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, USA

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 three categories based on shared characteristics, namely library-based, region-based, and task-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 address the limitations of NGS-based methods and explore potential directions for further improvement.

Introduction

In the 1960s, foundational principles of gene transcriptional regulation were established through seminal work and subsequent studies[1], 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 malfunction of specific TRs[9,10], highlighting the importance of identifying TRs in biological and medical research. Discovering functional transcriptional regulators can aid the discovery of biomarker genes or potential therapeutic targets[11–13]. 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-defined 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)[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 transcriptional 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[23–25] providing deep insights into this domain. 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-defined gene set. Unlike motif-based methods, which screen the genome by

computationally inferred motifs to find possible binding sites, NGS data can directly measure the actual binding sites.

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].

Additionally, chromatin accessibility profiling techniques (e.g., DNase-seq, FAIRE-seq, and ATAC-seq) and Histone ChIP-seq techniques (e.g., H3K27ac ChIP-seq) can measure genome-wide accessible chromatin regions or active cis-regulatory elements[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 defined 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 this database to identify functional TRs. 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.

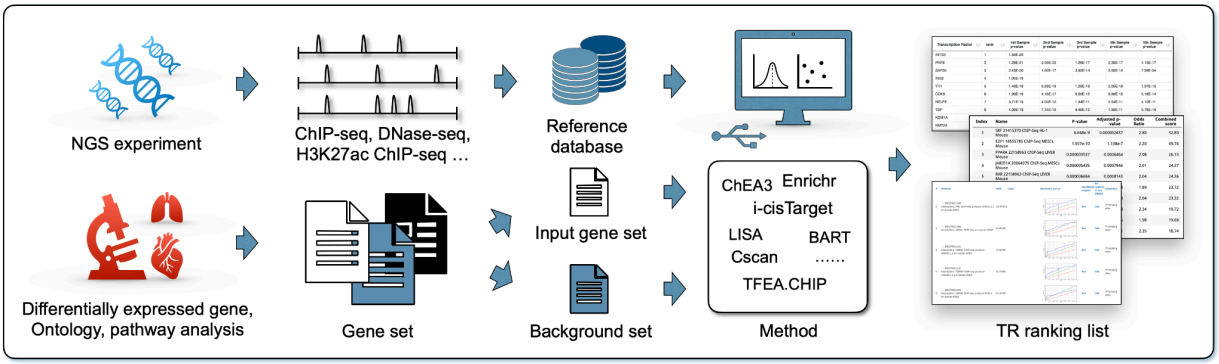

Figure 1. *A general workflow of computational methods for predicting transcriptional regulators. Query gene set is produced by differentially expressed gene analysis, ontology, and pathway analysis. The reference database consists of NGS data from TR ChIP-seq, DNase-seq, and H3K27ac ChIP-seq. Eleven 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 its own publication, a unified and systematic evaluation has not been conducted yet. Given that researchers

have widely used these eleven methods to identify TRs in thousands of publications so far[41–44], it is essential to elucidate their limitations by a benchmark study to avoid misinterpretation of the results by users and also guide the development of new methods in this field.

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 a benchmark evaluation using an extensive collection of gene sets derived from TR perturbation experiments. Third, it systematically assesses key aspects of the eleven methods, including accuracy, sensitivity, coverage and usability, as well as limitations and potential improvement directions.

Categorization of NGS-based Methods

The NGS-based methods reviewed in this article fall into two main categories, each with two sub-categories: library-based methods, dependent 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. (see Figure 2).

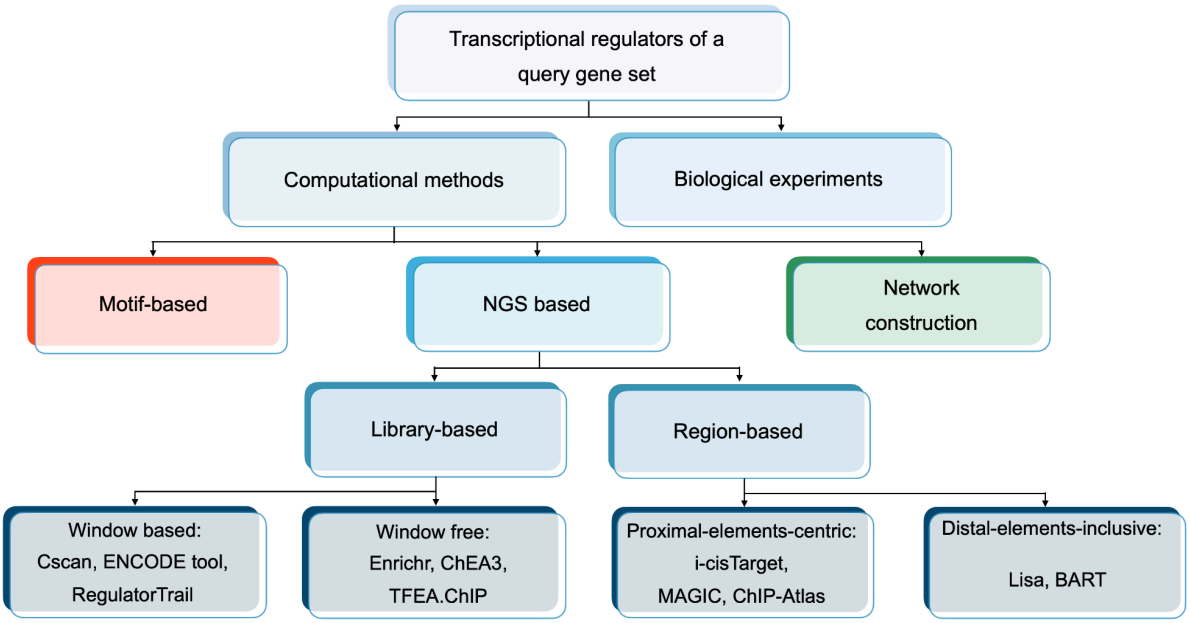

Figure 2. *Categorization of computational methods for prediction of TRs mainly based on their approaches.*

TRs or TR ChIP-seq datasets?

It is should be noted before the detailed discussion that while a 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, due to the uneven allocation of research resources across different TRs. Methods including Cscan, Enrichr, ENCODE ChIP-seq significance tool, i-cisTarget, Lisa and ChIP-Atlas generate a rank for each ChIP-seq dataset, 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

To predict the transcriptional regulators of an input gene set, the key step is building a connection between TR binding sites and gene symbols from the input and background set. Thus, there are two main approaches: (1) To convert the binding sites of TRs to target genes and perform enrichment analysis between gene sets, (2) To convert input gene symbols to their regulatory regions and decide which TR can have higher proportion peaks fall into these regions. The central of either conversion is the gene transcription start site (TSS), which are almost used in every method, as gene symbols must be mapped to exact locations on chromatin such that the linear epigenomic distance between a peak and the TSS can be measured, which is almost the sole information can be used to quantify the connection between TRs and genes directly.

Methods in this category include Cscan, ENCODE ChIP-seq significance tool, Enrichr, RegulatorTrail, ChEA3, and TFEA.ChIP follow the first approach to create a TG library for each TR or TR dataset (Fig 3). A TG library contains many target gene sets. For Cscan, ENCODE ChIP-seq significance tool, Enrichr and ChEA3, the library contains one target gene set for every TR ChIP-seq dataset in the database; for RegulatorTrail and TFEA.ChIP, target gene sets from multiple ChIP-seq datasets of the same TR are integrated into one so that each TR has only one target gene set in the library. 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 decide whether the target gene set is enriched with the input genes. Each TR or dataset associated with the target gene set is then ranked based on the statistical significance derived from the test.

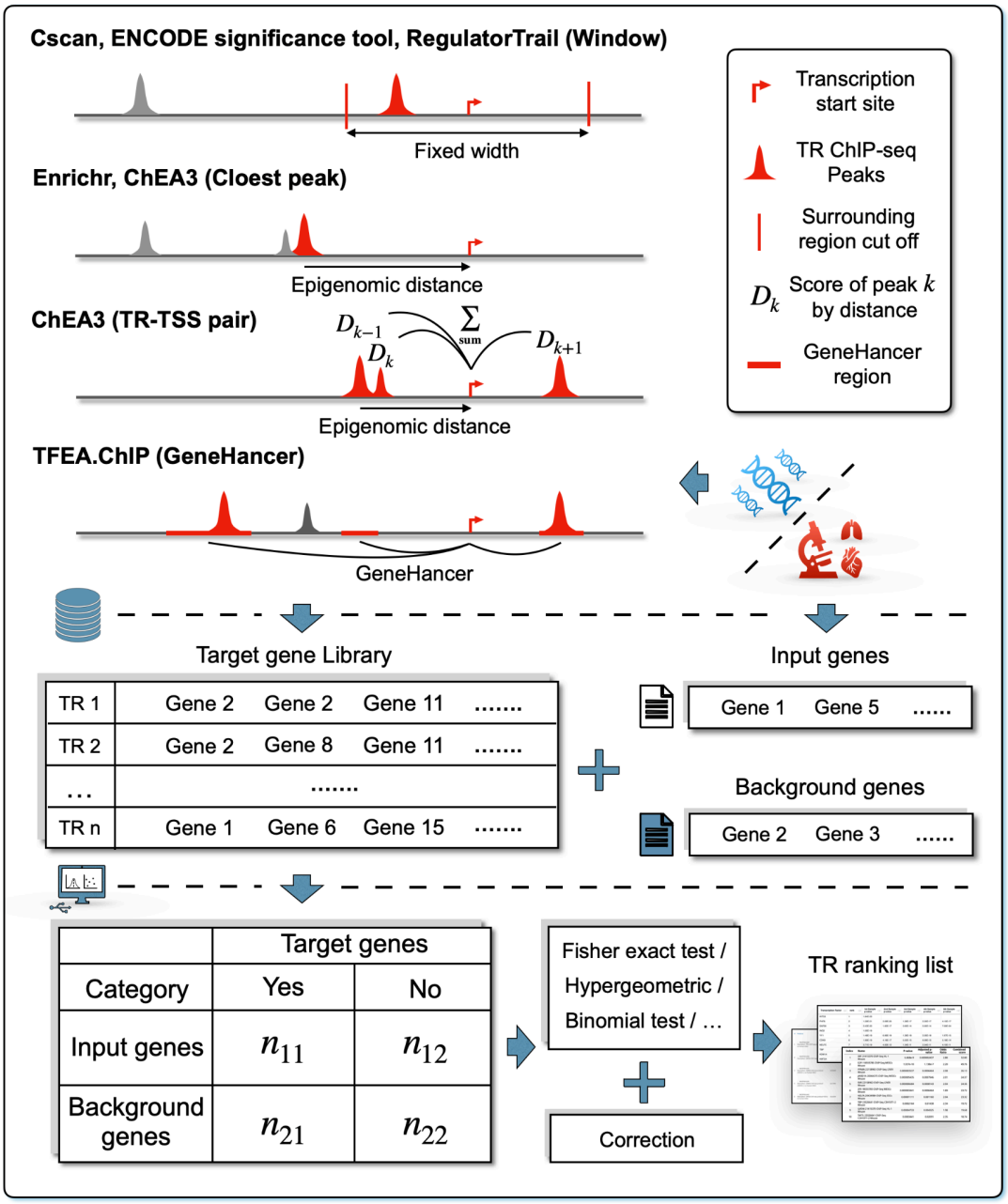

Figure 3. *The common workflow of library-based methods. To set up a target gene library, TR ChIP-seq peaks are associated with the nearest genes, usually by measuring the genomic distance linearly to the transcription start site (TSS) or whether they fall into the surrounding region. The library can also be derived from experimental evidence. Fisher exact test is commonly used to decide the enriched TR or dataset. The final output is a TR ranking list.*

The major difference among library-based methods is the algorithm used to assign the target genes of individual TRs. For example, Cscan, ENCODE ChIP-seq significance tool and RegulatorTrail define a window for each gene, which begins with the gene TSS. If a TR ChIP-seq dataset has peaks that fall within the window, then the corresponding gene is assigned as a target of the TR for this dataset. This type of approach is termed "window-based", where the window size is fixed across all genes and can vary from 50~10k bps.

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 genomic 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, where zero is assigned if no peak falls within the range nearby the TSS; then for each dataset, keep the top 5% TR-TSS pairs with non-zero scores and set the related genes as the 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 two key limitations to these six library-based methods. First, these methods have been using hard cutoffs to categorize genes into targets and non-targets, which can affect the sensitivity of predictive results. In Cscan, ENCODE ChIP-seq significance tool, and RegulatorTrail, the cutoff determines the window size and any peaks outside the window are not considered for the corresponding gene, while in Enrichr and ChEA3, the cutoff determines the number of genes retained in each ChIP-seq dataset. Since the cutoff is subjectively determined and there is no standard to guide its selection, different choices can result in different numbers of TGs and, subsequently, affect the prediction results.

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

It is worth noting that the enrichment analysis can differ between methods. First, methods such as RegulatorTrail and TFEA.ChIP can offer alternative options to the Fisher exact test, such as hypergeometric test, binomial test, or gene set enrichment analysis. Second, the methods can incorporate a correction method to control the false discovery rate (FDR), with the Benjamini–Hochberg method being the most commonly used, implemented in four methods (Cscan, RegulatorTrail, ChEA3 and TFEA.ChIP). Third, although the Fisher exact test is the primary method employed for the enrichment analysis, it may be biased by the varying gene set size and tends to rank TR or dataset associated with more target genes higher. Therefore, Enrichr computes z-scores by accounting for deviations of genes from their expected ranks, where the expected rank of each gene is generated by Fisher exact test conducted on many randomly generated gene sets.

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, as will be further explored in the benchmarking study below.

Region-based methods

The development of epigenomics now focus on the critical functions of cis-regulatory elements (CREs)[52], including promoters, enhancers, silencers, and insulators. These

CREs act as the binding sites for TRs and collaborate for regulating gene transcription[53], with promoters and enhancers are 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 the TG assignment algorithm and instead focus on the overall alignment between TR binding sites and the regions that can contain active CREs for the input genes.

Proximal-elements-centric methods include i-cisTarget, MAGIC, and ChIP-Atlas mainly use proximal CREs to predict TRs. Proximal CREs are cis-regulatory elements such as promoters that locate immediately upstream gene's TSS or gene body, which are essential for basic gene transcription and have a more straightforward regulatory role. For example, promoters can serve as the "on-off" switch to initiate gene transcription[54]. Past evidence suggests TRs bind to the proximal CREs usually regulate the nearby genes[55]. Therefore, the rationale behind these methods is that functional TRs should have a greater number of peaks or high signal value peaks align with the proximal CREs or fall into the surrounding regions of their regulated genes.

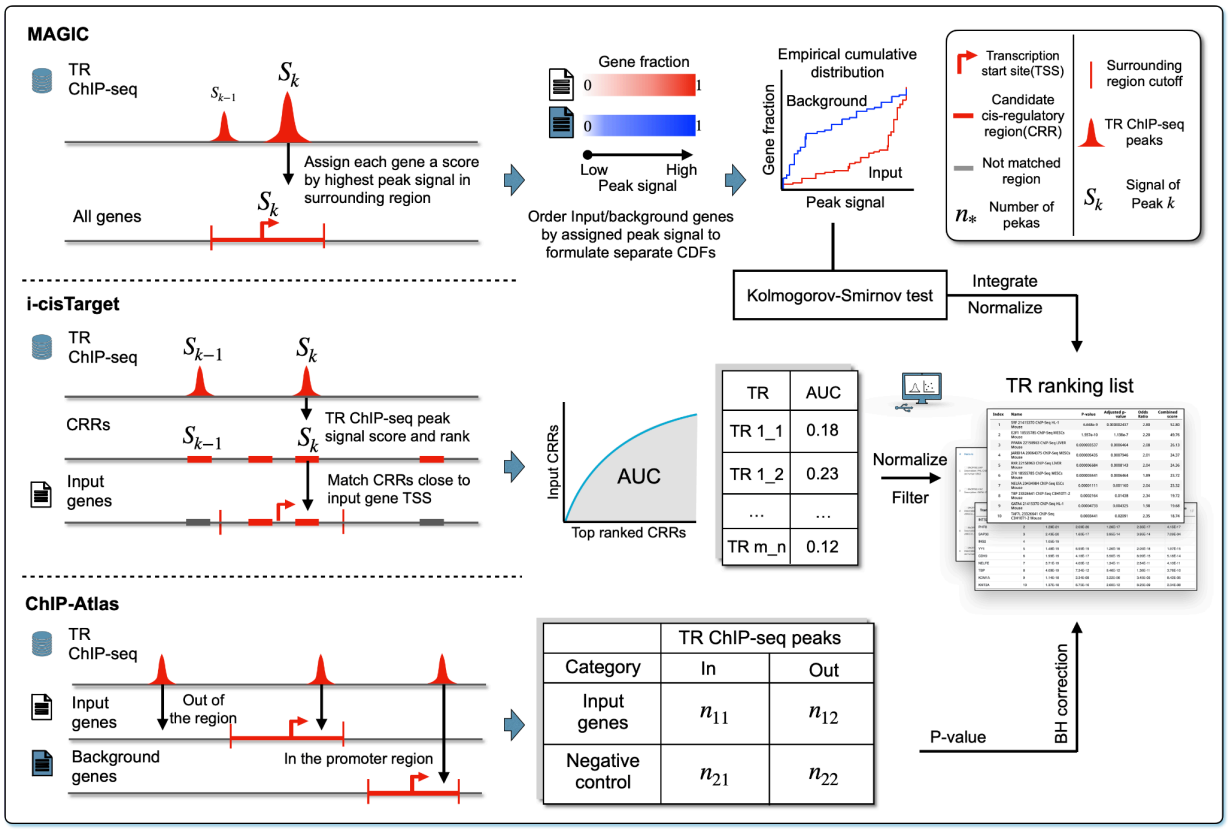

Figure 4. *Workflow of the proximal-elements-centric method. All three methods focus on the regions that surrounding the gene body or gene TSS. TR ChIP-seq datasets with high signal value peaks or more peaks enriched in the promoter regions would be ranked higher.*

i-cisTarget and MAGIC mainly use the high signal value peaks. Peak signal is calculated based on the number of sequenced reads aligned to peak position, a high signal suggests a strong TR-DNA interaction. i-cisTarget first predefined many candidate cis-regulatory regions (CRRs) 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 signal within each region. Then it calculates an area under curve (AUC) value for each dataset by counting how many top ranked CRRs that can match the

CRRs around the TSSes of input genes. Raw AUCs are filtered and normalized to obtain final scores to yield the final ranking list.

MAGIC assigns each gene in the input and background the highest signal value of peaks that fall into its surrounding region, 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 distributions (eCDFs) for each dataset. Kolmogorov-Smirnov tests are used to identify the TR ChIP-seq dataset that most effectively differentiates 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.

One major limitation with proximal-elements-centric methods is the unreliable signal values. Although peak signal can often positively correlate with binding affinity or functionality[56], it is not always a reliable measurement. For example, ChIP-seq peak signal is computed based on an average of millions of cells, peaks with low or medium signal might still be high-affinity sites in a small proportion of cells, while high signal peaks might be medium-affinity sites in all cells[57]. Also, there are evidence suggests low affinity sites can also be critical in regulating the transcription[58].

In addition, many CREs such as enhancers are located tens of thousands bps away from their target genes, known as distal CREs[59,60]. These distal regulatory elements can modulate the rate and timing of gene expression, thereby ensuring the precise control of gene transcription. Compared to proximal CREs, distal CREs are highly variable and are more sensitive to a wide range of environment factors. Including the distal CREs in the prediction of TRs can certainly improve the accuracy. However, as we mentioned earlier, measuring long-range interaction between distal CREs and their target directly is challenging due to the limited availability of such data.

The distal-elements-inclusive methods include BART and Lisa attempt to include the activity of distal CREs by using H3K27ac ChIP-seq datasets. H3K27ac is a type of Histone protein and is 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 few H3K27ac ChIP-seq datasets that can be predictive for the transcription of input genes. The rationale of the two methods is that TR should have a high number of peaks align with the high signal peaks in these selected samples.

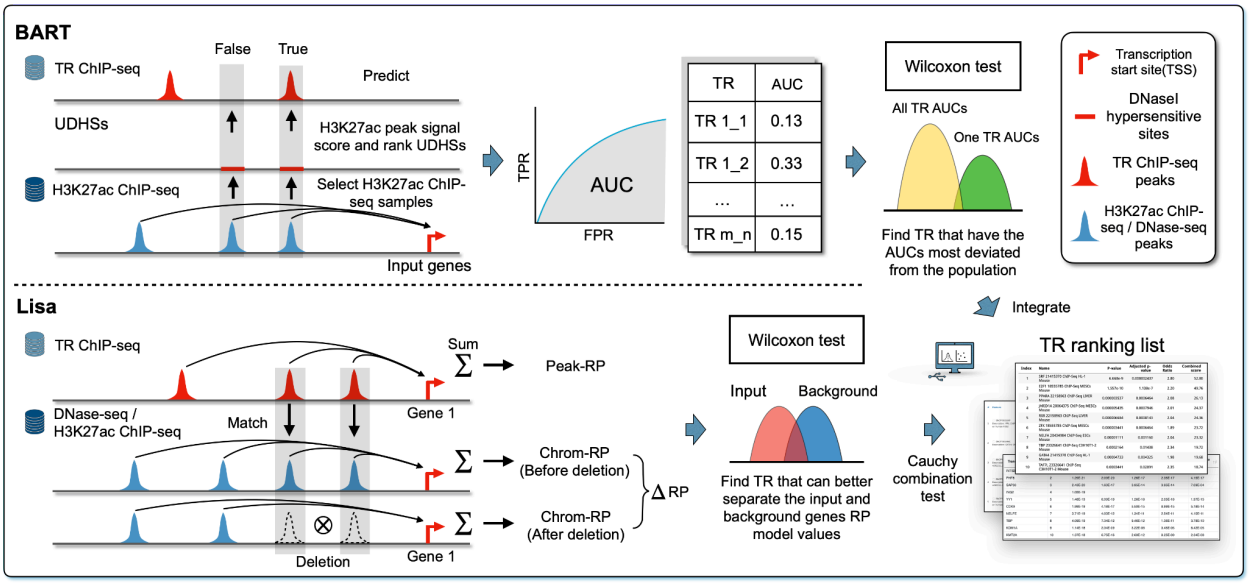

Figure 5. The core idea of BART and Lisa. Distal-elements inclusive methods first select predictive H3K27ac ChIP-seq / DNase-seq samples that quantify active distal cis-regulatory elements, which is achieved by summarizing the H3K27ac ChIP-seq peak signal around gene's TSS. Next, statistical methods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

In the selection step, a regulatory potential (RP) model is implemented in both methods, which counts the signals within the surrounding region of each gene and assign a RP score to each gene. Therefore, each H3K27ac dataset have different RP scores for each individual gene. Following this, predictive datasets are chosen using a stepwise logistic regression model. This model treats the presence of genes in the input as binary outcome variables and employs the RP scores of each gene in each dataset as independent variables.

After selecting predictive H3K27ac datasets, BART collects genome-wide DHSs and ranks these sites by the weighted linear combination of signals from selected datasets

on the same site, and then use the highly ranked sites to predict the peak presence in each TR ChIP-seq datasets and calculates the AUC value using false positive rate (FPR) and true positive rate (TPR). The AUC values of ChIP-seq datasets are grouped by the same TR and compared through the Wilcoxon rank sum test. Finally, BART assigns each TR a unique rank based on the average rank of AUC values, p-value from the test, and the standardized Wilcoxon statistic score.

In addition to H3K27ac datasets, Lisa also can select predictive DNase-seq datasets in the same manner; RP scores are referred as “chrom-RP” for the selected datasets in both cases. After the selection, Lisa matches the peaks from each TR ChIP-seq datasets with the selected datasets, deleting all matched signals to reproduce the chrom-RP for each gene. The coefficients from the fitted logistic regression model are used as weights to calculate a weighted linear combination of the paired chrom-RP changes before and after the deletion, denoted as ΔRP for each gene. Lisa further calculates a “Peak-RP” for each gene by using peaks from the TR ChIP-seq dataset. Therefore, each TR ChIP-seq dataset has both a ΔRP and a peak-RP for each gene in the input and background sets. Finally, ΔRP and peak-RP are used for two separate Wilcoxon tests, the results will be integrated by a Cauchy combination tests to derive final significance for each TR ChIP-seq datasets.

While both methods incorporate additional data from various NGS data types to assess the activity of distal CREs, they still assign greater weight to CREs located closer to genes. Additionally, there are conflicting evidence suggesting that the presence of H3K27ac alone may not adequately indicate the functionality of identified CREs[61].

This raises doubts about the assumption that more signals equate to higher functionality, similar to why relying on peak signals from TR ChIP-seq to determine the functionality of TR binding sites is problematic. Consequently, accurately quantifying the functionality of CREs to their target genes remains a significant challenge.

Besides, the uneven allocation of research resources across different TRs can cell lines resulting in a disproportionate depth of study for a select few TRs and cell lines compared to others. For example, only a limited number of TRs (e.g., CTCF and TP53) and cell lines (e.g., HeLa and K562) have undergone extensive molecular experimentation due to research preference or the availability of experimental materials. Given that these methods typically conduct thousands of separate ad-hoc statistical tests to assign a significance to each dataset, methods tend to rank TRs with more datasets more frequently to top positions than TRs with only few datasets.

Table 1 offers a comprehensive overview of these eleven NGS-based methods discussed in this study. Most NGS-based methods primarily rely on data from the ENCODE project[62], which serves as a trusted source for quality-controlled NGS data. However, a few methods, such as Lisa and ChIP-Atlas, expand their scope by NGS datasets gathered from individual publications and experiments. While these approaches provide broader coverage of TRs and cell lines, they may come with a trade-off in data quality assurance. Additionally, it is worth noting that the evolution of these published methods has followed a discernible trajectory—from library-based to region-based and eventually to more sophisticated approaches, which reflects that, in

the past two decades, the field of TR prediction has been evolving to adapt to new NGS technologies and tackling more and more complicated scientific questions.

Table 1. *Summary of published 11 NGS-based methods*

Method	NGS data collection	NGS data source	Support epigenomic regions as input	Implementation	Major method to associate TFs with genes	Primary statistical method	Method categorization	Year
Cscan	~900	ENCODE	No	Web tool	Peaks fall into the surrounding fixed width window of gene TSS	Fisher exact test	Library-based	2013
ENCODE ChIP-seq significance tool	~700	ENCODE	No	NA	Peaks fall into the surrounding fixed width window of gene TSS	Hypergeometric test	Library-based	2013
Enrichr	~2800	ENCODE	No	Web tool	Epigenomic distance between peaks and gene TSS	Fisher exact test / Corrected FET	Library-based	2013
i-cisTarget	~1500	ENCODE	Yes	Web tool	Peak signal to score pre-defined CRRs	Recovery analysis	Region-based	2015
RegulatorTrail	>1000	Multiple sources	Yes	Web tool	Peaks fall into the surrounding fixed width window of gene TSS	FET / Binomial test / Hypergeometric test	Library-based	2017
BART	~14000	CistromeDB	Yes	Web tool / Python	MARGE/adptive lasso to score pre-defined CRRs	Wilcoxon rank sum test	Region-based	2018
ChIP-Atlas	~50000	NCBI / DDBJ / ENA	Yes	Web tool	Proportion of peaks fall into gene promoter region	Fisher exact test	Region-based	2018
TFEA.ChIP	~2800	ENCODE / ReMap	No	Web tool / R	Peaks fall into the “elite” regions from GeneHancer	FET / Kolmogorov-Smirnov test / GSEA	Library-based	2019
ChEA3	~2800	ENCODE / ReMap / Literature	No	Web tool	Epigenomic distance between peaks and gene TSS	Corrected Fisher exact test	Library-based	2019
Lisa	~13000	CistromeDB	Yes	Web tool / Python	RP model with in silicon deletion method	Wilcoxon rank sum test / Cauchy combination test	Region-based	2020
MAGIC	~4200	ENCODE	No	Python	Peak signal to rank the genes	Kolmogorov-Smirnov test	Region-based	2020

Benchmarking

Overall study design

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

We used gene sets sourced from KnockTF, a comprehensive human gene expression profile database with TR knockdown/knockout experiments[64]. Gene sets derived from perturbation experiments often serve as a frequently used benchmark for evaluating the performance of TR-related computational methods[65]. In the database, each TR perturbation experiment has at least two control samples and two treatment samples, where Limma[66] was used to derive the differentially expressed (DE) gene set. In total, we retrieved 570 DE gene sets from the database, covering 308 known TRs. We further selected the top 200, 600, and 1000 most significant genes from each input gene set, respectively. We transformed the annotated genes to the required format following the

manual or instructions. For each method considered, we used the default settings when applicable, otherwise the parameters were chosen according to the documentation.

The accuracy of a method was evaluated by its ability to rank the perturbed TRs to the top positions within the TR ranking list, the ranking of first relevent term in the list is used. The metrics used in our evaluation include (1) the number of perturbed TRs can be ranked in top 10, top 50, and top 100, respectively, (2) the number of perturbed TRs can be ranked in top 1%, top 5%, and top 10%, respectively, within all the available terms in the TR ranking list, and (3) mean reciprocal rank (MRR) that values the first relevant term in a ranking list.

$$MRR_i = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{R_{ij}},$$

Where R_{ij} is the rank that method i assigns to the perturbed TR from which gene set j was derived, and N is the total number of gene sets used for evaluation. Note that only the first rank position was used in calculating MRR, and further relevant rankings were ignored. A larger MRR value suggests the method can better rank the relevant term to the top positions. Methods with small MRR values typically have low accuracy, and so cannot provide valuable insights into the prediction of TRs, impeding subsequent research.

Another major consideration is the coverage of a method, which relates to the coverage of TRs and cell lines by the NGS datasets collected in its internal database. If there is no available NGS data covering a specific TR, the name of the TR cannot even be

displayed in the final ranking list. Therefore, the method's coverage directly influences its performance. A higher coverage allows the inclusion of a wider range of cell lines and TRs, thereby increasing the chance of correctly identifying the TRs. It is hypothesized that this dimension favors methods that are regularly updated or most recently developed since they typically offer broader coverage by incorporating the latest data and discoveries.

In addition, it is commonly observed that methods with good performance can face challenges in their real-world implementation. Thus, the usability of a method is determined by its ease of implementation, along with the quality of documentation and code. We have defined "usability" based on seven criteria, each scored as either 0 or 1. These criteria include (1) availability of a web tool for direct usage, (2) presence of a CLI/API or R/Python package supporting large-scale analysis, (3) provision of a detailed manual, tutorial, and examples by the authors, (4) open-source code availability, (5) regular updates and maintenance of the method, (6) absence of unexpected errors during the evaluation process, and (7) ease of interpreting results. These seven criteria encompass the most important considerations for users when selecting a method.

Finally, this study evaluates the sensitivity of methods, focusing on their ability to produce consistent and reliable results. We primarily examine three types of biases which could potentially be addressed in future methodological advancements. We tested methods with $G = 200, 600, 1000$ genes that are most significant, respectively. As G increases, the number of less significant genes in the gene set tends to increase, which may negatively affect the performance. Next, we investigated whether TRs

associated with multiple datasets are significantly higher ranked compared to TRs only have a single dataset. Lastly, we examined if TRs with a higher number of peaks consistently achieve higher rankings, as peaks that fall into the proximial regions may not proportionally increase with the total number.

Besides the evaluation framework, it should be clarified that ENCODE ChIP-seq significance tool, were not evaluated for being perished, and the web tool is unable to access. Also, for Enrichr, the mean ranks from four libraries are used and for ChEA3 the integrated mean rank is recommended in the original publication and used in the benchmarking.

Accuracy

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

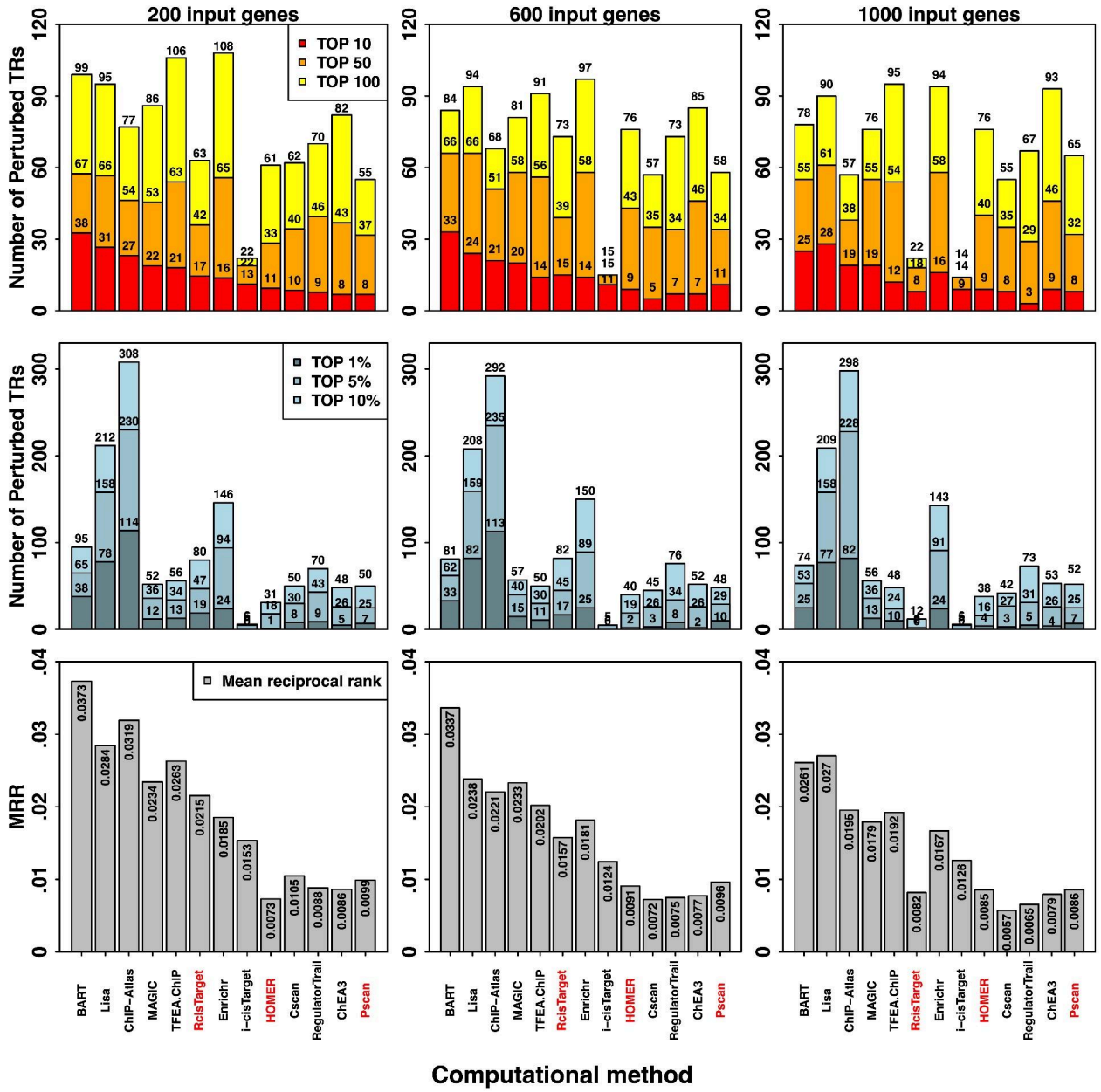

Figure 6.*Comparison of accuracy. Methods with red labels are motif-based. BART, Lisa, and ChIP-Atlas demonstrate the best overall performance across different scenarios. The NGS-based methods generally outperform the motif-based methods. All methods ranked fewer than 10% of perturbed TRs within the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

First of all, upon a closer examination of Figure 6, 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 more frequently rank perturbed TRs in the top 10 and top 1%. These methods also have higher MRR scores. The good performance is partially credited to the extensive collections of NGS data: ChIP-Atlas contains over 30,000 human TR ChIP-seq datasets, while BART and Lisa have around 7,000 human datasets, while other methods only have 1,000 to 3,000 datasets. Furthermore, the region-based approach might also play a pivotal role in enhancing their accuracy.

Second, TFEA.ChIP, Enrichr, and ChEA3 often rank relatively more perturbed TRs within the top 50 or top 100 positions. However, this might be due to the shorter TR ranking lists they produced. Recall that methods can produce TR ranking lists with different lengths, depending on the coverage of TRs and the collected NGS data. Absolute rankings can be in favor of methods that generate shorter TR ranking lists while to the disadvantage of those that produce longer lists. The mean reciprocal rank and top percentile ranking of perturbed TRs serve as complementary metrics to these absolute rankings, showing different patterns in the top panel of Figure 5.

Last, NGS-based methods outperform motif-based methods on average. No motif-based method can surpass the performance of the top-performing NGS-based methods (e.g., BART, ChIP-Atlas, and Lisa), regardless of input gene set sizes and metric used.

Combining results from various methods is believed to be more effective than relying on a single method due to differences in databases and algorithms used. We next explore the combination of two methods to determine the most optimal results. Figure 7 illustrates the number of perturbed TRs that are ranked in the top 10, 50, and 100 and the top 1%, 5%, and 10% in at least one TR ranking list when two methods are used (200 input genes used; see Figure S1 and S2 for 600 and 1000 input genes). We can see that combining Lisa with TFEA.ChIP, ChIP-Atlas, or BART leads to the best or close to the best performance among all pairings. In most cases, combining any of the two methods yields better performance than using individual methods alone.

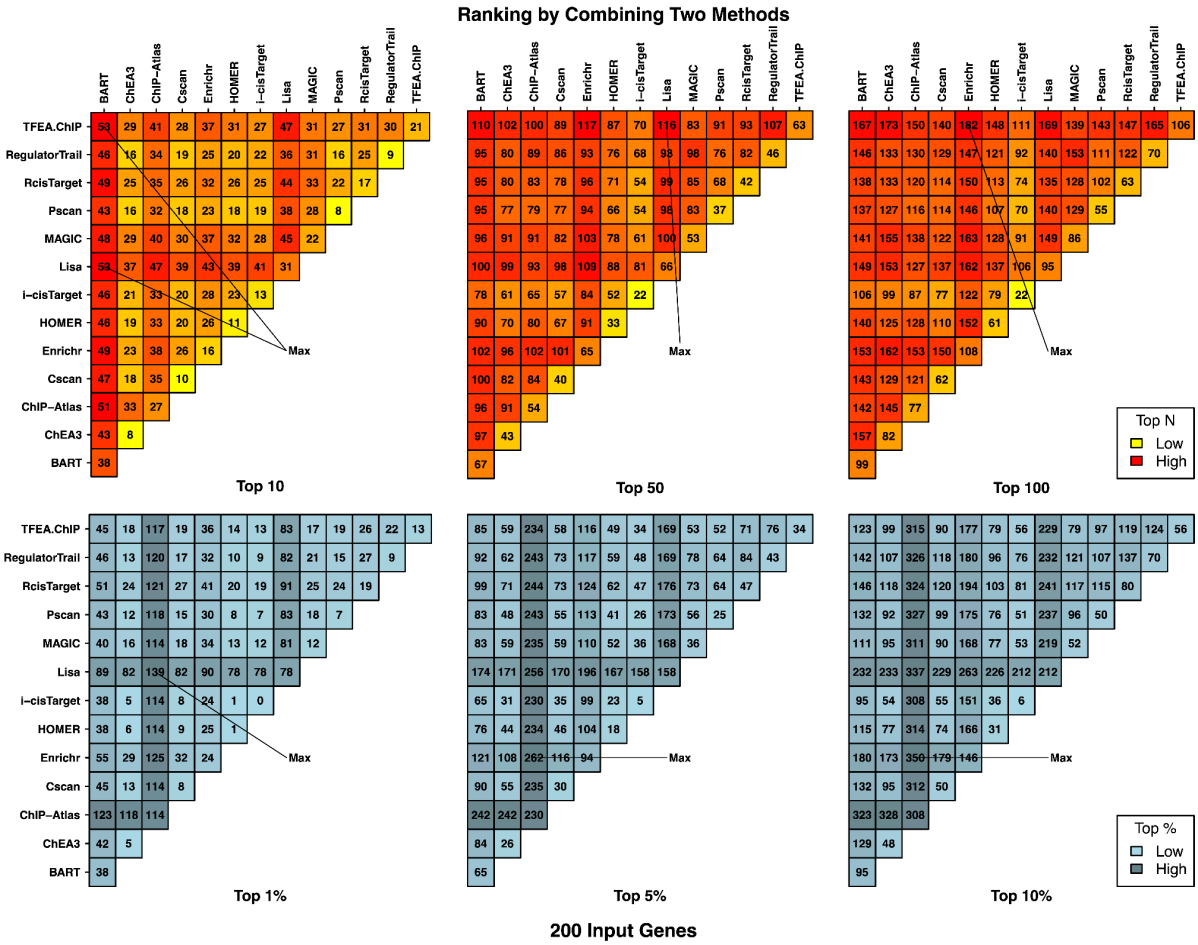

Figure 7. *The number of perturbed TRs ranked to top positions by either method in the combination of two methods (200 input genes used). Lisa paired with TFEA.ChIP, ChIP-Atlas, and BART offer the best performance in most scenarios.*

Figure 8 shows that many methods tend to rank the same TRs to the top regardless of the input gene set. It further shows that these TRs are rarely the perturbed TRs we aim to identify. Ranking certain irrelevant TRs higher may overshadow the true regulators, pushing them down in the ranking list. This problem is most severe for RegulatorTrail, Cscan and MAGIC. Thus, even if the top-ranked TRs match user expectations or prior

knowledge, it is recommended that users check the method's "favorite" TR list before selecting them for further analysis.

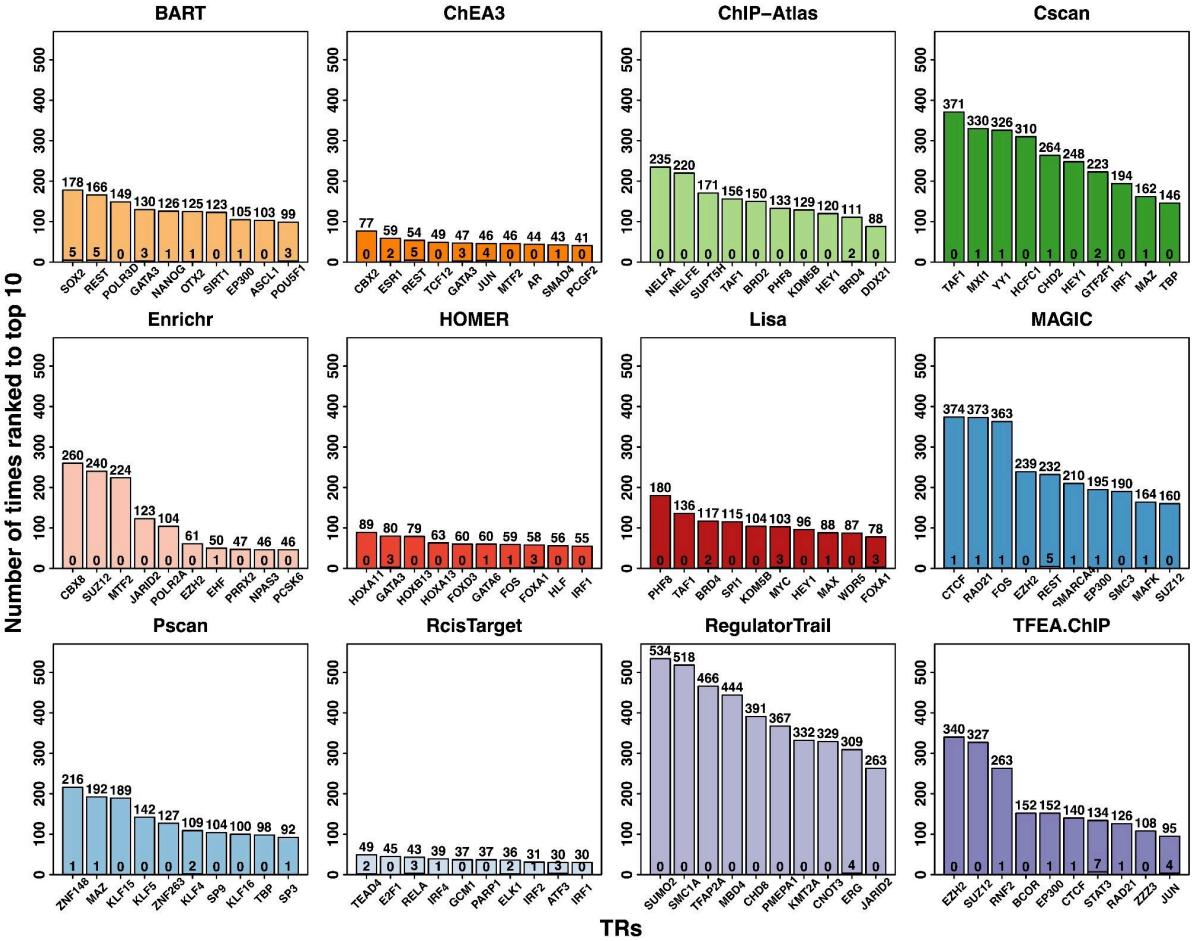

Figure 8. *TRs that are most commonly ranked to the top 10 positions by different methods. The number at the top of each bar indicates how many times the corresponding TR is ranked in the top 10 and that at the bottom indicates how many times that it is indeed the perturbed TR (out of 570). i-cisTarget was not included as the number of TRs in the final ranking list is often less than 10.*

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, as shown in the top panel of the figure. Among the top-performing methods (e.g.,Lisa, ChIP-Atlas, and BART), Lisa is less sensitive to the change of input size. The bottom panel uses the mean rank deviation to quantify the average change in the rank of the same perturbed TR when the input gene set size varies, using a baseline derived from the input of 200 genes. A positive deviation value indicates that the same TRs are placed in relatively lower ranks. Interestingly, methods such as ChIP-Atlas, ChEA3, and RcisTarget exhibit improvements based on this metric. When dividing the perturbed TRs into two subgroups based on whether their initial ranks fall within the top 50% of the TR ranking list, it reveals that these methods tend to push down the top-ranked TRs and elevate the bottom-ranked TRs as the input size expands, indicating more randomness in the rankings.

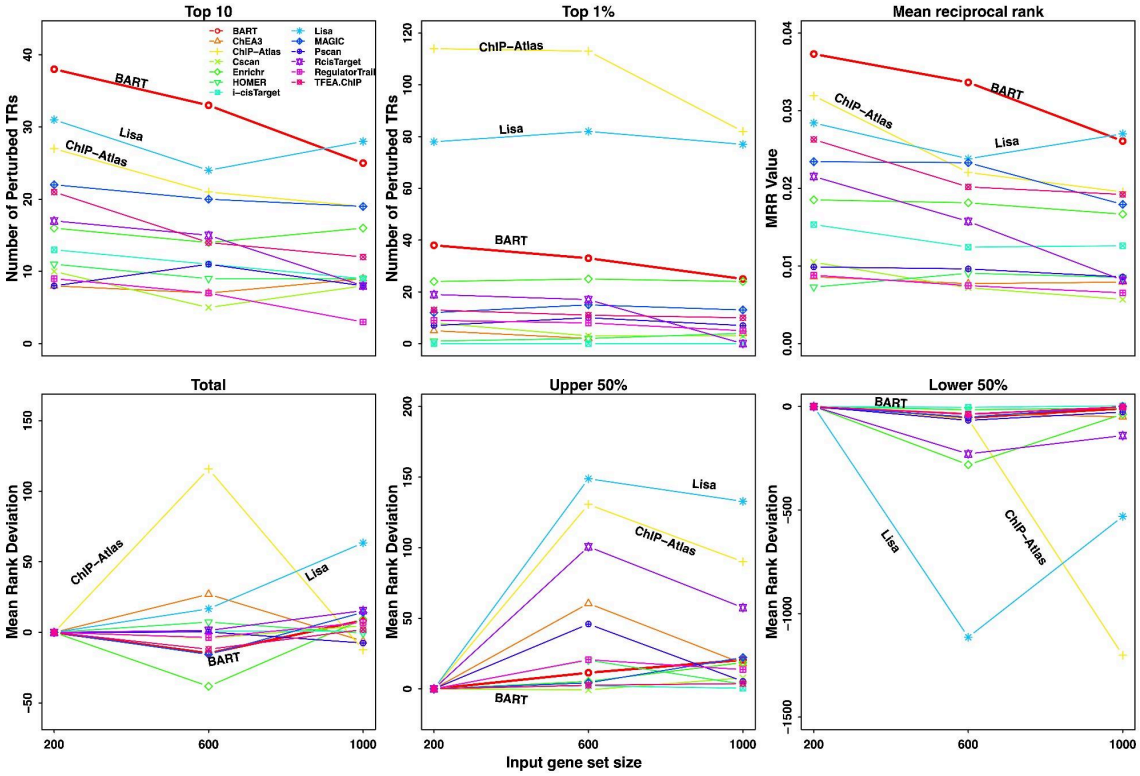

Figure 9. *As the size of the input gene set increases, the majority of methods demonstrate a decline in performance. This is observable in both the number of perturbed TRs appearing in the top ranks and the calculated mean reciprocal ranks (the top three plots). Moreover, as the input size increases, methods often demote high-ranking perturbed TRs while elevating lower-ranking ones, as quantified by the mean rank deviation, indicating overall performance deterioration.*

Except for the input gene set size, another major concern as aforementioned is the uneven allocation of research resources, as some TRs generally can be extensively investigated (Figure S3). We selected 20 TRs with multiple datasets and 20 TRs associated with just one dataset from the ENCODE database, based on their dataset count. For each gene set, we employed a one-sided Wilcoxon rank-sum test to determine if the 20 TRs with multiple datasets consistently ranked significantly higher

than those with a single dataset. Additionally, we excluded any gene set that featured a selected TR as the perturbation target, ensuring that each gene set could be considered a randomly generated set. The findings, detailed in Supplementary Table S1, reveal that most methods indeed tend to favor TRs associated with a larger number of datasets. ChEA3 is the only method that not show significance, which might be due to the integration step of using mean ranks of all datasets as the final rank. This bias might explain why certain TRs consistently appear in the top rankings as shown in Figure 8.

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

Coverage

With the detailed analysis of accuracy, it is evident that coverage plays a crucial role in the effectiveness of methods. The extent to which these methods can incorporate up-to-date NGS data can be a significant factor contributing to the performance. First,

Figure 10A and 10B show that the publication time has a positive correlation with the ranking performance and TR coverage, which confirms that the continuing efforts from generations of researchers have improved the prediction performance of algorithms to address the problem. Second, Figure 10C displays the number of perturbed TRs that are missing in the TR ranking list versus the publication date for each method. As we expect, more recently developed methods generally have a low number of missing TRs. Figure 10D plots the unique terms displayed in the TR ranking list of each method. Considering that the benchmark gene sets used in this study only cover a limited number of known TRs, a high number of missing TRs and a low number of available terms might suggest an insufficient coverage of TRs. Notably, ChIP-Atlas, Lisa, and BART demonstrated their relatively comprehensive coverage of NGS data, which resonates with their better accuracy observed before.

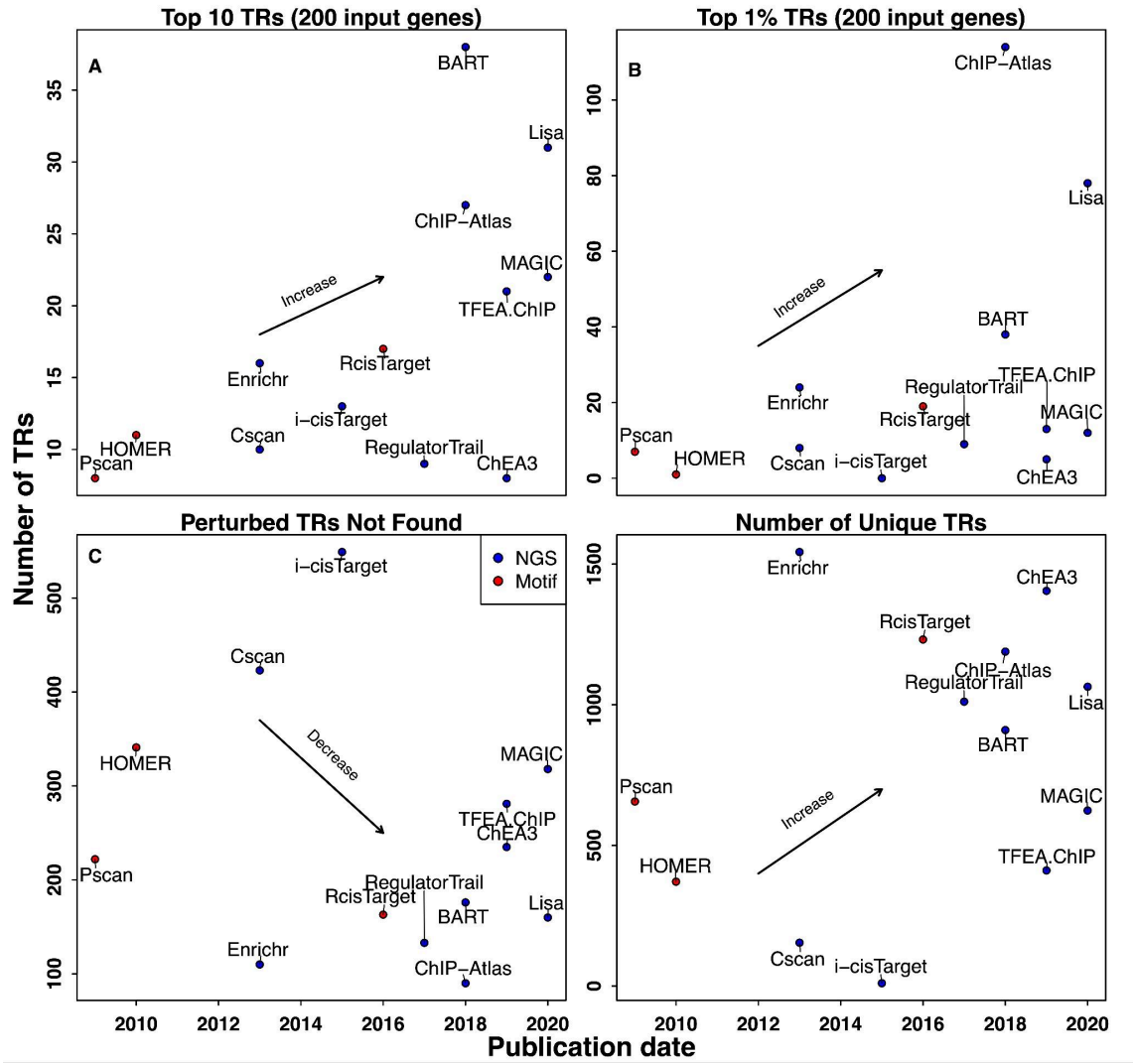

Figure 10. *An increasing trend in prediction performance and TR coverage, aligning with more recent publication dates. A. No. of perturbed TRs ranked within the top 10 positions versus method's publication date; B. No. of perturbed TRs ranked within top 1% versus method's publication date; C. No. of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

Usability

Table 2 provides an assessment of NGS-based methods using various criteria of usability, with a “Yes” indicating that the criterion is satisfied and a “No” otherwise. Enrichr, Lisa, ChIP-Atlas, and ChEA3 received good scores as these methods are generally user-friendly, easy to learn and use, and did not pose any issues during testing. We also found that the web tool version of BART was not functional during the evaluation process. 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.

Table 2. *The evaluation of usability*

Method	Web Tool	large-scale analysis	Documentation quality	Open-Source	Actively Maintained	No error in benchmarking	Easy Interpretation
Cscan	Yes	No	Yes	No	No	Yes	Yes
Enrichr	Yes	Yes	Yes	Yes	Yes	Yes	Yes
i-cisTarget	Yes	No	Yes	No	No	Yes	Yes
RegulatorTrail	Yes	Yes	Yes	No	No	No	Yes
BART	Yes	Yes	Yes	Yes	Yes	No	Yes
ChIP-Atlas	Yes	Yes	Yes	Yes	Yes	Yes	Yes
TFEA.ChIP	Yes	Yes	Yes	Yes	No	Yes	Yes
ChEA3	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Lisa	Yes	Yes	Yes	Yes	Yes	Yes	Yes
MAGIC	No	Yes	No	Yes	No	Yes	Yes

Summary

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 consistent and accurate predictions with perturbed TR gene sets. In addition, it is important to note that even the best-performing method may only rank less than 10% of perturbed TRs within the top 10 positions. Given the complexity of cellular processes and the limitations above, researchers should use caution when interpreting the results and should consider further validation with experimental approaches. Overall, recently developed NGS-based methods provide useful tools for generating hypotheses about transcriptional regulation. Further research is needed to fully understand their potential and limitations.

Discussion

This study evaluated the performance of various computational methods developed over the past decade for identifying TRs. While these NGS-based methods have been popular in the past due to the historical inertia of gene-centric research, they inevitably encounter limitations when associating epigenomic regulatory information with genes in the epigenomics era. Issues such as hard cutoffs, difficulties in differentiating functional/non-functional binding sites, unreliable binding affinity measurement, and failure to include distal cis-regulatory elements make it challenging to identify the correct

TRs. Therefore, new experimental and computational methods are urgently needed to address these challenges. Alternatively, the growing availability of epigenomics techniques makes it more attractive to directly combine epigenomics datasets to study the underlying mechanisms behind a cellular process. For example, the combination of DNase-seq/H3K27ac ChIP-seq data with the TR ChIP-seq data in Lisa has proved its feasibility[40].

Previous studies have implicated several possible directions to further 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[67]. However, this approach is limited to TRs with long residence times on chromatin, despite some TRs being known for lack of detectable footprints[68,69]. Second, including three-dimensional long-range interaction information from Hi-C data may also serve as a possible improvement direction[70–72], 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)[73]. In contrast to bulk sequencing data, these techniques sequence each cell separately and

provide a more granular view of the differences between individual cells[74]. This means that limitations such as unreliable 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 TRs should consider integrating single-cell data to improve their accuracy and resolution. In addition, with the rise of multi-omics experimental approaches[75,76], the availability of multi-omics data is no longer restricted to one or two types, and there is a growing need for novel multi-omics approaches that can leverage the full range of available data. In the coming decade, it is expected that more multi-omics-based methods will be developed, which will undoubtedly lead to further innovation and discovery in this field.

Key points

- An introduction to available computational methods for predicting functional TRs from a query gene set.
- A detailed walk-through along with the practical concerns and limitations.
- A systematic benchmark of thirteen methods by using 570 TR perturbation-derived gene sets, including accuracy, usability, coverage, and sensitivity.
- NGS-based methods perform better than motif-based methods, while methods with large databases and region-centric approaches perform better.

- BART, ChIP-Atlas, and Lisa are recommended as these methods have overall better performance in evaluated scenarios.

Data availability

Commands and scripts used to run each method can be accessed at https://github.com/ZeyuL01/Benchmark_MasterTR. Input genes are sourced from KnockTF[64] (https://bio.liclab.net/KnockTFv1/). NGS data are sourced from ENCODE[77] (https://www.encodeproject.org/). Top 200, 600, 1000 genes and TR ranking lists generated by each method are available at https://smu.box.com/s/hlofza4oe08ob4bb4i02nu3wl2zy39zj.

Funding

Reference

1. Jacob F, Monod J. Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol. 1961; 3:318–356

2. Latchman DS. Transcription factors: An overview. Int. J. Biochem. Cell Biol. 1997; 29:1305–1312

3. Vaquerizas JM, Kummerfeld SK, Teichmann SA, et al. A census of human transcription factors: function, expression and evolution. Nat. Rev. Genet. 2009; 10:252–263

4. Mullen AC, Orlando DA, Newman JJ, et al. Master Transcription Factors Determine Cell-Type-Specific Responses to TGF-β Signaling. Cell 2011; 147:565–576

5. Whyte WA, Orlando DA, Hnisz D, et al. Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. Cell 2013; 153:307–319

6. Hwang J-Y, Zukin RS. REST, a master transcriptional regulator in neurodegenerative disease. Curr. Opin. Neurobiol. 2018; 48:193–200

7. Thomas MC, Chiang C-M. The general transcription machinery and general cofactors. Crit. Rev. Biochem. Mol. Biol. 2006; 41:105–178

8. Chen T, Dent SY. Chromatin modifiers and remodellers: regulators of cellular differentiation. Nat. Rev. Genet. 2014; 15:93–106

9. Vandewalle C, Van Roy F, Berx G. The role of the ZEB family of transcription factors in development and disease. Cell. Mol. Life Sci. 2008; 66:773–787

10. Burda JE, O'Shea TM, Ao Y, et al. Divergent transcriptional regulation of astrocyte reactivity across disorders. Nature 2022; 606:557–564

11. Darnell Jr JE. Transcription factors as targets for cancer therapy. Nat. Rev. Cancer 2002; 2:740–749

12. Bhagwat AS, Vakoc CR. Targeting transcription factors in cancer. Trends Cancer 2015; 1:53–65

13. Bushweller JH. Targeting transcription factors in cancer—from undruggable to reality. Nat. Rev. Cancer 2019; 19:611–624

14. Frith MC. Detection of functional DNA motifs via statistical over-representation. Nucleic Acids Res. 2004; 32:1372–1381

15. Zambelli F, Pesole G, Pavesi G. Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes. Nucleic Acids Res. 2009; 37:W247–W252

16. McLeay RC, Bailey TL. Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. BMC Bioinformatics 2010; 11:

17. Heinz S, Benner C, Spann N, et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol. Cell 2010; 38:576–589

18. Jolma A, Yan J, Whitington T, et al. DNA-Binding Specificities of Human Transcription Factors. Cell 2013; 152:327–339

19. Lambert SA, Jolma A, Campitelli LF, et al. The Human Transcription Factors. Cell 2018; 172:650–665

20. Wooten DJ, Groves SM, Tyson DR, et al. Systems-level network modeling of Small Cell Lung Cancer subtypes identifies master regulators and destabilizers. PLOS

Comput. Biol. 2019; 15:e1007343

21. Su K, Katebi A, Kohar V, et al. NetAct: a computational platform to construct core transcription factor regulatory networks using gene activity. Genome Biol. 2022; 23:270

22. Kamimoto K, Stringa B, Hoffmann CM, et al. Dissecting cell identity via network inference and in silico gene perturbation. Nature 2023; 614:742–751

23. Delgado FM, Gómez-Vela F. Computational methods for Gene Regulatory Networks reconstruction and analysis: A review. Artif. Intell. Med. 2019; 95:133–145

24. Mercatelli D, Scalambra L, Triboli L, et al. Gene regulatory network inference resources: A practical overview. Transcr. Profiles Regul. Gene Netw. 2020; 1863:194430

25. Nguyen H, Tran D, Tran B, et al. A comprehensive survey of regulatory network inference methods using single cell RNA sequencing data. Brief. Bioinform. 2021; 22:bbaa190

26. Furey TS. ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. Nat. Rev. Genet. 2012; 13:840–852

27. Creyghton MP, Cheng AW, Welstead GG, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc. Natl. Acad. Sci. 2010; 107:21931–21936

28. Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. Nat. Rev. Genet. 2019; 20:207–220

29. Jiang S, Mortazavi A. Integrating ChIP-seq with other functional genomics data. Brief. Funct. Genomics 2018; 17:104–115

30. Zambelli F, Prazzoli GM, Pesole G, et al. Cscan: finding common regulators of a set of genes by using a collection of genome-wide ChIP-seq datasets. Nucleic Acids Res. 2012; 40:W510–W515

31. Auerbach RK, Chen B, Butte AJ. Relating genes to function: identifying enriched transcription factors using the ENCODE ChIP-Seq significance tool. Bioinformatics 2013; 29:1922–1924

32. Chen EY, Tan CM, Kou Y, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics 2013; 14:128

33. Imrichová H, Hulselmans G, Kalender Atak Z, et al. i-cisTarget 2015 update: generalized cis-regulatory enrichment analysis in human, mouse and fly. Nucleic Acids Res. 2015; 43:W57–W64

34. Kehl T, Schneider L, Schmidt F, et al. RegulatorTrail: a web service for the identification of key transcriptional regulators. Nucleic Acids Res. 2017; 45:W146–W153

35. Wang Z, Civelek M, Miller CL, et al. BART: a transcription factor prediction tool with query gene sets or epigenomic profiles. Bioinformatics 2018; 34:2867–2869

36. Oki S, Ohta T, Shioi G, et al. Ch IP -Atlas: a data-mining suite powered by full integration of public Ch IP -seq data. EMBO Rep. 2018; 19:

37. Puente-Santamaria L, Wasserman WW, del Peso L. TFEA.ChIP: a tool kit for transcription factor binding site enrichment analysis capitalizing on ChIP-seq datasets. Bioinformatics 2019; 35:5339–5340

38. Keenan AB, Torre D, Lachmann A, et al. ChEA3: transcription factor enrichment analysis by orthogonal omics integration. Nucleic Acids Res. 2019; 47:W212–W224

39. Roopra A. MAGIC: A tool for predicting transcription factors and cofactors driving gene sets using ENCODE data. PLOS Comput. Biol. 2020; 16:e1007800

40. Qin Q, Fan J, Zheng R, et al. Lisa: inferring transcriptional regulators through integrative modeling of public chromatin accessibility and ChIP-seq data. Genome Biol. 2020; 21:

41. Parolia A, Cieslik M, Chu S-C, et al. Distinct structural classes of activating FOXA1 alterations in advanced prostate cancer. Nature 2019; 571:413–418

42. Nami Yamashita, Mark Long, Atsushi Fushimi, et al. MUC1-C integrates activation of the IFN-γ pathway with suppression of the tumor immune microenvironment in triple-negative breast cancer. J. Immunother. Cancer 2021; 9:e002115

43. March-Diaz R, Lara-Ureña N, Romero-Molina C, et al. Hypoxia compromises the mitochondrial metabolism of Alzheimer's disease microglia via HIF1. Nat. Aging 2021; 1:385–399

44. Duplaquet L, Li Y, Booker MA, et al. KDM6A epigenetically regulates subtype plasticity in small cell lung cancer. Nat. Cell Biol. 2023; 25:1346–1358

45. Fishilevich S, Nudel R, Rappaport N, et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database 2017; 2017:

46. Kempfer R, Pombo A. Methods for mapping 3D chromosome architecture. Nat. Rev. Genet. 2020; 21:207–226

47. Bulger M, Groudine M. Functional and Mechanistic Diversity of Distal Transcription Enhancers. Cell 2011; 144:327–339

48. Rowley MJ, Corces VG. Organizational principles of 3D genome architecture. Nat.

Rev. Genet. 2018; 19:789–800

49. Gorkin DU, Leung D, Ren B. The 3D Genome in Transcriptional Regulation and Pluripotency. Cell Stem Cell 2014; 14:762–775

50. Zheng H, Xie W. The role of 3D genome organization in development and cell differentiation. Nat. Rev. Mol. Cell Biol. 2019; 20:535–550

51. Whitfield TW, Wang J, Collins PJ, et al. Functional analysis of transcription factor binding sites in human promoters. Genome Biol. 2012; 13:R50

52. Wittkopp PJ, Kalay G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nat. Rev. Genet. 2011; 13:59–69

53. Riethoven J-JM. Regulatory Regions in DNA: Promoters, Enhancers, Silencers, and Insulators. Comput. Biol. Transcr. Factor Bind. 2010; 33–42

54. Haberle V, Stark A. Eukaryotic core promoters and the functional basis of transcription initiation. Nat. Rev. Mol. Cell Biol. 2018; 19:621–637

55. Cramer P. Organization and regulation of gene transcription. Nature 2019; 573:45–54

56. Fisher WW, Li JJ, Hammonds AS, et al. DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in *Drosophila*. Proc. Natl. Acad. Sci. 2012; 109:21330–21335

57. Slattery M, Zhou T, Yang L, et al. Absence of a simple code: how transcription factors read the genome. Trends Biochem. Sci. 2014; 39:381–399

58. Kribelbauer JF, Rastogi C, Bussemaker HJ, et al. Low-Affinity Binding Sites and the Transcription Factor Specificity Paradox in Eukaryotes. Annu. Rev. Cell Dev. Biol. 2019;

35:357–379

59. Heintzman ND, Ren B. Finding distal regulatory elements in the human genome. Genomes Evol. 2009; 19:541–549

60. Ricci WA, Lu Z, Ji L, et al. Widespread long-range cis-regulatory elements in the maize genome. Nat. Plants 2019; 5:1237–1249

61. Zhang T, Zhang Z, Dong Q, et al. Histone H3K27 acetylation is dispensable for enhancer activity in mouse embryonic stem cells. Genome Biol. 2020; 21:45

62. Consortium TEP. An integrated encyclopedia of DNA elements in the human genome. Nature 2012; 489:57–74

63. Aibar S, González-Blas CB, Moerman T, et al. SCENIC: single-cell regulatory network inference and clustering. Nat. Methods 2017; 14:1083–1086

64. Feng C, Song C, Liu Y, et al. KnockTF: a comprehensive human gene expression profile database with knockdown/knockout of transcription factors. Nucleic Acids Res. 2019; 48:D93–D100

65. Garcia-Alonso L, Holland CH, Ibrahim MM, et al. Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Res. 2019; 29:1363–1375

66. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015; 43:e47–e47

67. Sung M-H, Guertin MJ, Baek S, et al. DNase Footprint Signatures Are Dictated by Factor Dynamics and DNA Sequence. Mol. Cell 2014; 56:275–285

68. Sung M-H, Baek S, Hager GL. Genome-wide footprinting: ready for prime time? Nat.

Methods 2016; 13:222–228

69. Baek S, Goldstein I, Hager GL. Bivariate Genomic Footprinting Detects Changes in Transcription Factor Activity. Cell Rep. 2017; 19:1710–1722

70. Belton J-M, McCord RP, Gibcus JH, et al. Hi–C: A comprehensive technique to capture the conformation of genomes. Methods 2012; 58:268–276

71. Xu H, Zhang S, Yi X, et al. Exploring 3D chromatin contacts in gene regulation: The evolution of approaches for the identification of functional enhancer-promoter interaction. Comput. Struct. Biotechnol. J. 2020; 18:558–570

72. Lohia R, Fox N, Gillis J. A global high-density chromatin interaction network reveals functional long-range and trans-chromosomal relationships. Genome Biol. 2022; 23:

73. Stuart T, Satija R. Integrative single-cell analysis. Nat. Rev. Genet. 2019; 20:257–272

74. Kulkarni A, Anderson AG, Merullo DP, et al. Beyond bulk: a review of single cell transcriptomics methodologies and applications. Curr. Opin. Biotechnol. 2019; 58:129–136

75. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017; 18:

76. Huang S, Chaudhary K, Garmire LX. More Is Better: Recent Progress in Multi-Omics Data Integration Methods. Front. Genet. 2017; 8:

77. Luo Y, Hitz BC, Gabdank I, et al. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. Nucleic Acids Res. 2020; 48:D882–D889