The adapted Activity-By-Contact-model for enhancer-gene assignment and its combination with transcription factor affinities in single cell data

Dennis Hecker[1,2,3], Fatemeh Behjati Ardakani[1,2,3], Marcel H. Schulz[1,2,3*]

1 Institute of Cardiovascular Regeneration, Goethe University and University Hospital, 60590 Frankfurt am Main, Germany,
2 Cardio-Pulmonary Institute, Goethe University, Frankfurt am Main, Germany
3 German Center for Cardiovascular Research, Partner site Rhein-Main, 60590 Frankfurt am Main, Germany

*** marcel.schulz@em.uni-frankfurt.de**

Abstract

Identifying regulatory regions in the genome is of great interest for understanding the epigenomic landscape in cells. One fundamental challenge in this context is to find the target genes whose expression the regulatory regions affect. A recent successful method is the Activity-By-Contact (ABC) model [17] which scores enhancer-gene interactions based on enhancer activity and the contact frequency of an enhancer to its target gene. It requires two types of assays to measure enhancer activity, which limits the applicability. The ABC score describes regulatory interactions entirely from a gene's perspective, and does not account for all the candidate target genes of an enhancer. Moreover, there is no implementation available which would allow for an integration with transcription factor (TF) binding information or an efficient analysis of single-cell data. We show that the ABC-model is equally accurate with only one assay to measure enhancer activity. Further, we demonstrate that it can yield a higher accuracy by adapting the enhancer activity according to the number of contacts the enhancer has to its candidate target genes. We combined our adapted ABC-model with TF binding information and illustrate an analysis of a single-cell ATAC-seq data set of the human heart. We were able to characterise cell type-specific regulatory interactions as well as to prioritise candidate TFs that drive cell type-specific expression. All executed processing steps are incorporated into our new computational pipeline STARE. The software is available at `https://github.com/schulzlab/STARE`

Introduction

Unravelling the mechanisms behind gene expression regulation is one of the central tasks in epigenomics. Enhancers are key players in this process. They are accessible regions in the genome, which can be bound by transcription factors (TFs) in a sequence-specific fashion. Those TFs in turn can have a variety of functions: they recruit other cofactors, remodel chromatin conformation, cause changes in epigenetic modifications or directly interact with the transcriptional complex, ultimately affecting gene expression (46; 21; 33). There is a variety of methods to annotate enhancers, for example by measuring openness of chromatin with assays like DNase-seq, ATAC-seq or NOMe-seq (61; 9; 30). Histone modifications associated with enhancer activity like H3K27ac or H3K4me1 can also aid in enhancer annotation (22; 10). Besides of a plethora of methods to define enhancers, another ongoing challenge is to identify target genes of enhancers, which is essential for understanding their function. These enhancer-gene interactions can span large distances and are insufficiently explained by linear proximity (68; 58). Several approaches have been developed to predict target genes of enhancers. Some methods rely on correlation of enhancer activity and gene expression across samples (63; 18; 56), while others associate deregulation of enhancers to genes in context of diseases (66; 72). Another possibility is to include chromatin contact data, such as Hi-C [37]. A common procedure is to call chromatin

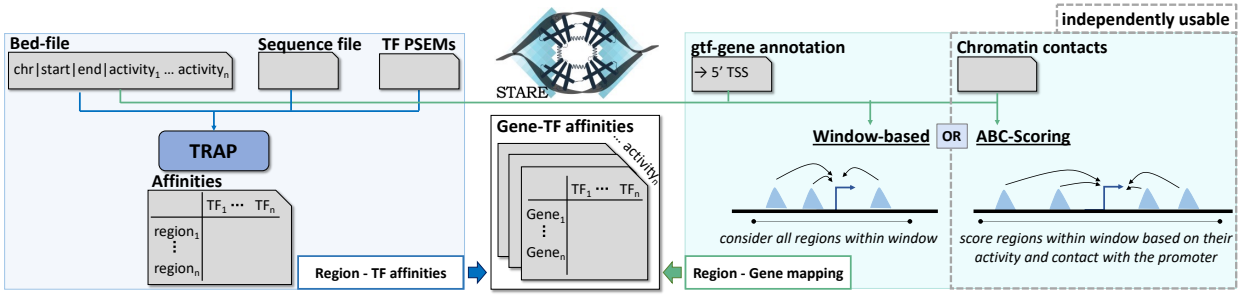

Figure 1. Overview of the STARE pipeline. It determines affinities of TFs to candidate enhancer regions (left panel) via TRAP [51] and combines it with a mapping of candidate enhancers to their target genes (right panel), using either a window-based approach or an implementation of the ABC-score [17]. The result is a matrix summarising TF affinities per gene. The principle is based on TEPIC [53]. STARE is adapted to run on multiple cell types with the same candidate enhancers but varying activity, represented by activity columns. PSEM: position-specific energy matrix.

loops for annotating enhancer-gene links (48; 55; 69). Although loops correlate with gene expression, their anchors only cover a fraction of active promoters and enhancers and their removal impacts the expression only for a few genes (44; 49; 58). [17] combined measurements of enhancer activity with chromatin contact data and proposed the Activity-By-Contact (ABC) model. The assumption is that enhancers that are active and frequently contact a gene's promoter are more likely to affect a gene's regulation. Moreover, the original ABC-model formulation is entirely gene-centric, which means it does not take the candidate target genes of an enhancer into account.

We implemented the ABC-model and could show that, instead of using both DNase-seq and H3K27ac ChIP-seq data, one assay for measuring the openness of enhancers is sufficient. Further, we propose an adapted calculation of the ABC-score, which describes enhancer activity in a gene-specific manner without requiring any additional data. We validated these adaptations on two data sets of experimentally tested enhancer-gene interactions in K562 cells (17; 19). We combined our implementation of the ABC-score with an approach to quantify TF binding affinity into STARE: a framework to derive TF affinities to genes. In addition, we designed STARE for potential application on single-cell data. To illustrate STARE's usability, we analysed a single-cell data set of the human heart consisting of eight defined cell types [24]. We exploited single-cell enhancer activity measurements and bulk chromatin contact data. Despite having only one data modality in single-cell resolution, we were able to identify cell type-specific regulatory interactions and to prioritise TFs for expression regulation.

Methods

1 Methods

1.1 Activity-By-Contact score

Throughout the manuscript we will use the terms enhancer and regulatory region interchangeably. The principle of the ABC-model is that an enhancer, which is highly active and has a high contact frequency with a gene, is also likely to have a regulatory impact on that gene [17]. The ABC-score represents the relative contribution of an enhancer r to the regulation of gene g, measured by the enhancer's activity A_r and its contact frequency $C_{r,g}$ with the promoter of gene g. For each candidate enhancer the activity is multiplied with the contact frequency and this product is then taken relative to the sum over all candidate enhancers R_g in a predefined window around the 5' transcription start site (TSS) of a gene:

$$ABC(r,g) = \frac{A_r \cdot C_{r,g}}{\sum\limits_{i \in R_g} A_i \cdot C_{i,g}} \,. \tag{1}$$

By definition the scores per gene sum up to 1. In practice, a cut-off is used to select valid interactions. The ABC-model allows a many-to-many relationship: a gene can be linked to multiple enhancers and an enhancer can be linked to multiple genes. In this work, we tested three types of epigenomic assays to approximate A_r by the amount of sequencing reads in the enhancer region: DNase-seq, ATAC-seq, and H3Kac27 ChIP-seq. The contact $C_{r,g}$ is taken from a normalised chromatin contact frequency matrix. We add a pseudocount to each $C_{r,g}$, so that all candidates R_g are taken into account and none of the $A_r \cdot C_{r,g}$ products is zero (see Supplementary Material for details).

1.2 Adapted Activity-By-Contact-score

We present an adaptation of the ABC-score, to account for all candidate target genes of an enhancer. The activity of an enhancer conceptually represents all regulatory interactions an enhancer has. However, an enhancer can interact with different genes and not all genes are equally likely to be brought into vicinity of the enhancer. We assume that an enhancer's regulatory input is a function of the number of contacts with all potential target genes. Target genes that are often in contact with the enhancer would receive more regulatory input than genes with fewer contacts. Thus, the activity A_r could be denoted in a relative manner:

$$A_r = \sum_{g \in G_r} A_{r,g} \,, \tag{2}$$

where $A_{r,g}$ denotes the relative regulatory activity of enhancer r towards gene g and G_r denotes the set of all genes that are within a predefined window around r. Since the values of $A_{r,g}$ are not known, we propose an approximation by using chromatin contacts:

$$A_{r,g} \approx A_r \cdot \frac{C_{r,g}}{\sum_{j \in G_r} C_{r,j}} \,. \tag{3}$$

In this formulation we approximate the regulatory activity $A_{r,g}$ by the relative fraction of its contact $C_{r,g}$ to the sum over the contacts to all genes G_r in a window around the enhancer. In consequence, the adapted ABC-score $\overline{ABC}_{r,g}$ can be written as:

$$\overline{ABC}_{r,g} = \frac{A_{r,g} \cdot C_{r,g}}{\sum\limits_{i \in R_g} A_{i,g} \cdot C_{i,g}} \,. \tag{4}$$

In comparison to the regular ABC score in Eq. (1), the activity A_r of an enhancer is replaced with its gene-specific relative activity $A_{r,g}$, as defined in Eq. (3). Thus, the adapted $\overline{\text{ABC}}$-score does not only function in a gene-centric way, by considering the relative number of contacts for each gene to its candidate enhancers, but also in an enhancer-centric way, by adapting the activity to the relative number of contacts of an enhancer's candidate target genes. This adapted $\overline{\text{ABC}}$-score therefore uses a reduced enhancer activity estimate, in particular for enhancers that are in contact with many genes, and prevents them from being accounted with their full activity for all genes within the scoring window.

1.3 Validation on experimentally measured enhancer-gene interactions

To validate different assays for measuring enhancer activity A_r and to test our adapted $\overline{\text{ABC}}$-score, we examined the performance on experimentally validated enhancer-gene interactions. We made use of two CRISPR-screens for K562 cells, where enhancer regions were perturbed followed by measuring changes in gene expression: [17] and [19]. Unlike [17], we do not divide interactions into enhancer-gene and promoter-gene pairs, as promoters can act in a similar way as enhancers [3], and thus should be included as potential regulatory interactions. We also do not exclude interactions where the expression decreased after enhancer perturbation. We took interactions from both studies with a false discovery rate (FDR) of $\leq 5\%$ as our validation data set. We used the same definition of candidate enhancers as [17]: each DNase I hypersensitive site was extended to a total length of 500 bp and in addition, all promoter regions of 500 bp around the TSS of all genes were considered. Only interactions that intersected at least one K562 candidate enhancer could

be considered. The sequencing reads in the candidate enhancers for DNase-seq and H3K27ac ChIP-seq were already provided in the work of [17]. ATAC-seq reads were retrieved from ENCODE data (see Sec. 4; 62; 12). In addition to the three aforementioned assays for measuring enhancer activity, we also tested taking the geometric mean of DNase-seq and H3K27ac ChIP-seq reads, as proposed by [17]. For each activity setup we evaluated the regular ABC- and the adapted $\overline{\text{ABC}}$-scoring approach. To construct precision-recall curves the cut-off was incremented by 10,000 steps in between the smallest and largest score. For each step, the interactions with a score above the cut-off were classified as positive set and compared against the validated interactions. If multiple candidate enhancers overlapped a validated perturbed region, we took the sum of the individual enhancers' score.

1.4 Transcription factor affinities and summarisation on gene-level

In addition to enhancer-gene interactions, STARE aims to retrieve a representation of a TF's regulatory impact on a gene. There are two steps required. First, genomic regions that influence the regulation of a gene have to be identified. This can be done via ABC-scoring as described before, or in a more simplistic approach by taking all open regions within a defined window around the gene's TSS into account. Utilising the ABC-model allows STARE to function with any desired window size around the TSS. Second, the affinities of TFs to the identified regions have to be quantified. Instead of relying on calling TF binding sites, we use the tool TRAP, which calculates relative binding affinities for TFs in a genomic region. TRAP describes TF binding with a biophysical model to predict the number of TF molecules that bind in a sequence [51]. The higher the affinity, the more likely a TF is to bind. For all analyses presented in this manuscript we used a collection of 726 human TF position-specific energy matrices (PSEMs), which is provided in the TEPIC 2 software [54] and consists of TF motifs from JASPAR, HOCOMOCO, and the work of Kheradpour and Kellis (52; 32; 31).

In order to summarise the TF affinities per gene, the TF affinities in the regulatory regions are combined with the predicted enhancer-gene interactions. The summarisation depends on how the region-gene mapping was done. For the window-based approach, the TF affinities in all regions around the gene's TSS are summed in the following fashion:

$$af_{g,tf} = \sum_{r \in R_g} \frac{af_{r,tf}}{ml_{tf}} \cdot A_r \cdot e^{-\frac{d_{r,g}}{d_0}} , \tag{5}$$

where $af_{g,tf}$ is the affinity of TF tf summarised for gene g. R_g is the set of all regions r that were located within the window around gene g. $af_{r,tf}$ is the affinity of tf in r, ml_{tf} is the motif length of tf and A_r is the activity for r. The affinity is corrected for the distance $d_{r,g}$ of r to the TSS of g by an exponential decay function, as proposed by [45], where d_0 is set to a constant of 5000 bp.

When the $\overline{\text{ABC}}$-score was used to assign regions to genes, the summarisation changes, as there is more epigenomic information available. Regions close to the TSS (≤ 2500 bp) are always included, independent of their $\overline{\text{ABC}}$-score, as the promoter region is very informative for the expression regulation of a gene [54]. They are also scaled differently, resulting in the formulation:

$$af_{g,tf} = \sum_{r \in R_g} \frac{af_{r,tf}}{ml_{tf}} \cdot \begin{cases} A_r \cdot e^{-\frac{d_{r,g}}{d_0}}, & \text{if } d_{r,g} \leq 2500\ bp \\ A_{r,g}, & \text{otherwise} \end{cases} , \tag{6}$$

where R_g is a set of regions that was linked to the gene with the $\overline{\text{ABC}}$-score and $A_{r,g}$ is the adapted activity (Eq. 3). For regions close to the TSS the base activity A_r is corrected with the exponential decay function, as the contact frequency would likely be the contact of the region with itself and thus $A_{r,g}$ could be erroneous. When not using the adapted $\overline{\text{ABC}}$-score, the affinity scaling changes as follows:

$$af_{g,tf} = \sum_{r \in R_g} \frac{af_{r,tf}}{ml_{tf}} \cdot \begin{cases} A_r \cdot e^{-\frac{d_{r,g}}{d_0}}, & \text{if } d_{r,g} \leq 2500\ bp \\ A_r \cdot \frac{C_{r,g}}{C_{max}}, & \text{otherwise} \end{cases} . \tag{7}$$

The regions close to the TSS are scaled in the same way as before. For all other regions we divide the contact frequency $C_{r,g}$ by the maximum contact that was measured for all region-gene pairs C_{max}. The reasoning is

to incorporate the contact frequency and to have both multipliers for A_r in the range of $[0, 1]$. Essentially, the activity multipliers for $\overline{\text{ABC}}$ and ABC differ only in how the contact is scaled: for $\overline{\text{ABC}}$ relative to all gene contacts of the respective enhancer and for ABC to the contacts of all enhancer-gene pairs.
In addition to the TF affinities, we report two additional gene features: the number of regions considered per gene, as well as their average base pair length (*AvgPeakSize*), following [57].

1.5 Application to single-cell data

As an example application, we ran STARE on a single-cell data set of the human heart from [24]. In this study, cell types were clustered based on single-nuclei (sn) ATAC-seq, as well as snRNA-seq. The candidate enhancers were pooled and ATAC-seq RPKM was measured for each cell type. For chromatin contacts we used H3K27ac HiChIP data of the left ventricle [4]. Regions known to accumulate an anomalous amount of sequencing reads were excluded (2, 62). As we had enhancer activity as well as enhancer contact data at hand, we determined enhancer-gene interactions for each cell type with the $\overline{\text{ABC}}$-score (cut-off 0.02, 5 MB window around the 5' TSS), and summarised TF affinities for each gene in each cell type based on those interactions.

1.5.1 Regulation of cell type-specific genes

To further illustrate the usability of STARE and to test how much of gene expression variation can be explained by enhancer activity, we looked more detailed into the regulation of cell type-specific (CS) genes. Based on snRNA-seq from [24], we selected genes where the z-score on expression in between cell types was ≥ 2 and TPM ≥ 0.5. For each set of CS genes we ran STARE with $\overline{\text{ABC}}$-scoring to derive a gene-TF affinity matrix, followed by INVOKE, a gene expression prediction tool from the TEPIC software [53]. INVOKE learns a linear regression model based on gene-TF scores, allowing to derive those TFs that are most predictive for gene expression. We limited the gene-TF matrix to those TFs that were expressed in a cell type (TPM ≥ 0.5). We also tested INVOKE on the whole gene set in each cell type. [55] applied INVOKE on similar data types for bulk data, but restricted information of distant enhancers to those connected to promoters via loops. With the ABC-approach we have a finer resolution of regulatory interactions and are able to integrate the contact frequency into the summarisation of TF affinities. Unlike [55], we did not split TF affinities in enhancer and promoters into separate features.

1.6 Implementation details of STARE

We combined our implementation of the ABC-scoring with an adaptation of the TEPIC pipeline [53] and created a new framework, called STARE (Fig. 1). The $\overline{\text{ABC}}$-scoring calculation can also be called independently. STARE was designed under the assumption that cell type specificity is mainly driven by enhancer activity. Thus, for analysing single-cell data it would be sufficient to define candidate enhancers and measure their activity in individual cells, or summarising activity over clusters of cells or cell types. We call these activity representations activity columns, one for each cell unit. A cell unit can represent any level of aggregation: from individual single cells to metacells up to summarised cell types. STARE was built to leverage this concept of shared candidate enhancers with multiple activity columns. It produces individual results for each activity column, but runs the calculations which are independent of enhancer activity only once. The same holds true for its ABC-scoring implementation. STARE consists of C++ programs connected by a main shell script, where runtime is reduced by utilising parallel computing.

2 Results

2.1 One activity assay suffices for accurate ABC-score

The original formulation of the ABC-score requires two types of assays to measure enhancer activity, namely DNase-seq and H3K27ac ChIP-seq [17]. We tested the performance of the ABC-scoring principle on two validation data sets (17; 19), using different assays for measuring enhancer activity. On both data sets the

combination of DNase-seq and H3K27ac ChIP-seq performs better than either of them alone (Fig. 2a+b left panels). However, the combination only yields a marginal improvement over the usage of DNase-seq data alone (AUPRC increase of 0.05 each; Tab 1). Quantifying enhancer activity with ATAC-seq resulted in a worse performance on the Fulco validation data, but in a clear improvement on the larger Gasperini data set, surpassing the combination of DNase-seq and H3K27ac ChIP-seq.

2.2 Adapted $\overline{\text{ABC}}$-score improves prediction of regulatory interactions

We propose an adapted version of the $\overline{\text{ABC}}$-score (Eq. 4), where the activity of an enhancer is described in a gene-specific manner (Fig. 2c, Eq. 3). On both validation data sets and for all combinations of activity measurements the $\overline{\text{ABC}}$-score outperforms its respective regular counterpart (Fig. 2a+b right panels; Tab. 1). The adapted activity prevents enhancers to be considered for all genes in the scoring window with their full activity. We could reproduce the higher accuracy of the $\overline{\text{ABC}}$-score in a direct comparison to the implementation of [17] (Fig. S1). The gene *CD58* is an example, where the $\overline{\text{ABC}}$-score is better for identification of enhancer-gene interactions. Out of nine candidate enhancers that were perturbed by [19], one showed a significant impact on gene expression. While the $\overline{\text{ABC}}$-score correctly classifies that particular interaction, the ABC-score calls four false positive interactions (both using a score cut-off of 0.02; Fig. 2d). Although it does not require additional information, the $\overline{\text{ABC}}$-score performs better than the regular ABC-score.

Table 1. AUPRC (area under the precision recall curve) values using ABC and $\overline{\text{ABC}}$ for identifying regulatory interactions on two validation data sets, using varying input assays for measuring enhancer activity, see Fig. 2 for examples. For setup details see Sec. 1.3.

Validation data	Fulco et al. (2019) 158 valid out of 3,999 interactions				Gasperini et al. (2019) 745 valid out of 35,809 interactions			
Assay(s) for enhancer activity	DNase & H3K27ac	DNase	H3K27ac	ATAC	DNase & H3K27ac	DNase	H3K27ac	ATAC
ABC	0.460	0.445	0.399	0.404	0.469	0.464	0.406	0.513
$\overline{\text{ABC}}$	0.466	0.453	0.440	0.421	0.515	0.517	0.457	0.551

2.3 Single-cell data analysis with STARE

We applied STARE on a single-cell data set of the human heart, analysing eight individual clusters of cell types [24]. The candidate enhancers of the cell types were pooled to a set of 286,777 regions with a summarised ATAC-seq measurement for each defined cell type (Fig. 3a). First, we looked into the regulatory interactions identified by the $\overline{\text{ABC}}$-score as part of STARE (score ≥ 0.02, 5 MB scoring window). On average each cell type had 416,813 enhancer-gene interactions (SD $\approx$ 12,575). Approximately 23.9% of a cell type's interactions were shared with all other cell types, meaning that the same candidate enhancer was assigned to the same gene. Each cell type also featured unique interactions, although this was highly variable with nervous cells (NR) having 18.0% of their interactions unique, while atrial cardiomyocytes (aCM) possessed 5.2% unique interactions. aCM and ventricular cardiomyocytes (vCM) formed the largest intersection of interactions found in only two cell types, consisting of 41,472 interactions (Fig. 3b). The average median of enhancers per expressed gene (TPM ≥ 0.5) across the cell types was 4.75 (SD ≈ 0.43) (Fig. S2a). We ranked genes by their number of assigned enhancers under the assumption that more enhancers indicate a more complex regulation. Among the highest ranked genes were many long noncoding intergenic RNA (lincRNA) genes, but also TF genes like *RUNX1*, *DMRT1* or *TCF4* and a number of different Zinc finger (ZNF) genes (Fig. 3c). However, we did not find a correlation between gene expression and the number of assigned enhancers. We performed gene ontology (GO) enrichment analysis [50] for the 1,000 genes with the highest number of assigned enhancers in each cell type, and found mainly unspecific terms (Fig. S2b). Altogether, despite all cell types having the same candidate enhancers and a shared contact measurement, their predicted enhancer-gene interactions appeared to be considerably distinct.

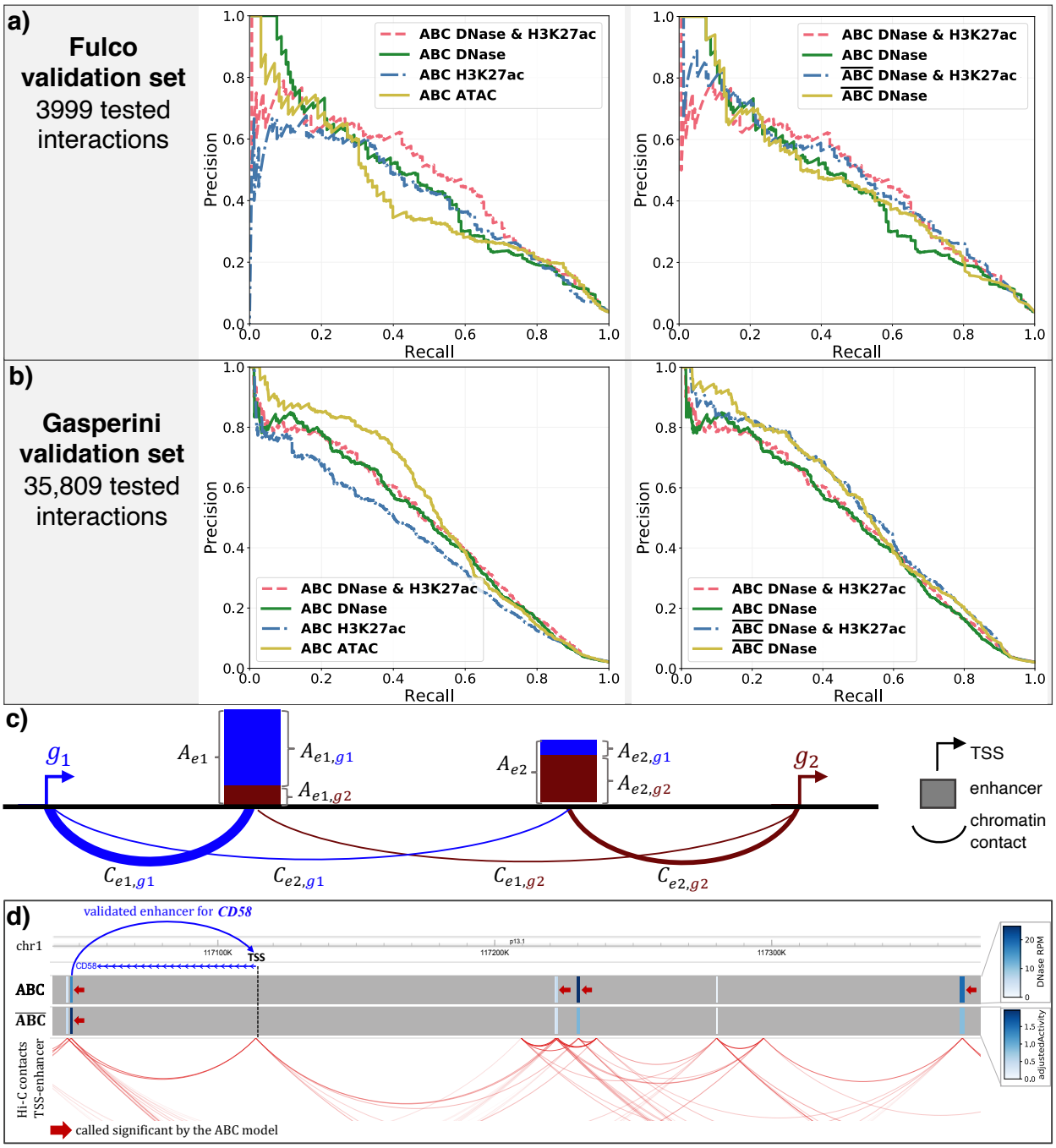

Figure 2. Performance of the ABC-score in K562 cells with varying assays for measuring enhancer activity and comparison to the $\overline{\text{ABC}}$-score. **(a, b)** Precision-recall (PR) curves of different ABC-scoring variants on two sets of experimentally validated enhancer-gene links (**(a)** [17]; **(b)** [19]). The respective areas under the PR curves can be found in Tab 1. **(c)** Schema for the adapted $\overline{\text{ABC}}$-score, which distributes the activity of an enhancer among its scored genes, dependent on contact frequencies. **(d)** Example gene *CD58* where the adapted $\overline{\text{ABC}}$ performs better than the regular ABC, based on interactions from [19]. The activity, or respectively the adapted activity, of the candidate enhancers is represented by their colour. The score cut-off was set to 0.02 for both approaches. Visualisation was done with the WashU Epigenome Browser [36].

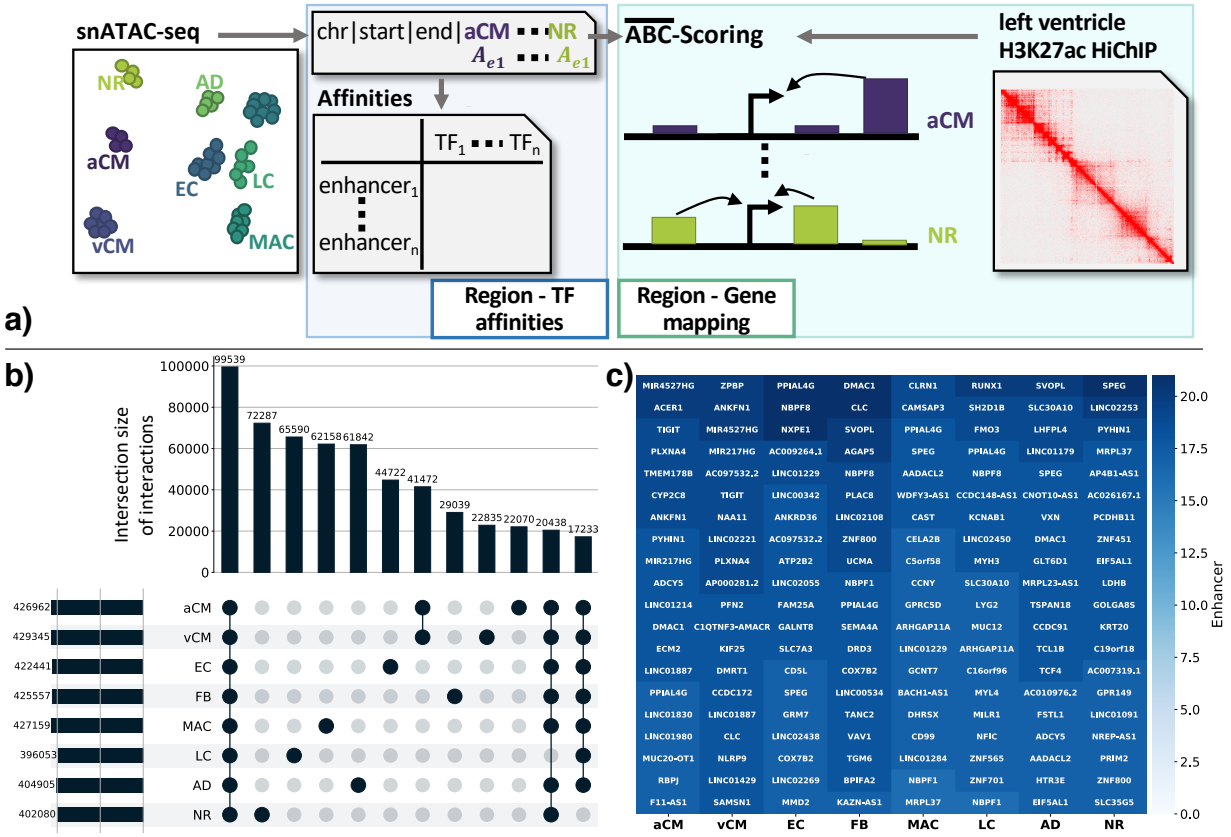

Figure 3. Application of STARE on single-cell data. **(a)** Schema of data processing. Single-nucleus ATAC-seq (snATAC-seq) from [24] was used to identify candidate enhancers and their activity in the annotated cell types. For each cell type interactions were called via $\overline{\text{ABC}}$-scoring. Enhancer-gene contacts were retrieved from left ventricle H3K27ac HiChIP data [4]. **(b)** Upset plot of the enhancer-gene interactions called in each cell type ($\overline{\text{ABC}}$-score ≥ 0.02). Only the 12 largest intersections are shown. **(d)** Top 20 expressed genes (TPM ≥ 0.5) with the highest number of called enhancers per cell type. **aCM**: atrial cardiomyocytes; **vCM**: ventricular cardiomyocytes; **EC**: endothelial cells; **FB**: fibroblasts; **MAC**: macrophages; **LC**: lymphocytes; **AD**: adipocytes; **NR**: nervous cells.

2.3.1 Linking epigenetic features to cell type-specific expression

Based on the snRNA-seq from [24] we defined a set of CS genes for each cell type (z-score on expression over all cell types ≥ 2 and TPM ≥ 0.5). Due to the high z-score cut-off, the CS genes were mostly unique to a cell type. Lymphocytes (LC) had the largest set with 2,617 CS genes, while fibroblasts (FB) had the smallest set with a size of 837 (Fig. 4a). The two types of cardiomyocytes formed the largest intersection of CS genes shared by more than one cell type. GO term enrichment on the CS genes yielded terms which appear to fit to the respective cell types, like 'lymphocyte activation' for macrophages (MC) and LC, or 'extracellular structure organization' for FB (Fig. S3a). To further characterise the sets of CS genes, we examined additional attributes and compared CS genes versus Non-CS genes (Fig. 4b; Fig. S3b). Across all cell types CS genes were depleted of housekeeping genes. Except in MAC, CS genes were also depleted for protein coding genes, but enriched for long noncoding RNA genes (lncRNA). We found that the Non-CS genes were more conserved, as assessed by the average phastCons100 score in the gene body [59]. Additionally, we examined attributes related to the enhancer-gene interactions called by the $\overline{\text{ABC}}$-score. CS genes had more assigned enhancers than Non-CS genes in all cell types except for LC. This matches the finding of [17] who describe a higher number of enhancers for tissue-specific genes than for ubiquitously expressed genes. Furthermore, CS genes tended to have a higher percentage of unique interactions, meaning the fraction of interactions of a gene that were exclusively found in that cell type was higher. These unique interactions could be responsible for the cell type-specific expression. By comparing the average enhancer activity between CS and Non-CS genes, we observed that most cell types had a slightly higher average activity of enhancers linked to CS genes, except for aCM and FB. There was no clear trend visible for the average contact frequency or TSS distance of the assigned enhancers.

We ran the gene expression prediction tool INVOKE on STARE's gene-TF affinity matrix and compared its performance on the regular ABC- and the adapted $\overline{\text{ABC}}$-scoring approach (Fig. 4c). The two versions differ in the identified enhancer-gene interactions, as well as in their summarisation of the TF affinities on gene level (see section 1.4). The gene sets used for training consisted of 1,404 genes on average (SD ≈ 582). We limited the TF affinities to those TFs that were expressed in a cell type (TPM ≥ 0.5). The Pearson and Spearman correlations on the predicted versus measured gene expression varied across cell types, but they were almost always higher for the $\overline{\text{ABC}}$-approach. We additionally trained INVOKE on the expression of all genes, which resulted in a more accurate prediction in comparison to training only on CS genes (Fig. S2c). The $\overline{\text{ABC}}$-approach outperformed the regular ABC-version on the larger training set. We also carried out a third window-based INVOKE run on the whole gene set, where the TF affinities were summarised over all regions within the scoring window. The resulting expression prediction was less accurate than the $\overline{\text{ABC}}$-based one.

INVOKE returns a regression coefficient for each TF, indicating how informative a TF was for the expression prediction. Although the genes on which INVOKE was trained were largely unique to cell types, many TFs with a high coefficient were shared. HLTF (human helicase-like transcription factor) was the most predictive TF for a high gene expression in all cell types (Fig. 4d). HLTF is not particularly described as a general transcriptional activator. Its functions include promotion of replication through DNA lesions [64], to slow replication fork progression and facilitate fork regression (8; 5). However, HLTF was also linked to developmental functions in the murine heart and collagen biogenesis (23; 13). The TFs NFATC1 and NFATC2 (nuclear factor of activated T-cell 1/2) were correlated with a high expression in aCM and vCM. The NFAT-calcineurin signalling pathway plays a role in regulating hypertrophic growth response in cardiomyocytes (41; 67). We also found multiple SOX factors among the TFs with a high coefficient. The family of SOX TFs are known as key regulators in cell fate decisions and because of those reprogramming abilities they are used to differentiate cells into a variety of lineages (28; 25). Specifically SOX15 was ranked highly in aCM and vCM. SOX15 is suggested to play an important role in skeletal muscle regeneration and regulation of myogenic progenitor cells (35; 39).

Some of the TFs with a negative correlation coefficient were reported to repress gene expression or at least fail to affect transcription at their target sites, like HSF4 (42; 16) or MAFG (29, 14).

Overall, with STARE we were able to characterise cell type-specific gene expression regulation and with the downstream example analysis INVOKE could suggest TF candidates which orchestrate this expression regulation.

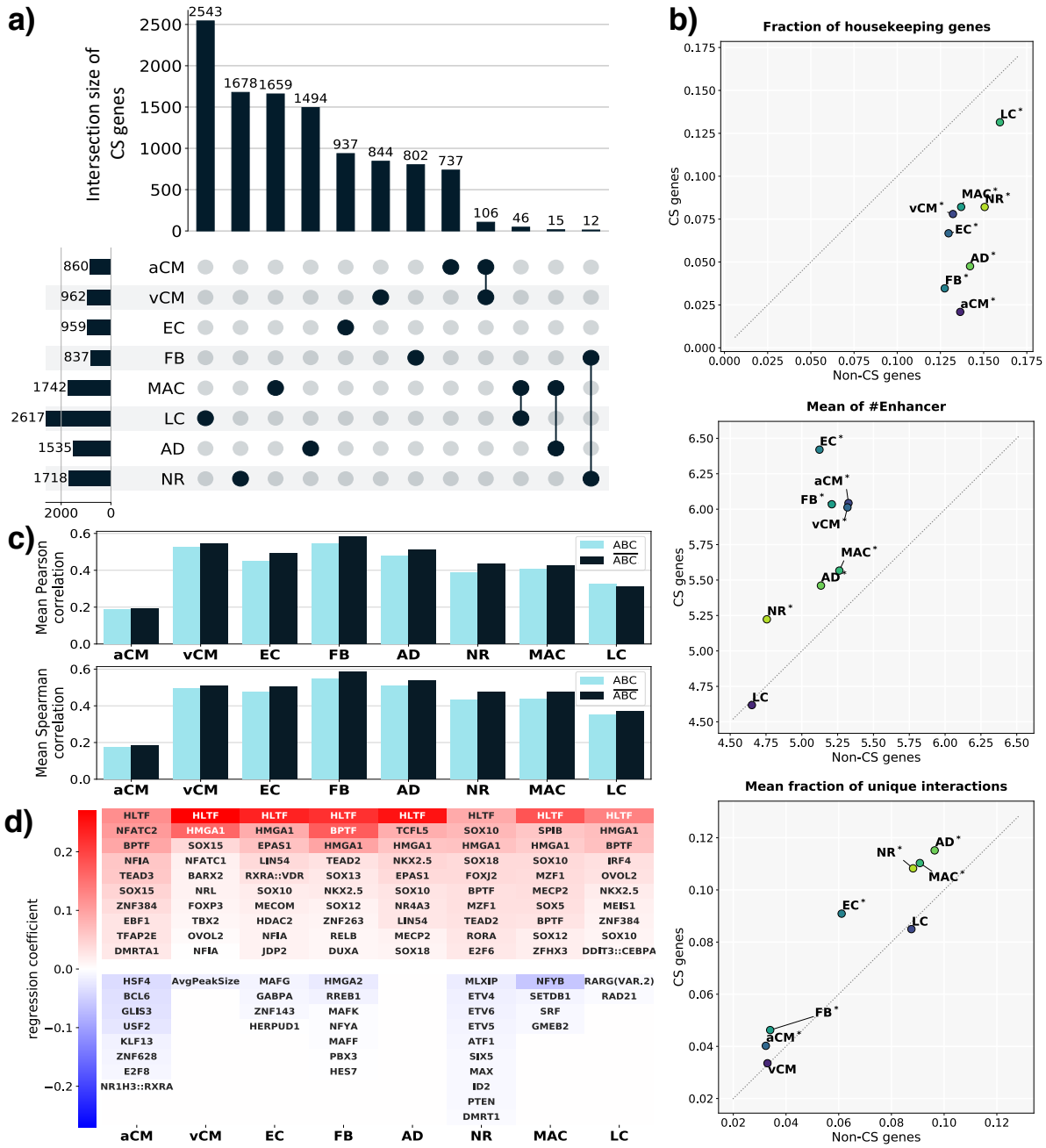

Figure 4. Application of STARE to analyse cell type-specific (CS) genes (TPM ≥ 0.5 and z-score ≥ 2). **(a)** Upset plot of CS genes, limited to the 12 largest intersections. **(b)** Comparison of attributes between CS and not cell type-specific (Non-CS) genes. An asterisk indicates a significant difference between CS and Non-CS genes (adjusted p-value ≤ 0.05). Binary attributes like housekeeping genes were tested with Fisher's exact test, numeric attributes like the average phastCons100 score with Mann-Whitney U test. False discovery correction was performed using the Benjamini-Hochberg procedure [7]. **(c)** Accuracy of CS gene expression prediction with INVOKE [53] using STARE's output. See Eq. 7 and Eq. 6 for the differences between ABC and $\overline{\text{ABC}}$. The Pearson and Spearman correlations are shown as average over a three-fold outer cross-validation. **(d)** TFs with the highest absolute regression coefficients in INVOKE's expression prediction. A positive coefficients indicates a positive correlation with gene expression and vice versa. Most cell types had less than ten negatively correlated TFs. **AvgPeakSize**: average base pair length of the regions assigned to a gene. See Fig. 3 for cell type abbreviations.

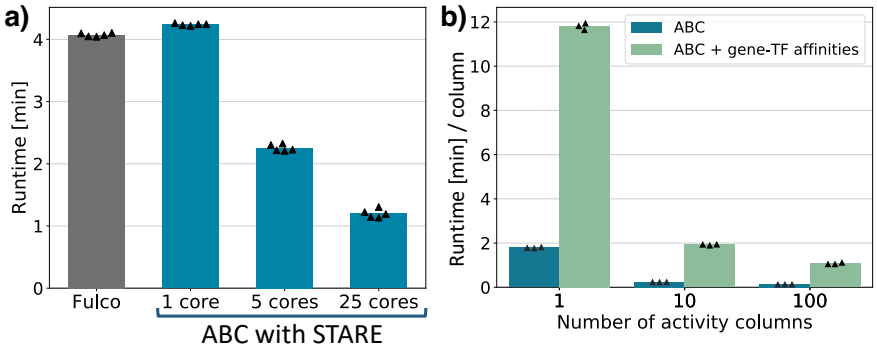

Figure 5. Runtime of the STARE pipeline. **(a)** Runtime of the original ABC pipeline by (Fulco et al.) and our ABC-scoring implementation with different number of cores. Any writing of output files was omitted. Calculation was done on data of K562 cells, scoring 155,679 candidate enhancers for 24,734 genes in a 10 MB window. The bars show the mean of five runs. **(b)** Runtime divided by the number of activity columns for the ABC-scoring alone and when additionally calculating the gene-TF affinity matrix. Writing and compression of output files is included. Calculations were done on single-cell heart data from [24] with 286,777 candidate enhancers, for 57,278 genes in a 5 MB window. 726 TFs were assessed. The bars' height represent the mean of three runs.

2.4 Runtime of STARE

We optimised STARE's runtime by allowing multiple steps to be run in parallel (Fig. 5a) and to omit redundant calculations, if multiple cell entities with differing activity representations are processed. The runtime per activity column decreases when multiple columns are handled in the same run, allowing large data sets to be processed in a few minutes (Fig. 5b).

3 Discussion

The ABC-model showed a better accuracy for detecting validated enhancer-gene links in its original study than other approaches [17]. We present a variation of identifying regulatory enhancer-gene interactions with the ABC-model. We were able to demonstrate that unlike suggested in the original work, one assay for measuring enhancer activity works similarly well. Specifically, using DNase-seq or ATAC-seq without H3K27ac ChIP-seq data yields a reasonable performance. This finding could broaden the applicability of the ABC-score to all data sets where only one measurement of enhancer activity is available. Especially single-cell epigenomics lacks large resources for different types of data and it is challenging to measure multiple modalities in the same cell.

Further, we propose an adapted $\overline{\text{ABC}}$-score where the activity of an enhancer is described in a gene-specific manner by taking it relative to the number of contacts the enhancer has to its candidate target genes. This adaptation resulted in an improvement in identifying experimentally validated enhancer-gene interactions and a more accurate gene expression prediction. It is important to note that any identification of enhancer-gene interactions is still only a prediction and likely not capturing the whole regulatory complexity of genes. The ABC-model requires two data types, which makes it applicable in a range of scenarios, but it might also miss out relevant epigenetic information. Further, the model assumes that all genes are regulated in the same way. In the context of interaction validation, it is debatable whether CRISPR-screens are able to detect all regulatory interactions, as true enhancers with small effect sizes could potentially be neglected. Moreover, perturbations of individual enhancers are presumably not capable of accounting for shadow enhancers which have redundant functionality (58; 60).

We present STARE as a framework to derive gene-TF affinities. Together with the mapping of candidate enhancers to genes, either with the aforementioned ABC-score or in a window-based approach, STARE summarises TF affinities on a gene level. Unlike other methods aiming to determine regulatory relations of TFs to genes (34; 65), STARE does not require scarce TF ChIP-seq data. It uses a motif-based biophysical

model [51] to determine TF affinities in open chromatin regions. As consequence, unlike other methods [38], it is able to retain low affinity binding information. [47] work with similar data as STARE and rely on a graph-based approach. However, they do not incorporate active regulatory regions and their analysis is limited to a window marked by the most distant CTCF peaks within 50 kb of the gene body. With the ABC-model, STARE specifically determines candidate enhancers within any selected window size. It should be noted that our model assumes an additive influence of TFs which is not likely to accurately capture biological reality [70]. Furthermore, there is potential redundancy on two levels: the aforementioned functional redundancy of enhancers and the redundancy of TF binding motifs (20; 11).

We applied our framework to a single-cell data set consisting of eight clustered cell types of the human heart [24], where the activity of a uniform set of candidate enhancers was measured for each cell type. The assumption was that cell type-specificity is mainly driven by activity of regulatory regions and less by spatial chromatin conformation. Although chromatin contacts were also found to change upon cell differentiation (15; 71), the 3D conformation of the genome is described as less dynamic and more as a scaffold to enable and stabilise regulatory interactions (58; 27). We were able to unravel differences in regulatory interactions across cell types and to characterise regulation of cell type-specific genes, despite using bulk chromatin contact data. In support of the feasibility to use bulk contact data, [17] reported a high accuracy of their ABC-model with averaged Hi-C data which was constructed from 10 data sets. In this context, we demonstrated a downstream application example of STARE with the expression prediction tool INVOKE [53] which allows to identify candidate regulatory TFs for further evaluation. Considering the small training sets and that INVOKE assumes the same regulation for all genes, the expression prediction yielded a reasonable performance with meaningful prioritisation of TFs. [55] also used INVOKE and incorporated distant enhancers. However, their work relies on annotated loops, which are not likely to cover all relevant regulatory interactions. In addition, their approach is not able to integrate the contact frequency into TF affinities, nor to derive cell type-specific interactions from bulk contact data. There are other tools predicting gene expression explicitly in individual cells that incorporate TF information, such as SCENIC [1], ACTION [40] or TRIANGULATE [6], but none of them considers long-range enhancer-gene interactions. A compelling approach would be to combine these expression prediction tools with information on distant enhancers.

STARE represents a currently unique form of deriving TF affinities on a gene level: it combines enhancer-gene links called by the ABC-score with a non hit-based TF annotation. Prospectively, we would like to apply STARE on individual cells instead of clustered cell types, which would require additional steps to account for the drastic sparsity of most single-cell measurements. It would also be highly interesting to further investigate the importance of chromatin contacts in cell type-specificity, once the resolution and availability for such data are advanced.

4 Genome version and data sources

The presented results are provided via Zenodo (`https://doi.org/10.5281/zenodo.5841992`). All data is in hg19. The validated interactions and predictions for K562 cells of [17] were taken from the original publication and their OSF repository (`https://osf.io/uhnb4/`). Validated interactions from [19] were taken from the 'at-scale' data set (GEO: GSE120861). We used K562 Hi-C data from [48] (GEO: GSE63525). Bam files of ATAC-seq data for K562 cells was retrieved from ENCODE: ENCFF128WZG, ENCFF534DCE, ENCFF077FBI. Single-cell data of the human heart was taken from [24] (GEO: GSE165839). The H3K27ac HiChIP of the left ventricle was kindly provided by [4]. The collection of TF PSEMs can be found in the GitHub repository of TEPIC: `https://github.com/SchulzLab/TEPIC/blob/master/PWMs/2.0/human_jaspar_hoc_kellis.PSEM`. The annotation of housekeeping genes was derived from the HRT Atlas [26]. For the phastCons100 score we used the track provided by the UCSC genome browser [43].

Acknowledgements

We thank Nina Baumgarten for advice and discussions on the implementation. We acknowledge the ENCODE Consortium and the lab of Michael Snyder in Stanford who generated the ATAC-seq data sets.

Funding

This work has been supported by the DZHK (German Centre for Cardiovascular Research, 81Z0200101) and the Cardio-Pulmonary Institute (CPI) [EXC 2026] ID: 390649896, and the DFG SFB (TRR 267) Noncoding RNAs in the cardiovascular system, Project-ID 403584255.

References

1. S. Aibar, C. B. González-Blas, T. Moerman, V. A. Huynh-Thu, H. Imrichova, G. Hulselmans, F. Rambow, J.-C. Marine, P. Geurts, J. Aerts, J. van den Oord, Z. K. Atak, J. Wouters, and S. Aerts. SCENIC: single-cell regulatory network inference and clustering. *Nature Methods*, 14(11):1083–1086, Nov. 2017.

2. H. M. Amemiya, A. Kundaje, and A. P. Boyle. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Scientific Reports*, 9(1):9354, Dec. 2019.

3. R. Andersson. Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model. *BioEssays*, 37(3):314–323, 2015.

4. C. G. Anene-Nzelu, W. L. W. Tan, C. J. M. Lee, Z. Wenhao, A. Perrin, A. Dashi, Z. Tiang, M. I. Autio, B. Lim, E. Wong, H. S. Tan, B. Pan, M. P. Morley, K. B. Margulies, T. P. Cappola, and R. S.-Y. Foo. Assigning Distal Genomic Enhancers to Cardiac Disease–Causing Genes. *Circulation*, 142(9):910–912, Sept. 2020.

5. G. Bai, C. Kermi, H. Stoy, C. J. Schiltz, J. Bacal, A. M. Zaino, M. K. Hadden, B. F. Eichman, M. Lopes, and K. A. Cimprich. HLTF Promotes Fork Reversal, Limiting Replication Stress Resistance and Preventing Multiple Mechanisms of Unrestrained DNA Synthesis. *Molecular Cell*, 78(6):1237–1251, June 2020.

6. F. Behjati Ardakani, K. Kattler, T. Heinen, F. Schmidt, D. Feuerborn, G. Gasparoni, K. Lepikhov, P. Nell, J. Hengstler, J. Walter, and M. H. Schulz. Prediction of single-cell gene expression for transcription factor analysis. *GigaScience*, 9(11):1–14, Oct. 2020.

7. Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

8. A. Blastyák, I. Hajdú, I. Unk, and L. Haracska. Role of Double-Stranded DNA Translocase Activity of Human HLTF in Replication of Damaged DNA. *Molecular and Cellular Biology*, 30(3):684–693, Feb. 2010.

9. J. D. Buenrostro, B. Wu, H. Y. Chang, and W. J. Greenleaf. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Current Protocols in Molecular Biology*, 109(1), Jan. 2015.

10. M. P. Creyghton, A. W. Cheng, G. G. Welstead, T. Kooistra, B. W. Carey, E. J. Steine, J. Hanna, M. A. Lodato, G. M. Frampton, P. A. Sharp, L. A. Boyer, R. A. Young, and R. Jaenisch. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *SYSTEMS BIOLOGY*, 107(50):21931–21936, 2010.

11. D. A. Cusanovich, B. Pavlovic, J. K. Pritchard, and Y. Gilad. The Functional Consequences of Variation in Transcription Factor Binding. *PLOS Genetics*, 10(3):13, 2014.

12. C. A. Davis, B. C. Hitz, C. A. Sloan, E. T. Chan, J. M. Davidson, I. Gabdank, J. A. Hilton, K. Jain, U. K. Baymuradov, A. K. Narayanan, K. C. Onate, K. Graham, S. R. Miyasato, T. R. Dreszer, J. Strattan, O. Jolanki, F. Y. Tanaka, and J. Cherry. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Research*, 46(D1):D794–D801, Jan. 2018.

13. M. Elserafy, A. A. Abugable, R. Atteya, and S. F. El-Khamisy. Rad5, HLTF, and SHPRH: A Fresh View of an Old Story. *Trends in Genetics*, 34(8):574–577, Aug. 2018.

14. M. Fang, J. Ou, L. Hutchinson, and M. Green. The BRAF Oncoprotein Functions through the Transcriptional Repressor MAFG to Mediate the CpG Island Methylator Phenotype. *Molecular Cell*, 55(6):904–915, Sept. 2014.

15. J. Fraser, M. Rousseau, S. Shenker, M. A. Ferraiuolo, Y. Hayashizaki, M. Blanchette, and J. Dostie. Chromatin conformation signatures of cellular differentiation. *Genome Biology*, 10(4):R37, 2009.

16. M. Fujimoto, K. Oshima, T. Shinkawa, B. B. Wang, S. Inouye, N. Hayashida, R. Takii, and A. Nakai. Analysis of HSF4 Binding Regions Reveals Its Necessity for Gene Regulation during Development and Heat Shock Response in Mouse Lenses. *Journal of Biological Chemistry*, 283(44):29961–29970, Oct. 2008.

17. C. P. Fulco, J. Nasser, T. R. Jones, G. Munson, D. T. Bergman, V. Subramanian, S. R. Grossman, R. Anyoha, B. R. Doughty, T. A. Patwardhan, T. H. Nguyen, M. Kane, E. M. Perez, N. C. Durand, C. A. Lareau, E. K. Stamenova, E. L. Aiden, E. S. Lander, and J. M. Engreitz. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nature Genetics*, 51(12):1664–1669, Dec. 2019.

18. T. Gao and J. Qian. EAGLE: An algorithm that utilizes a small number of genomic features to predict tissue/cell type-specific enhancer-gene interactions. *PLOS Computational Biology*, 15(10):e1007436, Oct. 2019.

19. M. Gasperini, A. J. Hill, J. L. McFaline-Figueroa, B. Martin, S. Kim, M. D. Zhang, D. Jackson, A. Leith, J. Schreiber, W. S. Noble, C. Trapnell, N. Ahituv, and J. Shendure. A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell*, 176(1-2):377–390, Jan. 2019.

20. A. Gitter, Z. Siegfried, M. Klutstein, O. Fornes, B. Oliva, I. Simon, and Z. Bar-Joseph. Backup in gene regulatory networks explains differences between binding and knockout results. *Molecular Systems Biology*, 5(1):276, Jan. 2009.

21. D. H. Gonzalez. Introduction to Transcription Factor Structure and Function. In *Plant Transcription Factors*, pages 3–11. Elsevier, 2016.

22. N. D. Heintzman, G. C. Hon, R. D. Hawkins, P. Kheradpour, A. Stark, L. F. Harp, Z. Ye, L. K. Lee, R. K. Stuart, C. W. Ching, K. A. Ching, J. E. Antosiewicz-Bourget, H. Liu, X. Zhang, R. D. Green, V. V. Lobanenkov, R. Stewart, J. A. Thomson, G. E. Crawford, M. Kellis, and B. Ren. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243):108–112, May 2009.

23. R. A. Helmer, R. Martínez-Zaguilán, J. S. Dertien, C. Fulford, O. Foreman, V. Peiris, and B. S. Chilton. Helicase-Like Transcription Factor (Hltf) Regulates G2/M Transition, Wt1/Gata4/Hif-1a Cardiac Transcription Networks, and Collagen Biogenesis. *PLoS ONE*, 8(11):e80461, Nov. 2013.

24. J. D. Hocker, O. B. Poirion, F. Zhu, J. Buchanan, K. Zhang, J. Chiou, T.-M. Wang, Q. Zhang, X. Hou, Y. E. Li, Y. Zhang, E. N. Farah, A. Wang, A. D. McCulloch, K. J. Gaulton, B. Ren, N. C. Chi, and S. Preissl. Cardiac cell type–specific gene regulatory programs and disease risk association. *Science Advances*, 7(20):eabf1444, May 2021.

25. L. Hou, Y. Srivastava, and R. Jauch. Molecular basis for the genome engagement by Sox proteins. *Seminars in Cell & Developmental Biology*, 63:2–12, Mar. 2017.

26. B. W. Hounkpe, F. Chenou, F. de Lima, and E. De Paula. HRT Atlas v1.0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets. *Nucleic Acids Research*, 49(D1):D947–D955, Jan. 2021.

27. E. Ing-Simmons, R. Vaid, X. Y. Bing, M. Levine, M. Mannervik, and J. M. Vaquerizas. Independence of chromatin conformation and gene regulation during Drosophila dorsoventral patterning. *Nature Genetics*, 53(4):487–499, Apr. 2021.

28. L. M. Julian, A. C. McDonald, and W. L. Stanford. Direct reprogramming with SOX factors: masters of cell fate. *Current Opinion in Genetics & Development*, 46:24–36, Oct. 2017.

29. K. Kataoka, K. Igarashi, K. Itoh, K. T. Fujiwara, M. Noda, M. Yamamoto, and M. Nishizawa. Small Maf proteins heterodimerize with Fos and may act as competitive repressors of the NF-E2 transcription factor. *Molecular and Cellular Biology*, 15(4):2180–2190, Apr. 1995.

30. T. K. Kelly, Y. Liu, F. D. Lay, G. Liang, B. P. Berman, and P. A. Jones. Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Research*, 22(12):2497–2506, Dec. 2012.

31. P. Kheradpour and M. Kellis. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Research*, 42(5):2976–2987, Mar. 2014.

32. I. V. Kulakovskiy, Y. A. Medvedeva, U. Schaefer, A. S. Kasianov, I. E. Vorontsov, V. B. Bajic, and V. J. Makeev. HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Research*, 41(D1):D195–D202, Jan. 2013.

33. S. A. Lambert, A. Jolma, L. F. Campitelli, P. K. Das, Y. Yin, M. Albu, X. Chen, J. Taipale, T. R. Hughes, and M. T. Weirauch. The Human Transcription Factors. *Cell*, 172(4):650–665, Feb. 2018.

34. X. Lan, H. Witt, K. Katsumura, Z. Ye, Q. Wang, E. H. Bresnick, P. J. Farnham, and V. X. Jin. Integration of Hi-C and ChIP-seq data reveals distinct types of chromatin linkages. *Nucleic Acids Research*, 40(16):7690–7704, Sept. 2012.

35. H.-J. Lee, W. Göring, M. Ochs, C. Mühlfeld, G. Steding, I. Paprotta, W. Engel, and I. M. Adham. Sox15 Is Required for Skeletal Muscle Regeneration. *Molecular and Cellular Biology*, 24(19):8428–8436, Oct. 2004.

36. D. Li, S. Hsu, D. Purushotham, R. L. Sears, and T. Wang. WashU Epigenome Browser update 2019. *Nucleic Acids Research*, 47(W1):W158–W165, July 2019.

37. E. Lieberman-Aiden, N. L. van Berkum, L. Williams, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, E. S. Lander, and J. Dekker. Comprehensive mapping of long range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, Oct. 2009.

38. R. C. McLeay, T. Lesluyes, G. Cuellar Partida, and T. L. Bailey. Genome-wide in silico prediction of gene expression. *Bioinformatics*, 28(21):2789–2796, Nov. 2012.

39. A. P. Meeson, X. Shi, M. S. Alexander, R. S. Williams, R. E. Allen, N. Jiang, I. M. Adham, S. C. Goetsch, R. E. Hammer, and D. J. Garry. Sox15 and Fhl3 transcriptionally coactivate Foxk1 and regulate myogenic progenitor cells. *The EMBO Journal*, 26(7):1902–1912, Apr. 2007.

40. S. Mohammadi, V. Ravindra, D. F. Gleich, and A. Grama. A geometric approach to characterize the functional identity of single cells. *Nature Communications*, 9(1):1516, Dec. 2018.

41. J. D. Molkentin, J.-R. Lu, C. L. Antos, B. Markham, J. Richardson, J. Robbins, S. R. Grant, and E. N. Olson. A Calcineurin-Dependent Transcriptional Pathway for Cardiac Hypertrophy. *Cell*, 93(2):215–228, Apr. 1998.

42. A. Nakai, M. Tanabe, Y. Kawazoe, J. Inazawa, R. I. Morimoto, and K. Nagata. HSF4, a new member of the human heat shock factor family which lacks properties of a transcriptional activator. *Molecular and Cellular Biology*, 17(1):469–481, Jan. 1997.

43. J. Navarro Gonzalez, A. S. Zweig, M. L. Speir, D. Schmelter, K. Rosenbloom, B. J. Raney, C. C. Powell, L. R. Nassar, N. Maulding, C. M. Lee, B. T. Lee, A. Hinrichs, A. Fyfe, J. Fernandes, M. Diekhans, H. Clawson, J. Casper, A. Benet-Pagès, G. P. Barber, D. Haussler, R. Kuhn, M. Haeussler, and W. Kent. The UCSC Genome Browser database: 2021 update. *Nucleic Acids Research*, 49(D1):D1046–D1057, Jan. 2021.

44. E. P. Nora, A. Goloborodko, A.-L. Valton, J. H. Gibcus, A. Uebersohn, N. Abdennur, J. Dekker, L. A. Mirny, and B. G. Bruneau. Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell*, 169(5):930–944, May 2017.

45. Z. Ouyang, Q. Zhou, and W. H. Wong. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proceedings of the National Academy of Sciences*, 106(51):21521–21526, Dec. 2009.

46. C. O. Pabo and R. T. Sauer. Transcription Factors: Structural Families and Principles of DNA Recognition. *Annu. Rev. Biochem*, 61:1053–95, 1992.

47. N. Patel and W. S. Bush. Modeling transcriptional regulation using gene regulatory networks based on multi-omics data sources. *BMC Bioinformatics*, 22(1):200, Dec. 2021.

48. S. Rao, M. Huntley, N. Durand, E. Stamenova, I. Bochkov, J. Robinson, A. Sanborn, I. Machol, A. Omer, E. Lander, and E. Aiden. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, 159(7):1665–1680, Dec. 2014.

49. S. S. Rao, S.-C. Huang, B. Glenn St Hilaire, J. M. Engreitz, E. M. Perez, K.-R. Kieffer-Kwon, A. L. Sanborn, S. E. Johnstone, G. D. Bascom, I. D. Bochkov, X. Huang, M. S. Shamim, J. Shin, D. Turner, Z. Ye, A. D. Omer, J. T. Robinson, T. Schlick, B. E. Bernstein, R. Casellas, E. S. Lander, and E. L. Aiden. Cohesin Loss Eliminates All Loop Domains. *Cell*, 171(2):305–320, Oct. 2017.

50. U. Raudvere, L. Kolberg, I. Kuzmin, T. Arak, P. Adler, H. Peterson, and J. Vilo. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, 47(W1):W191–W198, July 2019.

51. H. G. Roider, A. Kanhere, T. Manke, and M. Vingron. Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, 23(2):134–141, Jan. 2007.

52. A. Sandelin. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research*, 32(90001):91D–94, Jan. 2004.

53. F. Schmidt, N. Gasparoni, G. Gasparoni, K. Gianmoena, C. Cadenas, J. K. Polansky, P. Ebert, K. Nordström, M. Barann, A. Sinha, S. Fröhler, J. Xiong, A. Dehghani Amirabad, F. Behjati Ardakani, B. Hutter, G. Zipprich, B. Felder, J. Eils, B. Brors, W. Chen, J. G. Hengstler, A. Hamann, T. Lengauer, P. Rosenstiel, J. Walter, and M. H. Schulz. Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. *Nucleic Acids Research*, 45(1):54–66, Jan. 2017.

54. F. Schmidt, F. Kern, P. Ebert, N. Baumgarten, and M. H. Schulz. TEPIC 2—an extended framework for transcription factor binding prediction and integrative epigenomic analysis. *Bioinformatics*, 35(9):1608–1609, May 2019.

55. F. Schmidt, F. Kern, and M. H. Schulz. Integrative prediction of gene expression with chromatin accessibility and conformation data. *Epigenetics & Chromatin*, 13(1):4, Dec. 2020.

56. F. Schmidt, A. Marx, N. Baumgarten, M. Hebel, M. Wegner, M. Kaulich, M. Leisegang, R. Brandes, J. Göke, J. Vreeken, and M. Schulz. Integrative analysis of epigenetics data identifies gene-specific regulatory elements. *Nucleic Acids Research*, 49(18):10397–10418, Oct. 2021.

57. F. Schmidt and M. H. Schulz. On the problem of confounders in modeling gene expression. *Bioinformatics*, 35(4):711–719, 2019.

58. S. Schoenfelder and P. Fraser. Long-range enhancer–promoter contacts in gene expression control. *Nature Reviews Genetics*, 20(8):437–455, Aug. 2019.

59. A. Siepel. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8):1034–1050, Aug. 2005.

60. D. Singh and S. V. Yi. Enhancer Pleiotropy, Gene Expression, and the Architecture of Human Enhancer–Gene Interactions. *Molecular Biology and Evolution*, 38(9):3898–3909, Aug. 2021.

61. L. Song and G. E. Crawford. DNase-seq: A High-Resolution Technique for Mapping Active Gene Regulatory Elements across the Genome from Mammalian Cells. *Cold Spring Harbor Protocols*, 2010(2):pdb.prot5384, Feb. 2010.

62. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, Sept. 2012.

63. The FANTOM Consortium, R. Andersson, C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt, M. Boyd, Y. Chen, X. Zhao, C. Schmidl, T. Suzuki, E. Ntini, E. Arner, E. Valen, K. Li, L. Schwarzfischer, D. Glatz, J. Raithel, B. Lilje, N. Rapin, F. O. Bagger, M. Jørgensen, P. R. Andersen, N. Bertin, O. Rackham, A. M. Burroughs, J. K. Baillie, Y. Ishizu, Y. Shimizu, E. Furuhata, S. Maeda, Y. Negishi, C. J. Mungall, T. F. Meehan, T. Lassmann, M. Itoh, H. Kawaji, N. Kondo, J. Kawai, A. Lennartsson, C. O. Daub, P. Heutink, D. A. Hume, T. H. Jensen, H. Suzuki, Y. Hayashizaki, F. Müller, A. R. R. Forrest, P. Carninci, M. Rehli, and A. Sandelin. An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–461, Mar. 2014.

64. I. Unk, I. Hajdu, K. Fatyol, J. Hurwitz, J.-H. Yoon, L. Prakash, S. Prakash, and L. Haracska. Human HLTF functions as a ubiquitin ligase for proliferating cell nuclear antigen polyubiquitination. *Proceedings of the National Academy of Sciences*, 105(10):3768–3773, Mar. 2008.

65. S. Wang, H. Sun, J. Ma, C. Zang, C. Wang, J. Wang, Q. Tang, C. A. Meyer, Y. Zhang, and X. S. Liu. Target analysis by integration of transcriptome and ChIP-seq data with BETA. *Nature Protocols*, 8(12):2502–2515, Dec. 2013.

66. Z. Wang, Q. Zhang, W. Zhang, J.-R. Lin, Y. Cai, J. Mitra, and Z. D. Zhang. HEDD: Human Enhancer Disease Database. *Nucleic Acids Research*, 46(D1):D113–D120, Jan. 2018.

67. B. J. Wilkins and J. D. Molkentin. Calcineurin and cardiac hypertrophy: Where have we been? Where are we going? *The Journal of Physiology*, 541(1):1–8, May 2002.

68. L. Yao, B. P. Berman, and P. J. Farnham. Demystifying the secret mission of enhancers: linking distal regulatory elements to target genes. *Critical Reviews in Biochemistry and Molecular Biology*, 50(6):550–573, Nov. 2015.

69. X. Yi, Z. Zheng, H. Xu, Y. Zhou, D. Huang, J. Wang, X. Feng, K. Zhao, X. Fan, S. Zhang, X. Dong, Z. Wang, Y. Shen, H. Cheng, L. Shi, and M. J. Li. Interrogating cell type-specific cooperation of transcriptional regulators in 3D chromatin. *iScience*, 24(12):103468, Dec. 2021.

70. J. Zeitlinger. Seven myths of how transcription factors read the cis-regulatory code. *Current Opinion in Systems Biology*, 23:22–31, Oct. 2020.

71. C. Zhang, Z. Xu, S. Yang, G. Sun, L. Jia, Z. Zheng, Q. Gu, W. Tao, T. Cheng, C. Li, and H. Cheng. tagHi-C Reveals 3D Chromatin Architecture Dynamics during Mouse Hematopoiesis. *Cell Reports*, 32(13):108206, Sept. 2020.

72. G. Zhang, J. Shi, S. Zhu, Y. Lan, L. Xu, H. Yuan, G. Liao, X. Liu, Y. Zhang, Y. Xiao, and X. Li. DiseaseEnhancer: a resource of human disease-associated enhancer catalog. *Nucleic Acids Research*, 46(D1):D78–D84, Jan. 2018.