Assessing NGS-based computational methods for predicting transcriptional regulators with query gene sets

Zeyu Lu[1], Xue Xiao[2], Qiang Zheng[4], Xinlei Wang[3,4]* & Lin Xu[2,5]*

1. Department of Statistics and Data Science, Moody School of Graduate and Advanced Studies, Southern Methodist University, Dallas, TX, USA

2. Quantitative Biomedical Research Center, Peter O'Donnell Jr. School of Public Health, University of Texas Southwestern Medical Center, Dallas, TX, U∫SA

3. Department of Mathematics, University of Texas at Arlington, Arlington, TX 76019, USA

4. Center for Data Science Research and Education, College of Science, University of Texas at Arlington, Arlington, TX 76019, USA

5. Department of Pediatrics, Division of Hematology/Oncology, University of Texas Southwestern Medical Center, Dallas, TX, USA

Corresponding author:

Lin Xu (lin.xu@utsouthwestern.edu)

Xinlei Wang (xinlei.wang@uta.edu)

Abstract: This article provides an in-depth review of computational methods for predicting transcriptional regulators with query gene sets. Identification of transcriptional regulators is of utmost importance in many biological applications, including but not limited to elucidating biological development mechanisms, identifying key disease genes, and predicting therapeutic targets. Various computational methods based on next-generation sequencing (NGS) data have been developed in the past decade, yet no systematic evaluation of NGS-based methods has been offered. We classified these methods into two categories based on shared characteristics, namely library-based and region-based methods. We further conducted benchmark studies to evaluate the accuracy, sensitivity, coverage, and usability of NGS-based methods with molecular experimental datasets. Results show that BART, ChIP-Atlas, and Lisa have relatively better performance. Besides, we point out the limitations of NGS-based methods and explore potential directions for further improvement.

Key words: transcriptional regulator, benchmarking, query gene set, next-generation sequencing, prediction

30 **Key points**

31 - An introduction to available computational methods for predicting functional TRs from a
32 query gene set.
33 - A detailed walk-through along with practical concerns and limitations.
34 - A systematic benchmark of NGS-based methods in terms of accuracy, sensitivity,
35 coverage, and usability, using 570 TR perturbation-derived gene sets.
36 - NGS-based methods outperform motif-based methods. Among NGS methods, those
37 utilizing larger databases and adopting region-centric approaches demonstrate favorable
38 performance. BART, ChIP-Atlas, and Lisa are recommended as these methods have
39 overall better performance in evaluated scenarios.

Introduction

In the 1960s, foundational principles of gene transcriptional regulation were established through seminal work [1] and subsequent studies, which highlighted the pivotal role of transcriptional regulators (TRs) in controlling the transcription of hundreds or even thousands of genes [2, 3]. TRs are proteins encoded by genes and their regulation of transcription directly influences the expression levels of genes, playing a crucial role in various cellular processes including cell growth, differentiation, morphogenesis, and death [4-6]. A significant portion of these TRs are transcription factors (TFs), which bind to specific DNA sequences. The remainder of TRs consists of cofactors recruited by TFs and chromatin regulators that can alter the structure and function of chromatin [7, 8].

Evidence has linked multiple known diseases to the dysfunction of specific TRs [9, 10], highlighting the importance of identifying functional TRs in biological and medical research, which can aid the discovery of biomarker genes or potential therapeutic targets [11-13]. However, accurately determining the interactions between TRs and their targets requires significant computational and experimental efforts. Amid these challenges, one intriguing question that draws many researchers' attention is how to accurately predict the TRs that regulate a user-provided gene set that can be derived from sources such as differentially expressed gene (DEG) analysis, gene ontology analysis, and pathway enrichment analysis.

In the past, a variety of computational approaches have been proposed to answer the aforementioned question based on microarrays and sequencing technologies. Among them, motif enrichment analysis was initially developed by associating over-represented transcription factor binding motifs (TFBMs) with cis-regulatory elements (e.g., promoters and enhancers), which regulate the transcription of nearby genes [14]. TFBMs are short DNA subsequences that have

high affinities for specific TFs. If the TFBMs are over-represented, the related TFs are claimed to be functional regulators. Many methods have been developed following this rationale [15-17]. However, these methods share the following limitations: First, the high similarities of TFBMs among distinct TFs often make it challenging to decide which TFs actually bind the genomic loci [18]. Second, many TFs lack known binding motifs [19], which could not be analyzed by these published motif-based methods. Third, TFBMs are unable to capture the context-specific binding profile of a TF, which leads to a high false-positive rate in determining the actual binding sites in a particular cellular process. Therefore, motif enrichment analysis is not an ideal tool to search for TRs.

We further mention that the identification of TRs can serve as a downstream application of network construction analysis as well [20-22]. The realm of gene regulatory network reconstruction stands as a distinct research field, with comprehensive reviews providing deep insights into this domain [23-25]. Thus, these network construction methods were excluded from this review.

This review instead focuses on computational methods that can leverage recent advancement of next-generation sequencing (NGS) techniques, especially epigenomic techniques (e.g., ChIP-seq, ATAC-seq, and DNase-seq) to identify TRs that regulate a user-provided gene set. These methods are referred as NGS-based method in this review. Unlike motif-based methods, which screen the genome by computationally inferred motifs to find possible binding sites, using NGS data can provide evidence of actual binding sites within the genome.

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) including TR ChIP-seq and Histone ChIP-seq is a widely accepted technique to examine genome-wide DNA-protein interactions. TR ChIP-seq datasets serve as the foundation of all NGS-based methods. A TR

ChIP-seq dataset can accurately record the binding sites of one specific TR [26], and therefore makes it possible to establish the TR-gene connection.

Additionally, histone ChIP-seq techniques (e.g., H3K27ac ChIP-seq) and chromatin accessibility profiling techniques (e.g., DNase-seq, FAIRE-seq, and ATAC-seq) can measure genome-wide active cis-regulatory elements or accessible chromatin regions [27, 28]. Unlike TR ChIP-seq data required by all methods, data generated by these techniques cannot be directly associated with one specific TR, and so are only included by some methods for improving the prediction accuracy.

The interesting features extracted from these data are commonly termed as “peaks” [29], which are short epigenomic regions that show a significant enrichment of epigenetic marks or factors compared to the background levels. Peaks can indicate different features depending on the NGS techniques, such as TR binding sites (TR ChIP-seq), accessible chromatin regions (ATAC-seq, DNase-seq, FAIRE-seq), or sites of active promoters and enhancers (H3K27ac ChIP-seq).

Until now, over ten computational methods have been developed to predict TRs using NGS data. These methods are Cscan [30], ENCODE ChIP-seq significance tool[31], Enrichr [32], i-cisTarget [33], RegulatorTrail [34], BART [35], ChIP-Atlas [36], TFEA.ChIP [37], ChEA3 [38], MAGIC [39], and Lisa [40]. **Figure 1** illustrates the common workflow of these methods. The input is typically a query set of genes provided by a user. The background set comprises other genes in the same process that generates the query gene set (e.g., the query gene set consists of DEGs and the background set consists of non-DEGs). The reference database, composed of NGS data, functions as a unique 'fingerprint' repository for TRs and genes. Each TR ChIP-seq dataset in the database represents a distinct TR 'fingerprint'. Users can compare a 'fingerprint' from a new cellular study against ‘fingerprints’ in this database and identify functional TRs based on their similarity. The methods used integrate input, background, and reference data to produce a TR

ranking list, where higher ranks indicate a greater likelihood of the TR regulating the input gene set. In this sense, the methods discussed in this review mainly act as a ranking algorithm to rank TRs covered by their internal database based on the user input.

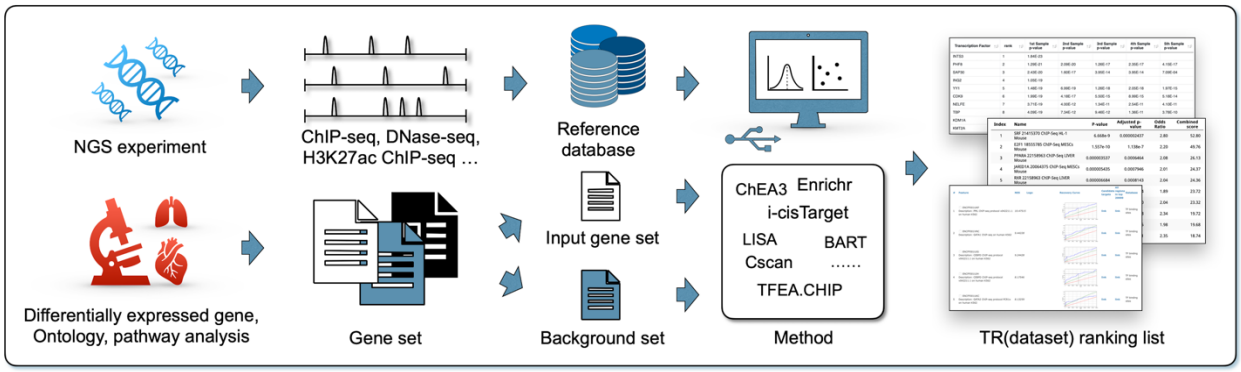


Figure 1. *A general workflow of computational methods for predicting transcriptional regulators. Query gene set is produced by differentially expressed gene analysis, ontology, and pathway analysis. The reference database consists of NGS data from TR ChIP-seq, DNase-seq, and H3K27ac ChIP-seq. Eleven NGS-based computational methods so far have been developed to identify TRs. The final output is typically a TR ranking list.*

Although each of these methods has been shown to perform well in certain applications, a unified and systematic evaluation has not been conducted yet. Considering their extensive use in thousands of publications so far [41-44], it is essential to elucidate their strengths and limitations by a benchmark study to avoid misinterpretation and inform the development of improved methods. The contribution of this review is three-fold. First, it offers a detailed overview of the NGS-based computational methods, highlighting their differences and similarities. Second, it provides an objective evaluation using an extensive collection of gene sets derived from TR perturbation experiments. Third, it systematically assesses key aspects of the methods, including

accuracy, sensitivity, coverage, and usability, as well as limitations and directions for potential improvement.

Categorization of NGS-based Methods

The NGS-based methods reviewed in this article fall into two main categories, each with two sub-categories: (i) library-based methods, relying on enrichment analysis of target gene (TG) libraries, subdivided into window-based and window-free approaches based on the main TG assignment algorithm, and (ii) region-based methods, which use cis-regulatory regions for TR prediction, further categorized into proximal-elements-centric and distal-elements-inclusive methods (**Figure 2**).

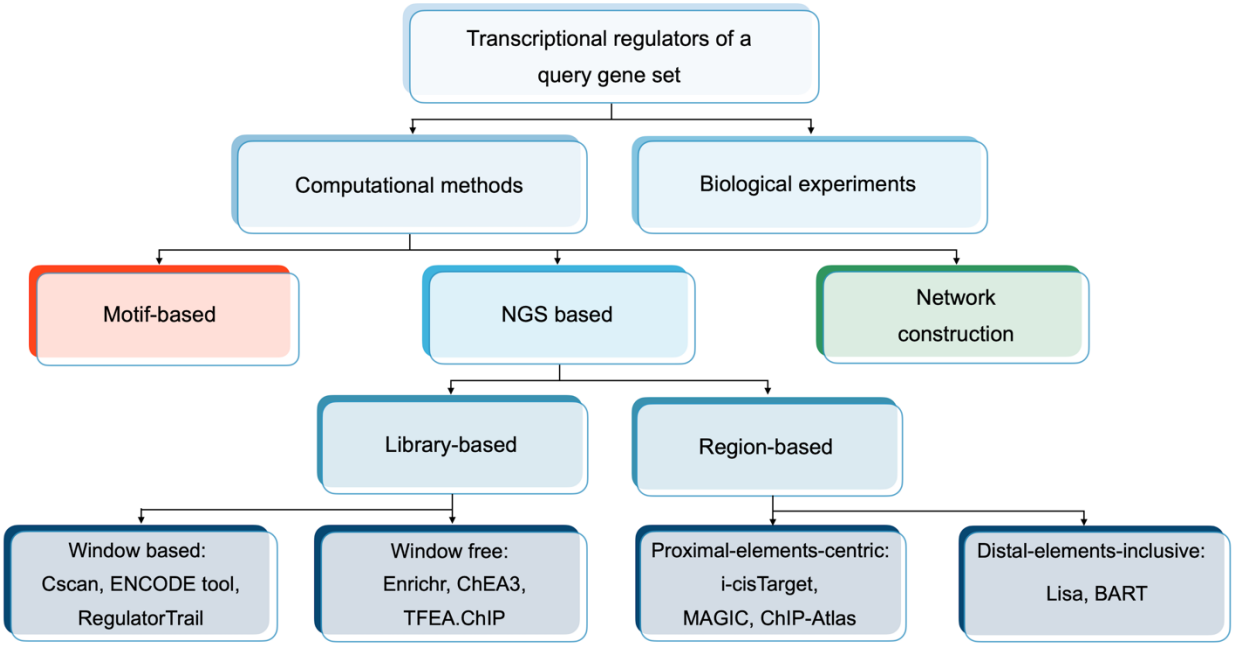


Figure 2. *Categorization of computational methods for prediction of TRs mainly based on their approaches.*

TRs with multiple ChIP-seq datasets

Due to the uneven allocation of research resources across different TRs, while a TR ChIP-seq dataset is only linked to one specific TR, a single TR may be associated with multiple ChIP-seq datasets in an epigenomic database used. Methods including Cscan, Enrichr, ENCODE ChIP-seq significance tool (ENCODE tool), i-cisTarget, Lisa, and ChIP-Atlas generate a rank for each of the ChIP-seq datasets, even when they are linked to the same TR. Therefore, a final ranking list from any of these methods can contain multiple ranks for any TR with multiple ChIP-seq datasets. In contrast, the rest methods have an integration step to assign one unique rank to each TR. This integration can be achieved through various approaches, such as calculating the mean rank, selecting the highest rank, or employing statistical tests.

Library-based methods

Predicting TRs that regulate a specific gene set requires establishing a link between TR binding sites and input gene symbols. Two main approaches exist: 1) converting TR binding sites to target gene sets followed by enrichment analyses to determine whether each target set is enriched by the input genes compared to the background set of genes, and 2) converting input gene symbols to their regulatory regions and identifying TRs with the highest proportion of binding peaks within these regions. Crucial to both strategies are the gene transcription start site (TSS), which serves as a reference point for mapping gene symbols to chromatin locations. This allows the calculation of the linear epigenomic distance between a TR binding peak and the TSS, a key measure for quantifying the TR-gene connection.

Library-based methods, including Cscan, ENCODE ChIP-seq significance tool, Enrichr, RegulatorTrail, ChEA3, and TFEA.ChIP, follow the first approach to create a TG library for TRs or TR datasets (**Figure 3**), where the library contains target gene sets identified from individual TR ChIP-seq experiments (Cscan, ENCODE tool, Enrichr, ChEA3) or integrated from multiple

experiments for the same TR (RegulatorTrail, TFEA.ChIP). Genes in the input and background sets are dichotomized according to each target gene set in the library, and a statistical test, usually the Fisher exact test, is then applied to compute the statistical significance of each target gene set enriched with the input genes. Each TR or dataset is then ranked based on the statistical significance of their associated target gene set.

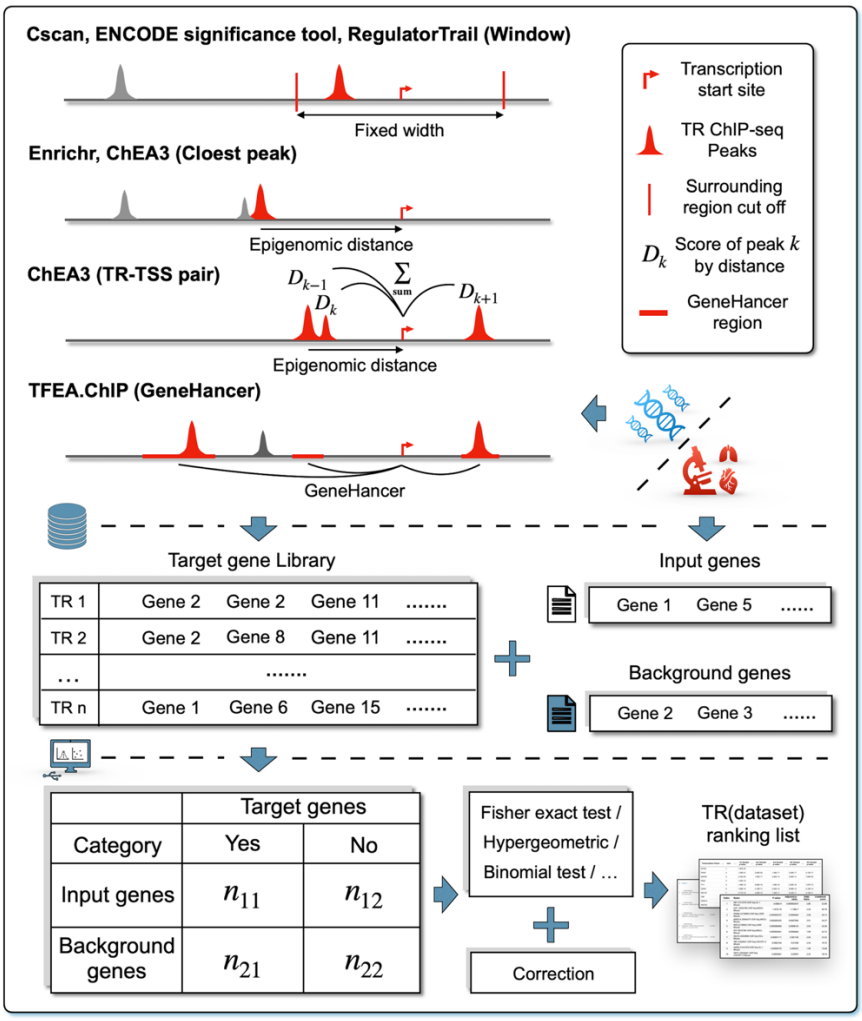


Figure 3. *The common workflow of library-based methods. To set up a target gene library, TR ChIP-seq peaks are associated with the nearest genes, usually by measuring the genomic distance linearly to the transcription start site (TSS) or whether they fall into the surrounding region. The library can also be derived*

from experimental evidence. Fisher exact test is commonly used to decide whether the target gene set corresponding to each TR or dataset in the library is enriched with the input genes. The final output is a TR (dataset) ranking list based on the test significance.

Library-based methods mainly differ in their target gene assignment algorithms. Cscan, ENCODE tool, and RegulatorTrail, for example, employ a "window-based" approach. They define a fixed window (typically 50 bp to 10 kb) around each gene's TSS. If a TR ChIP-seq dataset has peaks within this window for a particular gene, that gene is then classified as a target gene.

The other three methods mostly use a “window-free” approach. For example, Enrichr and ChEA3 focus on the proximity of peaks to genes. These methods sort peaks in a TR ChIP-seq dataset based on their linear epigenomic distances to the TSS of nearby genes and retain a certain number of genes with the closest peaks as targets. ChEA3 can further consider the density of peaks by employing a different method: first calculate a score for each TR-TSS pair by aggregating all the peaks of a TR ChIP-seq dataset within a specific range based on their distance to the TSS. Genes associated with TR-TSS pairs having the top 5% highest non-zero scores in each dataset are then designated as targets.

TFEA.ChIP is another “window-free” method, using the GeneHancer database that stores cis-regulatory elements of genes inferred from multiple data sources [45], such as FANTOM5 Erna-gene expression correlation. These cis-regulatory elements are defined as “elite” regions, and TFEA.ChIP assigns the target genes of a TR based on whether any ChIP-seq dataset of the TR has peaks in these “elite” regions.

There are three key limitations to these six library-based methods. First, these methods use hard cutoffs to identify target genes. This binary approach may be overly simplistic. In Cscan, ENCODE ChIP-seq significance tool, and RegulatorTrail, the cutoff determines the window size, and any

194 peaks outside the window are excluded from consideration. Similarly, Enrichr and ChEA3 use a
195 cutoff to determine the number of genes retained in each ChIP-seq dataset. Since the selection
196 of cutoff is subjective, different choices can result in different numbers of TGs and, subsequently,
197 affect the prediction results.

198 Second, it is widely recognized that the 3D chromatin distance between an epigenomic peak and
199 a gene promoter region, which can be measured by Hi-C or DNA-FISH [46], estimates their
200 interaction more accurately than the one-dimensional linear distance between two loci (on a
201 straight line) [47, 48]. However, most library-based methods use the linear distance to estimate
202 the interactions, neglecting that the genome is organized in a complex three-dimensional
203 structure, allowing distant locus to come in close physical contact [49, 50]. In addition, a significant
204 fraction of binding sites has been shown to be neutral or non-functional in regulating gene
205 expression even when located more proximal to a gene [51]. Thus, the dichotomization of genes
206 based solely on the linear distance might lead to misleading conclusions.

207 Finally, library-based methods lack an approach to quantify the interaction between TRs and their
208 target genes. Simply categorizing genes as targets and non-targets treats each TR-gene
209 interaction as equal in importance. This neglects the fact that TRs can have different degrees of
210 influence on gene expression. To fully understand how TRs regulate gene expression,
211 researchers need methods that can not only identify target genes but also quantify the strength
212 and functional significance of these interactions.

213 While the Fisher exact test is the primary statistical test used for enrichment analysis in library-
214 based methods, it is important to note some variations. Methods like RegulatorTrail and
215 TFEA.ChIP offer alternative statistical tests such as the hypergeometric test, binomial test, or
216 gene set enrichment analysis. Additionally, several methods (Cscan, RegulatorTrail, ChEA3, and

TFEA.ChIP) incorporate FDR correction approaches like the Benjamin-Hochberg method to control false positives. Notably, Enrichr discussed the potential bias in the Fisher exact test due to varying gene set sizes and addressed it by employing z-scores that account for deviations from expected ranks, which are generated through simulations.

In summary, the performance of library-based methods heavily depends on the efficiency of TG assignment algorithms. However, there is no universally accepted algorithm. The enrichment analysis can further complicate this, with different statistical methods giving different ranking lists. In addition, the limitations mentioned above can severely hinder the performance of these library-based methods.

Region-based methods

The development of epigenomics now focuses on critical functions of cis-regulatory elements (CREs) [52], including promoters, enhancers, silencers, and insulators. These CREs can act as the binding sites for TRs and collaborate to regulate gene transcription [53], with promoters and enhancers being the two most important ones. Therefore, unlike library-based methods which convert the TR binding sites to target genes, region-based methods avoid using a TG assignment algorithm and instead focus on the enrichment of TR binding sites in regions that contain active CREs for the input genes and use additional information to quantify the TR-gene interaction.

Proximal-elements-centric methods, including i-cisTarget, MAGIC, and ChIP-Atlas, mainly use proximal CREs, which are essential for basic gene transcription, to predict TRs (**Figure 4**). Proximal CREs are regulatory sequences found close to the gene they regulate. Promoters, for instance, are a type of proximal CRE located immediately upstream of a gene's TSS, which can serve as the “on-off” switch to initiate gene transcription [54]. Due to their proximity to the genes that they regulate, proximal CREs often have a more direct and immediate impact on gene

expression compared to distal regulatory elements located farther away from the gene locus. Past evidence suggests TRs bound to the proximal CREs usually regulate the nearby genes [55]. Therefore, the rationale behind these methods is that functional TRs should have more peaks or high-signal peaks that match the proximal CREs of their regulated genes.

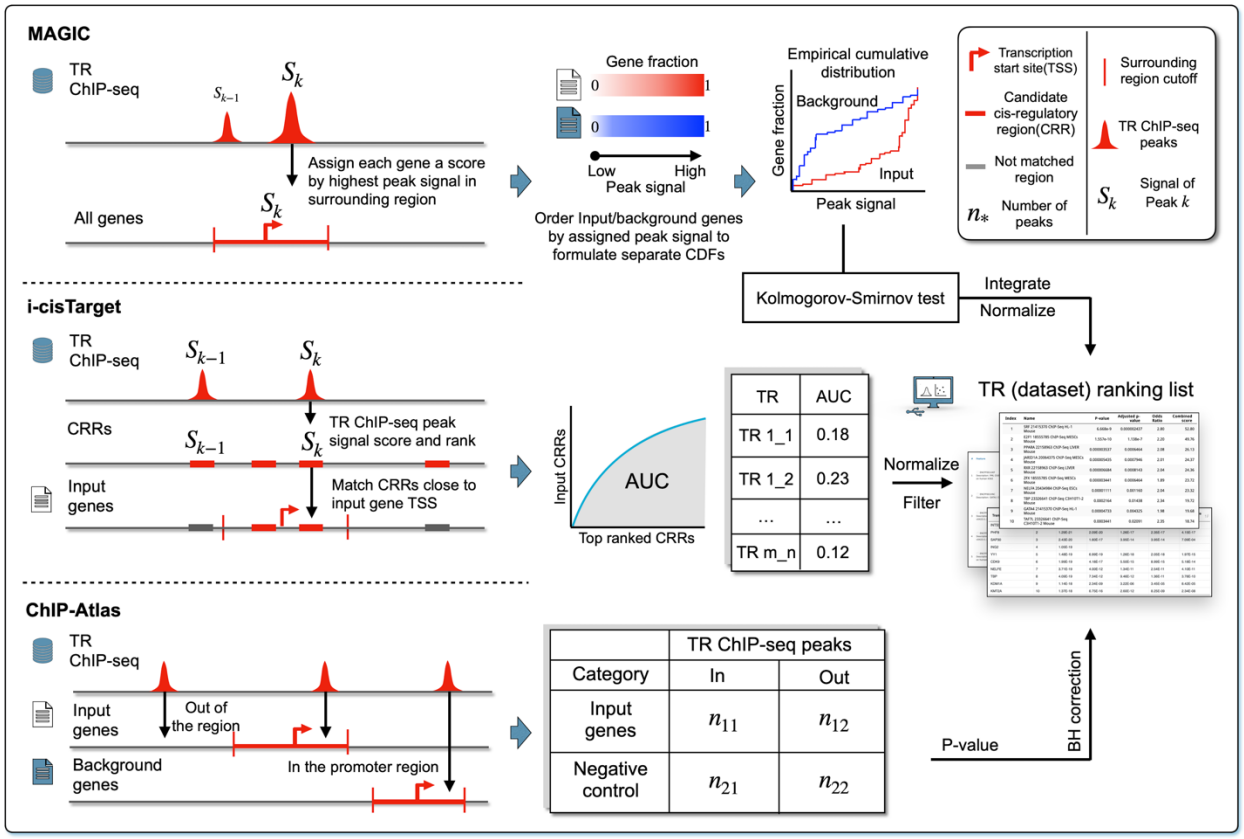


Figure 4. *Workflow of the proximal-elements-centric method. All three methods focus on the regions that surround the gene body or gene's TSS. TR ChIP-seq datasets with more peaks enriched in these regions would be ranked higher.*

i-cisTarget and MAGIC focus on peaks with high signal values. The peak signal is calculated based on the number of sequenced reads aligned to the peak position; a high signal value suggests a strong TR-DNA interaction. i-cisTarget first predefined many candidate cis-regulatory regions (CRRs) that may contain CREs based on public sources, mainly DNaseI hypersensitive

sites (DHSs). i-cisTarget scores and ranks these CRRs for each TR ChIP-seq dataset based on peak signals within the regions. Then, it calculates an area under the curve (AUC) value for each dataset by counting how many top-ranked CRRs can match the CRRs around the TSSs of input genes. Raw AUCs for the TR ChIP-seq datasets are filtered according to a threshold value and then normalized to obtain final scores that yield the final ranking list of these datasets.

MAGIC assigns each gene in the input and background the highest signal value among peaks that fall into its surrounding region, and the step is repeated for each TR ChIP-seq dataset. Next, the input and background genes are ranked based on these values to create two separate empirical cumulative distribution functions (eCDFs) for each dataset. Kolmogorov-Smirnov test is used to identify the TR ChIP-seq datasets that most effectively differentiate between the two eCDFs.

ChIP-Atlas does not use signal values. Instead, it aims to identify which TR ChIP-seq dataset has a higher proportion of peaks enriched in the promoter regions of the input genes than in those of the background genes. Even though a one-tailed test is more appropriate here, ChIP-Atlas applies a two-tailed Fisher exact test to determine the statistical significance for each TR ChIP-seq dataset, which is used to obtain the final ranking list.

A significant limitation of these methods lies in the unreliability of using signal strength to measure functionality. While strong peak signals often correlate with stronger binding affinity or functionality [56], they aren't a perfect measure. This is because ChIP-seq peak signals are computed based on an average across millions of cells. Peaks with low or medium signal could represent high-affinity binding sites in a specific subpopulation of cells, while high signal peaks might reflect only moderate affinity sites present in all cells [57]. Furthermore, emerging evidence suggests that

even low-affinity binding sites can play crucial roles in regulating gene transcription [58]. Therefore, a strong signal alone does not guarantee a functionally important regulatory element.

In addition to proximal CREs, distal regulatory elements like enhancers also can play crucial roles in gene transcription regulation [59, 60]. Located tens of thousands of base pairs away from their target genes, these distal CREs modulate the rate and timing of gene expression, ensuring precise control. Distal CREs are highly variable and more sensitive to environmental factors compared to proximal CREs. While incorporating distal CREs into TR prediction models could improve accuracy, directly measuring the long-range interactions between these elements and their targets remains challenging due to limited data availability.

The distal-elements-inclusive methods, including BART and Lisa, attempt to incorporate the activity of distal CREs by using H3K27ac ChIP-seq datasets (**Figure 5**). H3K27ac is a type of Histone protein used to mark active enhancers and promoters [27]. Instead of directly identifying which TR has more peaks enriched in the proximal regions, the two methods first select a few H3K27ac ChIP-seq datasets that can be predictive for the transcription of input genes, which serve as genome-wide CREs activity landscape. The rationale of the two methods is that a functional TR should have a higher number of peaks aligned with the high signal peaks in these selected H3K27ac ChIP-seq samples.

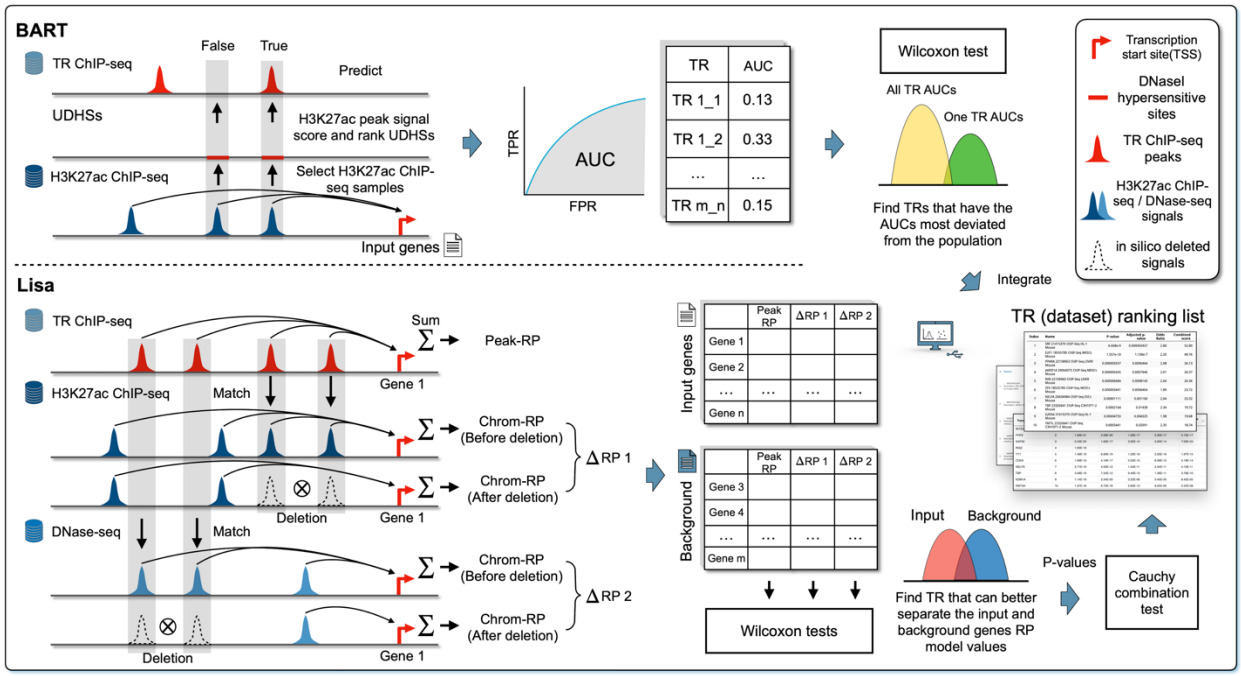


Figure 5. The core idea of BART and Lisa. Distal-elements inclusive methods first select predictive H3K27ac ChIP-seq (or DNase-seq) samples that quantify active distal cis-regulatory elements, by summarizing the H3K27ac ChIP-seq (or DNase-seq) peak signals around a gene's TSS. Next, statistical methods are applied to examine which TRs align more frequently with the high signal peaks from these selected samples.

In the selection step, a regulatory potential (RP) model is implemented in both methods, which counts the H3K27ac signals within the surrounding region of each gene to assign an RP score. Therefore, each H3K27ac dataset has a unique RP score for each individual gene. Following this, predictive H3K27ac datasets are chosen using stepwise logistic regression. That is, the two methods model the presence of a gene in the input list as a binary outcome (yes/no). The gene's RP scores across all H3K27ac datasets are used as potential predictors. Then, using a stepwise selection method, the two methods choose a subset of these datasets that best explains whether the gene is included in the input list.

305 After selecting predictive H3K27ac datasets, BART collects genome-wide DHSs and for each
306 DHS, computes a score by the weighted combination of signals from the selected datasets, and
307 then uses these scores to predict the peak presence in each TR ChIP-seq dataset and calculates
308 the AUC value using false positive rate (FPR) and true positive rate (TPR). The AUC values of
309 the ChIP-seq datasets are grouped by TR and each TR's AUC values are compared to the
310 distribution of AUC values from all TRs through the Wilcoxon rank sum test. Finally, BART assigns
311 each TR a unique rank by averaging the ranks of the maximum AUC value, the p-value from the
312 test, and the standardized Wilcoxon statistic score that is derived using TR-specific mean and
313 standard deviation calculated by running BART on more than 500 pre-collected gene sets.

314 In addition to H3K27ac datasets, Lisa selects predictive DNase-seq datasets as well; RP scores
315 are referred to as "chrom-RP" for the selected datasets of either type. After the selection, for each
316 type, Lisa matches the peaks from each TR ChIP-seq dataset with those from the selected
317 datasets, deleting all matched signals to reproduce the chrom-RP for each gene. Then the fitted
318 logistic regression model is used to calculate the chrom-RP change before and after the deletion,
319 denoted as ∆RP for each gene. Lisa further calculates a "Peak-RP" for each gene by using peaks
320 from the TR ChIP-seq dataset. Therefore, each TR ChIP-seq dataset has two ∆RPs, one from
321 selected H3K27ac samples and the other from selected DNase-seq samples, and one peak-RP
322 for each gene in the input and background sets. Finally, for each TR ChIP-seq dataset, ∆RPs and
323 peak-RPs are analyzed by three separate Wilcoxon tests to test whether there is a significant
324 difference between the input and background, and the p-values are integrated by Cauchy
325 combination test, the significance of which is used to derive a final ranking list of all TR ChIP-seq
326 datasets.

327 While both BART and Lisa integrate data from other NGS techniques to analyze distal CREs, they
328 still prioritize CREs closer to genes. Also, the presence of the H3K27ac histone mark alone might

329 not reliably indicate CRE activity [61]. This is similar to the limitations of using peak signals from
330 TR ChIP-seq, where a strong signal cannot guarantee functional binding. Consequently,
331 accurately quantifying CRE activity or functionality of TR binding sites to their target genes
332 remains a significant challenge.

333 Finally, the uneven allocation of research resources across different TRs also leads to a bias
334 where a select few TRs and cell lines have been studied in much greater depth than others. This
335 favoritism towards certain TRs (like CTCF and TP53) and cell lines (like HeLa and K562) often
336 stems from research preferences or the availability of experimental materials. Since these
337 methods rely on multiple statistical tests to assign significance to each dataset, they tend to favor
338 TRs with extensive data coverage over those with fewer datasets.

339 **Table 1** provides a summary of the eleven NGS-based methods examined in this study. The
340 majority rely on data from the ENCODE [62], a trusted source for high-quality NGS data. Methods
341 like LISA and ChIP-Atlas expand their scope by incorporating NGS datasets from individual
342 studies. This approach offers wider TR and cell line coverage but may introduce potential
343 variability in data quality. The evolution of these methods reflects a clear trajectory. Initially library-
344 based, they shifted to region-based, and now employ increasingly sophisticated strategies. This
345 highlights the field's continuous adaptation to new NGS technologies and its ability to address
346 increasingly complex scientific questions.

347 **Table 1. The summary of evaluated eleven NGS-based methods.**

Method (Year)	NGS data source	Implementation	Major method to associate TFs with genes	Primary statistical method	Method categorization
Cscan (2013)	ENCODE	Web tool	Peaks fall into the surrounding fixed width window of gene TSS	Fisher exact test	Library-based
ENCODE ChIP-seq significance tool (2013)	ENCODE	NA	Peaks fall into the surrounding fixed width window of gene TSS	Hypergeometric test	Library-based
Enrichr (2013)	ENCODE / ReMap / Literature	Web tool	Epigenomic distance between peaks and gene TSS	Fisher exact test / Corrected FET	Library-based
i-cisTarget (2015)	ENCODE	Web tool	Peak signal to score pre-defined CRRs	Recovery analysis	Region-based
RegulatorTrail (2017)	Multiple sources	Web tool	Peaks fall into the surrounding fixed width window of gene TSS	FET / Binomial test / Hypergeometric test	Library-based
BART (2018)	CistromeDB	Web tool / Python	RP model to score pre-collected DHSs	Wilcoxon rank sum test	Region-based
ChIP-Atlas (2018)	NCBI / DDBJ / ENA	Web tool	Proportion of peaks fall into gene promoter region	Fisher exact test	Region-based
TFEA.ChIP (2019)	ENCODE / ReMap	Web tool / R	Peaks fall into the “elite” regions from GeneHancer	FET / Kolmogorov Smirnov test / GSEA	Library-based
ChEA3 (2019)	ENCODE / ReMap / Literature	Web tool	Epigenomic distance between peaks and gene TSS	Corrected Fisher exact test	Library-based
Lisa (2020)	CistromeDB	Web tool / Python	RP model with in-silico deletion method	Wilcoxon rank sum test / Cauchy combination test	Region-based
MAGIC (2020)	ENCODE	Python	Peak signal to rank the genes	Kolmogorov-Smirnov test	Region-based

348

349 # Methods

350 ## Benchmarking study design

351 We systematically compared the performance of the NGS-based methods reviewed in the
352 previous section under a unified evaluation framework encompassing four key dimensions:
353 accuracy, sensitivity, coverage, and usability. Also included are three motif-based methods,
354 Pscan [15], HOMER [17], and RcisTarget [63] that can accept a gene set as input and prioritize
355 potential TRs through motif enrichment analysis. These methods use motif libraries of known TRs
356 and assign a higher rank to a TR whose associated motifs are determined to be over-represented.
357 By including motif-based methods, it is possible to assess whether using NGS data can indeed
358 enhance the prediction accuracy.

359 We used gene sets sourced from KnockTF, a comprehensive human gene expression profile
360 database generated from TR knockdown/knockout experiments [64]. Differentially expressed
361 genes derived from perturbation experiments often serve as a benchmark for evaluating the
362 performance of TR-related computational methods [65]. In the database, each TR perturbation
363 experiment has at least two control samples and two treatment samples, where Limma [66] was
364 used to derive the differentially expressed (DE) gene set. In total, we retrieved 570 DE gene sets
365 from the database, covering 308 known TRs. We further selected the top 200, 600, and 1000
366 most significant genes from each input gene set, respectively. The top 200 genes are used to
367 evaluate accuracy as these genes should be highly correlated with the perturbed TR. The other
368 two are used to evaluate sensitivity of the methods against noisy input. We transformed the
369 annotated genes to the required format following the manual or instructions. For each method
370 considered, we used the default settings when applicable, otherwise the parameters were chosen
371 according to the documentation.

To assess a method's accuracy, we focused on its ability to rank known, perturbed TRs at the top of its generated ranking list. This aligns with the user's primary interest: identifying the most relevant TRs at the forefront of the results. We first selected a range of threshold values (K) for "top K positions" (K=10, 50, 100). For each K value, we then applied four widely recognized metrics designed for evaluating ranking algorithms to assess the ranking results from the different methods [67]. Given gene lists $i = 1,2,\ldots,N$, N is the total number of evaluated gene lists. The terms ranked above threshold are ordered as $k = 1,2,\ldots,K$ by their ranks. The relevance of each term in the ranking list is then denoted as R_{iK}, where $R_{ik} = 1$ if the term is indeed the perturbed TR, and 0 otherwise. The total number of relevant terms in the ranking list i is denoted as n_{iK}, where $n_{iK} = \sum_{k=1}^{K} R_{ik}$ does not exceed the number of TR ChIP-seq data sets associated with the perturbed TR. The metrics used in our evaluation include (1) Hit rate at K, which calculates the proportion of ranking lists that contain at least one relevant term above the threshold,

$$H_K = \frac{1}{N}\sum_{i=1}^{N} I\{n_{iK} > 0\}.$$

(2) Mean reciprocal rank (MRR) at K, which values the first relevant term in each ranking list,

$$MRR_K = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{k_i^*}, k_i^* = \text{argmin}_k(R_{ik} = 1).$$

(3) Mean average precision (MAP) at K, which measures a method's capability of ranking relevant terms at the top, penalizing deviations. For a ranking list i, the metric first calculates the precision $P_i\,(k)$ at every k. It then sums over all k, further divided by n_{iK} for list i. Finally, it averages across all the ranking lists to get the overall precision.

$$MAP_K = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{n_{iK}}\sum_{k=1}^{K}P_i(k)\right], P_i(k) = \begin{cases} \sum_{j=0}^{k-1}\frac{R_{i,k-j}}{k}, & \text{if } R_{ik} = 1. \\ 0, & \text{otherwise.} \end{cases}$$

(4) Mean normalized discounted cumulative gain (mNDCG) at K, which also considers the position of relevant terms in a ranking list while giving more weights to the terms placed higher.

$$mNDCG_K = \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{k=1}^{K}\frac{R_{ik}}{\log_2(k+1)} / \sum_{k=1}^{n_{iK}}\frac{1}{\log_2(k+1)}\right].$$

Each of the four metrics has a different focus. The hit rate measures in general whether a method can identify the perturbed TRs. While MRR, MAP and mNDCG consider the relevant positions of perturbed TRs in the top, MRR focuses exclusively on the ranking of the first relevant term. In case a method ranks multiple datasets linked to the same TR above the threshold, MAP and mNDCG evaluate the rankings of all relevant terms but assign different weights and penalties. In addition to the ranking quality metrics that consider the absolute positions of perturbed TRs, we also counted the number of ranking lists that have perturbed TRs shown in the top 1%, top 5%, and top 10%. The accuracy of each method is ranked by their average performance over all metrics and thresholds.

We also evaluated the sensitivity of methods to noisy input, focusing on their ability to produce consistent and reliable results. We tested methods with $G = 200, 600, 1000$ genes that are most significant, respectively. The size of the input gene set is the major parameter that can be controlled by users. As G increases, the number of non-significant genes in the gene set tends to increase, which may negatively affect the performance. In addition, we investigated whether TRs with multiple ChIP-seq datasets are significantly ranked higher compared to TRs with only one

410 dataset by using Wilcoxon test. The p-values derived from same size gene set are integrated by
411 Stouffer's Z-score method.

412 Beyond accuracy and sensitivity, a method's coverage is crucial. This refers to the breadth of TRs
413 and cell lines represented in its internal NGS database. Without relevant data, a method cannot
414 rank a specific TR, excluding it from results. Therefore, coverage directly impacts performance –
415 broader coverage allows for wider analysis and a greater chance of identifying relevant TRs.
416 Methods with continually updated databases tend to have better coverage, reflecting the latest
417 discoveries.

418 It is worth noting that even high-performing methods may face practical implementation
419 challenges. Therefore, usability is critical and depends on ease of implementation, quality of
420 documentation, and code clarity. To assess usability, we defined nine criteria, each scored as 0
421 (not met) or 1 (met). These criteria address key user considerations when selecting a method: To
422 assess usability, we defined nine criteria, each scored as 0 (not met) or 1 (met). These criteria
423 address key user considerations when selecting a method: (1) a detailed manual or tutorial
424 providing clear instructions, (2) illustrative examples helping users understand the analysis
425 workflow, (3) availability of a CLI/API or R/Python package facilitating large-scale analyses, (4) a
426 user-friendly web tool allowing direct method execution, (5) open-source code allowing for
427 customization and transparency, (6) active maintenance ensuring the method remains functional,
428 (7) visualization enhancing user comprehension of analysis results, (8) data import accepting
429 gene symbols without requiring complex conversions, and (9) minimal unexpected errors during
430 use ensuring a smooth user experience.

431 Within this evaluation framework, we benchmarked thirteen methods. Note that we excluded the
432 ENCODE ChIP-seq significance tool due to being outdated and inaccessible. For Enrichr, we

433 used the highest ranks from four libraries. For ChEA3, we followed the original publication's
434 recommended approach of utilizing the integrated mean rank. Additionally, it's important to note
435 that i-cisTarget web tool employs a filtering step, therefore only datasets that pass this filter were
436 included in our evaluation.

437 # Results

438 ## Accuracy

439 **Figure 6** provides an overview of the performance of thirteen methods evaluated for ranking
440 perturbed TRs. In summary, no single method stood out as the best in all evaluation metrics.
441 However, actively updated or recently developed methods were observed to perform better,
442 perhaps due to their larger accumulation of NGS data and more sophisticated algorithm design.
443 Some general trends and potential issues were also observed, as detailed below.

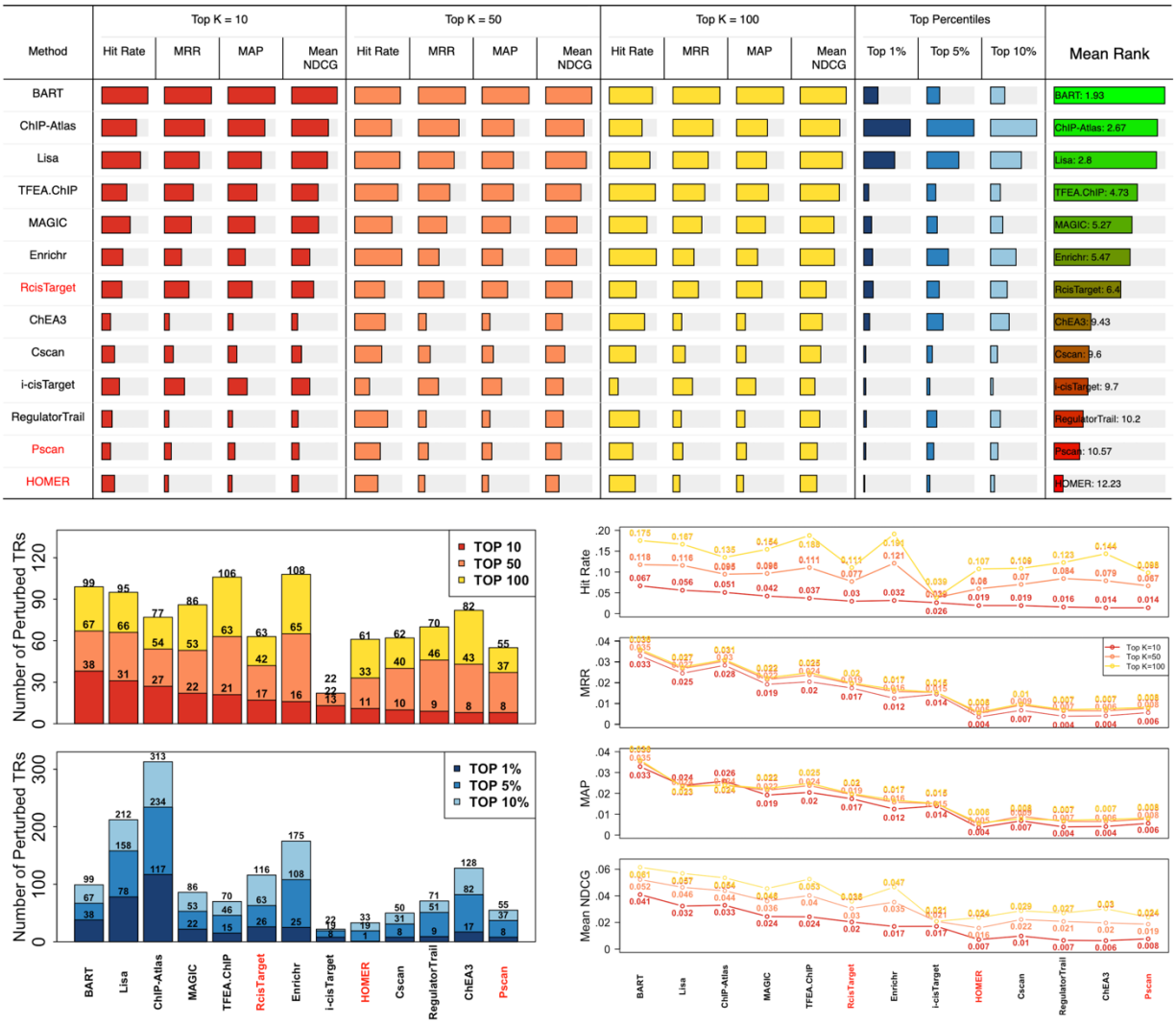


Figure 6. *Comparison of accuracy. Methods with red labels are motif-based. BART, ChIP-Atlas, and Lisa demonstrate the best overall performance across different scenarios. The NGS-based methods generally outperform the motif-based methods. All methods ranked fewer than 10% of perturbed TRs within the top 10 positions, suggesting a room for further improvement.*

First, BART, ChIP-Atlas, and Lisa are recommended in general as they are the top three performing methods in most scenarios. For example, BART, ChIP-Atlas, and Lisa have higher hit rate, MRR, MAP, and NDCG values for three selected thresholds. They also performed well by

452 considering the top percentile rankings. The good performance is partially credited to the
453 extensive collections of NGS data: ChIP-Atlas contains over 30,000 human TR ChIP-seq
454 datasets, BART and Lisa have around 7,000 human datasets, while other methods only have
455 1,000 to 3,000 datasets. Furthermore, the region-based approach might also play a pivotal role
456 in enhancing their accuracy. TFEA.ChIP, MAGIC, Enrichr, and RcisTarget also show good
457 performance under certain metrics. Besides, NGS-based methods outperform motif-based
458 methods on average. No motif-based method evaluated can surpass the performance of the top-
459 performing NGS-based methods (e.g., BART, ChIP-Atlas, and Lisa), regardless of the metrics
460 and thresholds used.

461 Combining results from various methods is believed to be more effective than relying on a single
462 method due to differences in databases and algorithms used. We next explore combining two
463 methods to determine the most optimal results. **Figure 7** illustrates the number of perturbed TRs
464 that are ranked in the top 10, 50, and 100 and the top 1%, 5%, and 10% in at least one TR ranking
465 list when two methods are used (200 input genes used; see **Figure S1** and **Figure S2** for 600
466 and 1000 input genes). We noticed that Lisa, TFEA.ChIP, BART, and ChIP-Atlas can be
467 combined to capture more perturbed TRs. In most cases, combining any of the two methods might
468 yield better performance than using individual methods alone.

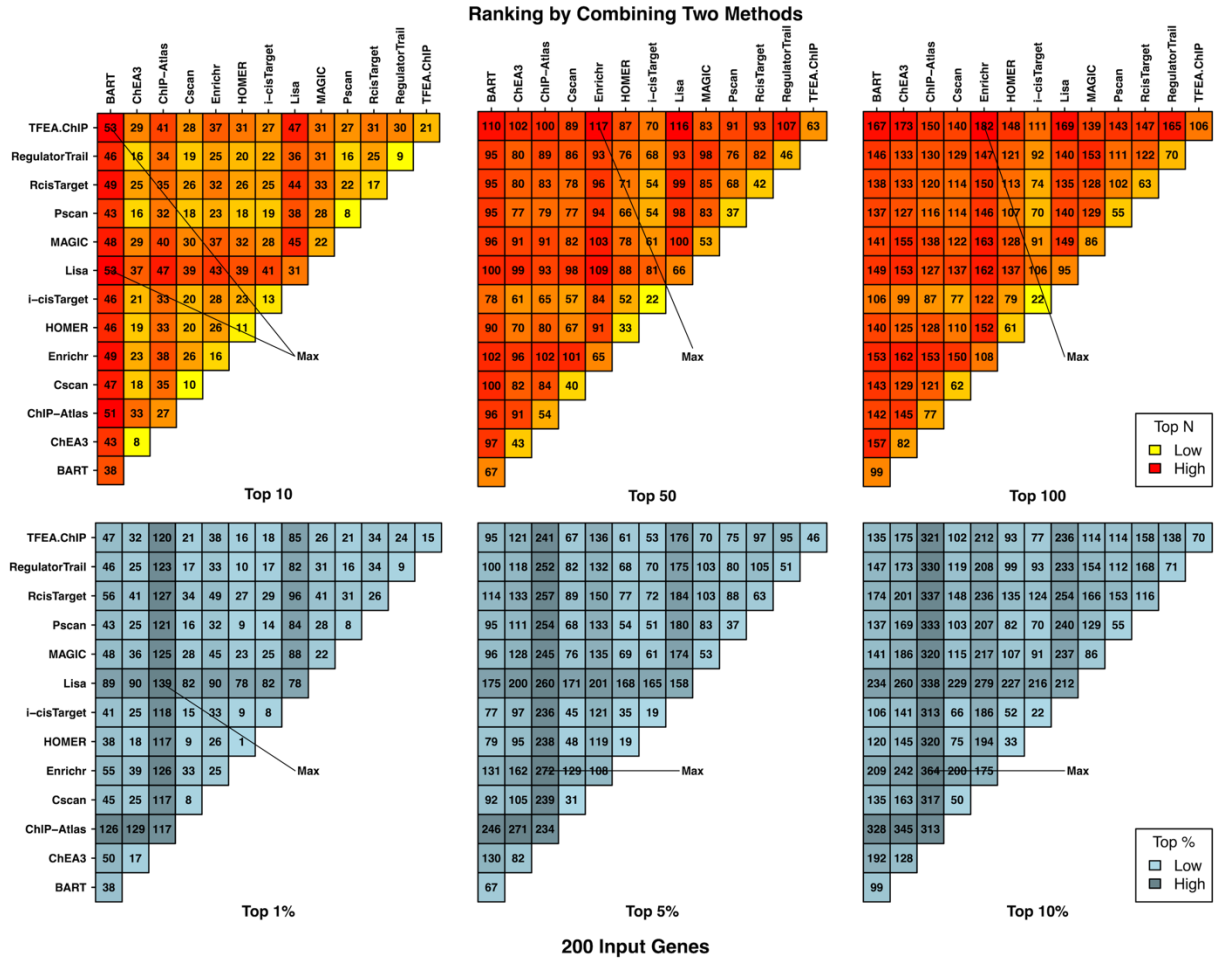


Figure 7. *The number of perturbed TRs ranked to top positions by either method in the combination of two methods (200 input genes used). The combination of any two from BART, ChIP-Atlas, Lisa, Enrichr, and TFEA.ChIP can achieve the best or close to the best performance.*

Figure 8 reveals a tendency among many methods to consistently rank certain TRs highly, regardless of the input gene set. However, these "favorite" TRs are rarely the experimentally perturbed TRs we want to identify. Prioritizing irrelevant TRs can overshadow true functional regulators, pushing them lower in the ranking list. This issue is most pronounced in RegulatorTrail, Cscan, and MAGIC. Thus, even if top-ranked TRs seem expected or familiar, we recommend

users cross-check them against a method's "favorite" TR list before proceeding with further analysis.

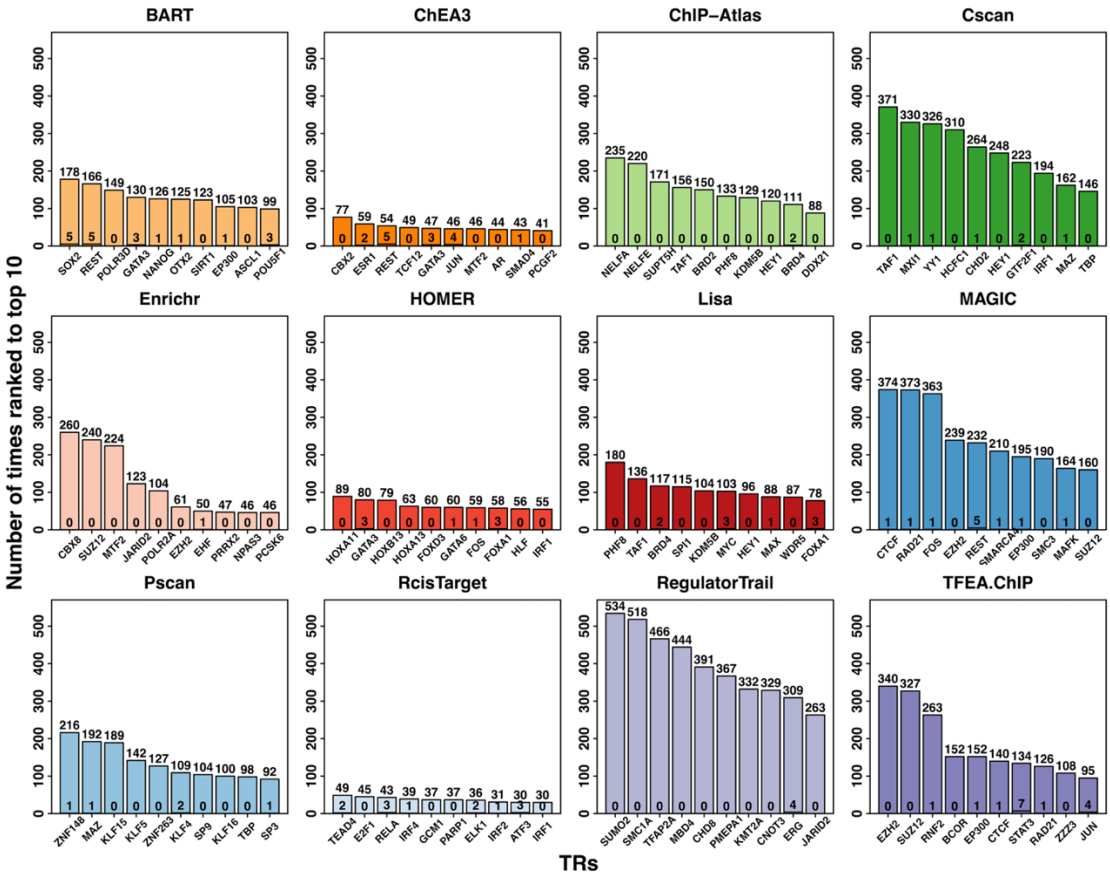


Figure 8. *TRs that are most frequently ranked to the top 10 positions by different methods. The number at the top of each bar indicates how many times the corresponding TR is ranked in the top 10 and that at the bottom indicates how many times that it is indeed the perturbed TR (out of 570). i-cisTarget was excluded as the number of TRs in the final ranking list is often less than 10 due to the threshold.*

Sensitivity

Figure 9 presents results from the sensitivity analysis with three input gene set sizes. As the input size increases, most methods have reduced performance as we expect, indicating that the

inclusion of more genes into the input set will lead to decreased performance of most methods. Therefore, it should be cautious when selecting genes used as input for these methods. The number of input genes itself can be a factor affecting the method performance. It is noted that HOMER, Pscan, and ChEA3 show relatively stable performance compared to others. However, given that accuracy remains the major consideration, among the top-performing methods (e.g., Lisa, ChIP-Atlas, and BART), Lisa is less sensitive to the change of input size.

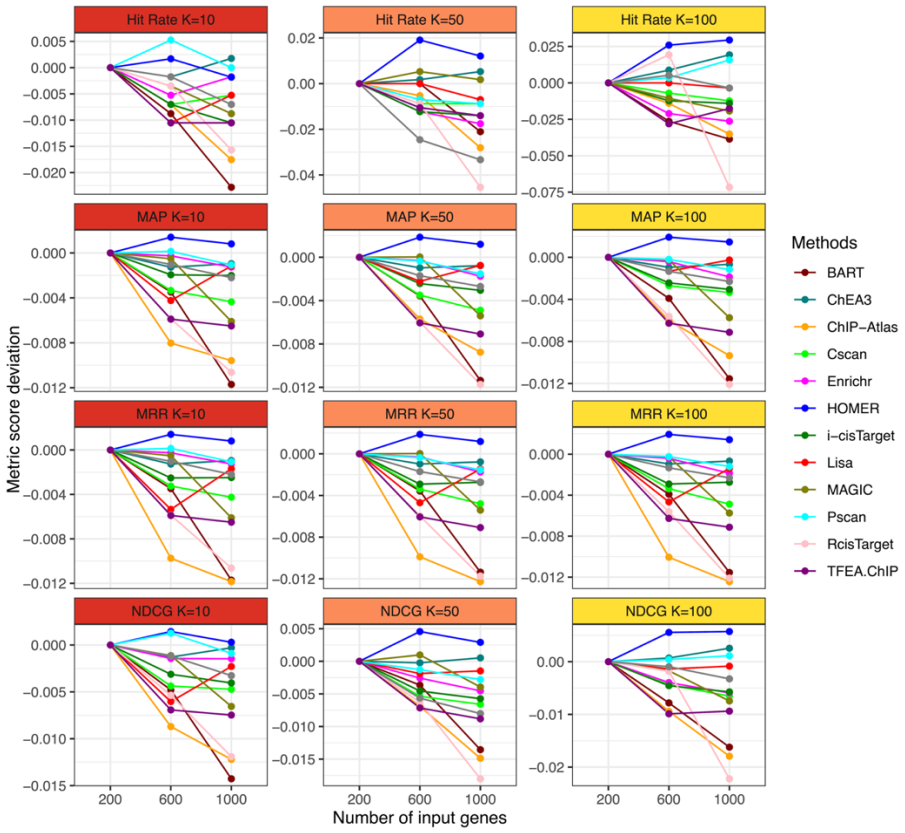


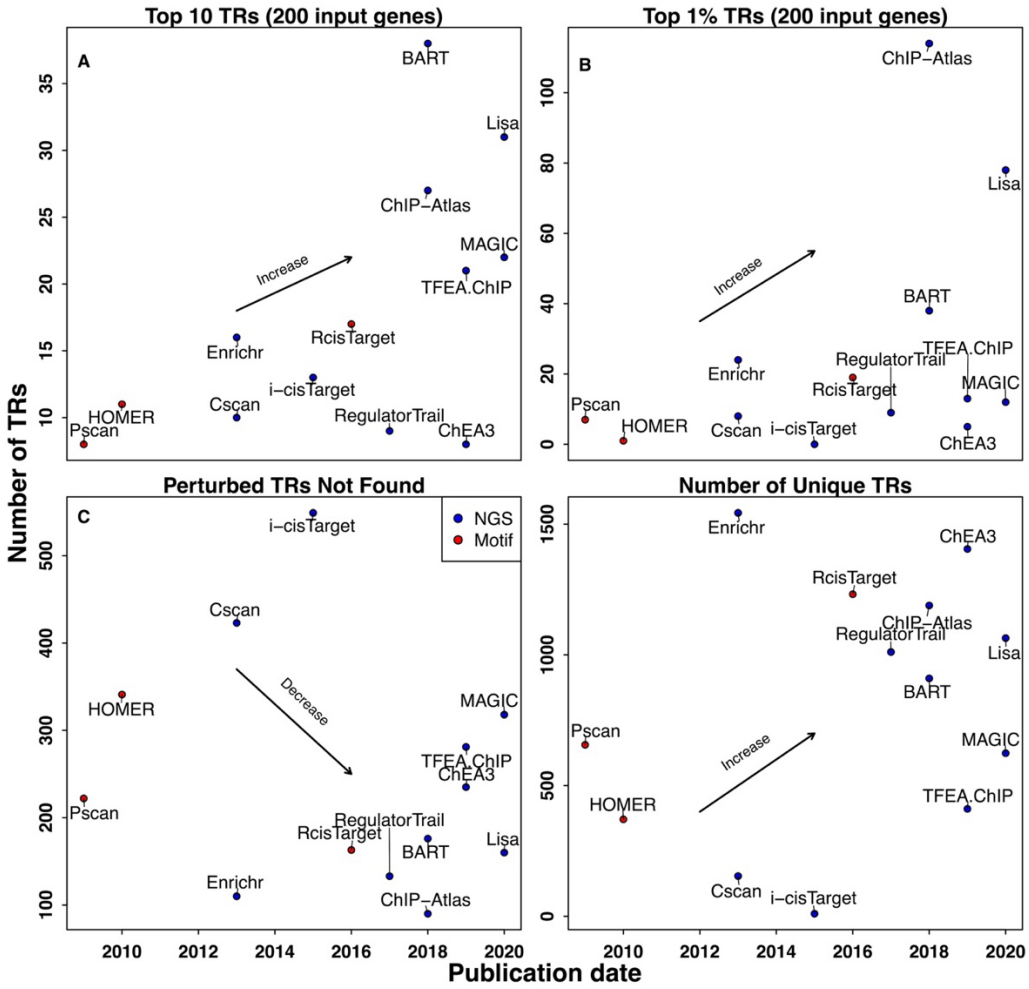
Figure 9. *As the size of the input gene set increases, the majority of methods demonstrate a decline in performance. Notably, HOMER, Pscan and ChEA3 show relatively stable or even better performance with increased number of input genes. For the top 3 performing methods in the evaluation of accuracy, Lisa displays the most consistent performance compared to BART and ChIP-Atlas.*

499 Except for the input gene set size, another major concern is the uneven allocation of research
500 resources, as some TRs have been extensively investigated (**Figure S3**). We selected 20 TRs
501 with the largest numbers of ChIP-seq datasets and 20 TRs associated with just one dataset
502 randomly from the ENCODE database. For each gene set, we employed a one-sided Wilcoxon
503 rank-sum test to determine if the 20 TRs with multiple datasets tend to be ranked higher than
504 those with a single dataset. Additionally, we excluded gene sets that feature selected TRs as the
505 perturbation targets, ensuring that each gene set could be considered a randomly generated set
506 that is irrelevant to any of the selected TRs. The findings, detailed in **Table S1**, reveal that most
507 methods indeed tend to favor TRs associated with a larger number of datasets. ChEA3 is the only
508 method that does not give significant results, which might be due to the integration step of using
509 mean ranks of all libraries as the final rank. This bias might explain why certain TRs consistently
510 appear at the top as shown in Figure 8.

511 Finally, we illustrated how the value of a cutoff can affect the rankings of individual perturbed TRs
512 that are expected to be ranked high by any appropriate methods (**Figure S4 - S5**). We sourced
513 ~2000 TR ChIP-seq datasets from the ENCODE project and implemented similar workflows as
514 Cscan and Enrichr, both library-based methods, to identify target genes, followed by a one-sided
515 Fisher exact test for each TR to produce the rank based on its significance. To assign target
516 genes, Cscan uses windows of a predetermined width and Enrichr retains a prefixed number of
517 genes based on the genomic distance, as discussed before. The figures show that, no matter
518 which quantity is used as the cutoff, the individual ranks of some perturbed TRs can change
519 drastically as we vary the cutoff, which again confirms that the use of a hard cutoff may pose a
520 challenge in these methods.

521 **Coverage**

522 With the detailed analysis of accuracy, it is evident that coverage plays a crucial role in the
523 effectiveness of methods. The extent to which these methods can incorporate up-to-date NGS
524 data can be a significant factor contributing to the performance (**Figure 10**). First, **Figure 10 A**
525 and **B** show that the publication time has a positive correlation with the ranking performance and
526 TR coverage, which confirms that the continuing efforts from generations of researchers have
527 improved the prediction performance of algorithms to address the problem. Second, **Figure 10 C**
528 displays the number of perturbed TRs that are missing in the TR ranking list versus the publication
529 date for each method. As we expect, more recently developed methods generally have lower
530 numbers of missing TRs. **Figure 10 D** plots the number of unique terms displayed in the TR
531 ranking list of each method. Considering that the benchmark gene sets used in this study only
532 cover a limited number of known TRs, a high number of missing TRs coupled with a low number
533 of available terms might suggest an insufficient coverage of TRs. Notably, ChIP-Atlas, Lisa, and
534 BART demonstrated their relatively comprehensive coverage of NGS data, which resonates with
535 their better accuracy observed before, indicating the method performance is positively correlated
536 with the number of covered TRs in method's database.



537

538 **Figure 10.** *An increasing trend in prediction performance and TR coverage, aligning with more recent*
539 *publication dates. (**A**) No. of perturbed TRs ranked within the top 10 positions versus method's*
540 *publication date; (**B**) No. of perturbed TRs ranked within top 1% versus method's publication date; (**C**) No.*
541 *of perturbed TRs that cannot be found in each method's TR ranking list versus its publication date; (**D**)*
542 *No. of unique TRs in the TR ranking list of each method versus its publication date.*

543 **Usability**

544 **Figure 11** provides an assessment of NGS-based methods using various criteria of usability, with
545 a "Yes" indicating that the criterion is satisfied and a "No" otherwise. Enrichr, TFEA.ChIP, BART,

ChIP-Atlas, ChEA3, and Lisa received good scores as these methods are generally user-friendly, easy to learn and use, and did not pose any issues during testing. In summary, we recommend refraining from using methods that are not open-source, actively updated, or well-maintained. Such methods often rely on outdated databases and could pose compatibility issues.

	Documentation		Implementation		Maintenance		Usage			
Method	Manual	Example	CLI/API/ Package	Web-Tool	Open Source	Update	Visual	Import	No-Error	Score
Enrichr	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Enrichr: 9
TFEA.ChIP	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	TFEA.ChIP: 9
BART	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	BART: 8
ChIP-Atlas	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	ChIP-Atlas: 8
ChEA3	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	ChEA3: 8
Lisa	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Lisa: 8
i-cisTarget	Yes	Yes	No	Yes	No	No	Yes	Yes	Yes	i-cisTarget: 6
MAGIC	Yes	No	Yes	No	Yes	No	Yes	Yes	Yes	MAGIC: 6
RegulatorTrail	Yes	Yes	Yes	Yes	No	No	No	Yes	No	RegulatorTrail: 5
Cscan	Yes	Yes	No	Yes	No	No	No	No	Yes	Cscan: 4

Figure 11. *The summary of usability evaluated by using nine criteria covering most user considerations for method usage. "Yes" indicates the method satisfied the criterion and a "No" otherwise. Enrichr, TFEA.ChIP, BART, ChIP-Atlas, ChEA3, and Lisa received good scores.*

Conclusion

Our evaluation of NGS-based methods in four dimensions leads us to recommend Lisa, ChIP-Atlas, and BART, either individually or in combination, for predicting TRs of query gene sets. These methods have demonstrated relatively better performance in providing accurate predictions with perturbed TR gene sets. The evaluated methods have three major limitations in general: (1) Peaks have to be associated with genes based on linear epigenomic distance. (2) There is no ideal approach to quantify the functionality or activity of binding sites or CREs. (3)

561 The internal database of reviewed methods can have a highly skewed distribution of the number
562 of datasets for each TR. In addition, it is important to note that even the best-performing method
563 may only rank less than 10% of perturbed TRs within the top 10 positions. Given the complexity
564 of cellular processes and the limitations above, researchers should use caution when interpreting
565 the results and should consider further validation with experimental approaches. Nevertheless,
566 recently developed NGS-based methods provide useful tools for generating hypotheses about
567 transcriptional regulation. Further methodological research is needed to fully utilize the wealth of
568 NGS data and address the limitations of current methods.

Discussion

This study evaluated the performance of various computational methods developed over the past decade for predicting TRs given a query gene set. Though NGS data can provide evidence of actual binding sites, multiple limitations make it challenging for these methods to accurately identify the functional TRs. Therefore, new experimental and computational methods are urgently needed to address these challenges. Previous studies have implicated several possible directions to improve these existing methods. First, further processing of the NGS data by using "footprinting" methods to identify binding sites in chromatin accessibility data was reported before [68]. However, this approach is limited to TRs with long residence times on chromatin, despite some TRs being known for lack of detectable footprints [69, 70]. Second, including three-dimensional long-range interaction information from Hi-C data may also serve as a possible improvement direction [71-73], but the sparse nature and low resolution of Hi-C data might also cause misleading conclusions and therefore should be used with caution and further validation.

With these limitations in mind, we also highlight that the breakthrough of single-cell sequencing techniques may lead to the development of new methods. While bulk sequencing data has dominated epigenomics research in the past decade, the limitations of bulk sequencing techniques have prompted the development of single-cell sequencing techniques such as single-cell RNA sequencing (scRNA-seq), single-cell ATAC sequencing (scATAC-seq), and single-cell ChIP-sequencing (scChIP-seq) [74]. In contrast to bulk sequencing data, these techniques sequence each cell separately and provide a more granular view of the differences between individual cells [75]. This means that limitations such as non-ideal measurement of binding affinities caused by averaging over the cells can now be addressed. In the next decade, there will likely be a surge in single-cell data, and most computational methods that aim to predict functional

592 TRs should consider integrating single-cell data to improve their accuracy and resolution. In
593 addition, with the rise of multi-omics experimental approaches [76, 77], the availability of multi-
594 omics data is no longer restricted to one or two types, and there is a growing need for novel multi-
595 omics approaches that can leverage the full range of available data. In the coming decade, it is
596 expected that more multi-omics-based methods will be developed, which will undoubtedly lead to
597 further innovation and discovery in this field.

Data availability statement

Commands and scripts used to run each method can be accessed on GitHub (https://github.com/ZeyuL01/Benchmark_NGSmethods). TR perturbation derived differentially expressed genes are sourced from KnockTF [64] (https://bio.liclab.net/KnockTFv1/). ChIP-seq data are sourced from ENCODE [62] (https://www.encodeproject.org/). Top 200, 600, 1000 genes and ranking lists generated by each method are available on SMU BOX (https://smu.box.com/s/hlofza4oe08ob4bb4i02nu3wl2zy39zj).

Funding

This work was supported by the following funding: the Rally Foundation, Children's Cancer Fund (Dallas), the Cancer Prevention and Research Institute of Texas (RP180319, RP200103, RP220032, RP170152 and RP180805), and the National Institutes of Health funds (R21CA259771, P30CA142543, HG011996, and R01HL144969) (to L.X.)

(R01CA258584)(to X.W.)

Zeyu Lu is a PhD candidate in the Department of Statistics and Data Science at Southern Methodist University, who focuses on statistics and epigenomics studies.

Xue Xiao is a Data Scientist in the Peter O'Donnell Jr. School of Public Health at University of Texas Southwestern Medical Center, who focuses on epigenomics studies.

Qiang Zheng is an Assistant Professor of Research in the Department of Mathematics at University of Texas at Arlington, who focuses on statistics and epigenomics studies.

Xinlei Wang is Jenkins-Garrett Professor of Statistics and Data Science in the Department of Mathematics and Director of Division of Data Science for Research in College of Science, University of Texas at Arlington, who focuses on development of statistical methods and computational tools for biomedical data.

Lin Xu is an Assistant Professor in the Peter O'Donnell Jr. School of Public Health at University of Texas Southwestern Medical Center, who focuses on epigenomics studies.

623 **Reference**

624 1. Jacob F, Monod J. Genetic regulatory mechanisms in the synthesis of proteins, J Mol
625 Biol 1961;3:318-356.
626 2. Latchman DS. Transcription factors: an overview, Int J Biochem Cell Biol 1997;29:1305-
627 1312.
628 3. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. A census of human
629 transcription factors: function, expression and evolution, Nat Rev Genet 2009;10:252-263.
630 4. Mullen AC, Orlando DA, Newman JJ et al. Master transcription factors determine cell-
631 type-specific responses to TGF-β signaling, Cell 2011;147:565-576.
632 5. Whyte WA, Orlando DA, Hnisz D et al. Master transcription factors and mediator
633 establish super-enhancers at key cell identity genes, Cell 2013;153:307-319.
634 6. Hwang JY, Zukin RS. REST, a master transcriptional regulator in neurodegenerative
635 disease, Curr Opin Neurobiol 2018;48:193-200.
636 7. Thomas MC, Chiang CM. The general transcription machinery and general cofactors,
637 Crit Rev Biochem Mol Biol 2006;41:105-178.
638 8. Chen T, Dent SY. Chromatin modifiers and remodellers: regulators of cellular
639 differentiation, Nat Rev Genet 2014;15:93-106.
640 9. Vandewalle C, Van Roy F, Berx G. The role of the ZEB family of transcription factors in
641 development and disease, Cell Mol Life Sci 2009;66:773-787.
642 10. Burda JE, O'Shea TM, Ao Y et al. Divergent transcriptional regulation of astrocyte
643 reactivity across disorders, Nature 2022;606:557-564.
644 11. Darnell JE, Jr. Transcription factors as targets for cancer therapy, Nat Rev Cancer
645 2002;2:740-749.

646 12. Bhagwat AS, Vakoc CR. Targeting Transcription Factors in Cancer, Trends Cancer
647 2015;1:53-65.
648 13. Bushweller JH. Targeting transcription factors in cancer - from undruggable to reality,
649 Nat Rev Cancer 2019;19:611-624.
650 14. Frith MC, Fu Y, Yu L et al. Detection of functional DNA motifs via statistical over-
651 representation, Nucleic Acids Res 2004;32:1372-1381.
652 15. Zambelli F, Pesole G, Pavesi G. Pscan: finding over-represented transcription factor
653 binding site motifs in sequences from co-regulated or co-expressed genes, Nucleic Acids Res
654 2009;37:W247-252.
655 16. McLeay RC, Bailey TL. Motif Enrichment Analysis: a unified framework and an
656 evaluation on ChIP data, BMC Bioinformatics 2010;11:165.
657 17. Heinz S, Benner C, Spann N et al. Simple combinations of lineage-determining
658 transcription factors prime cis-regulatory elements required for macrophage and B cell identities,
659 Mol Cell 2010;38:576-589.
660 18. Jolma A, Yan J, Whitington T et al. DNA-binding specificities of human transcription
661 factors, Cell 2013;152:327-339.
662 19. Lambert SA, Jolma A, Campitelli LF et al. The Human Transcription Factors, Cell
663 2018;172:650-665.
664 20. Wooten DJ, Groves SM, Tyson DR et al. Systems-level network modeling of Small Cell
665 Lung Cancer subtypes identifies master regulators and destabilizers, PLoS Comput Biol
666 2019;15:e1007343.
667 21. Su K, Katebi A, Kohar V et al. NetAct: a computational platform to construct core
668 transcription factor regulatory networks using gene activity, Genome Biol 2022;23:270.

669 22. Kamimoto K, Stringa B, Hoffmann CM et al. Dissecting cell identity via network inference
670 and in silico gene perturbation, Nature 2023;614:742-751.
671 23. Delgado FM, Gómez-Vela F. Computational methods for Gene Regulatory Networks
672 reconstruction and analysis: A review, Artif Intell Med 2019;95:133-145.
673 24. Mercatelli D, Scalambra L, Triboli L et al. Gene regulatory network inference resources:
674 A practical overview, Biochim Biophys Acta Gene Regul Mech 2020;1863:194430.
675 25. Nguyen H, Tran D, Tran B et al. A comprehensive survey of regulatory network
676 inference methods using single cell RNA sequencing data, Brief Bioinform 2021;22.
677 26. Furey TS. ChIP-seq and beyond: new and improved methodologies to detect and
678 characterize protein-DNA interactions, Nat Rev Genet 2012;13:840-852.
679 27. Creyghton MP, Cheng AW, Welstead GG et al. Histone H3K27ac separates active from
680 poised enhancers and predicts developmental state, Proc Natl Acad Sci U S A 2010;107:21931-
681 21936.
682 28. Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory
683 epigenome, Nat Rev Genet 2019;20:207-220.
684 29. Jiang S, Mortazavi A. Integrating ChIP-seq with other functional genomics data, Brief
685 Funct Genomics 2018;17:104-115.
686 30. Zambelli F, Prazzoli GM, Pesole G, Pavesi G. Cscan: finding common regulators of a set
687 of genes by using a collection of genome-wide ChIP-seq datasets, Nucleic Acids Res
688 2012;40:W510-515.
689 31. Auerbach RK, Chen B, Butte AJ. Relating genes to function: identifying enriched
690 transcription factors using the ENCODE ChIP-Seq significance tool, Bioinformatics
691 2013;29:1922-1924.

692 32. Chen EY, Tan CM, Kou Y et al. Enrichr: interactive and collaborative HTML5 gene list
693 enrichment analysis tool, BMC Bioinformatics 2013;14:128.
694 33. Imrichová H, Hulselmans G, Atak ZK et al. i-cisTarget 2015 update: generalized cis-
695 regulatory enrichment analysis in human, mouse and fly, Nucleic Acids Res 2015;43:W57-64.
696 34. Kehl T, Schneider L, Schmidt F et al. RegulatorTrail: a web service for the identification
697 of key transcriptional regulators, Nucleic Acids Res 2017;45:W146-w153.
698 35. Wang Z, Civelek M, Miller CL et al. BART: a transcription factor prediction tool with
699 query gene sets or epigenomic profiles, Bioinformatics 2018;34:2867-2869.
700 36. Oki S, Ohta T, Shioi G et al. ChIP-Atlas: a data-mining suite powered by full integration
701 of public ChIP-seq data, EMBO Rep 2018;19.
702 37. Puente-Santamaria L, Wasserman WW, Del Peso L. TFEA.ChIP: a tool kit for
703 transcription factor binding site enrichment analysis capitalizing on ChIP-seq datasets,
704 Bioinformatics 2019;35:5339-5340.
705 38. Keenan AB, Torre D, Lachmann A et al. ChEA3: transcription factor enrichment analysis
706 by orthogonal omics integration, Nucleic Acids Res 2019;47:W212-w224.
707 39. Roopra A. MAGIC: A tool for predicting transcription factors and cofactors driving gene
708 sets using ENCODE data, PLoS Comput Biol 2020;16:e1007800.
709 40. Qin Q, Fan J, Zheng R et al. Lisa: inferring transcriptional regulators through integrative
710 modeling of public chromatin accessibility and ChIP-seq data, Genome Biol 2020;21:32.
711 41. Parolia A, Cieslik M, Chu SC et al. Distinct structural classes of activating FOXA1
712 alterations in advanced prostate cancer, Nature 2019;571:413-418.
713 42. Yamashita N, Long M, Fushimi A et al. MUC1-C integrates activation of the IFN-γ
714 pathway with suppression of the tumor immune microenvironment in triple-negative breast
715 cancer, J Immunother Cancer 2021;9.

716 43. March-Diaz R, Lara-Ureña N, Romero-Molina C et al. Hypoxia compromises the
717 mitochondrial metabolism of Alzheimer's disease microglia via HIF1, Nat Aging 2021;1:385-399.
718 44. Duplaquet L, Li Y, Booker MA et al. KDM6A epigenetically regulates subtype plasticity in
719 small cell lung cancer, Nat Cell Biol 2023;25:1346-1358.
720 45. Fishilevich S, Nudel R, Rappaport N et al. GeneHancer: genome-wide integration of
721 enhancers and target genes in GeneCards, Database (Oxford) 2017;2017.
722 46. Kempfer R, Pombo A. Methods for mapping 3D chromosome architecture, Nat Rev
723 Genet 2020;21:207-226.
724 47. Bulger M, Groudine M. Functional and mechanistic diversity of distal transcription
725 enhancers, Cell 2011;144:327-339.
726 48. Rowley MJ, Corces VG. Organizational principles of 3D genome architecture, Nat Rev
727 Genet 2018;19:789-800.
728 49. Gorkin DU, Leung D, Ren B. The 3D genome in transcriptional regulation and
729 pluripotency, Cell Stem Cell 2014;14:762-775.
730 50. Zheng H, Xie W. The role of 3D genome organization in development and cell
731 differentiation, Nat Rev Mol Cell Biol 2019;20:535-550.
732 51. Whitfield TW, Wang J, Collins PJ et al. Functional analysis of transcription factor binding
733 sites in human promoters, Genome Biol 2012;13:R50.
734 52. Wittkopp PJ, Kalay G. Cis-regulatory elements: molecular mechanisms and evolutionary
735 processes underlying divergence, Nat Rev Genet 2011;13:59-69.
736 53. Riethoven JJ. Regulatory regions in DNA: promoters, enhancers, silencers, and
737 insulators, Methods Mol Biol 2010;674:33-42.
738 54. Haberle V, Stark A. Eukaryotic core promoters and the functional basis of transcription
739 initiation, Nat Rev Mol Cell Biol 2018;19:621-637.

740 55. Cramer P. Organization and regulation of gene transcription, Nature 2019;573:45-54.
741 56. Fisher WW, Li JJ, Hammonds AS et al. DNA regions bound at low occupancy by
742 transcription factors do not drive patterned reporter gene expression in Drosophila, Proc Natl
743 Acad Sci U S A 2012;109:21330-21335.
744 57. Slattery M, Zhou T, Yang L et al. Absence of a simple code: how transcription factors
745 read the genome, Trends Biochem Sci 2014;39:381-399.
746 58. Kribelbauer JF, Rastogi C, Bussemaker HJ, Mann RS. Low-Affinity Binding Sites and the
747 Transcription Factor Specificity Paradox in Eukaryotes, Annu Rev Cell Dev Biol 2019;35:357-
748 379.
749 59. Heintzman ND, Ren B. Finding distal regulatory elements in the human genome, Curr
750 Opin Genet Dev 2009;19:541-549.
751 60. Ricci WA, Lu Z, Ji L et al. Widespread long-range cis-regulatory elements in the maize
752 genome, Nat Plants 2019;5:1237-1249.
753 61. Zhang T, Zhang Z, Dong Q et al. Histone H3K27 acetylation is dispensable for enhancer
754 activity in mouse embryonic stem cells, Genome Biol 2020;21:45.
755 62. An integrated encyclopedia of DNA elements in the human genome, Nature
756 2012;489:57-74.
757 63. Aibar S, González-Blas CB, Moerman T et al. SCENIC: single-cell regulatory network
758 inference and clustering, Nat Methods 2017;14:1083-1086.
759 64. Feng C, Song C, Liu Y et al. KnockTF: a comprehensive human gene expression profile
760 database with knockdown/knockout of transcription factors, Nucleic Acids Res 2020;48:D93-
761 d100.
762 65. Garcia-Alonso L, Holland CH, Ibrahim MM et al. Benchmark and integration of resources
763 for the estimation of human transcription factor activities, Genome Res 2019;29:1363-1375.

764 66. Ritchie ME, Phipson B, Wu D et al. limma powers differential expression analyses for
765 RNA-sequencing and microarray studies, Nucleic Acids Res 2015;43:e47.
766 67. Valcarce D, Bellogín A, Parapar J, Castells P. Assessing ranking metrics in top-N
767 recommendation, Information Retrieval Journal 2020;23:411-448.
768 68. Sung MH, Guertin MJ, Baek S, Hager GL. DNase footprint signatures are dictated by
769 factor dynamics and DNA sequence, Mol Cell 2014;56:275-285.
770 69. Sung MH, Baek S, Hager GL. Genome-wide footprinting: ready for prime time?, Nat
771 Methods 2016;13:222-228.
772 70. Baek S, Goldstein I, Hager GL. Bivariate Genomic Footprinting Detects Changes in
773 Transcription Factor Activity, Cell Rep 2017;19:1710-1722.
774 71. Belton JM, McCord RP, Gibcus JH et al. Hi-C: a comprehensive technique to capture the
775 conformation of genomes, Methods 2012;58:268-276.
776 72. Xu H, Zhang S, Yi X et al. Exploring 3D chromatin contacts in gene regulation: The
777 evolution of approaches for the identification of functional enhancer-promoter interaction,
778 Comput Struct Biotechnol J 2020;18:558-570.
779 73. Lohia R, Fox N, Gillis J. A global high-density chromatin interaction network reveals
780 functional long-range and trans-chromosomal relationships, Genome Biol 2022;23:238.
781 74. Stuart T, Satija R. Integrative single-cell analysis, Nat Rev Genet 2019;20:257-272.
782 75. Kulkarni A, Anderson AG, Merullo DP, Konopka G. Beyond bulk: a review of single cell
783 transcriptomics methodologies and applications, Curr Opin Biotechnol 2019;58:129-136.
784 76. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease, Genome Biol
785 2017;18:83.
786 77. Huang S, Chaudhary K, Garmire LX. More Is Better: Recent Progress in Multi-Omics
787 Data Integration Methods, Front Genet 2017;8:84.

788