# The adapted Activity-By-Contact model for enhancer-gene assignment and its application to single-cell data

Dennis Hecker[1,2,3], Fatemeh Behjati Ardakani[1,2,3], Alexander Karollus[4], Julien Gagneur[4,5,6,7], 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 Centre for Cardiovascular Research, Partner site Rhein-Main, 60590 Frankfurt am Main, Germany**
**4 School of Computation, Information and Technology, Technical University of Munich, 85748 Garching, Germany,**
**5 Institute of Human Genetics, Technical University of Munich, 81675 Munich, Germany,**
**6 Computational Health Center, Helmholtz Center Munich, 85764 Neuherberg, Germany and**
**7 Munich Data Science Institute, Technical University of Munich, 85748 Garching, 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 is affected by the regulatory regions. A recent successful method is the Activity-By-Contact (ABC) model (Fulco et al., 2019) which scores enhancer-gene interactions based on enhancer activity and the contact frequency of an enhancer to its target gene. However, it describes regulatory interactions entirely from a gene's perspective, and does not account for all the candidate target genes of an enhancer. In addition, the ABC-model requires two types of assays to measure enhancer activity, which limits the applicability. Moreover, there is no implementation available that could allow for an integration with transcription factor (TF) binding information nor an efficient analysis of single-cell data. We demonstrate that the ABC-score can yield a higher accuracy by adapting the enhancer activity according to the number of contacts the enhancer has to its candidate target genes and also by considering all annotated transcription start sites of a gene. Further, we show that the model is comparably accurate with only one assay to measure enhancer activity. We combined our generalised ABC-model (gABC) with TF binding information and illustrate an analysis of a single-cell ATAC-seq data set of the human heart, where we were able to characterise cell type-specific regulatory interactions and predict gene expression based on transcription factor affinities. All executed processing steps are incorporated into our new computational pipeline STARE. The software is available at `https://github.com/schulzlab/STARE`.

## 1 Introduction

Unravelling the mechanisms behind gene expression regulation is a central task 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 have a variety of functions: recruit other cofactors, remodel chromatin conformation, cause changes in epigenetic modifications or directly interact with the transcription machinery, affecting gene expression (Pabo and Sauer, 1992; Gonzalez, 2016; Lambert et al., 2018). Many methods exist for annotating enhancers, *e.g.*, using open-chromatin assays like DNase-, ATAC- or NOMe-seq (Song and Crawford, 2010; Buenrostro et al., 2015; Kelly et al., 2012). Histone modifications

associated with enhancer activity like H3K27ac or H3K4me1 also aid enhancer annotation (Heintzman et al., 2009; Creyghton et al., 2010). Besides 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 (Yao et al., 2015; Schoenfelder and Fraser, 2019). Many approaches exist to predict target genes of enhancers, *e.g.*, using correlation of enhancer activity and gene expression (The FANTOM Consortium et al., 2014; Gao and Qian, 2019; Schmidt et al., 2021), or correlation of accessibility across samples (Pliner et al., 2018). Another possibility is to include chromatin contact data, *e.g.* Hi-C (Lieberman-Aiden et al., 2009), to call chromatin loops for annotating enhancer-gene links (Rao et al., 2014; Schmidt et al., 2020; Yi et al., 2021). Although loops correlate with gene expression, their anchors only cover a fraction of active promoters and enhancers and their removal impacts expression of only few genes (Nora et al., 2017; Rao et al., 2017; Schoenfelder and Fraser, 2019). Fulco et al. (2019) combined measurements of enhancer activity with chromatin contact data and proposed the Activity-By-Contact (ABC) model. The assumption is that active enhancers that frequently contact a gene's promoter are more likely to affect a gene's regulation. The ABC-model requires DNase-seq, H3K27ac ChIP-seq data and a Hi-C matrix. The Hi-C matrix can be substituted with a matrix averaged over multiple cell types, or with a quantitative function describing the distance-contact relationship (Fulco et al., 2019). 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 propose a generalised ABC-score with two adaptations: firstly it describes enhancer activity in a gene-specific manner, and secondly it uses the information of all annotated transcription start sites (TSSs) of a gene. Further, we could show that, instead of using both DNase- and H3K27ac ChIP-seq data for the ABC-model, one assay for measuring enhancer activity yields a similar accuracy. We validated the adaptations on three data sets of experimentally tested enhancer-gene interactions in K562 cells and on expression quantitative trait loci (eQTL) data from different tissues. We developed STARE, a fast implementation of the ABC-score with an approach to quantify TF binding affinity for genes. STARE can compute enhancer-gene interactions from single-cell chromatin accessibility data, illustrated on data of the human heart. With only one data modality at single-cell resolution, we identified and characterised cell type-specific regulatory interactions.

# 2 Methods

## 2.1 Activity-By-Contact score

We 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 likely to regulate it (Fulco et al., 2019). 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 $g$. For each candidate enhancer the activity is multiplied with the contact frequency and this product is taken relative to the sum over all candidate enhancers $R_g$ in a predefined window around the 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 may link to multiple enhancers and an enhancer may link to multiple genes. We tested three types of epigenomic assays to approximate $A_r$ by counting sequencing reads in the enhancer: DNase-seq, ATAC-seq, or H3K27ac ChIP-seq. The contact $C_{r,g}$ is taken from a normalised chromatin contact frequency matrix. A pseudocount is added to each $C_{r,g}$, so that all candidates $R_g$ are taken into account (see Supplementary Material).

## 2.2 Generalised Activity-By-Contact-score

We present a two-fold adaptation of the ABC-score, to account for all candidate target genes of an enhancer and for all transcription start sites (TSSs) of a gene. Regarding the former, 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 located within a predefined window around $r$. Since the values of $A_{r,g}$ are not known, we propose an approximation by using chromatin contacts:

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

We approximate the regulatory activity $A_{r,g}$ by the relative fraction of its contact $C_{r,g}$ to the TSS of $g$ to the sum over the contacts to all genes $G_r$ in a window around the enhancer. Thus, the ABC-score becomes:

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

In comparison to Eq. (1), the activity $A_r$ of an enhancer is replaced with its gene-specific relative activity $A_{r,g}$, see Eq. (3). Thus, this adapted score does not only function in a gene-centric way, 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 adaptation therefore uses a reduced enhancer activity estimate, in particular for enhancers in contact with many genes, and prevents them from being accounted with their full activity for all genes within the window.
In addition, instead of considering only one TSS per gene, e.g. the 5' TSS, we propose to include all TSSs of a gene in the following fashion:

$$gABC_{r,g} = \frac{\sum\limits_{t \in TSS_g} A_{r,t} \cdot C_{r,t}}{\sum\limits_{i \in R_g} \sum\limits_{t \in TSS_g} A_{i,t} \cdot C_{i,t}} \,, \tag{5}$$

where $TSS_g$ are all annotated TSSs of gene $g$. That allows to include the contact information to more potentially relevant transcription sites and omits the selection of an individual TSS. We name the score with these two changes the generalised ABC (gABC) score.

## 2.3 Validation on CRISPR-screens

To validate our generalised ABC-score and to test different assays for measuring enhancer activity, we examined the performance on experimentally validated enhancer-gene interactions. We made use of three CRISPRi-screens for K562 cells. Gasperini et al. (2019) used a single-cell CRISPRi-screen, introducing guide RNAs at a high multiplicity of infection, followed by single-cell RNA-seq. Schraivogel et al. (2020) developed targeted Perturb-seq (TAP-seq), which promises to be more sensitive by targeting genes of interest for the transcriptomic readout, and was established by a screening on two chromosomes. Fulco et al. (2019) used their CRISPRi-FlowFISH approach and collected data from other CRISPR-based studies. Unlike Fulco et al. (2019), we neither divide interactions into enhancer-gene and promoter-gene pairs nor exclude interactions, where the expression decreased after enhancer perturbation. We tested four different set-ups for measuring enhancer activity: 1) DNase-seq, 2) H3K27ac ChIP-seq, 3) the geometric mean of DNase-seq and H3K27ac ChIP-seq, 4) and ATAC-seq. In addition, we assessed a K562 Hi-C matrix, a Hi-C matrix averaged across 10 cell types constructed by Fulco et al. (2019), and a contact estimate (inverse of the linear distance) based on a fractal globule model (Lieberman-Aiden et al., 2009). For each combination we evaluated the ABC- and

gABC-scoring approach and calculated precision-recall curves. In addition, we tested the significance of the pairwise difference between the area under the receiver operate characteristic (ROC) curves (DeLong et al., 1988, Robin et al., 2011) (more details in Supplementary Material).
We also compared gABC to Enformer, a sequence-based deep learning model predicting gene expression and chromatin states, whose characteristic is an increased information flow between distal sequence positions (Avsec et al., 2021). To quantify enhancer-gene interactions with Enformer we compared the expression estimate upon *in silico* mutagenesis of the enhancer region, by either replacing 2 kb centred at the enhancer with neutral nucleotides, or by shuffling the sequence for 25 iterations (Karollus et al., 2022). Due to the size of Enformer's receptive field, we limited the comparison to interactions with $\leq$96 kb distance.

## 2.4 TF affinities and summarisation on gene-level

In addition to enhancer-gene interactions, STARE aims to describe a TF's regulatory impact on a gene (Fig. S1). 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. 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 (Roider et al., 2007). The higher the affinity, the more likely a TF is to bind. Retaining low binding affinities can hold valuable information (Schmidt et al., 2016, Kribelbauer et al., 2019) and omits selection of an arbitrary threshold. For all analyses presented in this manuscript we used a non-redundant collection of 818 human TF motifs in the form of position frequency matrices (PFMs) from JASPAR 2022, HOCOMOCO v11, and the work of Kheradpour and Kellis (Castro-Mondragon et al., 2022; Kulakovskiy et al., 2018; Kheradpour and Kellis, 2014). When converting PFMs to position-specific energy matrices required by TRAP, we take the average nucleotide content of the candidate regulatory regions as background.
To summarise the TF affinities in enhancers per gene, we combine them with the predicted enhancer-gene interactions. The summarisation depends on how the region-gene mapping was done. For the window-based approach, TF affinities in all open regions around the gene's TSS are summed:

$$af_{g,tf} = \sum_{r \in R_g} \frac{af_{r,tf}}{ml_{tf}} \cdot A_r \cdot e^{-\frac{d_{r,g}}{d_0}} , \tag{6}$$

where $af_{g,tf}$ is the affinity of TF $tf$ summarised for gene $g$. $R_g$ is the set of all open regions $r$ that were located within the window around $g$. $af_{r,tf}$ 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 Ouyang et al. (2009), where $d_0$ is set to a constant of 5000 bp.
When the gABC-score was used to assign regions to genes, the summarisation changes, as there is more epigenomic information available. Regions close to the TSS ($\leq$2500 bp) are always included, independently of their gABC-score, as they are very informative for the expression regulation of a gene (Schmidt et al., 2019). They are also scaled differently:

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

$R_g$ is the set of regions that was linked to the gene with the gABC-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 using the regular ABC-score, the affinity scaling changes as follows:

$$af_{g,tf} = \sum_{r \in R_g} \frac{af_{r,tf}}{ml_{tf}} \cdot \begin{cases} A_r \cdot e^{-\frac{d_{r,g}}{d_0}}, & \text{if } d_{r,g} \leq 2500\ bp \\ A_r \cdot \frac{C_{r,g}}{C_{max}}, & \text{otherwise} \end{cases} . \tag{8}$$

The regions close to the TSS are scaled in the same way as before. For all the 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 gABC and ABC differ only in how the contact is scaled: for gABC 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 three additional gene features: the number of regions considered per gene, the regions' average distance to the TSS, and the regions' average base pair length, as all three can be predictive of gene expression (Schmidt and Schulz, 2019).

## 2.5 Application to single-cell data

To test the capability of gABC-scored enhancer-gene interactions combined with TF affinities to capture regulatory cell type-specific information, we analysed a single-cell data set of the human heart from Hocker et al. (2021), providing single-nuclei (sn) ATAC-seq, as well as snRNA-seq data. The candidate enhancers were pooled and ATAC-seq RPKM was measured for each cell type. For chromatin contacts we tested H3K27ac HiChIP data of the left ventricle (Anene-Nzelu et al., 2020), as well as an average Hi-C matrix (Fulco et al., 2019). Regions known to accumulate an anomalous amount of sequencing reads were excluded (Amemiya et al., 2019, The ENCODE Project Consortium, 2012). As we had enhancer activity and enhancer contact data at hand, we determined enhancer-gene interactions for each cell type with the gABC-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. To assess the predictability of gene expression by gene-TF affinities, we used INVOKE (Schmidt et al., 2016), which implements a linear regression model based on gene-TF scores, and selects TFs that are most predictive of gene expression. We trained prediction models for multiple set-ups for each cell type to compare their prediction accuracy. We repeated this process for cell type-specific (CS) genes, defined as genes where the z-score of expression between cell types was $\geq 2$ and TPM $\geq 0.5$. The gene-TF matrices were limited to TFs expressed in a cell type (TPM $\geq 0.5$). Schmidt et al. (2020) 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 can integrate the contact frequency into the summarisation of TF affinities.

### 2.5.1 Comparison to co-accessibility analysis

A common approach to identify regulatory interactions in single-cell ATAC-seq data is to call co-accessible regions. Hocker et al. (2021) ran Cicero (Pliner et al., 2018) on their snATAC-seq data to derive pairs of regions with correlated accessibility, limited to a distance of 250 kb. Whenever either side of a co-accessible region pair overlapped a 400 bp window around any annotated TSS of a gene, we considered it as an enhancer-gene interaction for that gene. We tested how informative the resulting 62,384 co-accessible interactions are for our gene expression prediction model. Summarisation of TF affinities was done according to Eq. 6, and the affinities of regions close to the TSS ($\leq 2500$ bp) were included, as described in Sec. 2.4.

### 2.5.2 Intersection with eQTLs

Further, we compared the agreement of the regular ABC- and gABC-score with eQTL data. We intersected ABC-scored interactions from four different heart cell types (Hocker et al., 2021) and K562 cells with eQTL-gene pairs of matching samples from the GTEx portal (The GTEx Consortium, 2020). We used high confidence eQTL-gene pairs from three different fine-mapping approaches, namely CAVIAR, CaVEMaN and DAP-G (Hormozdiari et al., 2014, Brown et al., 2017, Wen et al., 2016). For each set of eQTLs, we defined the enhancer-gene pairs that were supported by the eQTLs, meaning all candidate enhancers of a cell type with a variant, where the affected target gene was within the chosen ABC window size. Then, we compared the fraction of those eQTL-supported enhancer-gene pairs that we could also find among a variable number of highest scored ABC/gABC interactions (Recall). As we also had the co-accessibility analysis on the heart data, we examined how many eQTL-gene pairs the resulting interactions recover.

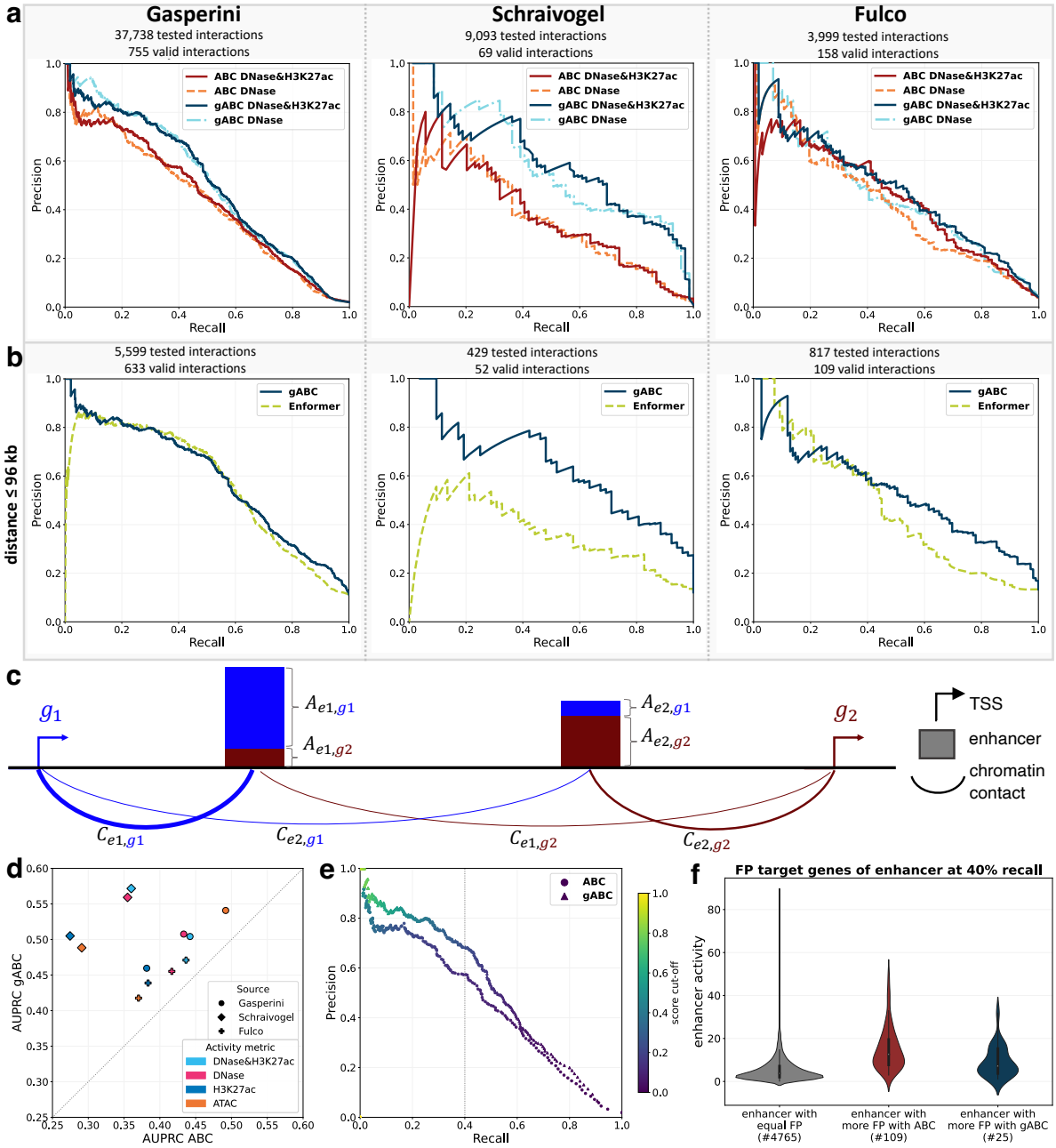


**Figure 1.** Performance comparison of the ABC- and gABC-score on CRISPRi-screens in K562 cells using different epigenomic assays. For chromatin contacts a K562 Hi-C matrix (5 kb resolution) was used (Rao et al., 2014). **(a)** Precision-recall (PR) curves of both ABC-scores on experimentally validated enhancer-gene links. The AUPRC values can be found in Tab. 1. **(b)** PR curves comparing gABC with Enformer on interactions with a distance of $\leq$96 kb. Out of four tested calculations for the predicted expression difference of Enformer the best one is shown. The AUPRC values are listed in Tab. S4. DNase-seq and H3K27ac ChIP-seq was used as activity for gABC. **(c)** Schema for the gene-specific enhancer activity, which distributes the activity of an enhancer among its scored genes, dependent on contact frequencies. **(d)** Direct comparison of the AUPRC for ABC and gABC on different CRISPRi-screens and with different assays for enhancer activity. **(e)** PR curve coloured by ABC/gABC score respectively with DNase&H3K27ac as activity on the CRISPRi-screen of Gasperini et al. (2019). The dotted grey line marks the position at 40% recall. **(f)** Distribution of the activity of enhancers (geometric mean of read counts of DNase-seq and H3K27ac ChIP-seq) separated by the number of false positive (FP) target genes called by each method at 40% recall on the screen of Gasperini et al. (2019). 'equal FP' contains all enhancers where the number of FP target genes is the same for both scores (0 FP included). 'more FP' means that either of the scores called more FP target genes for that enhancer. The number of enhancers in each category is shown below the x-axis labels.

# 3 Results

## 3.1 Generalised ABC-score improves interaction prediction

We propose a generalised ABC-score (Eq. 5), where the activity of an enhancer is described in a gene-specific manner (Fig. 1c, Eq. 3), and all annotated TSSs of a gene are considered. On all validation data sets and for all combinations of activity measurements the gABC-score outperformed the regular ABC-score (p-value≈0.0005 Wilcoxon signed-rank test) (Fig. 1a+d; Tab. 1). The difference was more pronounced in the Gasperini and Schraivogel validation data. Each of the two adaptations of the gABC-score individually increased the area under the precision recall curve (AUPRC) compared to the regular ABC across activity assays, with the gene-specific activity providing an average gain of 0.026, and including all TSSs giving an average improvement of 0.053. Taken together, the gABC-score yielded on average a 0.107 higher AUPRC (Tab. S1). The areas under the ROC curves for gABC were significantly higher in ten out of twelve pairwise comparisons across CRISPRi-screens and activity assays (Tab. S5). Using an average Hi-C matrix changed the accuracy marginally for both ABC-scores (Fig. S2a; Tab. 1). When using the fractal globule module to estimate contact frequency only based on distance, gABC achieved less improvement over the regular ABC with an average AUPRC gain of 0.048 (Tab. S1). We could reproduce the higher accuracy of the gABC-score in a direct comparison to the implementation of Fulco et al. (2019) (Tab. S2). Further, we examined the correlation between each of the ABC-scoring approaches and the absolute change in gene expression as measured in the CRISPRi-screens. The gABC-score showed a higher Spearman correlation coefficient across all three data sets (Tab. S3).

To disentangle for which enhancers the gABC-score performs better than the regular ABC, we focused on the largest CRISPRi-screen from Gasperini et al. (2019) (Fig. S3a) and found that the regular ABC predicted more false positive target genes for enhancers with a high activity (Fig. 1e+f, Fig. S3b-e). There was a small subset of enhancers in gene-rich regions for which the gABC-score predicted more false positive interactions (Fig. S3f-i).

Enformer is a novel method that predicts gene expression and chromatin states directly from the DNA sequence using a complex neural network architecture outperforming other sequence-based models (Avsec et al., 2021). We compared the gABC-score and an Enformer model learned on K562 cells using *in silico* mutagenesis, where the strength of interactions was quantified by the predicted expression change upon sequence perturbation of the enhancer. gABC achieved a higher accuracy on all three validation data sets (Fig. 1b) testing alternative ways for the *in silico* mutagenesis (Tab. S4). This was also reflected in significantly higher areas under the ROC curves (all p-values $\leq 0.05$, DeLong et al., 1988, Robin et al., 2011).

Although it uses the same information, the gABC-score performed better than the regular ABC-score and outperformed the accurate sequence-based Enformer model.

## 3.2 One activity assay yields similar accuracy

The original formulation of the ABC-score requires two types of assays to measure enhancer activity, namely DNase-seq and H3K27ac ChIP-seq (Fulco et al., 2019). We tested the performance of the gABC-scoring principle using different assays for measuring enhancer activity (Fig. 1a+d; Tab. 1; Fig. S2b). On the data sets of Schraivogel and Fulco the combination of DNase-seq and H3K27ac ChIP-seq performed better than either of them alone, although the performance drop with only DNase-seq was small. On the Gasperini data, DNase-seq without H3K27ac ChIP-seq was slightly better. Quantifying enhancer activity with ATAC-seq resulted in a worse performance on the Schraivogel and Fulco validation data, but a clear improvement on the Gasperini data set, surpassing the combination of DNase-seq and H3K27ac ChIP-seq.

## 3.3 Regulatory interactions in single-cell data

Using just one epigenome assay for the ABC-score enables direct application on high-resolution snATAC-seq data, which has the potential to improve the prediction of cell type-specific interactions. As a proof of concept, we analysed a human heart snATAC-seq data set, comprising eight cell type clusters (Hocker et al., 2021). The candidate enhancers of the cell types were pooled to a set of 286,777 regions with a summarised

**Table 1.** AUPRC using ABC and gABC for identifying regulatory interactions on three validation data sets, with different assays for enhancer activity and contact data. The highest AUPRC within each column is written in bold.

| Validation data | Gasperini et al. (2019) 755 valid / 37,738 interactions | | | | Schraivogel et al. (2020) 69 valid / 9,093 interactions | | | | Fulco et al. (2019) 158 valid / 3,999 interactions | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Enhancer activity assay | DNase& H3K27ac | DNase | H3K27ac | ATAC | DNase& H3K27ac | DNase | H3K27ac | ATAC | DNase& H3K27ac | DNase | H3K27ac | ATAC |
| ABC K562 Hi-C | 0.4421 | 0.4333 | 0.3816 | 0.4919 | 0.3600 | 0.3549 | 0.2746 | 0.2909 | 0.4365 | 0.4167 | 0.3835 | 0.3702 |
| gABC K562 Hi-C | **0.5042** | **0.5076** | 0.4596 | **0.5408** | **0.5717** | **0.5592** | **0.505** | **0.4884** | **0.471** | **0.4552** | **0.4388** | **0.4176** |
| ABC avg Hi-C | 0.4486 | 0.4395 | 0.3898 | 0.4929 | 0.3554 | 0.3661 | 0.2806 | 0.2945 | 0.436 | 0.4085 | 0.386 | 0.3663 |
| gABC avg Hi-C | 0.5038 | 0.5072 | **0.4605** | 0.5364 | 0.5552 | 0.5507 | 0.4985 | 0.475 | 0.452 | 0.4387 | 0.4259 | 0.4015 |

ATAC-seq measurement for each defined cell type (Fig. 2a). On average we predicted 408,846 gABC-scored interactions (SD $\approx$ 12,200) over all cell types. Approximately 23.6% of a cell type's interactions were shared with all other cell types (Fig. 2b; Fig. S4a). Each cell type also featured unique interactions although this was highly variable ($\mu \approx$11.6%, SD$\approx$5.1%). Atrial cardiomyocytes (aCM) and ventricular cardiomyocytes (vCM) formed the largest intersection of interactions found in only two cell types, consisting of 39,707 interactions. The average median of enhancers per expressed gene (TPM $\geq$ 0.5) across cell types was 4.75 (SD $\approx$ 0.43) (Fig. S4b). Despite all cell types having the same candidate enhancers and a shared contact measurement, their predicted enhancer-gene interactions appeared to be considerably distinct.

### 3.3.1 gABC-score recovers more eQTL-gene pairs

We examined how many eQTLs from different tissues are recovered by ABC interactions from matching cell types, including four heart cell types from Hocker et al. (2021) and K562 cells. We took high confidence eQTL-gene pairs from three different fine-mapping methods and compared which fraction of enhancer-gene interactions supported by eQTLs were also found by the 300,000 highest scored ABC- and gABC-interactions (Fig. 2c). The gABC-interactions recovered significantly more eQTL-gene pairs across all fine-mapping methods and eQTL data sets (p-value$\approx$0.0005 Wilcoxon signed-rank test). This finding was reproduced for the 100,000 and 200,000 highest scored interactions. The gABC-interactions also captured more eQTL-gene pairs than interactions derived via co-accessibility analysis (Fig. S4c).

### 3.3.2 Linking epigenetic features to cell type-specific expression

We trained cell type-specific gene expression prediction models based on gene-TF affinity matrices, constructed with different approaches (Fig. 3a). Using the gABC-score performed best across all cell types (Pearson correlation coefficient $\mu \approx$0.563), with similar results using the average Hi-C matrix ($\mu \approx$0.556). The regular ABC-score had a slightly lower performance ($\mu \approx$0.538), whereas interactions identified via co-accessibility resulted in the lowest performance ($\mu \approx$0.505). Since the co-accessible interactions were limited to a distance of 250 kb, we also tested the gABC-score in a 500 kb window, which marginally decreased the performance in the prediction model ($\mu \approx$0.562, Fig. S5a).
We defined a set of CS genes (TPM $\geq$ 0.5 and z-score $\geq$ 2) for each cell type ($\mu$=1,404 genes, SD$\approx$582 genes) and analysed those in more detail. The CS genes were mostly unique to a cell type (Fig. S4d), and GO term enrichment returned cell-type appropriate terms (Fig. S4e). To further characterise the sets of CS genes, we examined additional attributes and compared CS genes to non-CS genes, both sets restricted to expressed genes (TPM $\geq$ 0.5) (Fig. 3b; Fig. S4f). CS genes had more assigned enhancers than non-CS genes in all cell types except for LC. This matches the finding of Fulco et al. (2019), who described a higher number of enhancers for tissue-specific 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. Most cell types had a slightly higher average activity in enhancers linked to CS genes than in enhancers linked to non-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. To exclude that the differences were solely caused by the CS genes' higher expression, we repeated the comparisons but restricted

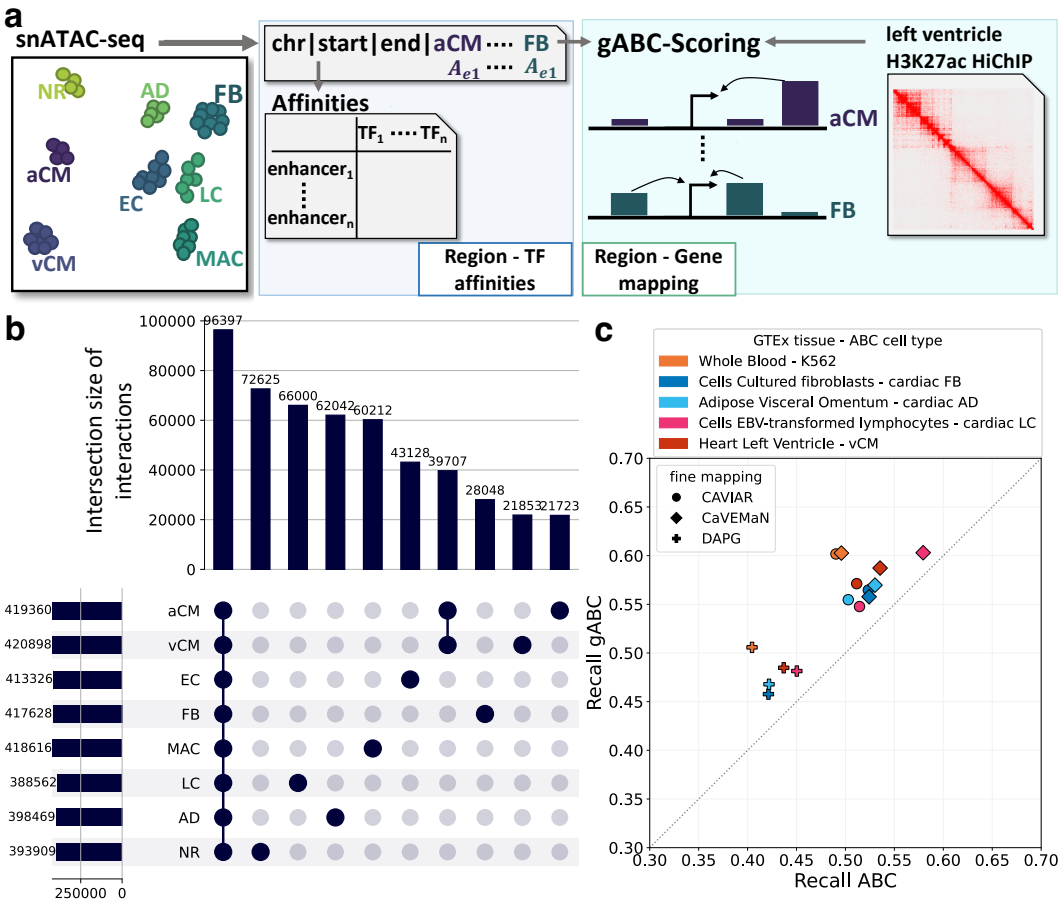


**Figure 2.** Enhancer-gene interactions called in single-cell heart data. **(a)** Schema of data processing. snATAC-seq from Hocker et al. (2021) was used to identify candidate enhancers and their activity in the annotated cell types. For each cell type gABC-interactions were called. Enhancer-gene contacts were retrieved from left ventricle H3K27ac HiChIP data (Anene-Nzelu et al., 2020). **(b)** Upset plot of the enhancer-gene interactions called in each cell type. Only the 10 largest intersections are shown. **aCM**: atrial cardiomyocytes; **vCM**: ventricular cardiomyocytes; **EC**: endothelial cells; **FB**: fibroblasts; **MAC**: macrophages; **LC**: lymphocytes; **AD**: adipocytes; **NR**: nervous cells. **(c)** Intersection of eQTL-gene pairs from different GTEx samples with the ABC- and gABC-interactions. Recall is the fraction of enhancer-gene pairs found by each score out of all pairs where the enhancer contained an eQTL whose target gene was within the window size used for ABC-scoring. The 300,000 highest scored interactions of ABC and gABC were used.

them to CS and non-CS genes from the upper quartile of TPM values. The results were highly similar across all features (data not shown). We repeated the training of gene expression models but limited it to CS genes and observed an overall decrease in prediction accuracy (Fig. S5b). The gABC-score still performed better than the ABC-score. The prediction model returns a regression coefficient for each TF, indicative of TF relevance, and allows to investigate which TFs might drive gene expression in which cell type (Fig. S5c). Overall, we were able to characterise cell type-specific expression regulation based on single-cell chromatin accessibility and bulk chromatin contact data.

## 3.4 Runtime of STARE

We optimised STARE's runtime by allowing multiple steps to be run in parallel (Fig. 4a) and to omit redundant calculations, if multiple cell types/metacells/individual cells with a respective activity measurement 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. 4b).

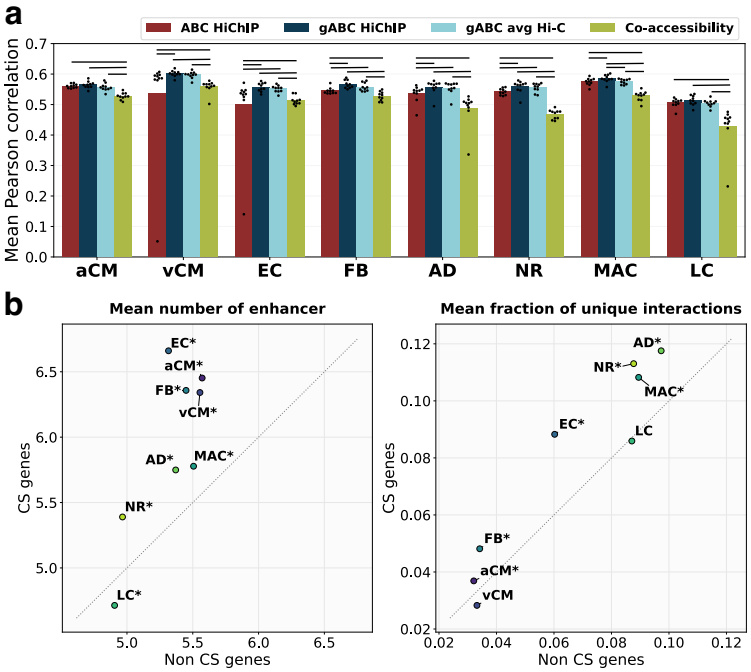


**Figure 3.** Gene expression prediction on single-cell data and characterisation of CS genes. **(a)** Accuracy of a gene expression prediction model based on different gene-TF affinity matrices. The model was trained on all genes with available expression values. The Pearson correlation coefficient is shown as average over a 10-fold outer cross-validation. A horizontal line above the bars indicate significance (p-value $\leq 0.05$, Mann-Whitney U test). snATAC-seq data was used as enhancer activity for all approaches. ABC/gABC H3K27ac HiChIP: regular/generalised ABC scoring with H3K27ac HiChIP as contact data; gABC avg Hi-C: gABC with an average Hi-C matrix as contact data; Co-accessibility: enhancer-gene links defined by Cicero (Pliner et al., 2018), see Sec. 2.5.1. The respective Spearman correlation coefficients, as well as additional approaches and their performance in a training on CS genes only, are presented in Fig. S5. **(b)** Comparison of attributes between CS (TPM $\geq 0.5$ and z-score $\geq 2$) and not cell type-specific (non-CS) genes (TPM $\geq 0.5$) (*: p-value $\leq 0.05$, Mann-Whitney U test). See Fig. 2 for cell type abbreviations.

# 4 Discussion

We present a variation of identifying regulatory enhancer-gene interactions with the ABC-model. In its original study the ABC-model already showed a better accuracy for detecting validated enhancer-gene links than other approaches (Fulco et al., 2019). We propose a generalised ABC-score where the activity of an enhancer is described in a gene-specific manner by weighting it relative to the number of enhancer-gene contacts and where all TSSs of a gene are considered. This generalisation resulted in an improvement in identifying experimentally validated enhancer-gene interactions, both compared to the regular ABC, as well as to the deep learning model Enformer. It should be noted however, that Enformer was built to predict gene expression, and not to identify enhancer-gene links (Avsec et al., 2021). The accuracy of different scoring approaches was partially inconsistent between validation data sets, especially when using ATAC-seq for enhancer activity. Although all data sets are based on a dCas9-KRAB system, there are differences in the experimental set-ups and scope. Gasperini et al. (2019) introduced multiple guide RNAs and measured expression via single-cell RNA-seq, allowing quantification of interactions for over 10,000 genes. Schraivogel et al. (2020) presented their method TAP-seq, while Fulco et al. (2019) published CRISPRi-FlowFISH and collected previous results, both evaluating interactions for less than one hundred genes. Details in candidate enhancer selection, filtering steps, and processing might lead to biases or varying sensitivity in the identification of enhancer-gene pairs, which is already indicated by the different fractions of significant interactions. Further, it is debatable whether CRISPRi-screens are able to detect all regulatory interactions, as true enhancers with small effect sizes may be overlooked. Moreover, perturbations of individual enhancers are presumably not capable of accounting for shadow enhancers with redundant functionality (Schoenfelder and Fraser, 2019; Singh and Yi, 2021). More large-scale validation data could consolidate and contextualise our findings, but are currently unavailable. Importantly, any model to annotate enhancer-gene interactions is only a prediction and likely not capturing the whole regulatory complexity of genes. The ABC-model

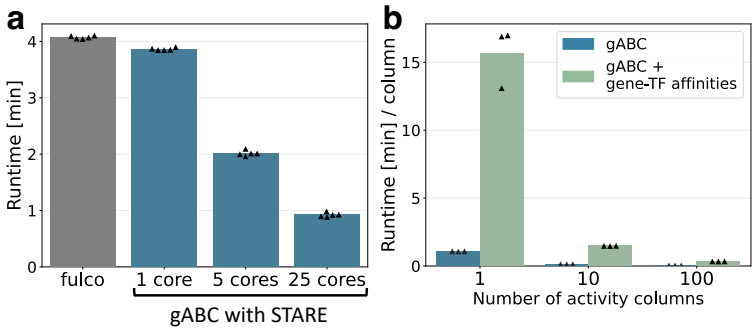


**Figure 4.** Runtime of the STARE pipeline. **(a)** Runtime of the original ABC pipeline by (Fulco et al.) and our ABC-scoring implementation with a different number of cores. Any writing of output files was omitted. Calculations were done for K562 cells, scoring 155,976 candidate enhancers for 24,586 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 (Hocker et al., 2021) with 286,777 candidate enhancers, for 55,765 genes in a 5 MB window. 818 TFs were assessed. The bars' height represent the mean of three runs.

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 all genes are regulated in the same way.

In context of required data types, we were able to demonstrate that, unlike suggested in the original work, one assay for measuring enhancer activity works similarly well, specifically using DNase- or ATAC-seq without H3K27ac ChIP-seq data. Further, using averaged contact data yielded a high performance as well, which broadens the applicability of the ABC-score to all data sets, where a measurement of enhancer activity is available. Especially for single-cell epigenomics it is challenging to measure multiple modalities in the same cell.

We present the STARE framework to derive gene-TF affinities. After mapping candidate enhancers to genes, using either the ABC-score or a window-based approach, STARE summarises TF affinities on a gene level. Unlike other methods aiming to determine regulatory relations of TFs to genes (Lan et al., 2012; Wang et al., 2013), STARE does not require scarce TF ChIP-seq data. It uses a motif-based biophysical model (Roider et al., 2007) to determine TF affinities in accessible regions. As consequence, unlike other methods (McLeay et al., 2012), it is able to retain low affinity binding information. Patel and Bush (2021) use similar data as STARE and rely on a graph-based approach, but they do not incorporate active regulatory regions and 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. Notably, our model assumes an additive influence of TFs, which is not likely to accurately capture biological reality (Zeitlinger, 2020). Furthermore, there is potential redundancy on two levels: the aforementioned functional redundancy of enhancers and the redundancy of TF binding motifs (Gitter et al., 2009; Cusanovich et al., 2014).

We applied our framework to single-cell data with clustered cell types of the human heart (Hocker et al., 2021), where the activity of a unified 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 (Fraser et al., 2009; Zhang et al., 2020), the 3D conformation of the genome is described as less dynamic and more as a scaffold to enable and stabilise regulatory interactions (Schoenfelder and Fraser, 2019; Ing-Simmons et al., 2021). 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. We demonstrated a downstream application example of STARE with a linear expression model, which allows to identify candidate regulatory TFs for further evaluation. Considering the small training sets and that the model assumes the same regulation for all genes, the prediction yielded a reasonable performance. Schmidt et al. (2020) used a similar expression prediction approach and incorporated distant enhancers. However, their work relies on annotated loops, which are not likely to cover all relevant regulatory interactions. In addition, their strategy is not able to integrate contact frequencies into TF affinities, nor to derive cell type-specific interactions from bulk contact data. There are other tools for predicting gene expression explicitly in individual cells

that incorporate TF information, such as SCENIC (Aibar et al., 2017), ACTION (Mohammadi et al., 2018) or TRIANGULATE (Behjati Ardakani et al., 2020), 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.

## Availability and genome version

The presented results are provided via Zenodo (`https://doi.org/10.5281/zenodo.5841991`). All data is in hg19. For details on the used data see Supplementary Material.


## 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), the Cardio-Pulmonary Institute (CPI) [EXC 2026] ID: 390649896, the DFG SFB (TRR 267) Noncoding RNAs in the cardiovascular system, Project-ID 403584255 and by MERGE: German Bundesministerium für Bildung und Forschung (BMBF) through the Model Exchange for Regulatory Genomics project MERGE (031L0174A).


## References


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

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

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

Avsec, , Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P., and Kelley, D. R. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18(10):1196–1203.

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

Brown, A. A., Viñuela, A., Delaneau, O., Spector, T. D., Small, K. S., and Dermitzakis, E. T. (2017). Predicting causal variants affecting expression by using whole-genome sequencing and RNA-seq from multiple human tissues. *Nature Genetics*, 49(12):1747–1751.

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

Castro-Mondragon, J. A., Riudavets-Puig, R., Rauluseviciute, I., Berhanu Lemma, R., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Manosalva Pérez, N., Fornes, O., Leung, T., Aguirre, A., Hammal, F., Schmelter, D., Baranasic, D., Ballester, B., Sandelin, A., Lenhard, B., Vandepoele, K., Wasserman, W. W., Parcy, F., and Mathelier, A. (2022). JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 50:D165–D173.

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

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

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3):837.

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

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., Nguyen, T. H., Kane, M., Perez, E. M., Durand, N. C., Lareau, C. A., Stamenova, E. K., Aiden, E. L., Lander, E. S., and Engreitz, J. M. (2019). Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nature Genetics*, 51(12):1664–1669.

Gao, T. and Qian, J. (2019). 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.

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., Trapnell, C., Ahituv, N., and Shendure, J. (2019). A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell*, 176(1-2):377–390.

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

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

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

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

Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B., and Eskin, E. (2014). Identifying Causal Variants at Loci with Multiple Signals of Association. *Genetics*, 198(2):497–508.

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

Karollus, A., Mauermeier, T., and Gagneur, J. (2022). Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. *bioRxiv*.

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

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

Kribelbauer, J. F., Rastogi, C., Bussemaker, H. J., and Mann, R. S. (2019). Low-affinity binding sites and the transcription factor specificity paradox in eukaryotes. *Annual Review of Cell and Developmental Biology*, 35(1):357–379. PMID: 31283382.

Kulakovskiy, I. V., Vorontsov, I. E., Yevshin, I. S., Sharipov, R. N., Fedorova, A. D., Rumynskiy, E. I., Medvedeva, Y. A., Magana-Mora, A., Bajic, V. B., Papatsenko, D. A., Kolpakov, F. A., and Makeev, V. J. (2018). HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-seq analysis. *Nucleic Acids Research*, 46:D252–D259.

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

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

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

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

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

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

Ouyang, Z., Zhou, Q., and Wong, W. H. (2009). 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.

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

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

Pliner, H. A., Packer, J. S., McFaline-Figueroa, J. L., Cusanovich, D. A., Daza, R. M., Aghamirzaie, D., Srivatsan, S., Qiu, X., Jackson, D., Minkina, A., Adey, A. C., Steemers, F. J., Shendure, J., and Trapnell, C. (2018). Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. *Molecular Cell*, 71(5):858–871.e8.

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

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

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). pROC: an open-source package for r and s+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12(1):77.

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

Schmidt, F., Gasparoni, N., Gasparoni, G., Gianmoena, K., Cadenas, C., Polansky, J. K., Ebert, P., Nordström, K., Barann, M., Sinha, A., Fröhler, S., Xiong, J., Dehghani Amirabad, A., Behjati Ardakani, F., Hutter, B., Zipprich, G., Felder, B., Eils, J., Brors, B., Chen, W., Hengstler, J. G., Hamann, A., Lengauer, T., Rosenstiel, P., Walter, J., and Schulz, M. H. (2016). Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. *Nucleic Acids Research*, 45(1):54–66.

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

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

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

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

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

Schraivogel, D., Gschwind, A. R., Milbank, J. H., Leonce, D. R., Jakob, P., Mathur, L., Korbel, J. O., Merten, C. A., Velten, L., and Steinmetz, L. M. (2020). Targeted perturb-seq enables genome-scale genetic screens in single cells. *Nature Methods*, 17(6):629–635.

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

Song, L. and Crawford, G. E. (2010). 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.

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

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

The GTEx Consortium (2020). The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science*, 369:1318–1330.

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

Wen, X., Lee, Y., Luca, F., and Pique-Regi, R. (2016). Efficient integrative multi-SNP association analysis via deterministic approximation of posteriors. *The American Journal of Human Genetics*, 98(6):1114–1129.

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

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

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

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