**, 1311.
66. Duren,Z., Chen,X., Jiang,R., Wang,Y. and Wong,W.H. (2017) Modeling gene regulation from paired expression and chromatin accessibility data. *Proc. Natl Acad. Sci. U.S.A.*, **114**, E4914–E4923.