Nucleic Acids Research, 2021 **1**
https://doi.org/10.1093/nar/gkab547

Leveraging three-dimensional chromatin architecture for effective reconstruction of enhancer–target gene regulatory interactions

Elisa Salviato [1,*], Vera Djordjilović[2], Judith Mary Hariprakash[1], Ilario Tagliaferri[1], Koustav Pal[1] and Francesco Ferrari [1,3,*]

[1]IFOM, the FIRC Institute of Molecular Oncology, Milan 20139, Italy, [2]Department of Economics, Ca' Foscari University of Venice, Venice 30100, Italy and [3]Institute of Molecular Genetics "Luigi Luca Cavalli-Sforza", National Research Council, Pavia 27100, Italy

Received February 16, 2021; Revised June 07, 2021; Editorial Decision June 09, 2021; Accepted June 17, 2021

ABSTRACT

A growing amount of evidence in literature suggests that germline sequence variants and somatic mutations in non-coding distal regulatory elements may be crucial for defining disease risk and prognostic stratification of patients, in genetic disorders as well as in cancer. Their functional interpretation is challenging because genome-wide enhancer–target gene (ETG) pairing is an open problem in genomics. The solutions proposed so far do not account for the hierarchy of structural domains which define chromatin three-dimensional (3D) architecture. Here we introduce a change of perspective based on the definition of multi-scale structural chromatin domains, integrated in a statistical framework to define ETG pairs. In this work (i) we develop a computational and statistical framework to reconstruct a comprehensive map of ETG pairs leveraging functional genomics data; (ii) we demonstrate that the incorporation of chromatin 3D architecture information improves ETG pairing accuracy and (iii) we use multiple experimental datasets to extensively benchmark our method against previous solutions for the genome-wide reconstruction of ETG pairs. This solution will facilitate the annotation and interpretation of sequence variants in distal non-coding regulatory elements. We expect this to be especially helpful in clinically oriented applications of whole genome sequencing in cancer and undiagnosed genetic diseases research.

INTRODUCTION

Distal non-coding regulatory elements (enhancers) are crucial players in the control of gene expression. These are also the genomic features carrying the most marked epigenetic differences across cell types, thus constituting a fundamental component of the molecular and genetic mechanisms defining cell identity (1,2). Enhancer activity status is itself regulated by epigenetics, chromatin accessibility and its three-dimensional (3D) architecture (3). In fact, the formation of chromatin loops allows distal regulatory regions to come in close physical proximity to their target gene promoters to regulate transcription (4). Their importance for human physiology is attested by their enrichment in polymorphisms associated to genetic diseases and cancer risk (5,6). More mechanistic studies have shown the functional role of enhancer alteration in several pathologies, sometime collectively termed enhanceropathies (7,8). Therefore, a genome-wide definition of the regulatory network constituted by enhancers and their target genes would be a valuable resource in biomedical research. For example, it would be instrumental for the annotation and interpretation of non-coding somatic mutations or germline sequence variants, to understand their effect on the broader gene regulatory network, in basic biology as well as in more translational studies.

Despite its importance, the reconstruction of a comprehensive network of enhancer–target gene (ETG) pairs remains elusive, especially because enhancers position with respect to the target genes is highly variable. Indeed, they can regulate one or more genes that appear distant in the linear sequence of the genome but may be in close physical proximity in the 3D chromatin organisation (9).

In this context, the development of molecular biology methods to study the 3D chromatin organization has been pivotal for achieving a better understanding of distal regulatory elements. In particular, the methods based on ligation by proximity, i.e. Chromosome Conformation Capture (3C) (10) and its high-throughput derivatives (11–13) (e.g. 4C, 5C and Hi-C), allow quantifying the frequency of physical interactions between distant chromatin regions

*To whom correspondence should be addressed. Tel: +39 02 57430 3830; Fax: +39 02 574303088; Email: francesco.ferrari@ifom.eu
Correspondence may also be addressed to Elisa Salviato. Email: elisa.salviato@ifom.eu

(chromatin loops). Hi-C is the high-throughput genome-wide version of this technique, allowing researchers to map the contact frequency between virtually any pair of genomic loci (14).

In principle, Hi-C data could be used for the genome-wide identification of specific points of contact, such as ETG loops. However, Hi-C data is generally analysed by binning read counts at a resolution of few kilobases (kb), with the highest coverage datasets available to date reaching 1 kb (15–17). This resolution level is lower than what is needed to map ETG pairs when multiple enhancers are close to each other, or close to promoters. In all these cases, a distance smaller than 2 bins would not allow discriminating the interacting partners. Even the most recent literature, based on ENCODE3 data, reported that using Hi-C interaction calls to directly map ETG contacts is not a valuable strategy (18) to annotate distal regulatory elements, due to the resolution limit. Another challenge in this approach, is that different algorithms for calling point interactions in Hi-C data have generally discordant results and are influenced by the sequencing coverage (19). Additional experimental approaches aimed to define physical pairing of ETG with higher resolution include capture Hi-C (cHi-C) (20–23), HiChIP (24) and ChIA-PET (25). These techniques have been instrumental to define experimentally validated ETG physical contacts. However, these experimental methods are generally considered cell-type specific.

Thus, chromatin 3D architecture has not been optimally incorporated in the ETG network reconstruction algorithms proposed in literature so far. Some publications marginally used chromatin conformation data to define true positive contacts (26–29), despite the resolution and methodological shortcomings discussed above. Moreover, these approaches have been applied to a limited number of cell types, due to the reduced availability of cell-type specific experimental datasets. Alternatively, chromatin structural domains have been used only to restrict the initial search space of ETG pairs (30–33).

We hypothesised that the effective incorporation of chromatin 3D architecture data would enhance the accuracy of a generalizable genome-wide definition of ETG pairs. To this concern, here we introduce a change of perspective based on the current biological knowledge. Namely, there is a general consensus in the field about enhancer–target gene interactions occurring within the insulated boundaries of the so-called Topologically Associated Domains (TADs) (34,35), which are relatively insulated domains enriched in local interactions. Moreover, several studies reported that TADs are largely conserved across different cell types (36–38). On the other hand, it is generally accepted that TADs can be defined at different levels of resolution, i.e. there is a hierarchy of TADs (17,39,40). More recent literature indicates that alternative TADs structures may indeed co-exist within a cell population, and the stochastic dynamics of active loop extrusion mechanisms could explain their formation and the patterns detected in Hi-C data (41–43). Therefore, we use multi-resolution TAD definitions as prior-knowledge to inform the selection of ETG pairs.

In this work, (i) we develop a computational and statistical framework to reconstruct a comprehensive map of ETG pairs leveraging functional genomics data. Namely, we use a large panel of epigenomics datasets to define enhancer activity across multiple cell and tissue types, along with high resolution Hi-C data. (ii) Then we demonstrate that the incorporation of chromatin 3D architecture information improves the accuracy in defining ETG pairs. In this context, we compute a score encoding the multiscale hierarchical structure of chromatin and use it as side information for controlling false discoveries and achieving high statistical power. (iii) Finally, we extensively benchmark our method against previous solutions for the genome-wide reconstruction of ETG pairing. We show that our method is a valuable general-purpose solution, providing good ETG pairing performances for both long- and mid-range interactions.

MATERIALS AND METHODS

Definition of reference set of enhancer and gene promoter regions

We defined a reference set of enhancer regions using epigenomics datasets based on high-throughput sequencing across a compendium of cell and tissue types. In particular, we used ChIP-seq (Chromatin ImmunoPrecipitation followed by high-throughput sequencing) data for specific histone modifications, as detailed in the relevant results sections, as well as chromatin accessibility data based on DNase I hypersensitive sites (DHS), identified with DNase-seq. We downloaded histone H3 lysine 27 acetylation (H3K27ac) ChIP-seq and DHS narrow peaks (based on MACS v2.0.20 calls by the Roadmap Epigenomic consortium) called for 44 uniformly processed and consolidated cell and tissue types from Roadmap Epigenomic portal (https://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidated/narrowPeak/). H3K27ac is a post-translational histone modification associated with active enhancer and promoter regions, whereas DNase-seq allows assessing chromatin accessibility. We focused on the subset of cells and tissue types for which both H3K27ac ChIP-seq and DNase-seq were available (Supplementary Table S1A). We further filtered the results for subsequent analyses considering only peaks with strong significant enrichment, i.e. $-\log_{10}$(adj.*P*-value) ≥ 5. Both the number (Supplementary Figure S1A) and size (Supplementary Figure S1B) of DNase-seq and H3K27ac ChIP-seq peaks vary across cell and tissue types. Namely, DNase-seq peaks were 123 400 on average, with average size 358 bp. Conversely, H3K27ac peaks were fewer (53 721 on average) and larger (940 bp average size).

To obtain a comprehensive list of cis-regulatory elements we conducted a two-step procedure. Firstly, for each cell type, the intersection between H3K27ac and DHS peaks with overlapping regions ($\geq$1 bp) were used to define cell-specific enhancers. Additional filters were applied ex-post, such as the removal of interval portions overlapping annotated exons (for both coding and non-coding genes) and the removal of intervals shorter than 10 bp or larger than 2.5 kb.

Secondly, cell-specific enhancers with overlapping intervals across different cell types were merged (union) together to define a consensus set of enhancer regions. This set was further annotated with respect to the transcription start site (TSS) as promoter-proximal (within 3.5 kb upstream and 1.5 kb downstream of TSS) or distal, and only the

promoter-distal ones were retained as reference list of enhancer elements, hereinafter referred as *enhancer catalogue* ($n = 347\,777$). This is meant to be a comprehensive reference set of regulatory regions, that can be active enhancers in at least one of the cell and tissue types considered.

In the subsequent analyses to identify ETG pairs, we actually focused on the epigenetic status of the gene promoter regions, used as a proxy for the activity of the target genes. Thus, hereinafter we will refer to enhancer−promoter (EP) pairs when explicitly focusing on these genomic regions or epigenetic features, whereas we will refer to ETG pairs when focusing on the functional interaction to regulate the target gene. We defined reference promoters as 2 kb regions (1.5 kb upstream and 0.5 kb downstream) around the transcription start site (TSS) of annotated protein coding genes, based on RefSeq annotations in UCSC (refGene.txt.gz, May 2019, hg19 genome assembly). Non-canonical and Y chromosome were excluded. To reduce possible ambiguities, in case of multiple alternative transcripts for the same gene, only the most upstream TSS was maintained as reference for each gene. Moreover, promoters were merged in case of two close TSSes (±0.5 kb interval), e.g. divergent transcripts on opposite DNA strands, so as to obtain the final reference set of promoter regions ($m = 18\,027$).

To compare our enhancer catalogue with alternative functional genomic definitions we employed: (i) the atlas of active enhancers provided by FANTOM5 project (https://fantom.gsc.riken.jp/5/data/) (44), based on 808 human Cap Analysis of Gene Expression (CAGE) experiments (45) and (ii) the collection of *in vivo* validated enhancers coming from the VISTA Enhancer database (46), based on transgenic mice reporter assays in 23 tissues of mouse embryos (47).

We downloaded enhancer coordinates from FANTOM5 repository (https://fantom.gsc.riken.jp/5/datafiles/latest/extra/Enhancers/human_permissive_enhancers_phase_1_and_2.bed.gz), and we retrieved positive (i.e. elements that show consistent reporter gene expression among at least three embryos) 'Human only' enhancers from VISTA Enhancer Browser (date version: 12 February 2020). In line with the procedure used to define our enhancer catalogue, interval portions overlapping annotated exons and promoter proximal elements were removed to obtain the final set of enhancers from each of these alternative sources. These filtered FANTOM and VISTA enhancer sets were used in the subsequent analyses and are composed of 58 200 and 894 enhancers, respectively.

Hi-C dataset processing

We leveraged chromatin 3D architecture data from genome-wide chromosome conformation capture experiments based on high-throughput sequencing (Hi-C). Namely, we processed eleven Hi-C datasets (Supplementary Table S1B) covering different cell lines and primary tissues from a compendium of public datasets (17,48–52).

For each Hi-C dataset we retrieved the raw FASTQ files from the NIH SRA database (https://www.ncbi.nlm.nih.gov/sra). The sequencing reads were aligned with the iterative mapping procedure (single-end mode) as implemented in hiclib (https://github.com/mirnylab/hiclib-legacy)(version from gitHub commit d38f198, date: 28 September 2017) based on botwie2 (version 2.3.4.3) aligner (53) (https://github.com/BenLangmead/bowtie2). Briefly, in this iterative alignment procedure reads were truncated at 30 bp and aligned to the reference genome (hg19). Reads that were not uniquely aligned were elongated (5 bp) and the alignment procedure repeated, with additional iterations until full read length or successful alignment is achieved. For each FASTQ file the information on uniquely mapped reads were stored in a HDF5 (Hierarchical Data Format) file. Biological or technical replicates belonging to the same dataset were merged in a single HDF5 file (hdf5 library, version 2.9.0). We filtered read pairs with a sum of distances from the downstream restriction site not compatible with the expected fragment size: i.e. events originating from non-canonical enzyme activity or non-enzymatic physical breakage. The distance cut-off was estimated for each dataset based on the frequency distribution of distances and the expected fragment length. Duplicated read pairs, as well as read pairs derived from unligated or circularized fragments, were also removed.

Finally, the genome was binned at 10 kb bin size, and the raw read counts were summarized in a Hi-C contact matrix for each chromosome, accounting for intra-chromosomal interactions. To allow comparability among all tissues and cell types and correct for technical biases, chromosome-wise iterative correction (ICE) with default parameters (54,55) was applied (using cooler version 0.8.5, https://github.com/open2c/cooler). This procedure returned a balanced matrix of relative contact probabilities, in which each row (excluding the elements in the first two removed diagonals) summed up to one. The output files (cool format) were converted to txt files and compressed.

Hierarchical contact score

To account for the 3D spatial proximity of regulatory elements, we devised a score proportional to the likelihood of enhancer−promoter (EP) pairs co-localization, named Hierarchical Contact (HC) score (Figure 1B). HC accounts for the TADs hierarchical structure across multiple tissue and cell types. For HC definition we relied on the Local Score Differentiator (LSD) TAD borders calling procedure (56), as implemented in the HiCBricks (version 1.8.0) Bioconductor package (57). We defined TADs as regions between two consecutive domain boundaries. LSD is based on the directionality index (DI) score originally proposed by Dixon *et al.* (36). Among the user defined parameters in this algorithm, the DI-window (i.e. the number of upstream and down-stream bins over which the DI score is computed) influences the scale of the TAD domains that are identified: the larger the DI-window, the larger the average resulting TAD size.

The HC score is thus defined by considering a collection $\mathcal{D}$ of Hi-C contact matrices (all binned at the same bin size) and an ensemble of TADs boundaries for each $D \in \mathcal{D}$, denoted by $TAD_D(w)$, where $w \in W$ is the DI-window (Figure 1B, top panel). We considered $W = [5, 10, 20, 50]$ for the DI-window size. Thus, each $TAD_D(W)$ represents the

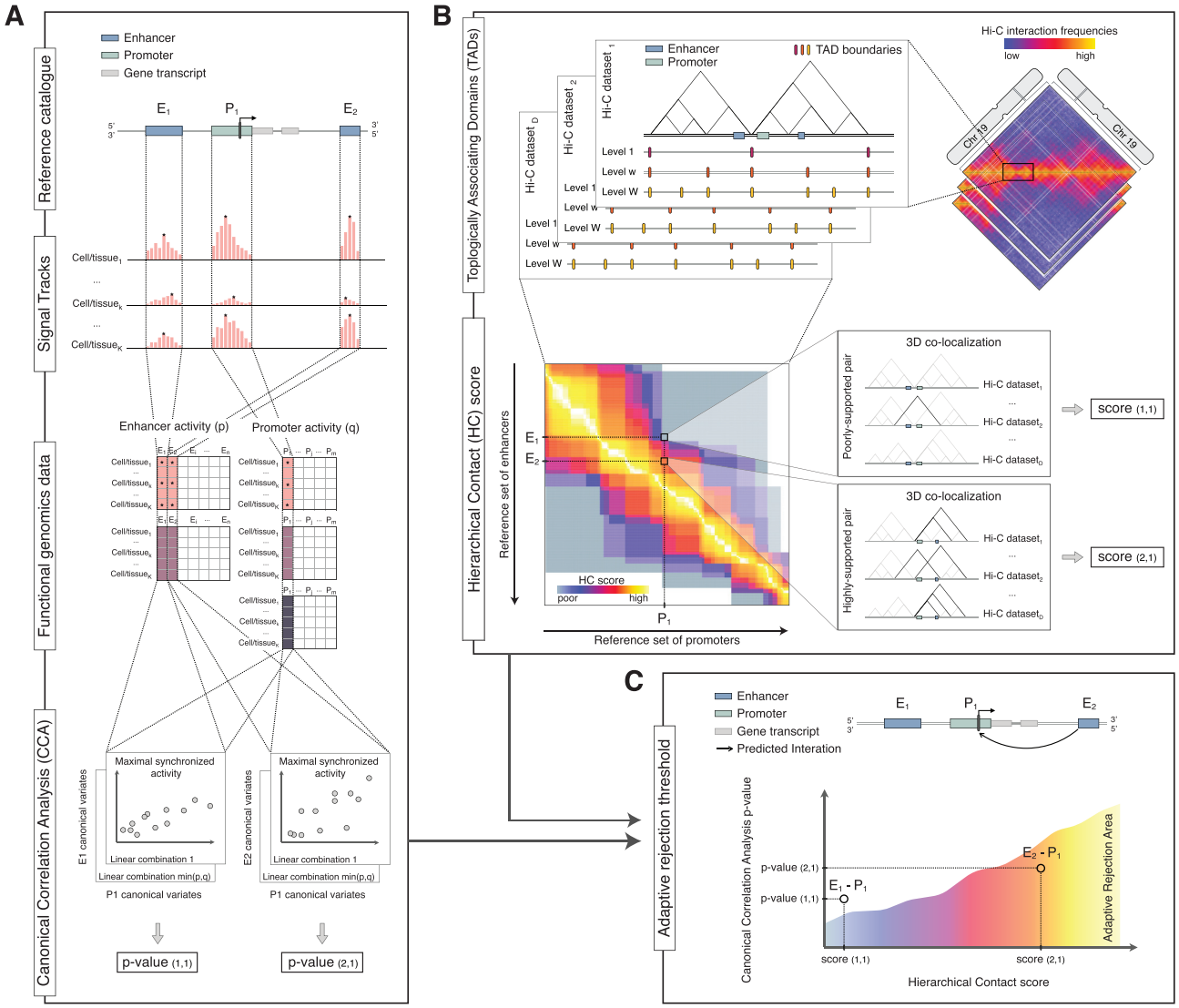

Figure 1. Methodological framework overview. Schematic illustration of the workflow of our methodological framework. Throughout this figure panels the (*i*,*j*) labels refer to enhancer (*i*) and promoter (*j*) pairs. (**A**) Starting from a reference catalogue of enhancer and promoter regions, it is possible to quantify their respective activity status using two sets of *p* and *q* functional genomic data (e.g. ChIP-seq data for chromatin marks), respectively. Then, the Canonical Correlation Analysis (CCA) is used to investigate the synchronised activity of each enhancer–promoter (EP) pair across *k* cell and tissue types. The two original sets of chromatin marks are transformed through linear combinations that allow maximizing the relationship between the two sets, and the respective canonical correlation is tested. The procedure returns a *P*-value for each specific EP pair. (**B**) For each Hi-C dataset in the selected collection, the boundaries of Topologically Associating Domains (TADs) are identified across multiple levels of resolution. The resulting ensemble of boundaries represents the hierarchical structure of TADs for a specific cell or tissue type. Considering the occurrence of each EP pair within these ensembles called from *D* Hi-C datasets, we can describe their broader spatial co-localization pattern through the Hierarchical Contact (HC) score. A high score is associated to pairs supported by several combinations of Hi-C datasets and hierarchical levels (e.g. $E_2 - P_1$ pair). Conversely, a weak score is associated to pairs supported only in few combinations (e.g. $E_1 - P_1$ pair). (**C**) The 3D co-localization information encoded in the HC score is used to estimate an adaptive rejection threshold to control for FDR in the multiple testing hypothesis of EP pairs synchronisation. On similar equal nominal *p*-value (y-axis) a less conservative significance criterion is used for the EP pair showing higher HC score (x-axis and color gradient). Namely, even if one enhancer (E_1) will exhibit a stronger synchronization with a specific promoter (P_1), being at greater 3D distance will be less likely to regulate it than the closest one (E_2).

multi-resolution TADs structure for a specific Hi-C contact matrix in a specific cell and tissue type.

Given a list of n enhancers and m promoters, we can define a binary matrix $M_D(w)$, in which an element (i, j) is set to 1 if the *i*th enhancer and the *j*th promoter are within the same TAD belonging to the ensemble $\text{TAD}_D(w)$. The matrix $M_D(w)$ is thus an $n \times m$ co-occurrence matrix. To estimate the overall spatial relationships between enhancers and promoters over the hierarchical structure of each Hi-C contact matrix D, we propose the aggregate score:

$$S^{(D)} = \sum\nolimits_{w \in W} \lambda(w)\ M_D(w) \qquad (1)$$

where $\lambda(w) = \sqrt{\max(W)/w}$ is a scaling factor that gives higher weight to smaller TAD hierarchies according to the set *W*. Namely, we are setting the highest level to have a weight equal to one. To extend the score to the entire col-

lection $\mathcal{D}$ of Hi-C matrices, we define the HC score as:

$$S = \sum_{D \in \mathcal{D}} S^{(D)} \quad (2)$$

Each element of the matrix s_{ij} is meant to capture the broader spatial co-localization pattern of the ith enhancer and the jth promoter across both different layers of TADs hierarchy and tissue or cell types (Figure 1B, bottom panel). By definition, for the set $W = [5, 10, 20, 50]$ and using 11 Hi-C datasets, the lower and upper limits of the score are $\min(S) = 0$ and $\max(S) = 87.8$, respectively, where the maximum for each Hi-C contact matrix is equal to $\max(S^{(D)}) = 7.98$.

As the HC score can be calculated whenever a hierarchy of TADs is provided, for comparison purposes we considered TopDom (58), a public available tool meant to identify TADs at sub-mega base resolution. TopDom identifies TAD boundaries looking at significant local minima of the bin signal function, which is computed with a procedure similar to the previously proposed insulation score (59). Namely, the bin signal function is the average contact signal in the neighbourhood of each bin along the diagonal, considering a diamond-shape area of width $2\hat{w}$, where $\hat{w}$ is a tuneable parameter that defines the window size. We used TopDom R package (https://github.com/HenrikBengtsson/TopDom, version 0.8.1) to call TADs, defined as regions between two boundaries flagged as significant minima (local.ext = -1). For the $\hat{w}$ parameter we used the same set of values adopted for LSD DI-window sizes, i.e. $W = \hat{W} = [5, 10, 20, 50]$.

We also considered alternative definitions of the HC score that we show in specific analyses as indicated in the text and related Supplementary Figures. These were based on alternative definitions of the scaling factor $\lambda(w)$ in (1), including: unweighted sums with $\lambda(w) = 1$; weighted and rescaled sums with $\lambda(w) = \max(W)/w$; logarithm (base 10) of weighted and rescaled sums with $\lambda(w) = \log_{10}(\max(W)/w)$; weighted sums using the inverse of the DI-window with $\lambda(w) = 1/w$; square root of weighted and rescaled sums with $\lambda(w) = \sqrt{\max(W)/w}$ (default choice of our method in the manuscript).

Enhancer-promoter pairs synchronization analysis with Canonical Correlation

We adopted the Canonical-Correlation Analysis (CCA) (60) to quantify the strength of coordinated activity in each EP pair (Figure 1A). We considered enhancer and promoter regions separately, and quantified their respective activity status using two sets of epigenetic marks: we used the enrichment of DNase-seq and H3K27ac ChIP-seq ($p = 2$) for enhancers and DNase-seq, H3K27ac and H3K4me3 ($q = 3$) for the promoters.

Namely, we downloaded H3K27ac, H3K4me3 and DNase-seq consolidated fold-change enrichment signal tracks (bigwig format) from the Roadmap Epigenomic consortium web portal (https://egg2.wustl.edu/roadmap/data/byFileType/signal/consolidated/macs2signal/foldChange/) for all the cell and tissue types for which all the three epigenetic marks were available ($k = 44$) (Supplementary Table S1A). For each enhancer and promoter region, we computed the maximum signal from the proper bigwig genomic tracks, using rtracklayer R package (version 1.44.4) (61).

We then used CCA to investigate the inter-set correlation patterns (Figure 1A, bottom panels). More formally, let $X^{(i)}$ denote a p-dimensional random vector of quantitative features describing the activity of the ith enhancer. Let $Y^{(j)}$ denote a q-dimensional random vector of quantitative features describing the activity of the jth promoter. Our data consist of k independent observations of $X^{(i)}$ and $Y^{(j)}$ across k cell and tissue types. We are interested in testing the null hypothesis of independence between $X^{(i)}$ and $Y^{(j)}$, *i.e.,* the lack of synchronized activity:

$$H^{(ij)} : X^{(i)} \, and \, Y^{(j)}, \; are \, independent \quad (3)$$

against a general alternative. Assuming normality of $X^{(i)}$ and $Y^{(j)}$, the null hypothesis of interest can be equivalently expressed as:

$$\mathrm{H}^{(\mathrm{ij})} : \rho_1^{(\mathrm{ij})} = \ldots = \rho_{\min(\mathrm{p,q})}^{(\mathrm{ij})} = 0 \quad (4)$$

where $\rho_l^{(ij)}$ is the lth canonical correlation coefficient associated to $X^{(i)}$ and $Y^{(j)}$. Briefly, the canonical correlation coefficients measure the correlation over subsequent linear transformations of the original p and q variables, that allow maximizing the relationship between the two sets, while ensuring independence within each set. The maximum number of linear transformations is $\min(p, q)$, *i.e.,* two in our case. The key advantage of CCA is to reduce the dimensionality and the inter-confounding factors of each set, while extracting the major correlation patterns.

Following the CCA, we calculate the P-value for the null hypothesis (4) by testing the sequential hypotheses that the first canonical correlation coefficient, and all the following ones, are zero using the Wilk's lambda statistics (60):

$$\lambda = \prod_{l=1}^{\min(p,q)} \left(1 - \left(r_l^{(ij)}\right)^2\right) \quad (5)$$

where $r_l^{(ij)}$ is the estimated lth canonical correlation coefficient. To improve the accuracy for small sample sizes, we adopted the Rao's F-approximation (62). Namely, λ was transformed to an F-statistic using Rao's approximation as implemented in the candisc R package (version 0.8-3).

The procedure returned a single P-value $p_{(ij)}$ for the overall dependence of the jth promoter on the ith enhancer. $p_{(ij)}$ quantifies the amount of evidence provided by the data for the presence of the synchronized activity between a specific EP pair.

3D architecture integration in the enhancer−promoter pairs FDR control

The reconstruction of the EP pairs based on CCA, as described above, is based on testing millions of hypotheses (i.e. one for each EP pair), thus requiring some control over the number of false discoveries. In large scale testing problems of this kind, the typical goal is the control of the False Discovery Rate (FDR), defined as the expected fraction of false discoveries. The Benjamini-Hochberg (BH) (63) correction is a frequently used method for controlling FDR in

genomics data analyses. In this work, however, we aim to increase statistical power over the standard BH procedure by considering relevant contextual information. An example is provided by the 3D co-localization information encoded in the HC score. To include this information in the testing problem, we relied on the Adaptive *P*-value thresholding procedure (AdaPT), recently proposed by Li and colleagues (64) and implemented in the adaptMT R package (version 1.0.0). AdaPT estimates Bayes-optimal *P*-value rejection threshold based on user-defined side information, and controls FDR in finite samples. Five logistic-Gamma generalized linear models with natural cubic splines as candidate models were explored to identify the best threshold estimate, as proposed by default parameters of AdaPT implementation. This choice implies that the results of our method are robust with respect to different definitions of the scaling factor $\lambda(w)$ in (1) (e.g. unweighted sums or logarithmic transformation) or the HC score (e.g. usage of a smaller subsets of input Hi-C matrices), as long as a linear relationship among alternative score definitions is preserved.

Formally, we considered our collection of hypotheses $\{H^{(ij)},\ i = 1,\ \ldots,\ n,\ j = 1, \ldots,\ m\}$, as defined in (4) for which we computed (i) a *P*-value $p^{(ij)} \in [0, 1]$ quantifying the strength of evidence for the presence of a synchronized activity between the ith enhancer and the jth promoter and (ii) a score $s_{ij} \in \mathbb{R}$ capturing the likelihood of their 3D proximity. Then, we used AdaPT to determine a rejection threshold in function of the HC score (Figure 1C), such that the estimate FDR is bounded by α. A more detailed exposition of the theoretical framework can be found in the original article (64). AdaPT has been shown (64,65) to significantly increase statistical power in situations in which the considered side information provides a useful basis for prioritizing most promising hypotheses. Nevertheless, statistical guarantees regarding FDR control are preserved also when the side information is inaccurate or not relevant for the problem at hand: in this case, the weight given to the side information will be low and AdaPT will converge to the standard BH method.

Reference benchmarking datasets

Expression quantitative trait loci databases. Expression quantitative trait loci (eQTLs) are Single Nucleotide Polymorphisms (SNPs) associated to an alteration in the expression of a specific gene. We considered multiple eQTL datasets as reference for benchmarking the pairing of distal regulatory elements to their target gene. In particular, we considered eQTL data from i) the Genotype-Tissue Expression (GTEx) project (66), with eQTLs inferred from a panel of 15 201 samples in 48 tissue types; and ii) the pan-cancer eQTL (PanCanQTL) analysis (67), with eQTLs inferred from a panel of 9196 tumour samples in 33 cancer types from The Cancer Genome Atlas (TGCA).

Cis-eQTL files from the v8 GTEx data release for 48 tissue types (Supplementary Table S1C) were downloaded from GTEx portal (https://www.gtexportal.org/home/datasets). All *.sign_variant_gene_pairs.txt.gz files were converted to GenomicRanges (1.36.1, Bioconductor package) objects and merged maintaining only one eQTL in case of redundancy. The *.egenes.txt.gz files were used to convert Ensemble gene IDs to Gene Symbols. Genomic coordinates were converted from hg38 to hg19 genome build using liftOver tool (rtracklayer R package version 1.44.4 (61)).

Cis-eQTL files from PanCanQTL (Supplementary Table S1D) were downloaded from URL http://gong_lab.hzau.edu.cn/PancanQTL/cis (filenames with suffix *_tumor.cis_eQTL.xls). All Cis-eQTL files were converted in GenomicRanges objects and merged maintaining only one eQTL in case of redundancy.

An EP pair was considered supported by an eQTL (i.e. validated) if the corresponding SNP was located within an enhancer genomic region and associated with the expression of the cognate promoter. If multiple SNPs were within the same enhancer genomic region, they were considered only once. If the same eQTL was predicted in multiple tissue types to regulate a specific target gene, it was also considered only once.

Capture Hi-C datasets. We also considered nine capture Hi-C (cHi-C) experiments (Supplementary Table S1E) coming from seven different studies (68–74), specifically designed to identify DNA-DNA interaction between promoters and distal chromatin regions. All the downloaded interaction lists (washU format) were already pre-processed in the original articles, and CHiCAGO (Capture Hi-C Analysis of Genomic Organization) (70) algorithm was used to select significant interactions (CHiCAGO score $\geq$5). Genomic coordinates were converted from hg38 to hg19 genome build using liftOver tool (rtracklayer R package version 1.44.4 (61)), when needed.

An EP pair was considered supported by a cHi-C interaction if the promoter region overlaps ($\geq$1 bp) with the 'bait fragment' and the enhancer with the 'other end fragment', or vice versa. Ambiguous EP pairs due to the cHi-C resolution (i.e. pairs supported by the same cHi-C interaction) were not discarded.

CRISPR-based perturbation datasets. As alternative and independent functionally validated sets of ETG pairs, we used the datasets by Fulco *et al.* (75) and Gasperini *et al.* (76), two recent publications adopting CRISPR-based perturbation techniques coupled with single cell transcriptomic readout (Supplementary Table S1F). We focused on K562 human erythroleukemia cells, which is the most characterized cell line in both datasets.

For Fulco dataset we downloaded the *'Dataset of experimentally tested noncoding element-gene connections in k562 cells'* from the supplementary materials of the original paper (Supplementary Table S1F). From the complete list of candidates ETG pairs, we removed those involving non-coding elements classified as 'promoter'. Following the original authors suggestions, we considered as validated the pairs with an adjusted *P*-value lower than 0.05 and 0.8 power to detect 25% effects. This selection resulted in 3836 candidate pairs, of which 141 (3.7%) were validated.

For Gasperini dataset (Supplementary Table S1F), we downloaded the complete list of gRNAs-target gene pairs of the scaled multiplex enhancer-gene pair screen from the Gene Expression Omnibus repository ('*GSE120861_all_deg_results.at_scale.txt.gz*'). From this

list we maintained only gRNAs associated to DHS peaks, i.e. we removed those classified as promoter proximal elements ('TSS' and 'selfTSS') or positive and negative controls ('NTC' and 'positive_ctrl'). According to the filters applied by the original authors, we retained only gRNAs flagged as 'top_two' in the 'quality control' field and for which an adjusted empirical *P*-value was available. To identify validated gRNAs-target gene pairs we applied the recommended threshold of 0.1 at the adjusted empirical *P*-values. This selection resulted in 40 322 candidate pairs, of which 664 (1.6%) were validated.

BENGI benchmark. As an additional reference dataset, we considered the Benchmark of candidate Enhancer–Gene Interactions (BENGI) dataset (77). We downloaded *All-Pairs.Natural-Ratio* files from BENGI GitHub repository (https://github.com/weng-lab/BENGI/). This included a total of 21 lists of curated interactions (Supplementary Table S1G) supported by ChIA-PET, Hi-C, eQTL and CRISPR genome editing experiments and covering seven cell lines and six tissue types. For each file, only the enhancer-like signatures (i.e. the ones marked as high DNase and H3K27ac signal) were considered. In line with the previous section ('Definition of enhancer and gene promoter regions'), but using BENGI gene definitions (GENCODEv19-TSSs.bed.gz annotation file), we defined promoter intervals as 2 kb windows (1.5 kb upstream and 0.5 kb downstream) around the transcription start site (TSS). Only the most upstream TSS for each gene was preserved. Enhancer intervals were annotated using the hg19-cCREs.bed.gz file. All 21 lists of enhancers and promoters were pooled to perform the EP pairing analysis based on our framework and then split for the assessment. An EP pair was deemed as true positive if supported by a specific BENGI curated interaction (i.e. the internal flag was equal to 1).

ETG pairs by other tools

To benchmark our ETG pairing framework against other algorithms, we considered state-of-the-art methods among the 36 listed in a recent review (78). To overcome limitations related to the lack of user-friendly software, we considered only algorithms with publicly available ETG pairs lists, called as described in the original publications. Namely, the selection resulted in eight tools (Supplementary Table S1H): FOCS (FDR-corrected OLS with Cross-validation and Shrinkage) (79), JEME (Joint Effect of Multiple Enhancers) (27), RIPPLE (Regulatory Interaction Prediction for Promoters and Long-range Enhancers) (80), PETmodule (Predicting Enhancer Target by modules) (28), TargetFinder (29), DeepTACT (Deep neural networks for chromatin conTACTs prediction) (81), PreSTIGE (Predicting Specific Tissue Interactions of Genes and Enhancers) (82) and ABC (Activity-By-Contact model) (75).

For each algorithm, we downloaded the lists of ETG pairs for all the available cell and tissue types and we processed them to obtain a uniform format of annotations. Namely, for each ETG pair we stored: (i) enhancer genomic region coordinates (chr:start-end); (ii) promoter region or TSS genomic coordinates (chr:start-end), depending on the information reported by the authors; (iii) gene symbol; (iv) prediction flag (i.e. 1: predicted, 0: not predicted), as sometimes the original authors reported only the predicted pairs and sometimes also the entire set of initial candidates; (v) distance between enhancer mid-point and promoter mid-point (or TSS); and other optional information returned by the specific algorithm (e.g. score, etc.). Genomic coordinates were converted from hg38 to hg19 genome build using liftOver tool (rtracklayer R package version 1.44.4 (61)), when needed. Gene symbol and TSS coordinates were retrieved from BioMart database, through R Bioconductor interface (version 2.40.5, host = 'grch37.ensembl.org', path = 'biomart/martservice/', database = 'hsapiens_gene_ensembl'). ETG pairs associated to Ensemble gene IDs without any match with gene symbols were discarded. To make enhancers of the other tools comparable with our enhancer reference catalogue, we applied the filters described in section 'Definition of reference set of enhancer and gene promoter regions' with minor modifications. Namely, we removed interval portions overlapping annotated exons (for both coding and non-coding genes, RefSeq annotations in UCSC) and promoter proximal elements. If one or more exons were completely located within an enhancer interval, the enhancer was split and the pair duplicated in concordance with the number of resulting enhancers. A promoter proximal element was defined as a pair with distance between enhancer mid-point and promoter mid-point (or TSS) smaller than 3 kb.

ETG pairs in cell specific context

To evaluate the performances of our method in identifying cell-type specific ETG pairs, we performed a direct comparison with JEME, using the initial set of enhancers, gene and candidate ETG pairs, inferred from 127 cell and tissue types, collected by the Roadmap Epigenomic consortium (Supplementary Table S1H). In line with JEME annotations, we defined promoter intervals as 1 kb windows (0.5 kb upstream and 0.5 kb downstream) around the transcription start site (TSS). Interval portions of enhancers overlapping promoters were removed. Pairs with distance between enhancer and TSS smaller than 0.5 kb were considered as promoter-proximal and removed from the list of candidate pairs. All lists of enhancers and promoters were initially pooled together to perform the EP pairing analysis based on our framework, and then split for the assessment. The HC scores were calculated based on EP pairs co-localization and no cut-off was applied (i.e. we did not filter pairs with HC >11). Candidate pairs were sorted based on the 'confidence score' (descending order) and *P*-values (ascending order), for JEME and our framework, respectively.

To investigate the expression of the predicted target genes we downloaded all the available matched consolidated RNA-seq profiles (57 out of 127) from Roadmap Epigenomics portal (https://egg2.wustl.edu/roadmap/data/byDataType/rna/expression/57epigenomes.RPKM.pc.gz).

We further validated the results in the cell-type specific context by using the two CRISPR-based perturbations datasets on K562 cell line, described above (i.e. Fulco and Gasperini). Using the JEME initial set of candidates ETG pairs for the matched cell line ('en-

coderoadmap_lasso.121.csv' file), we retained only those for which the enhancers described in Fulco and Gasperini datasets overlaps (≥1 bp) with the JEME set of enhancers. Moreover, since Fulco dataset is the training set for the ABC predictor, we further filtered ETG pairs in Gasperini dataset, so as to allow a comparative assessment against ABC as well. Namely, we used 'K562.AllPredictions.txt' file for the ABC algorithm.

This selection resulted in 581 and 4185 candidates ETG pairs for Fulco and Gasperini, respectively, of which 18 (3.1%) and 62 (1.5%) were validated.

Assessment of predicted ETG pairs and other indices

For each ETG calling algorithm (Supplementary Table S1H) we define as Z the number of candidate pairs, i.e.,the initial number of input EP pairs for which a score or P-value was calculated by the original authors; V the number of true pairs, i.e. all the EP pairs contained in at least one of the reference benchmarking datasets; z the number of predicted pairs, i.e. the EP pairs that satisfied the selection criteria as applied by the original authors (e.g. P-value $\leq \alpha$); v the number of true predicted pairs, i.e. the predicted EP pairs contained in at least one of the reference benchmarking datasets.

We used four different indices for performance assessment: *Precision* (P), the percentage of true predicted pairs over the total number of predicted pairs (v/z); *Recall* (R), the percentage of true predicted pairs over the total number of true pairs (v/V); *F1 score*, the harmonic mean of precision and recall ($2 \times \frac{P \times R}{P+R}$); and *Relative Improvement* (RI), the improvement respect to random choice ($\frac{v/z}{V/Z}$). Precision-recall curves were computed by sorting EP pairs based on distance, HC score, canonical correlation or HC-based AdaPT corrected P-value, and calculating precision and recall for all the possible cut-offs (Z) of the candidate pairs list.

The Jaccard Index (JI) between two sets of genomic regions was calculated as i) the number of elements that overlap (≥1 bp) over the total number of elements in the two sets (i.e. JI on overlap); or ii) the total length of intersections divided by the total length of the union of the two sets (i.e. JI on coverage).

RESULTS

Methodological framework overview

Here we present a general framework for the definition of enhancer–target gene (ETG) pairs leveraging the current biological knowledge on chromatin 3D architecture and integrating heterogeneous functional genomics data into a rigorous statistical framework. Its three key features are:

Statistical framework for quantifying enhancer−promoter pairs synchronization. The method is flexible in terms of input, as it starts from user-defined sets of (i) enhancer and promoter regions and (ii) functional genomics data to quantify their activity (Figure 1A). This flexibility is ensured by the use of Canonical-Correlation Analysis (CCA) to quantify the synchronization of enhancer−promoter (EP) pairs activity across cell types. Moreover, it is designed to leverage multiple types of functional genomics data, also accounting for the correlation within sets of features.

Hierarchical contact (HC) score. It incorporates chromatin architecture as experimentally measured by Hi-C, to compute the HC score accounting for ETG pairs 3D co-localization (Figure 1B). Differently from previous methods, we leverage biological knowledge on TADs multi-scale hierarchical organization and their conservation across cell types.

Chromatin 3D architecture and functional genomics data integration. The information on chromatin 3D architecture is used to increase the statistical power to detect ETG pairs synchronization, while controlling false discoveries (Figure 1C). This is the first time that chromatin 3D architecture is directly integrated as side information in the statistical model for defining ETG pairs.

Definition of the reference enhancer catalogue

The first challenge in the definition of ETG pairs is the lack of a universal reference list of enhancer regions, as they do not have a univocal nucleotide sequence. A comprehensive definition of enhancers based on functional genomics data in principle would require analysing virtually every cell and tissue type. This is practically impossible, despite ambitious large-scale collaborative projects such as the ENCODE (83), FANTOM (84) and Roadmap Epigenomics consortia (1). However, the goal of our work was not to define the ultimate set of enhancers, but rather to verify if accounting for chromatin 3D organization can improve ETG pairing.

Thus, we primarily relied on Roadmap Epigenomics dataset as (i) it covers a broad range of cell and tissue types; (ii) it adopted shared protocols and quality standards, which is preferable to merging data from heterogenous sources; iii) the use of enhancers defined with epigenomics data facilitates the comparison against previously published algorithms for ETG pairing. More specifically, we used the peaks called by Roadmap Epigenomics for DNase-seq and H3K27ac ChIP-seq to define active enhancer regions for each of the selected 44 cell and tissue types (see Materials and Methods section and Supplementary Table S2A). The average number of cell-specific enhancers is 33 560 (Figure 2A), with average size 316 bp (Supplementary Figure S1B). Their pairwise comparison showed on average 50.6% of similarity (JI on overlap) (Figure 2B and Supplementary Table S2B). On the other hand, the mean JI for coverage was 24.4% (Supplementary Table S2A), due to the variable range of enhancer region sizes.

To define a comprehensive reference enhancer catalogue, we considered the union of genomic intervals for cell-specific enhancers, resulting in $n = 347\,777$ enhancer regions, with average size 416 bp (Supplementary Figure S1B), i.e. slightly higher than the cell-specific enhancers, as the final catalogue is derived from their union.

This reference enhancer catalogue can be considered exhaustive and representative also for other cell types. To

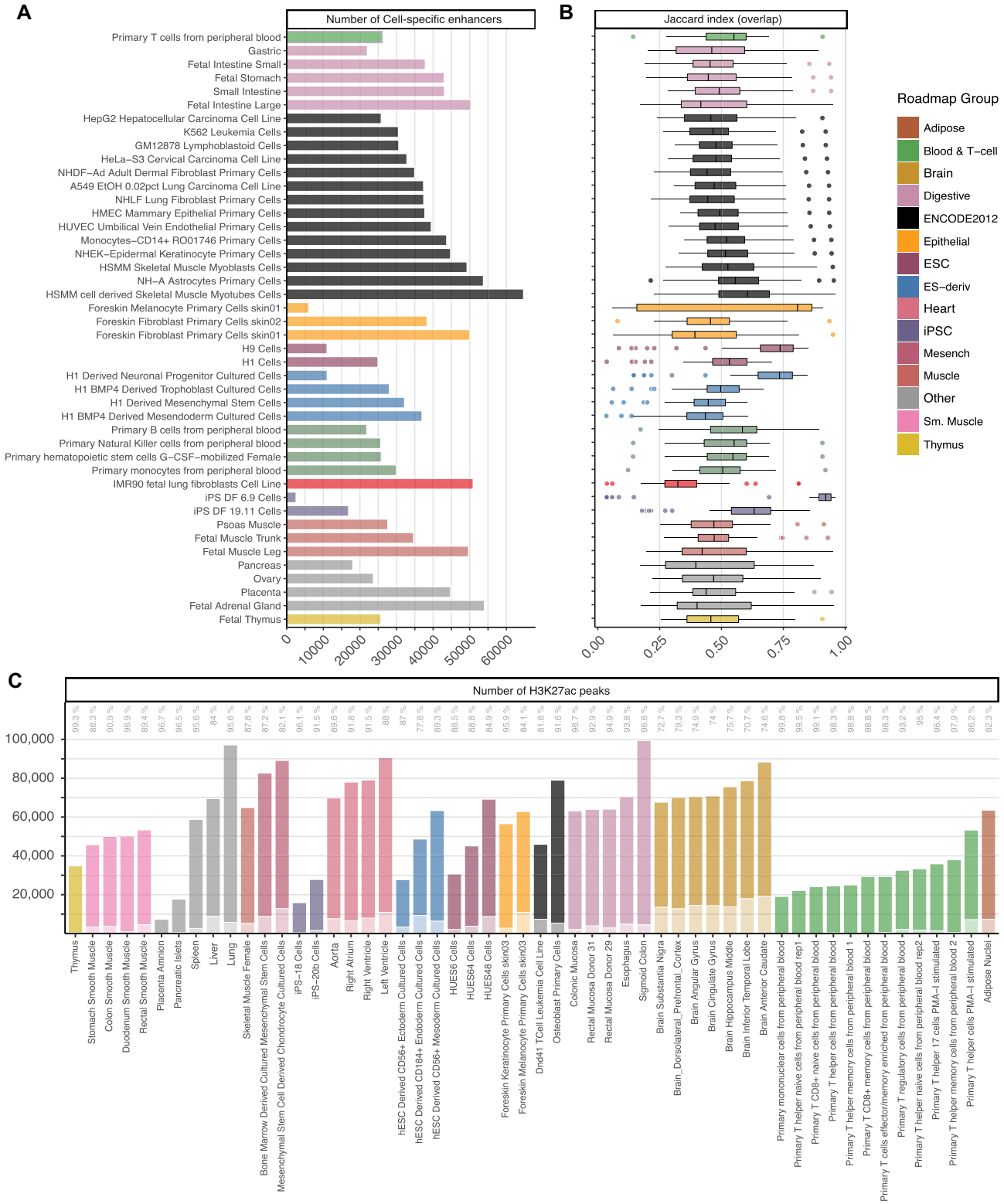

Figure 2. Definition of the reference enhancer catalogue. (**A**) Number of cell-specific enhancers resulting from the intersection of DNase-seq and H3K27ac ChIP-seq peaks in a selected set of 44 cell and tissue types collected by the Roadmap Epigenomics consortium, coloured by Roadmap groups. (**B**) Similarity (Jaccard index on overlap) among cell-specific set of enhancers. Each data point in the boxplots represents the ratio between the intersection of two cell-specific sets of enhancers over their union, taking as reference the group on the row. The median is marked with a line across each box, the box margins mark the interquartile range (IQR), the whiskers extend up to 1.5 IQR and individual data points are shown for outliers beyond this range. (**C**) Number of H3K27ac ChIP-seq peaks of 54 additional cell and tissue types from the Roadmap Epigenomics project that overlap (darker colour shade) or do not overlap (lighter colour shade) with the union of H3K27ac ChIP-seq merged peaks of the set of cell and tissue types used to define the reference catalogue of active enhancers. The percentages of total H3K27ac peaks extension overlap are reported as text labels on top.

this concern, we considered 54 additional Roadmap Epigenomics H3K27ac profiles, that were omitted from our enhancer catalogue definition because they lack a corresponding DNase-seq profile (Supplementary Table S2C). On average, 91.3% of the additional cell-specific H3K27ac peaks overlap to the union of H3K27ac peaks across the 44 cell types considered above. This overlap is 89.9% if we consider its extension over the total coverage in the respective cell type (Figure 2C). The large overlap can be considered indicative for the completeness of our catalogue.

We also compared our reference enhancer catalogue to enhancer definitions by CAGE, from the fifth release of the FANTOM project (Functional ANnoTation Of the Mammalian genome) (84). We found that 57% out of the 58 200 filtered FANTOM enhancers (median length 270 bp) were also represented in our catalogue (Supplementary Table S2D). It is noteworthy that FANTOM enhancer definitions were based on functional data from a much larger set of cell and tissue types including 432 primary cells, 135 tissue types and 241 cell lines (808 in total) (44). Thus, we deem our strategy a good compromise as FANTOM is based on 18 times more cell and tissue types.

Finally, we compared our catalogue to an *in vivo* validated set of enhancers coming from the VISTA Enhancer Browser database (46). Out of the starting 894 filtered VISTA database enhancers (median length 1676 bp) (Supplementary Figure S1C and Table S2D), 55% are present in our enhancer catalogue. It is worth remarking that VISTA is made of enhancers validated to be active in mouse embryos at development day 11.5. Thus, it is based on a different model organism and a very specific embryonic development stage, as opposed to our epigenomics datasets, which are derived from human samples, including several from differentiated tissues and cells from adult individuals. Moreover, superimposing the filtered FANTOM5 enhancers, only a minor residual number is detected in addition to our enhancer catalogue (Supplementary Figure S1C and Table S2D). This observation confirms that alternative functional genomics definitions of enhancers, such as the CAGE-based FANTOM5, are overall comparable to ours.

Enhancer–promoter interactions in the 3D context

Enhancer-promoter contacts are generally confined within the boundaries of TADs (85), i.e. structurally separated domains relatively insulated from surrounding regions. TAD boundaries are mostly stable across cell types, but their insulation is far from absolute. Moreover, it is possible to identify a hierarchy of TADs, as any given Hi-C contact matrix can be analysed at different scales to derive alternative definitions of insulated domains (39,86).

In order to account for these known features of chromatin 3D organization, we devised the HC score, which is proportional to the likelihood of 3D co-localization of EP pairs (see Materials and Methods) (Figure 1B). We used 11 high-coverage Hi-C datasets (on average 660 million aligned reads), covering 10 different cell and tissue types (Supplementary Table S1B) and binned at 10 kb resolution. We then applied the LSD algorithm to identify TAD boundaries at multiple scales, thus obtaining different segmentations of the genome that account for the hierarchy of structural domains. The number and size of TADs show a trend related to the LSD DI-window size parameter. Namely, we find fewer and larger TADs when increasing DI-window (hierarchy level) (Supplementary Figure S2A): with average number ranging from 18 254 to 8953, and average size from 183 to 525 kb (Supplementary Table S3A). This pattern is comparable across datasets, despite differences related to sequencing depth. The pairwise comparison of domain boundaries across datasets, and across hierarchy levels, shows an average JI (coverage) ranging from 44.9 (for DI-window 5) to 35.8 (for DI-window 50) (Supplementary Figure S2B). These results are in line with previous studies (87) and with the notion that several TADs are conserved across cell types.

We mapped our catalogue of enhancers (347 777) and reference set of promoters (18 027) to the inferred TADs (Supplementary Table S3B), which are expected to compartmentalize the interactions between distal regulatory elements and target genes, and then we computed the HC score for each EP pair. About 70% of EP pairs are within the same TAD in two or more Hi-C datasets, with a frequency distribution that is similar across all TAD hierarchy levels (Figure 3A). Nevertheless, the number of EP pairs grows as the hierarchy considered increases, as expected.

The EP pairs mapped within at least one TAD definition (i.e. HC score > 0) are 12 949 150 (Supplementary Table S3C) of which $\sim$75% have a weak score (HC score ≤ 11), i.e. they are supported only in few combinations of datasets and hierarchy levels. Possible scenarios above this threshold include EP pairs supported by all 11 datasets in at least one hierarchy level, or EP pairs supported by 2 or more datasets across multiple hierarchy levels. We deliberately designed the score to give comparable weight to these alternative situations.

This pattern, together with the overall trend observed in the HC score, is robust respect to the TADs calling algorithm that is adopted. For example, we used as alternative input the TADs called by TopDom (58), another method that allows calling TADs at different levels of resolution by adjusting a tuning parameter (see Materials and Methods). We observed a similar percentage of poorly-supported pairs (79.8% with HC score ≤ 11), as well as a per chromosome average 0.76 correlation (Spearman) between HC scores based on TopDom or LSD TADs (Supplementary Table S3D). While our observations are robust to different TADs calling algorithms, we noted that the number of candidate EP pairs based on LSD TADs is almost totally (98%) included in the set based on TopDom TAD definitions (Supplementary Figure S2C). To this concern, it should be considered that the total number of candidate EP pairs for TopDom is approximately three-fold higher (38,527,013 pairs). This disparity is partly explained by the variable TAD calling performance of the two algorithms across datasets, which tend to give more different results when the coverage is lower (Supplementary Figure S2D). Nonetheless, the comparison and benchmarking of TAD callers is beyond the scope of this work and has been extensively addressed in previous literature (19,88,89).

To ensure the robustness of the downstream results, we discarded poorly-supported EP pairs (HC score ≤ 11) from subsequent analyses, as they may be the consequence of

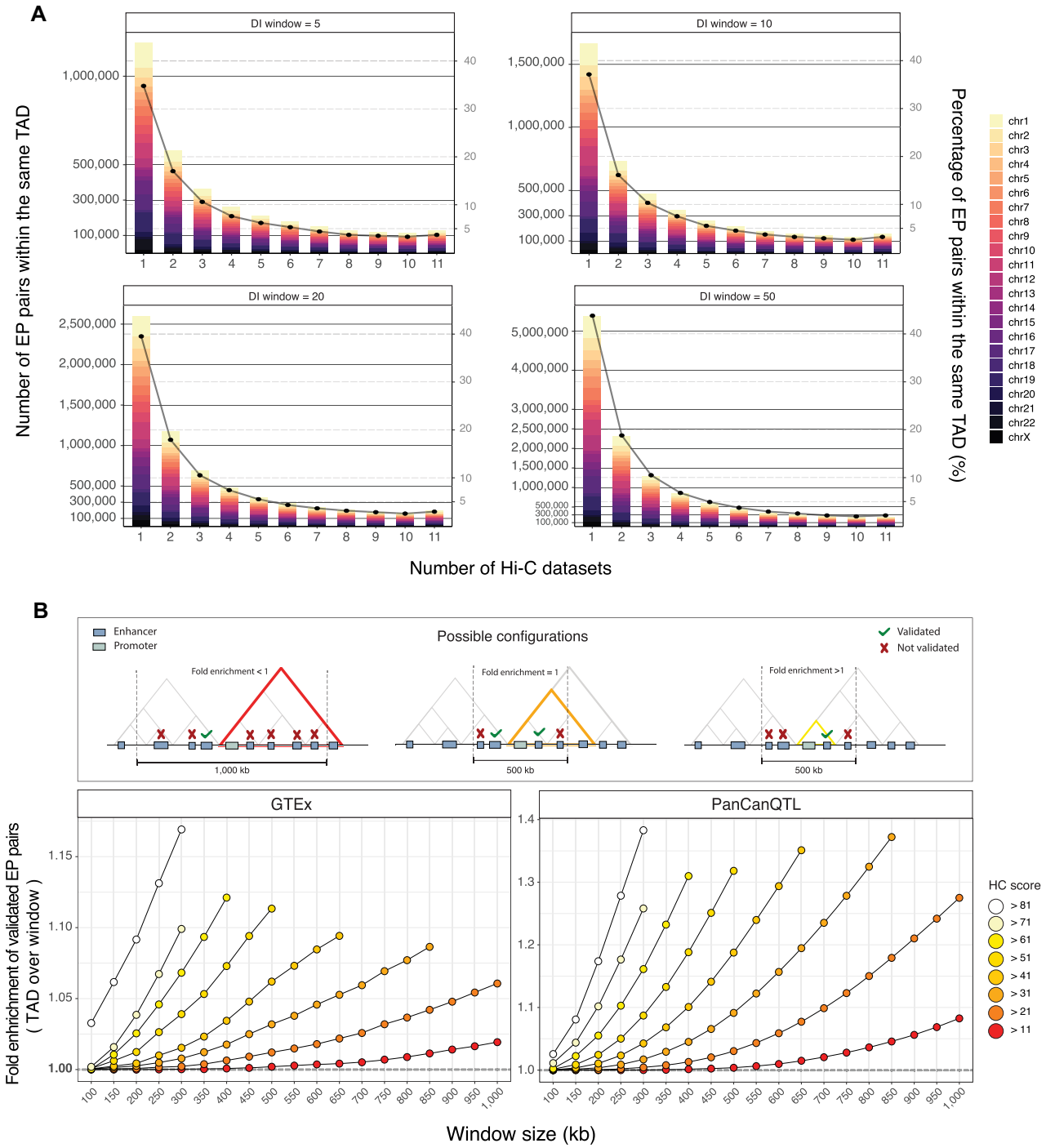

Figure 3. Enhancer-promoter interactions in the 3D context. (**A**) Number (bar, left y-axis) and percentage (point, right y-axis) of EP pairs located within the same TAD for one or multiple of eleven analysed Hi-C datasets (x-axis), considering different TADs hierarchical levels (i.e. DI-window, panels), grouped by chromosomes (colours). (**B**) Fold enrichment (y-axis) of eQTL-supported ETG pairs coming from two public datasets (GTEx, left panel; PanCanQTL, right panel) with respect to candidate pairs as defined with HC score over a fixed-width window around the promoter of the target gene, for different cut-offs (score, colours; window-width, x-axis). As illustrated in the cartoon on top of the figure, values greater than one imply an enrichment of TAD-based pairing over results obtained with a fixed-window.

noise in the data depending on technical variables (e.g. coverage). This filter resulted in a total of 3 192 806 candidate ETG pairs.

There is a consensus on the fact that enhancers may not target the closest gene, in terms of linear sequence of the genome. Nevertheless, previous literature on genome-wide reconstruction of ETG pairs often adopted a fixed-width window around TSS to restrict the EP pairs search space. A commonly adopted boundary is a 1 Mb window (±500 bp around the TSS) to define the initial set of candidate pairs (90). Likewise, also the literature on expression quantitative trait loci (eQTLs) adopted a similar simplification. Indeed, cis-eQTL are often defined as SNPs within a 1 Mb window around the TSS, as opposed to trans-eQTL if the SNP falls beyond that distance threshold (or beyond 5 Mb for some other studies) or in another chromosome (91).

To verify if the use of chromatin 3D architecture as incorporated in the HC score brought an advantage over the standard choice of a fixed-width window, we used a true positive set of ETG pairs based on eQTLs from the GTEx project (66) and PanCanQTL (67). Specifically, we verified the proportion of eQTL-supported ETG pairs with respect to the total number of considered pairs as defined with HC or fixed-width windows. Since eQTLs are explored only for SNPs at a maximum distance of 1 Mb from the candidate target genes, to make a fair comparison we removed from our list all candidate pairs more distant than this threshold for a total of 3 102 154 remaining candidates. This filter was applied also for any subsequent analysis in which eQTLs were considered for comparison. We observed that in both eQTLs datasets there is generally a higher frequency of validated pairs when accounting for the chromatin 3D architecture, even if varying the threshold on HC score and fixed-width window parameters (Figure 3B). These results highlight the existence of a stronger relationship between eQTLs and 3D distance, rather than linear distance.

Physical proximity increases power of detection

We then reconstructed the enhancer regulatory map by integrating information on physical co-localization of EP pairs (HC) and their activity synchronisation (CCA).

Enhancers are expected to show the properties of an active regulatory region in the specific cell context where they are contributing to activate a target gene (92). Thus, searching for 'synchronised' enhancer and promoter activity across multiple cell types is a commonly adopted strategy in ETG pairing literature (93), although there is no consensus regarding a measure that best conveys their synchronisation. Differently from previously published methods, we adopted CCA as a convenient statistical framework to assess the synchronisation between activity of enhancers and promoters across multiple cell types in a fast and efficient way. We chose CCA because (i) it is flexible with regard to the set of input functional genomics data used to estimate the activity level of enhancers and promoters; and (ii) it accounts for the confounding factor of multiple types of functional genomics data being correlated with each other (Figure 1A).

In our case, we used a combination of DNase-seq and ChIP-seq enrichment profiles to quantify the activity of enhancers and promoters. Namely, we used the maximum enrichment of DNase-seq and H3K27ac ChIP-seq for enhancers (347 777) and DNase-seq, H3K27ac and H3K4me3 for the promoters (18027), as described in Materials and Methods (Supplementary Figure S3A). To minimize the influence of possible outliers and make the distributions comparable across all cell types, we used $\log_2(x+1)$ transformed enrichment values and adopted a chromosome-wise quantile-normalisation, respectively. We also tested cycle loess and variance stabilizing normalisation (VSN) as alternatives. They all yield similar results (Supplementary Figure S3B-C), thus we selected the quantile normalisation as it preserves the original range of values.

To assess the association between the *i*th enhancer and the *j*th promoter, we performed the CCA considering the enrichment in these chromatin marks across the selected set of 44 cell and tissue types. The procedure returns a single *P*-value $p^{(ij)}$, for each EP pair under consideration (see Materials and Methods). We performed CCA on the subset of EP candidate pairs filtered by HC score >11, resulting in a total of $N = 3\ 192\ 806$ hypotheses to be tested (Supplementary Table S3C). For each chromosome we estimated an adaptive *P*-value rejection threshold using the AdaPT procedure (64) with side information derived from the physical proximity of enhancer−promoter (HC score) (Figure 1C). A representative example of the estimated thresholding rules for chromosome 19 is depicted in Figure 4A. We can see an increasing trend of the rejection curve in relation to HC, implying that on equal nominal *P*-value our framework uses a less conservative significance criterion for the EP pairs showing higher likelihood of 3D contact interaction across cell types. Similar trends are observed for all chromosomes.

To illustrate the improvement achieved by integrating the score on chromatin 3D architecture, we adjusted the CCA *P*-values using Benjamini-Hochberg (BH) and HC-based AdaPT correction (Figure 4B), and Bonferroni approach. Using the union of the GTEx and PanCanQTL datasets as reference true positive benchmark, we assessed the three methods in terms of Precision, Recall and F1 score (Supplementary Table S4A). In Figure 4C we observe that, regardless of the level of confidence chosen, integrating the HC in the *P*-value correction leads to an appreciable increase of the power (almost twice) without affecting the accuracy of the predictions. These results are highly robust with respect to different definitions of the score (e.g. unweighted sums, logarithmic transformation; Supplementary Figure S3D) and the input Hi-C datasets (e.g. usage of a smaller subsets of input Hi-C matrices; Supplementary Figure S3E), as long as a linear relationship among alternative score definitions is maintained.

The complete list of candidate enhancer−promoter pairs annotated with the HC score, corrected and uncorrected *P*-values, validations according to multiple reference datasets are publicly released (see Data Availability). For the subsequent analyses we maintained the EP pairs with HC-based AdaPT adjusted *P*-value ≤ 0.05 resulting in a total of 233 304 predicted pairs.

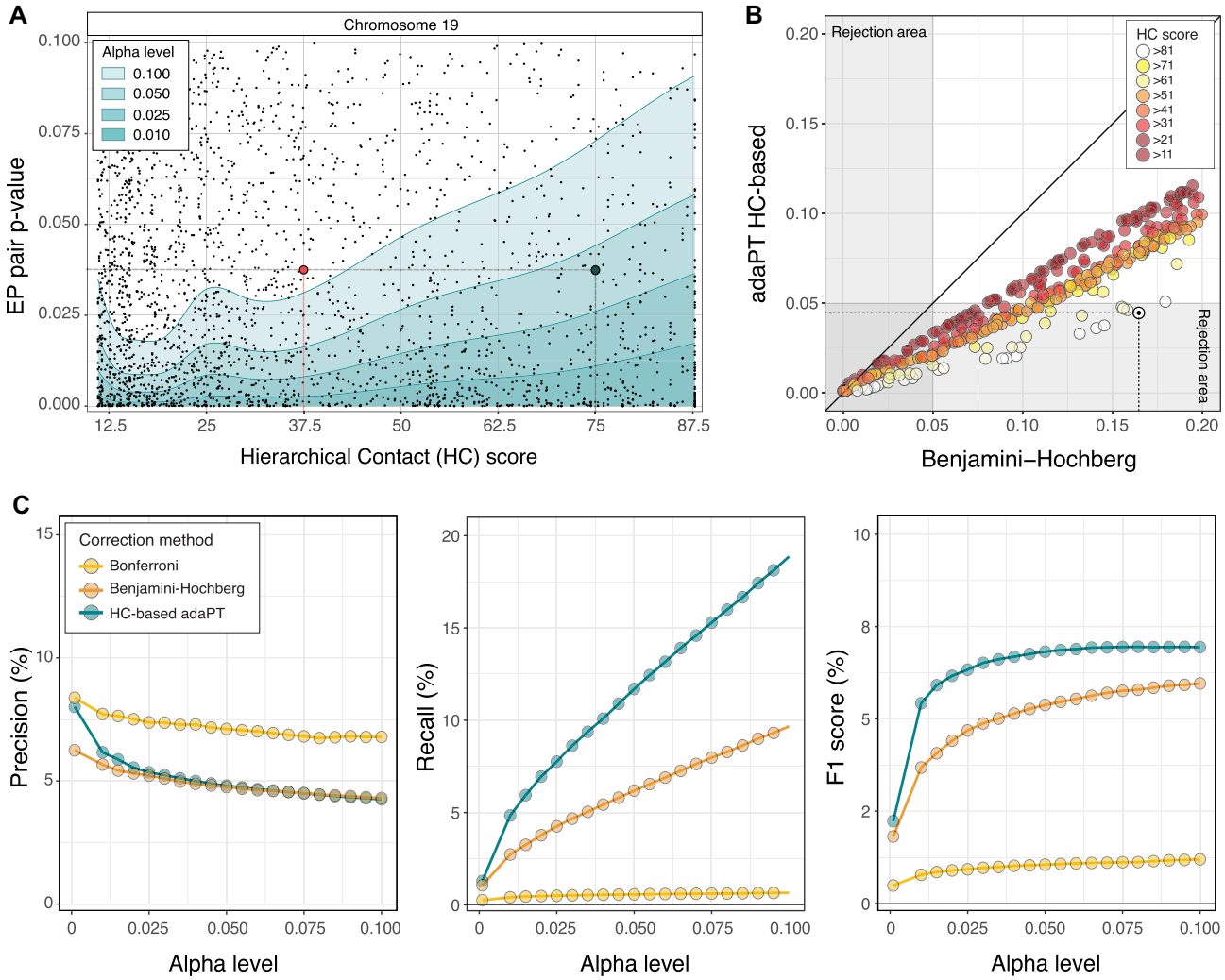

Figure 4. Physical proximity increases power of detection. (**A**) Estimated iterative adaptive rejection threshold (AdaPT) leveraging side information derived from HC scores (x-axis), for a representative subset of chromosome 19 at different alpha levels (green-shadow areas). Each point reports the CCA *P*-value for the synchronised activity between an EP pair (y-axis). Highlighted, two EP pairs with the same nominal *P*-values, but different HC scores, for which the null hypothesis of independence is rejected at a confidence level of 0.95 (green point, high score), and not rejected (red point, low score). (**B**) *P*-values associated to a representative subset of EP pairs for chromosome 19 adjusted with Benjamini-Hochberg (x-axis) and AdaPT (y-axis) approaches, coloured by HC score thresholds. Highlighted with a solid black dot, an example EP pair located within the same TADs for all cell and tissue types and hierarchies (*i.e.*, HC score>81), for which the null hypothesis of independence is rejected at a confidence level of 0.95 only by HC-based AdaPT approach. (**C**) Precision, recall and F1 score of predicted EP pairs based on eQTLs-supported ETG pairs (GTEX and PanCanQTL) over different alpha levels (x-axis), adopting three different multiple-testing correction approaches: Bonferroni (yellow), Benjamini-Hochberg (orange) and HC-based AdaPT (green). These same precision and recall curves are reported in Supplementary Figure S3D and S3E along with the curves obtained with alternative versions of HC score.

Benchmarking against other ETG pairing methods

To benchmark our ETG pairing framework against other methods, as described more in details in the Materials and Methods section, we selected eight algorithms (Supplementary Table S1H): FOCS (79), PreSTIGE (82), RIPPLE (80), PETmodule (28), TargetFinder (29), JEME (27), DeepTACT (81) and ABC (75). Overall, these approaches represent the evolution of ETG predictors proposed between 2014 and 2019, covering all the categories as defined by Hariprakash and Ferrari (94) (i.e. regression/correlation, supervised learning and distance/score-based methods).

It is worth noting that previous publications used different definitions for the reference set of enhancers and promoters, thus introducing heterogeneity in the input candidate pairs which we considered in our analyses.

The ETG pairs called by the selected tools showed prominent differences in terms of EP distance distributions (Supplementary Figure S4A). For this reason, in addition to the eQTLs datasets we also considered as true positives a set of nine capture Hi-C (cHi-C) datasets (68–74) (Supplementary Table S1E), designed to identify contacts between promoters and distal chromatin regions. Taken together, the two types of data provide a broader coverage of functional and physical interactions occurring at different distance ranges. Namely, eQTLs and cHi-C data are representative of mid-range (average distance of 82 kb) and long-range (average

distance of 326 kb) interactions, respectively (Supplementary Figure S4B).

Figure 5 and Supplementary Table S4B summarize the performance of algorithms grouped into two categories: i.e. methods that rely on (i) a unique set of promoters and enhancers as input (our framework and FOCS) or (ii) cell-type specific definitions of enhancers and promoters (JEME, RIPPLE, PETmodule, TargetFinder, DeepTACT, PreSTIGE and ABC). In the first group performances are assessed directly on the unique list of ETG pairs, whereas in the second group performances are reported as average across the cell-type specific lists.

Considering mid-range interactions (based on eQTLs, Figure 5A), our method ranked above most of the other algorithms with a precision ranging from 4.8% to 9.4%, depending on the HC score cut-off. Only ABC, JEME and PreSTIGE showed remarkable performances with 10.1%, 13.1% and 16.4% precision, respectively. However, our method exhibited recall values ranging from 12% to 27%, depending on the HC score cut-off, which were comparable to JEME (21.9%). Conversely, ABC obtained the poorest recall performance (4.4%). The recall cannot be computed for PreSTIGE and other methods not providing the starting set of candidate ETG pairs.

Instead, focusing on FOCS, which among the selected algorithms is the only other one with a unique EP list, our approach showed better performances for HC scores greater than 71, while a decline was observed for the remaining cut-offs. This pattern is mainly due to the use of a candidate EP pairs which is about 16 times larger than FOCS (3 099 004 versus 192,800 pairs) and the exploration of interactions over longer distances, up to 8 times more distant (average distance: 334 kb versus 42 kb). These two peculiarities result in a large imbalance in the initial proportion of true pairs, making their detection more challenging. Indeed, considering an index that is not affected by this bias (i.e. the relative improvement, RI), we estimated that the observed to expected ratio of true pairs (Supplementary Table S4B) in FOCS was equal to the random choice over the initial candidate pairs (0.99), contrary to our algorithm (from 1.15 to 1.59).

Instead, when considering long-range interactions (based on cHi-C, Figure 5B), we observed that precision for DeepTACT (75%) and TargetFinder (60%) were clearly above average, whereas the other methods had a comparable performance (with values around 13%), except for the slightly better PETmodule (17.9%) and ABC (18.1%). Although the precision for our approach was slightly lower than the other tools (ranging from 9.2% to 12.1% with different HC thresholds), the recall proved to be good (ranging from 10% to 24.5%). Moreover, it should be stressed that DeepTACT was trained on a large portion of long-interactions used for our validation (i.e. Javier *et al.* (69) dataset), which may affect the high precision measured in our benchmarking.

Interestingly, the worst-performing algorithms in the mid-range interactions (RIPPLE, PETmodule, TargetFinder and DeepTACT) include the best-performing ones in the long-range interactions. These tools are all based on supervised classifiers trained on physical interaction datasets (e.g. 3C, 5C or Hi-C experiments, cHi-C and ChIA-PET). The only exception over these classifiers is JEME that employed both eQTLs and physical interactions in the training process, reaching good precision and recall performances in both conditions, together with our approach.

Cell specificity of the predicted enhancer–promoter pairs

To further appraise the performance of our method, and investigate its behaviour in identifying cell-type specific ETG pairs, we performed an additional direct comparison with JEME. This choice was motivated by the consideration that JEME is the most comprehensive in terms of cell-type specific ETG lists (127 cell types from the Roadmap Epigenomics dataset, Supplementary Table S1H) and overall resulted as the best-performing among the eight algorithms selected for our benchmarking.

We used JEME initial set of enhancers, genes and candidate ETG pairs (see Materials and Methods section). This choice allows us to minimise the sources of heterogeneity in the comparison, and to test the flexibility of our algorithm with inputs other than those used as reference in our work. In particular, we considered all 127 cell-type specific lists of candidate ETG pairs provided by JEME authors.

It must be noted that JEME does not return a formal *P*-value, thus it is not possible to directly compare results based on a common threshold on statistical significance. Thus, we computed precision using increasing cut-offs on the top-ranked ETG pairs (1000, 3000 and 5000). As reported in Figure 6A and Supplementary Table S4C, the median precision over the 127 cell and tissue types for the mid-range interactions is higher in JEME (19.6–18% range across cutoffs) as opposed to our method (15.5–11.8%). However, in the long-range interactions we obtained comparable performances for JEME (9.8–10.4% range) and our method (8.6–10.6%). It is worth remarking that our choice of using JEME definitions of enhancers, genes and candidate ETG pairs, might put JEME in an advantageous position and render the comparison of the two approaches somewhat biased.

Surprisingly there is a limited overlap between the top ranked ETG pairs identified by JEME and our approach. Among the top 1,000 pairs on average only 0.5% and 0.2% of the mid- and long-range true pairs are identified by both approaches, respectively.

We noted that the majority of pairs predicted by JEME are in the distance range 5–50 kb (Figure 6B). On the contrary, our method exhibits a wider coverage over all the linear distances, with the frequency distribution across distances resembling the distribution of validated ETG pairs (Figure 6B and Supplementary Figure S5A).

When considering the target genes in the top ranked ETG, we noted that on average only 53% are common to both methods. To better characterise the differences in the sets of target genes in the top ranked ETG pairs, we used the matched gene expression data available for 57 out the 127 considered cell types (Figure 6C). Although JEME identified a slightly higher percentage of targets that are strictly cell-type specific (15% versus 8.8%, respectively), more than one third are generally low expressed genes (38.1% versus 24.8%). On the contrary, our framework leads to the identification of ubiquitously expressed targets (44.4% ver-

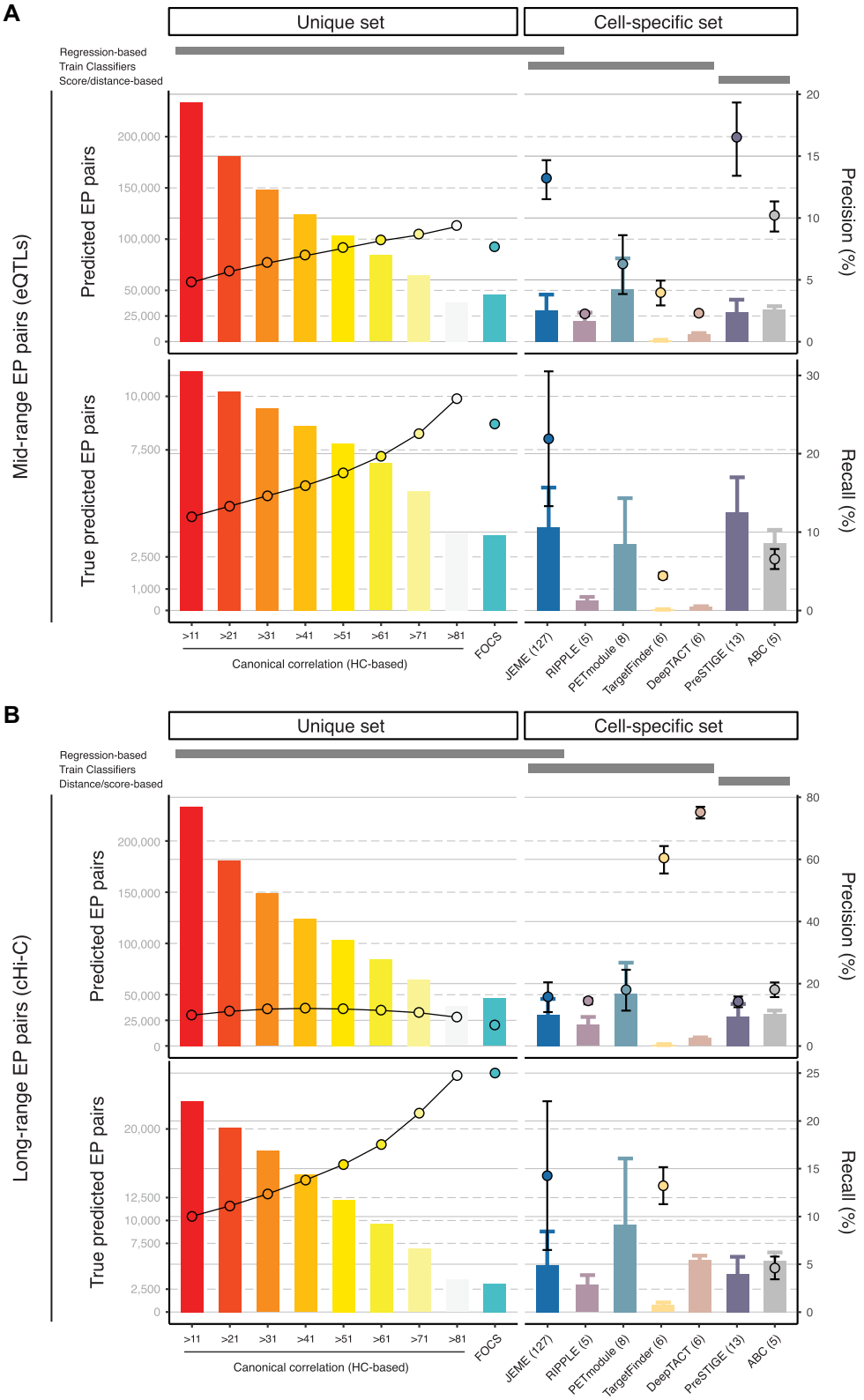

Figure 5. Benchmarking against other ETG pairing methods. Performances of our approach (evaluated at different HC score cut-offs) and other eight ETG pairing algorithms assessed based on mid-range (**A**, eQTLs supported) and long-range (**B**, cHi-C supported) true positive EP interactions. Bars (left y-axis) report the number of predicted (upper panel), and true predicted EP pairs (bottom panel). Points (right y-axis) report precision (upper panel) and recall (bottom panel). Recall is available only for tools for which the list of EP candidate pairs was released. Algorithms are grouped in two categories: methods that rely on a unique set of promoters and enhancers as input (left panels) or rely on cell-type specific definitions of these sets (right panels). For methods in this last category, the ±1 standard deviation (whiskers) and the number of evaluated cell and tissue types (numbers in brackets) are reported. An additional schematic annotation is reported on the top margin of each panel, to highlight the main categories of ETG pairing methods (i.e., regression/correlation, supervised learning and distance/score-based methods).

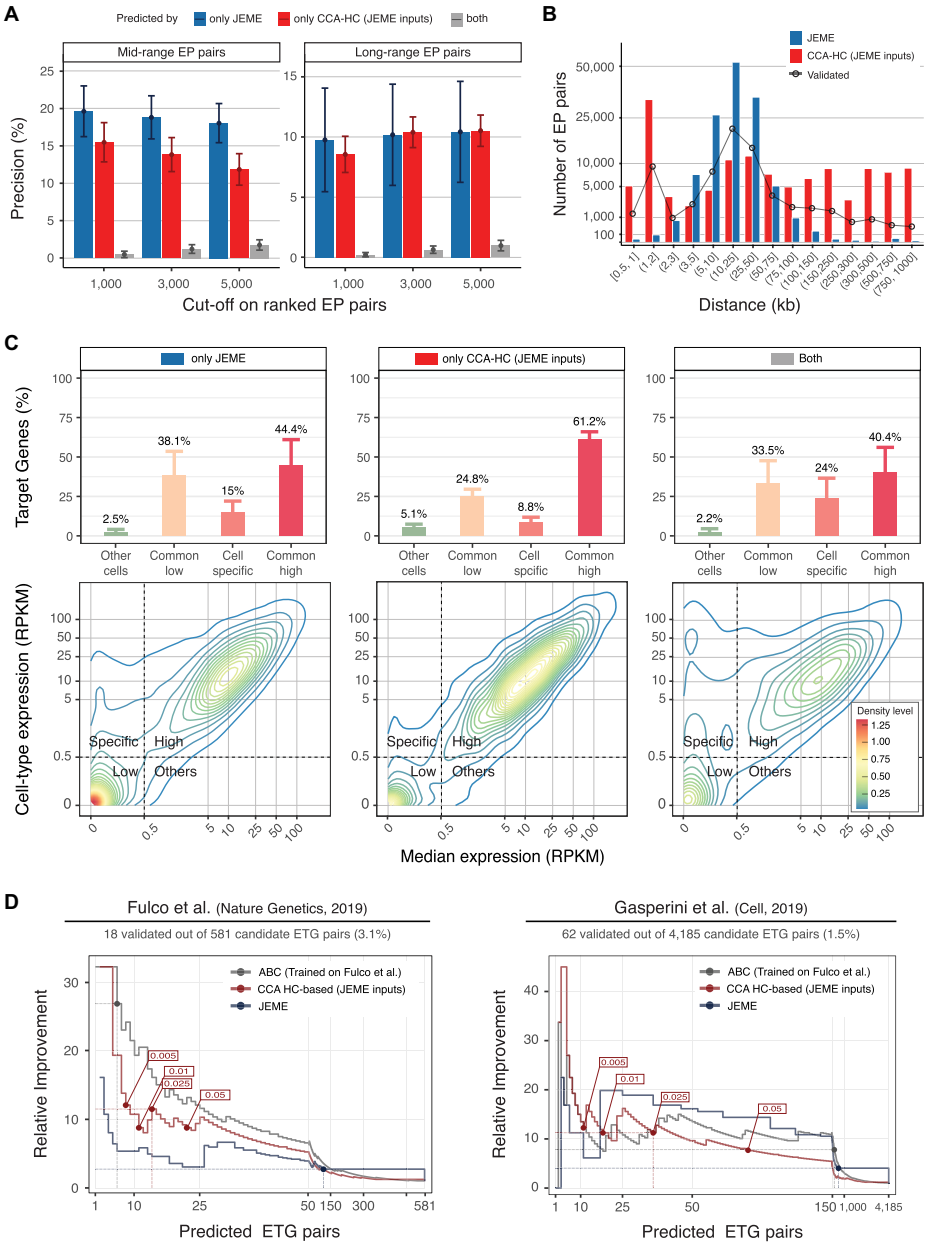

Figure 6. Cell specificity of the predicted enhancer–promoter pairs. (**A**) Average precision (bars) and ±1 standard deviation (whiskers) assessed based on mid-range (eQTLs supported, left panel) and long-range (cHi-C supported, right panel) true positive interactions for ranked predicted ETG pairs by only JEME (blue bars), only our framework (red bars) and both tools (grey bars) in 127 cell types coming from Roadmap Epigenomics datasets for different cuts-off of the ranked lists of predictions (x-axis). Initial sets of enhancers, genes and candidate ETG pairs are the ones described in the original publication of JEME. (**B**) Number (y-axis, logarithmic scale) of predicted ETG pairs by JEME (blue bars) and our method (red bars) considering the top 1000 for each of the 127 cell types coming from Roadmap Epigenomics datasets, grouped by enhancer–target gene distance classes (x-axis). The distance distribution of mid- and long-range true pairs predicted by at least one of the two methods, is reported with black line and dots. (**C**) Gene expression density (contour plots, bottom panels) and percentages (bars, upper panels) with ±1 standard deviation (whiskers) of target genes on top 1000 predicted interactions by only JEME (left panels), only our framework (middle panels) and both tools (right panels) in matched 57 out of 127 cell types considered in the original publication of JEME. Contour plots are calculated merging the set of predicted target genes (y-axis: expression of the target gene in the cell type considered; x-axis: median expression in all the cell types, both axes in logarithmic scale) for each of the 57 cell types. For each cell type, a target gene is classified based on its expression in the specific cell type versus the median expression profiles in all the cell types as: commonly low (salmon, common low) or highly expressed (dark pink, common high), expressed only in the cell type considered (light pink, cell specific) or expressed only in a small subgroup of other cell types (light green, other cells). The threshold used for the classification is highlighted with dotted grey lines in the contour plots. (**D**) Enrichment of validated ETG pairs with respect to the random choice (y-axis, Relative Improvement) over an increasing number of predicted interactions (x-axis), for our method (red lines), JEME (blue lines) and ABC algorithm (grey lines), in two datasets (left and right panels) of CRISPR-based enhancer perturbation experiments on K562 cell line. The cut-offs suggested within the original articles are reported as coloured points, and the associated performances are highlighted with coloured dotted lines. The initial sets of enhancers, genes and candidate ETG pairs (further filtered for compatibility with CRISPR-based datasets) are the ones described in the original publication of JEME, for our method and JEME. The predicted and filtered ETG pairs by ABC method, using Fulco dataset as training set, are used. The plot is reporting an expanded x-axis in the initial part of the curve (up to 50 pairs left panel, and up to 150 pairs in the right panel) to provide a more detailed visualization of the most informative part of the chart.

sus 61.2%). It may be worth remarking that this analysis is based on cell type specific enhancer lists, thus genes expressed in multiple cell types can be under the control of different enhancers in distinct cells.

We further validated the results in a cell-type specific context by using data from CRISPR-based enhancer perturbation experiments by two recent studies, hereinafter referred as Fulco (75) and Gasperini (76) datasets, which provide a more direct functional validation of interactions. We focused on the data for K562 cell line, which is characterized in both datasets, as well as in JEME, and we filtered the data to ensure comparability (see Materials and Methods). We also included the ABC method in this comparison, i.e. the algorithm proposed in the Fulco dataset article.

As the CRISPR-based benchmark datasets are not as comprehensive as the eQTL and cHi-C datasets, for a more robust comparisons among methods we focused on the Relative Improvement (RI) metric. RI evaluates the enrichment of validated ETG pairs over an increasing number of predicted interactions with respect to the random choice (Figure 6D).

In the Fulco dataset (Figure 6D, Supplementary Figure S5B left panels and Supplementary Table S4D) our method performed better than JEME for any threshold on its list of predicted pairs (blue dot). In the same settings our approach is slightly worse than the ABC model, which is, however, trained on this very same set of validated ETG interactions. In the Gasperini dataset (Figure 6D, Supplementary Figure S5B right panels and Supplementary Table S4E), our approach shows an RI better than or equal to ABC for the top ranked pairs up to FDR <0.025 and better than JEME for FDR <0.01. Notably, both ABC and JEME return a score which is not formally bound to FDR, thus compromising their applicability on independent datasets. Lastly, investigating the activity of predicted targets within the top ranked ETG pairs (Supplementary Figure S5C), our method showed the higher number of cell-specific genes, comprising almost all the ones predicted by the other tools (i.e. HBE1, RHAG, GATA1, ALAS2 and COL6A5, ordered by expression level).

Overall, these results confirmed the reliability and versatility of our proposal in detecting relevant ETG pairs along the whole search space with results that can be generalized to other cell types. Even if we may identify a greater portion of ubiquitously expressed genes compared to algorithms trained on a specific cell line, our method is also good at identifying target genes with cell-type specific expression (Figure 6C, right panel, and Supplementary Figure S5C).

Benchmarking against independent reference dataset

A recent publication proposed a curated benchmarking dataset for ETG pairing: the Benchmark of candidate Enhancer–Gene Interactions (BENGI) (77). This is not a method to identify ETG pairs, but rather a database of interactions, that can be used as independent reference to assess the performance of current and future algorithms. The BENGI database contains a collection of uniformly processed datasets that integrate the Registry of candidate *cis*-regulatory elements (cCREs) (83) with experimentally derived genomic interactions (Supplementary Table S1G). Thus, we used their enhancers, genes and candidate ETG pairs as input for our framework.

To appreciate the distinct features captured by the experimental datasets used to curate BENGI interactions, we computed precision-recall (PR) curves for the GM12878 cell line (a lymphoblastoid cell line), which is the most extensively surveyed one in BENGI (Figure 7A and Supplementary Figure S6A). The PR curves include: AdaPT corrected *P*-values (sorted in ascending order); canonical correlation (decreasing order); HC score (decreasing order); EP distance (increasing order). We noticed two scenarios: (i) in top-ranked pairs supported by ChIA-PET or 3C-derived methods, which include the majority of the BENGI validated interactions (1 706 837 pairs, 95%), the distance does not provide any insight in the identification of true EP pairs, resulting in close to random selection or worse; (ii) instead in eQTL datasets (87 982 pairs, 5%), the distance classifier remains consistently above the performance of our method, although CCA is initially aligned. It is also worth mentioning that a true positive EP pair in one BENGI experimental dataset list could be a true negative in another BENGI list.

To further clarify this observation, we considered the twenty individual experimental datasets used by BENGI and we computed the relative improvement (RI) achieved by our method or EP distance alone (Figure 7B and Supplementary Table S4F). As shown, the pattern described above is consistent across all validation datasets. Indeed, our method obtains robust and significantly enriched performances (i.e., higher than random classification) both for eQTLs (from 2.15 to 3.65) and ChIA-PET or 3C-derived interactions (from 1.36 to 3.14). Instead, the distance alone yields performances lower than what expected by chance in seven out of eight datasets of 3C-derived interactions.

DISCUSSION

Here, we present a new approach to refine the pairing of enhancers and target gene promoters. The three principles underlying our approach are: (i) the flexibility, with respect to all input data; (ii) the use of prior-knowledge, as we leverage 3D chromatin architecture to inform EP pairing; (iii) the robustness of the statistical framework, as the method does not require arbitrary parameter tuning decisions and guarantees statistical control of the false discovery rate.

The flexibility of the method allows the end users to provide any preferred definition for the reference set of enhancers, genes, and functional genomics data used to quantify their activity. This versatility is primarily guaranteed by the use of CCA which provides the foundation for a very general statistical framework. Indeed, other correlation tests, as well as several parametric tests (including ANOVA, linear or multivariate regression, discriminant analysis, and chi-squared test), can be described as special cases of CCA, as demonstrated in literature (95).

Then the prior-knowledge on chromatin 3D organization has been carefully considered in defining the HC score to quantify EP pairs physical proximity. We took into account the general consensus in literature that ETG pairs primarily occur within TAD domains, but they should not be considered as a hard constraint as EP interactions may also span TAD boundaries (69,96). We also considered that alter-

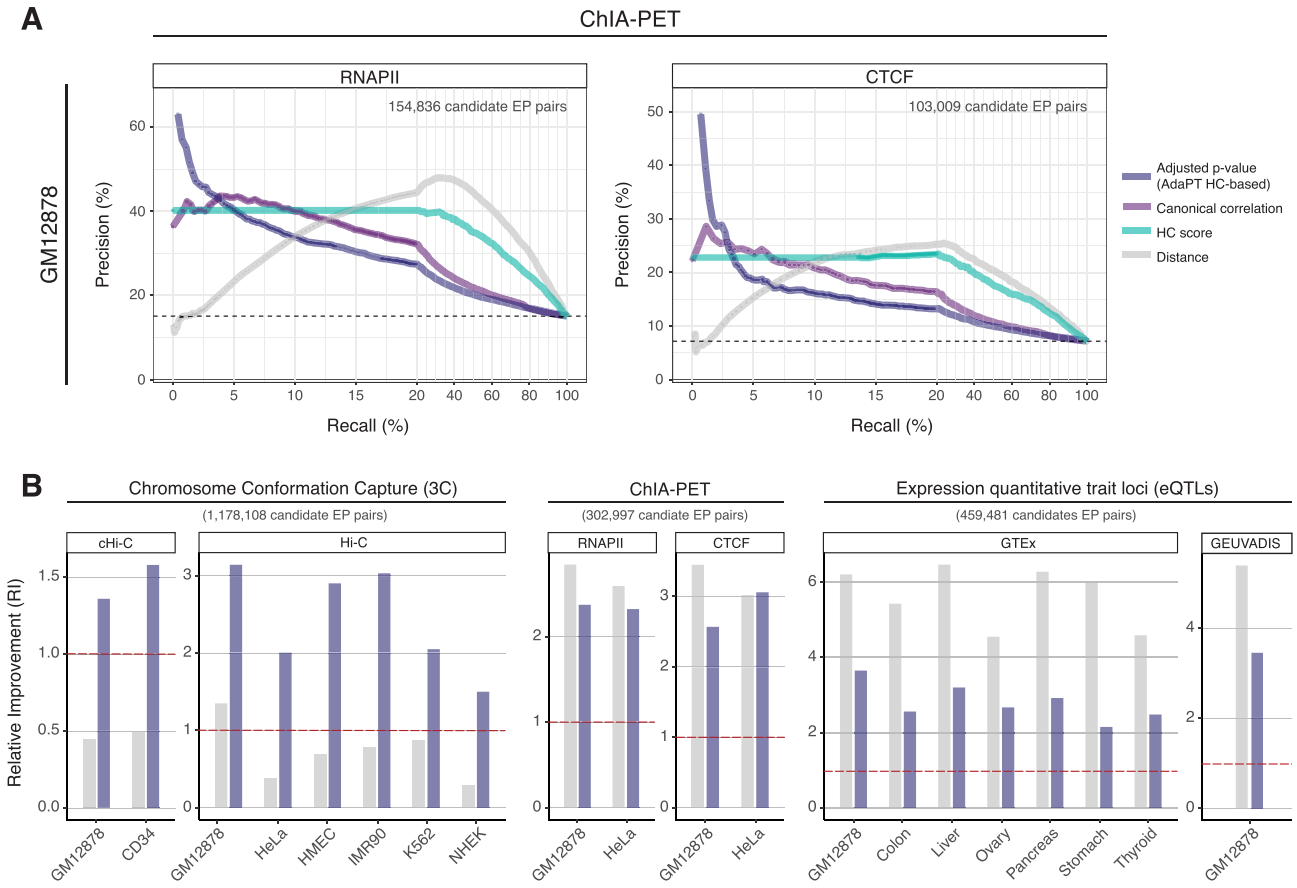

Figure 7. Benchmark against independent reference datasets. (**A**) Precision-recall curves for GM12878 cell line BENGI benchmark datasets, assessed employing two datasets of ChIA-PET supported true ETG positive interactions (RNAPII, left panel; CTCF, right panel). Performances are calculated for: AdaPT HC-based adjusted *P*-values (dark purple lines), canonical correlation (light purple lines), HC score (aquamarine lines) and linear EP distance (grey lines). The total number of ETG pair considered is reported on the upper right corner of each panel. The plot is reporting an expanded x-axis in the initial part of the curve (corresponding to recalls up to 20%) to provide a more detailed visualisation for the most informative part of the chart. For higher recall values, the precision-recall curves reported here tend to converge without further crossing each other. (**B**) Relative improvement (y-axis) for all twenty BENGI benchmark interactions, assessed employing different sources of true ETG positive interactions. Namely, chromosome conformation capture (left panels), ChIA-PET (middle panels) and expression quantitative trait loci (right panels). Random choice (RI = 1) is marked by red dashed lines. Performances are calculated for AdaPT HC-based adjusted *P*-values (dark purple bars) and linear EP distance (grey bars), based on the same cuts-off on FDR = 0.01 for adjusted *P*-values.

native TAD definitions at multiple scales are concurrently present in the cell population. Indeed, TADs should be intended as a probabilistic structure dynamically defined by loop extrusion mechanisms (41–43). This biology-derived knowledge has been directly used to compute the HC score integrated as side information in the adjustment of CCA *P*-values for each EP pair.

This may seem a counter-intuitive solution as opposed to directly using the EP loci contact frequency from the Hi-C matrix. However, using the multiscale TAD structure instead is a fundamental change of perspective that allows overcoming relevant technical limitations. First of all, Hi-C contact matrices are generally binned at few kb resolution, thus at a scale that does not allow distinguishing regulatory regions close to each other. Indeed, even the most recent ETG pairing attempted with this strategy could not go beyond 5 kb resolution (75). Moreover, Hi-C point interaction calling algorithms have been shown to yield very variable results even across biological replicates (19). Finally, Hi-C point interactions are generally considered to be very dynamic across cell types, thus they would always require a precise match between the cell and tissue type used for Hi-C and other genomics data required for the ETG pairing. Instead, TADs are expected to be more conserved across cell types, thus allowing to extend the applicability of the HC score to different cell contexts.

It is also worth remarking that our method is flexible for what concerns the hierarchy of TAD structural domains provided as input. As such, the end user may adopt the preferred algorithm for calling TADs at multiple scales.

Additionally, the robustness of the method is safeguarded by the use of AdaPT multiple testing correction which is combining the CCA *P*-value and the HC score for each EP pair. This solution increases statistical power by prioritizing most promising hypotheses based on side information. As a representative example attesting the importance of such strategy we noted the HBB-LCR region, which is a known complex distal regulatory region, responsible for the coordinated regulation during development of the human beta globin genes (HBE1, HBG2, HBG1, HBD and HBB). The

LCR region contains several enhancers that are paired to one or multiple beta globin genes in our list of EP pairs (see Data Availability). For several of them, the BH adjusted *P*-value is not significant, whereas the AdaPT adjustment is able to detect a significant association to the beta globin genes. In particular, the enhancer at chr11:5297767-5298471 has a significant AdaPT *P*-value for all of the five beta globin gene promoters, whereas the same enhancer would be significantly associated only to HBG1 and HBG2 based on BH correction.

We also note that our strategy of using HC as side information to control FDR is in principle a generalizable approach, that could also be applied to *P*-values for ETG pairs coming from other methods. Unfortunately, all the methods that we surveyed (78) did not really address the multiple testing correction problem, as most of them are either based on a classifier, or some other custom score, thus not returning an actual *P*-value for individual EP pairs.

Thus, we reconstructed the map of ETG regulatory interactions by applying our framework using genome-wide profiles of epigenetic marks for 44 cell and tissue types, together with multi-scale TAD calls derived from 11 high-coverage Hi-C datasets. To this concern, it may be worth remarking that we quantified the gene activity using the epigenetic marks at promoters, as opposed to the expression level of the gene transcripts. This is a commonly adopted choice in the literature of this field. The rationale behind this solution is based on the role of enhancers in triggering transcription: hence enhancers will show a synchronised activity with markers of transcription initiation in their targets. Instead, the actual transcripts abundance will depend also on multiple levels of co-transcriptional (e.g. RNA polymerase pausing, processivity in elongation, splicing and poly-adenylation) and post-transcriptional regulation (e.g. mRNA stability). All of these mechanisms will confound the synchronisation between enhancers activity and transcriptional output.

We identified a total of 233 304 EP pairs with adjusted *P*-value $\leq$ 0.05 and extensively benchmarked our results against eight pre-existing algorithms representing multiple categories of ETG pairing methods. We used multiple sources of true positive ETG pairs, including eQTLs, cHi-C, CRISPR-based perturbations and a recently published curated database (BENGI), which is relying as well on multiple sources of experimental data.

We observed consistent performances in both mid- and long-range interactions for our method, as opposed to previously published algorithms that generally perform better on one of the two distance ranges (Figure 5). Moreover, we showed that our method compares well also to algorithms capturing cell-type specific ETG pairs (Figure 6), even though we aim to provide a generalizable ETG map that can be extended to multiple cell types.

It is worth noting that some of the previous algorithms based on supervised methods were actually trained using eQTLs or cHi-C data as true positive sets, that we have used for the benchmarking as well. Thus, other methods may have an advantage in the benchmarking statistics presented here.

We also must note that the definition of true positive ETG pairs may suffer some limitations. Namely, cHi-C and similar techniques can confirm a physical proximity between specific genomic regions, but this is not always resulting in a functional regulatory interaction between them. Likewise, eQTLs confirm a correlation between a gene expression and a genetic variants (SNP), but the actual distal regulatory region may be in a different position within the SNP linkage disequilibrium block. As such, it may be argued that both data types provide only an indirect validation of interaction.

Considering these limitations, which are anyway affecting also the previous benchmarks of ETG pairing algorithms, we further dissected our method performances using CRISPR-based perturbation datasets. CRISPR-based strategies have been proposed to identify functional connections in ETG pairs, in particular in combination with single cell transcriptomics readout (97,98). Using two recently published datasets we confirmed the reliability and versatility of our method in detecting relevant ETG pairs, with good performances even if compared to algorithms specifically trained on these settings (Figure 6D and Supplementary Figure S5B, C).

However, CRISPR-based perturbation datasets cannot be considered as a comprehensive and generalizable benchmark, because even the latest and largest datasets are limited to a few enhancer and cell types. Therefore, we also assessed our method against the BENGI database, containing an independent curated reference benchmark for ETG pairs. Even in this case, our method confirmed consistent performances across all types of ETG supporting data, covering both mid- and long-range interactions (Figure 7B).

As discussed in detail by (99), performing a quantitative comparison of ETG pairing methods is a challenging task, where critical points should be considered such as (i) properly separating training and validation sets; (ii) considering the distance as a relevant feature affecting ETG pairing; (iii) paying attention to different definitions of enhancer and promoter windows adopted by distinct algorithms. Throughout our work we carefully took into account these critical points as discussed in details for each individual analysis. The resulting framework proved to yield coherent results across different test datasets and cell types, thus confirming its value as a generalizable approach for ETG regulatory interactions reconstruction. In order to facilitate reproducibility of results, and widespread adoption in the community, we are publicly releasing the code and input datasets used for this study (see Data Availability). This tool will provide a valuable resource especially for translational studies aiming to annotate the functional role of non-coding sequence variants in distal regulatory elements. In particular, we envision possible applications in clinical genomics studies of cancer and undiagnosed genetic diseases.

DATA AVAILABILITY

All public datasets used in this manuscript are described in Supplementary Table S1. The source code and the complete list of candidate enhancer–promoter pairs annotated with the HC score, corrected and uncorrected *P*-values, and validations according to multiple reference datasets are available at https://github.com/ElisaSalviato/3D-ETG. A web based-user friendly portal to browse and query our results

is also available at the URL https://bioinformatics.ifom.eu/3D-ETG.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Chiara Romualdi for advice on statistical analysis and feedbacks. We thank Raoul J.P. Bonnal and the IFOM IT team for help in setting up the web portal. We thank Marco Morelli and Gioacchino Natoli for critical feedback on the manuscript. We thank Vincenzo Corbo and Pietro Delfino for stimulating discussions. We thank Cristiano Petrini for precious help with pipelines. We are grateful to Orso Maria Romano for support and constructive inputs on models and graphs.

FUNDING

AIRC 'Sergio Bernardini' fellowship [2235 to E.S.]; AIRC 2015 Start-up grant [16841 to F.F.]; AIRC fellowship [21012 to K.P., 22416 to J.M.H.]. Funding for open access charge: AIRC 2015 Start-up grant [16841].
Conflict of interest statement. None declared.

REFERENCES

1. Roadmap,Epigenomics Consortium, Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–329.
2. Nott,A., Holtman,I.R., Coufal,N.G., Schlachetzki,J.C.M., Yu,M., Hu,R., Han,C.Z., Pena,M., Xiao,J., Wu,Y. *et al.* (2019) Brain cell type–specific enhancer–promoter interactome maps and disease-risk association. *Science*, **366**, 1134–1139.
3. De Laat,W. and Duboule,D. (2013) Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature*, **502**, 499–506.
4. Schoenfelder,S. and Fraser,P. (2019) Long-range enhancer–promoter contacts in gene expression control. *Nat. Rev. Genet.*, **20**, 437–455.
5. Gallagher,M.D. and Chen-Plotkin,A.S. (2018) The post-GWAS era: from association to function. *Am. J. Hum. Genet.*, **102**, 717–730.
6. Maurano,M.T., Humbert,R., Rynes,E., Thurman,R.E., Haugen,E., Wang,H., Reynolds,A.P., Sandstrom,R., Qu,H., Brody,J. *et al.* (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science*, **337**, 1190–1195.
7. Smith,E. and Shilatifard,A. (2014) Enhancer biology and enhanceropathies. *Nat. Struct. Mol. Biol.*, **21**, 210–219.
8. Sur,I. and Taipale,J. (2016) The role of enhancers in cancer. *Nat. Rev. Cancer*, **16**, 483–493.
9. Visel,A., Rubin,E.M. and Pennacchio,L.A. (2009) Genomic views of distant-acting enhancers. *Nature*, **461**, 199–205.
10. Dekker,J., Rippe,K., Dekker,M. and Kleckner,N. (2002) Capturing chromosome conformation. *Science*, **295**, 1306–1311.
11. de Laat,W. and Dekker,J. (2012) 3C-based technologies to study the shape of the genome. *Methods*, **58**, 189–191.
12. Kempfer,R. and Pombo,A. (2020) Methods for mapping 3D chromosome architecture. *Nat. Rev. Genet.*, **21**, 207–226.
13. Schmitt,A.D., Hu,M. and Ren,B. (2016) Genome-wide mapping and analysis of chromosome architecture. *Nat. Rev. Mol. Cell Biol.*, **17**, 743–755.
14. Lieberman-Aiden,E., Van Berkum,N.L., 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.
15. Pal,K., Forcato,M. and Ferrari,F. (2019) Hi-C analysis: from data generation to integration. *Biophys. Rev.*, **11**, 67–78.
16. Bonev,B., Mendelson Cohen,N., Szabo,Q., Fritsch,L., Papadopoulos,G.L., Lubling,Y., Xu,X., Lv,X., Hugnot,J.P., Tanay,A. *et al.* (2017) Multiscale 3D genome rewiring during mouse neural development. *Cell*, **171**, 557–572.
17. 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.
18. Zhang,J., Lee,D., Dhiman,V., Jiang,P., Xu,J., McGillivray,P., Yang,H., Liu,J., Meyerson,W., Clarke,D. *et al.* (2020) An integrative ENCODE resource for cancer genomics. *Nat. Commun.*, **11**, 3696.
19. Forcato,M., Nicoletti,C., Pal,K., Livi,C.M., Ferrari,F. and Bicciato,S. (2017) Comparison of computational methods for Hi-C data analysis. *Nat. Methods*, **14**, 679–685.
20. Hughes,J.R., Roberts,N., Mcgowan,S., Hay,D., Giannoulatou,E., Lynch,M., De Gobbi,M., Taylor,S., Gibbons,R. and Higgs,D.R. (2014) Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat. Genet.*, **46**, 205–212.
21. Jäger,R., Migliorini,G., Henrion,M., Kandaswamy,R., Speedy,H.E., Heindl,A., Whiffin,N., Carnicer,M.J., Broome,L., Dryden,N. *et al.* (2015) Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nat. Commun.*, **6**, 6178.
22. 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**, 598–606.
23. Sahlén,P., Abdullayev,I., Ramsköld,D., Matskova,L., Rilakovic,N., Lötstedt,B., Albert,T.J., Lundeberg,J. and Sandberg,R. (2015) Genome-wide mapping of promoter-anchored interactions with close to single-enhancer resolution. *Genome Biol.*, **16**, 156.
24. 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**, 919–922.
25. Fullwood,M.J., Liu,M.H., Pan,Y.F., Liu,J., Xu,H., Mohamed,Y. Bin, Orlov,Y.L., Velkov,S., Ho,A., Mei,P.H. *et al.* (2009) An oestrogen-receptor-α-bound human chromatin interactome. *Nature*, **462**, 58–64.
26. He,B., Chen,C., Teng,L. and Tan,K. (2014) Global view of enhancer−promoter interactome in human cells. *Proc. Natl. Acad. Sci. USA*, **111**, E2191–E2199.
27. 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.*, **49**, 1428–1436.
28. Zhao,C., Li,X. and Hu,H. (2016) PETModule: A motif module based approach for enhancer target gene prediction. *Sci. Rep.*, **6**, 30043.
29. 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**, 488–496.
30. Okonechnikov,K., Erkek,S., Korbel,J.O., Pfister,S.M. and Chavez,L. (2019) InTAD: chromosome conformation guided analysis of enhancer target genes. *BMC Bioinformatics*, **20**, 60.
31. Lin,C.Y., Erkek,S., Tong,Y., Yin,L., Federation,A.J., Zapatka,M., Haldipur,P., Kawauchi,D., Risch,T., Warnatz,H.J. *et al.* (2016) Active medulloblastoma enhancers reveal subgroup-specific cellular origins. *Nature*, **530**, 57–62.
32. Johann,P.D., Erkek,S., Zapatka,M., Kerl,K., Buchhalter,I., Hovestadt,V., Jones,D.T.W., Sturm,D., Hermann,C., Segura Wang,M. *et al.* (2016) Atypical Teratoid/Rhabdoid tumors are comprised of three epigenetic subgroups with distinct enhancer landscapes. *Cancer Cell*, **29**, 379–393.
33. Mack,S.C., Pajtler,K.W., Chavez,L., Okonechnikov,K., Bertrand,K.C., Wang,X., Erkek,S., Federation,A., Song,A., Lee,C. *et al.* (2018) Therapeutic targeting of ependymoma as informed by oncogenic enhancer profiling. *Nature*, **553**, 101–105.
34. Bonev,B. and Cavalli,G. (2016) Organization and function of the 3D genome. *Nat. Rev. Genet.*, **17**, 661–678.

35. Rowley,M.J. and Corces,V.G. (2018) Organizational principles of 3D genome architecture. *Nat. Rev. Genet.*, **19**, 789–800.
36. 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.
37. Dixon,J.R., Jung,I., Selvaraj,S., Shen,Y., Antosiewicz-Bourget,J.E., Lee,A.Y., Ye,Z., Kim,A., Rajagopal,N., Xie,W. *et al.* (2015) Chromatin architecture reorganization during stem cell differentiation. *Nature*, **518**, 331–336.
38. 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**, 381–385.
39. Weinreb,C. and Raphael,B.J. (2016) Identification of hierarchical chromatin domains. *Bioinformatics*, **32**, 1601–1609.
40. Fraser,J., Ferrai,C., Chiariello,A.M., Schueler,M., Rito,T., Laudanno,G., Barbieri,M., Moore,B.L., Kraemer,D.C.A., Aitken,S. *et al.* (2015) Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Mol. Syst. Biol.*, **11**, 852.
41. Sanborn,A.L., Rao,S.S.P., Huang,S.C., Durand,N.C., Huntley,M.H., Jewett,A.I., Bochkov,I.D., Chinnappan,D., Cutkosky,A., Li,J. *et al.* (2015) Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, E6456–E6465.
42. Nuebler,J., Fudenberg,G., Imakaev,M., Abdennur,N. and Mirny,L.A. (2018) Chromatin organization by an interplay of loop extrusion and compartmental segregation. *Proc. Natl. Acad. Sci. USA*, **115**, E6697–E6706.
43. Fudenberg,G., Imakaev,M., Lu,C., Goloborodko,A., Abdennur,N. and Mirny,L.A. (2016) Formation of chromosomal domains by loop extrusion. *Cell Rep.*, **15**, 2038–2049.
44. 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.
45. Kodzius,R., Kojima,M., Nishiyori,H., Nakamura,M., Fukuda,S., Tagami,M., Sasaki,D., Imamura,K., Kai,C., Harbers,M. *et al.* (2006) Cage: Cap analysis of gene expression. *Nat. Methods*, **3**, 211.
46. Visel,A., Minovitsky,S., Dubchak,I. and Pennacchio,L.A. (2007) VISTA enhancer browser - a database of tissue-specific human enhancers. *Nucleic Acids Res.*, **35**, D88–D92.
47. Bard,J.B., Kaufman,M.H., Dubreuil,C., Brune,R.M., Burger,A., Baldock,R.A. and Davidson,D.R. (1998) An internet-accessible database of mouse developmental anatomy based on a systematic nomenclature. *Mech. Dev.*, **74**, 111–120.
48. Lawlor,N., Márquez,E.J., Orchard,P., Narisu,N., Shamim,M.S., Thibodeau,A., Varshney,A., Kursawe,R., Erdos,M.R., Kanke,M. *et al.* (2019) Multiomic profiling identifies cis-regulatory networks underlying human pancreatic β cell identity and function. *Cell Rep.*, **26**, 788–801.
49. 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.
50. Barutcu,A.R., Lajoie,B.R., McCord,R.P., Tye,C.E., Hong,D., Messier,T.L., Browne,G., van Wijnen,A.J., Lian,J.B., Stein,J.L. *et al.* (2015) Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells. *Genome Biol.*, **16**, 214.
51. Bunting,K.L., Soong,T.D., Singh,R., Jiang,Y., Béguelin,W., Poloway,D.W., Swed,B.L., Hatzi,K., Reisacher,W., Teater,M. *et al.* (2016) Multi-tiered reorganization of the genome during B cell affinity maturation anchored by a germinal center-specific locus control region. *Immunity*, **45**, 497–512.
52. Schmitt,A.D., Hu,M., Jung,I., Xu,Z., Qiu,Y., Tan,C.L., Li,Y., Lin,S., Lin,Y., Barr,C.L. *et al.* (2016) A Compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell Rep.*, **17**, 2042–2059.
53. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
54. Imakaev,M., Fudenberg,G., McCord,R.P., Naumova,N., Goloborodko,A., Lajoie,B.R., Dekker,J. and Mirny,L.A. (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods*, **9**, 999–1003.
55. Abdennur,N. and Mirny,L.A. (2020) Cooler: Scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics*, **36**, 311–316.
56. Pal,K., Forcato,M., Jost,D., Sexton,T., Vaillant,C., Salviato,E., Mazza,E.M.C., Lugli,E., Cavalli,G. and Ferrari,F. (2019) Global chromatin conformation differences in the Drosophila dosage compensated chromosome X. *Nat. Commun.*, **10**, 5355.
57. Pal,K., Tagliaferri,I., Livi,C.M. and Ferrari,F. (2020) HiCBricks: building blocks for efficient handling of large Hi-C datasets. *Bioinformatics*, **36**, 1917–1919.
58. Shin,H., Shi,Y., Dai,C., Tjong,H., Gong,K., Alber,F. and Zhou,X.J. (2015) TopDom: An efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res.*, **44**, e70.
59. Crane,E., Bian,Q., McCord,R.P., Lajoie,B.R., Wheeler,B.S., Ralston,E.J., Uzawa,S., Dekker,J. and Meyer,B.J. (2015) Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature*, **523**, 240–244.
60. Mardia,K., Kent,J. and Bibby,J. (1979) Multivariate analysis. *Acad. Press Inc. London*, **15**, 518.
61. Lawrence,M., Gentleman,R. and Carey,V. (2009) rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics*, **25**, 1841–1842.
62. Rao,C.R. (1951) An asymptotic expansion of the distribution of Wilks' criterion. *Bull. Int. Stat. Inst.*, **33**, 177–180.
63. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.
64. Lei,L. and Fithian,W. (2018) AdaPT: an interactive procedure for multiple testing with side information. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **80**, 649–679.
65. Yurko,R., G'Sell,M., Roeder,K. and Devlin,B. (2020) A selective inference approach for false discovery rate control using multiomics covariates yields insights into disease risk. *Proc. Natl. Acad. Sci. U.S.A.*, **117**, 15028–15035.
66. Lonsdale,J., Thomas,J., Salvatore,M., Phillips,R., Lo,E., Shad,S., Hasz,R., Walters,G., Garcia,F., Young,N. *et al.* (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.
67. Gong,J., Mei,S., Liu,C., Xiang,Y., Ye,Y., Zhang,Z., Feng,J., Liu,R., Diao,L., Guo,A.Y. *et al.* (2018) PancanQTL: systematic identification of cis -eQTLs and trans -eQTLs in 33 cancer types. *Nucleic Acids Res.*, **46**, D971–D976.
68. Montefiori,L.E., Sobreira,D.R., Sakabe,N.J., Aneas,I., Joslin,A.C., Hansen,G.T., Bozek,G., Moskowitz,I.P., McNally,E.M. and Nóbrega,M.A. (2018) A promoter interaction map for cardiovascular disease genetics. *Elife*, **7**, e35788.
69. Javierre,B.M., Sewitz,S., Cairns,J., Wingett,S.W., Várnai,C., Thiecke,M.J., Freire-Pritchett,P., Spivakov,M., Fraser,P., Burren,O.S. *et al.* (2016) Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell*, **167**, 1369–1384.
70. Cairns,J., Freire-Pritchett,P., Wingett,S.W., Várnai,C., Dimond,A., Plagnol,V., Zerbino,D., Schoenfelder,S., Javierre,B.M., Osborne,C. *et al.* (2016) CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. *Genome Biol.*, **17**, 127.
71. Pan,D.Z., Garske,K.M., Alvarez,M., Bhagat,Y. V., Boocock,J., Nikkola,E., Miao,Z., Raulerson,C.K., Cantor,R.M., Civelek,M. *et al.* (2018) Integration of human adipocyte chromosomal interactions with adipose gene expression prioritizes obesity-related genes from GWAS. *Nat. Commun.*, **9**, 1512.
72. Beekman,R., Chapaprieta,V., Russiñol,N., Vilarrasa-Blasi,R., Verdaguer-Dot,N., Martens,J.H.A., Duran-Ferrer,M., Kulis,M., Serra,F., Javierre,B.M. *et al.* (2018) The reference epigenome and regulatory chromatin landscape of chronic lymphocytic leukemia. *Nat. Med.*, **24**, 868–880.
73. Choy,M.K., Javierre,B.M., Williams,S.G., Baross,S.L., Liu,Y., Wingett,S.W., Akbarov,A., Wallace,C., Freire-Pritchett,P., Rugg-Gunn,P.J. *et al.* (2018) Promoter interactome of human embryonic stem cell-derived cardiomyocytes connects GWAS regions to cardiac gene networks. *Nat. Commun.*, **9**, 2526.
74. Miguel-Escalada,I., Bonàs-Guarch,S., Cebola,I., Ponsa-Cobas,J., Mendieta-Esteban,J., Atla,G., Javierre,B.M., Rolando,D.M.Y.,

Farabella,I., Morgan,C.C. *et al.* (2019) Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes. *Nat. Genet.*, **51**, 1137–1148.
75. 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.
76. 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.
77. Moore,J.E., Pratt,H.E., Purcaro,M.J. and Weng,Z. (2020) A curated benchmark of enhancer-gene interactions for evaluating enhancer–target gene prediction methods. *Genome Biol.*, **21**, 17.
78. Xu,H., Zhang,S., Yi,X., Plewczynski,D. and Li,M.J. (2020) Exploring 3D chromatin contacts in gene regulation: The evolution of approaches for the identification of functional enhancer−promoter interaction. *Comput. Struct. Biotechnol. J.*, **18**, 558–570.
79. 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**, 56.
80. 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.
81. Li,W., Wong,W.H. and Jiang,R. (2019) DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning. *Nucleic Acids Res.*, **47**, e60.
82. Corradin,O., Saiakhova,A., Akhtar-Zaidi,B., Myeroff,L., Willis,J., Cowper-Sallari,R., Lupien,M., Markowitz,S. and Scacheri,P.C. (2014) Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res.*, **24**, 1–13.
83. Abascal,F., Acosta,R., Addleman,N.J., Adrian,J., Afzal,V., Aken,B., Akiyama,J.A., Jammal,O. Al, Amrhein,H., Anderson,S.M. *et al.* (2020) Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, **583**, 699–710.
84. Lizio,M., Abugessaisa,I., Noguchi,S., Kondo,A., Hasegawa,A., Hon,C.C., De Hoon,M., Severin,J., Oki,S., Hayashizaki,Y. *et al.* (2019) Update of the FANTOM web resource: expansion to provide additional transcriptome atlases. *Nucleic Acids Res.*, **47**, D752–D758.
85. Lupiáñez,D.G., Kraft,K., Heinrich,V., Krawitz,P., Brancati,F., Klopocki,E., Horn,D., Kayserili,H., Opitz,J.M., Laxova,R. *et al.* (2015) Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, **161**, 1012–1025.
86. Kruse,K., Hug,C.B., Hernández-Rodríguez,B. and Vaquerizas,J.M. (2016) TADtool: visual parameter identification for TAD-calling algorithms. *Bioinformatics*, **32**, 3190–3192.
87. Sauerwald,N. and Kingsford,C. (2018) Quantifying the similarity of topological domains across normal and cancer human cell types. *Bioinformatics*, **34**, i475–i483.
88. Zufferey,M., Tavernari,D., Oricchio,E. and Ciriello,G. (2018) Comparison of computational methods for the identification of topologically associating domains. *Genome Biol.*, **19**, 217.
89. Dali,R. and Blanchette,M. (2017) A critical assessment of topologically associating domain prediction tools. *Nucleic Acids Res.*, **45**, 2994–3005.
90. Gilad,Y., Rifkin,S.A. and Pritchard,J.K. (2008) Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet.*, **24**, 408–415.
91. Aguet,F., Brown,A.A., Castel,S.E., Davis,J.R., He,Y., Jo,B., Mohammadi,P., Park,Y.S., Parsana,P., Segrè,A. V. *et al.* (2017) Genetic effects on gene expression across human tissues. *Nature*, **550**, 204–213.
92. 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.
93. 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.
94. Hariprakash,J.M. and Ferrari,F. (2019) Computational biology solutions to identify enhancers-target gene pairs. *Comput. Struct. Biotechnol. J.*, **17**, 821–831.
95. Knapp,T.R. (1978) Canonical correlation analysis: a general parametric significance-testing system. *Psychol. Bull.*, **85**, 410–416.
96. Freire-Pritchett,P., Schoenfelder,S., Várnai,C., Wingett,S.W., Cairns,J., Collier,A.J., García-Vílchez,R., Furlan-Magaril,M., Osborne,C.S., Fraser,P. *et al.* (2017) Global reorganisation of cis-regulatory units upon lineage commitment of human embryonic stem cells. *Elife*, **6**, e21926.
97. Datlinger,P., Rendeiro,A.F., Schmidl,C., Krausgruber,T., Traxler,P., Klughammer,J., Schuster,L.C., Kuchler,A., Alpar,D. and Bock,C. (2017) Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods*, **14**, 297–301.
98. Jaitin,D.A., Weiner,A., Yofe,I., Lara-Astiaso,D., Keren-Shaul,H., David,E., Salame,T.M., Tanay,A., van Oudenaarden,A. and Amit,I. (2016) Dissecting Immune circuits by linking CRISPR-Pooled screens with Single-Cell RNA-Seq. *Cell*, **167**, 1883–1896.
99. Cao,F. and Fullwood,M.J. (2019) Inflated performance measures in enhancer–promoter interaction-prediction methods. *Nat. Genet.*, **51**, 1196–1198.