BIT: Bayesian Identification of Transcriptional Regulators from Epigenomics-Based Query Region Sets

Zeyu Lu[1,3], Lin Xu[2,5*], Xinlei Wang[3,4*]

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. Division of Data Science, 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 authors:

Xinlei Wang (xinlei.wang@uta.edu)

Lin Xu (lin.xu@utsouthwestern.edu)

Abstract

Transcriptional regulators (TRs) are master controllers of gene expression and play a critical role in both normal tissue development and disease progression. However, existing computational methods for identification of TRs regulating specific biological processes have significant limitations, such as relying on distance on a linear chromosome or binding motifs that have low specificity. Many also use statistical tests in ways that lack interpretability and rigorous confidence measures. We introduce BIT, a novel Bayesian hierarchical model for *in-silico* TR identification. Leveraging a comprehensive library of TR ChIP-seq data, BIT offers a fully integrated Bayesian approach to assess genome-wide consistency between user-provided epigenomic profiling data and the TR binding library, enabling the identification of critical TRs while quantifying uncertainty. It avoids estimation and inference in a sequential manner or numerous isolated statistical tests, thereby enhancing accuracy and interpretability. BIT successfully identified critical TRs in perturbation experiments, functionally essential TRs in various cancer types, and cell-type-specific TRs within heterogeneous cell populations, offering deeper biological insights into transcriptional regulation.

Key Words: ATAC-seq, Bayesian hierarchical model, chromatin accessibility profiling, epigenomics, Jaccard index, Pólya-Gamma data augmentation.

Introduction

Transcriptional regulators (TRs), comprising transcription factors, cofactors, chromatin modulators, and various regulatory molecules, play a key role in regulating gene transcription[1-3]. TRs are essential in diverse cellular processes, including growth, differentiation, morphogenesis, and apoptosis[4-6]. Dysfunction of TRs can lead to various diseases such as cancers[7-9], diabetes[10,11], and heart diseases[12,13]. Thus, gaining insights into the identity and function of TRs that regulate specific biological processes is essential in a broad range of biological and medical research, particularly in finding new biomarkers and therapeutic targets[14-16]. Because traditional wet lab experiments to identify TRs are labor-intensive and time-consuming, researchers are increasingly turning to efficient statistical and machine learning methods for TR identification, especially those leveraging next-generation sequencing (NGS) data.

The *in-silico* identification of TRs regulating a specific biological process typically requires two key components. The first component is input data that are often generated from cost-effective techniques, such as a set of genes identified from transcriptomics techniques (e.g., microarray, RNA-seq) or a set of accessible chromatin regions derived from epigenomic profiling techniques[17] (e.g., ATAC-seq). These Input data typically reflect the downstream effects of key TR activity or provide epigenomic features closely associated with TR binding sites. For example, differential gene expression between diseased and normal samples is often attributed to the change in key TR activity[18-20].

In our recent comprehensive review[21], we discussed the limitations of using gene sets as input. This approach requires assigning each gene a representative chromatin location, typically the transcription start site (TSS). The association between a TR and a gene is then assessed based on the linear distance between the binding sites of the TR and the TSS of this gene, which often underestimates long-range interactions, such as distant enhancers that locate thousands of base

58 pairs away from their target genes[22]. In contrast, alternative input such as chromatin accessibility
59 profiling data directly correlate with TR binding activity[23-25]. As accessible chromatin is a
60 prerequisite for TR binding, functional TRs can be inferred as those with binding sites enriched in
61 the input accessible chromatin regions, thereby avoiding the use of linear distance. In addition,
62 epigenomic profiling techniques such as ATAC-seq have gained increasing popularity over the
63 past decade. The abundance of such datasets has stimulated the development of computational
64 methods to uncover new insights into transcriptional regulation.

65 The second component of *in-silico* TR identification is a reference library consisting of TR ChIP-
66 seq data or TR binding motifs. TRs bind to specific cis-regulatory elements to exert their regulatory
67 function[2], and thus, data that accurately capture TR binding sites serve as a valuable reference.
68 TR ChIP-seq provides an accurate representation of in-vivo TR binding profile[26]. However, each
69 ChIP-seq dataset can only decipher one TR's binding profile under a specific condition. It is not
70 feasible to conduct thousands of ChIP-seq experiments for one single condition to measure every
71 TR's binding profile. Therefore, researchers often rely on computational methods to infer
72 functional TRs based on a library of binding profiles collected from thousands of existing ChIP-
73 seq experiments that cover many known TRs. In addition, TR binding sites can also be inferred
74 by TR binding motifs, which are DNA sequence patterns that reveal TR binding preference.

75 Despite tremendous successes in advancing biomedical research, existing computational
76 methods can suffer from various limitations, regardless of the reference type. (i) For methods (e.g.,
77 WhichTF[27] and HOMER[28]) that leverage TR binding motifs as reference, the highly similar motifs
78 shared among many TRs make accurate identification challenging[29]. In addition, motifs alone
79 cannot characterize any context-specific binding pattern, and the lack of well-defined motifs for
80 some TRs further reduces the reliability of using motifs for TR identification[2,30]. (ii) For methods
81 using TR ChIP-seq data (e.g., BART[31] and ChIP-Atlas[32]), each TR's binding profile can vary in
82 different cell types or environmental conditions. However, published methods have no rigorous

treatment of the within-TR heterogeneity when dealing with multiple binding profiles from the same TR. Further, TRs with more datasets can benefit from randomness and have a greater chance of being ranked higher in the final output, leading to biased results[21]. (iii) Most existing methods use frequentist approaches for measuring TR's statistical significance, lacking coherent probabilistic interpretation or uncertainty estimation. Given the dynamic nature of TR binding activity, the ability to quantify uncertainty is not just beneficial but essential.

To address these limitations, we propose a novel Bayesian hierarchical model for TR identification, named BIT (**B**ayesian **I**dentification of **T**ranscriptional regulators from epigenomics-based query regions sets). The input of BIT is a set of user-provided epigenomic regions containing the chromosome numbers and coordinates of the regions, typically referred to as "peaks", often derived from genome-wide epigenomic profiling of specific biological processes. Compared to existing methods, BIT offers the following advantages.

First, BIT leverages over 10,000 TR ChIP-seq datasets gathered from previous studies covering a large collection of TRs. This approach bypasses the need of computationally predicted motifs, which can lead to unreliable TR predictions in motif-based methods[28,33,34]. In addition, the context-specific binding patterns captured by TR ChIP-seq will not be neglected.

Second, through a unique hierarchical model setup, BIT integrates information across multiple TRs and across multiple ChIP-seq datasets of the same TRs, appropriately accounting for both between- and within-TR heterogeneity in binding profiles. This formal model-based approach enables BIT to estimate the uncertainty associated with varying context-specific binding patterns existing within different ChIP-seq datasets for the same TR, leading to a more reliable and robust prediction, as shown in the analysis results of this study. By contrast, many existing methods need to conduct thousands of separate *ad-hoc* statistical tests (e.g., BART[31] and ChIP-Atlas[32]), which often give higher rankings to TRs with more ChIP-seq datasets[21].

Third, BIT employs a Bayesian framework to handle its modeling and computational needs. Besides information pooling for improved prediction, a Bayesian approach is inherently attractive because (i) it offers the advantage of readily quantifying the estimation uncertainty, such as providing 95% credible intervals for any quantity of interest (e.g., an estimated parameter or predicted importance score); (ii) in situations when meaningful prior knowledge is available, it enables researchers to formally incorporate such knowledge through prior specification for improved results. In all other situations, with the default non-informative prior setting, it becomes a fully automated procedure that does not require any tuning parameter. In contrast, none of the existing methods can offer all these advantages altogether. The R package of BIT is available on GitHub (https://github.com/ZeyuL01/BIT) along with a detailed manual. We also provide an easy-to-use online portal (http://43.135.174.109:8080/).

Results

The integration of epigenomic profiling data has proven to be a feasible approach for identifying TRs in biomedical research[24,35,36]. It is well recognized that the binding sites of TRs critical to a specific biological process are primarily enriched in active regulatory regions of downstream genes that drive the process[37]. These epigenomic regions can be profiled by sequencing techniques[17] (e.g., ATAC-seq), which offer a 'snapshot' of genome-wide TR activity, akin to distinct 'fingerprints' at a crime scene, providing the clues about individuals involved. To link this TR fingerprints snapshot to specific TRs, we leverage TR ChIP-seq data. In humans and mice, there are approximately 1600 known TRs regulating a wide variety of biological processes, and tens of thousands of TR ChIP-seq experiments have been conducted, each creating a unique reference "fingerprint" for an individual TR (under some specific condition). To use this wealth of data, we pre-processed a substantial collection of TR ChIP-seq datasets, creating a comprehensive library of TR binding profiles (distinctive "fingerprints" for individual TRs) in human and mouse genomes, respectively. We utilized 10,140 human TR ChIP-seq datasets covering 988 TRs, and 5,681 mouse TR ChIP-seq datasets covering 607 TRs. These datasets were sourced from GTRD[38], one of the most comprehensive databases in epigenomic research (**Fig. S1**). Each library serves as a powerful reference. Users can compare their own snapshots of TR "fingerprints" (from ATAC-seq, etc.) with those in the corresponding library to identify TRs whose binding patterns show the highest consistency, indicating their functional relevance in the biological context under study.

BIT operates based on two key biological principles: (1) If a TR is truly involved in regulating a given biological process, its binding pattern across the genome should closely align with the user-provided input (active regulatory regions) more closely than a TR that is irrelevant. A stronger correlation between TR binding sites and these input regions suggests a greater likelihood that the TR regulates key genes, thus significantly impacting the biological process under investigation. (2) Each TR possesses a distinct binding pattern, enabling its identification even in the presence

143 of some variation. This means that the binding patterns of a specific TR will be more similar across
144 different conditions, compared to the binding patterns of other TRs (**Fig. S2**). This inherent
145 specificity allows for accurate recognition and differentiation of TRs within complex biological
146 systems.

147 These principles, also used in other bioinformatics tools[27,31,39], underscore that an active TR
148 involved in a cellular process will exhibit frequent DNA interactions, leading to captured
149 epigenomic regions near its binding sites. In addition, while variations exist in ChIP-seq data for
150 the same TR due to factors like measurement errors, data quality, tissue specificity, and cell-type
151 specificity, the overall consistency stemming from TR binding preferences allows for unique
152 binding characteristics to emerge. This uniqueness, when compared to input epigenomic data,
153 enables the differentiation of functional TRs from non-functional ones.

154 BIT only requires a set of epigenomic regions containing the chromosome numbers and
155 coordinates of the regions as input (e.g., Peaks called by MACS2 with the common format such
156 as bed/narrowPeak) and consists of two main steps: (i) data pre-processing and (ii) Bayesian
157 computation and posterior inference with a Gibbs sampler (see **Methods** and **Supplementary**
158 **Note 1**). First, BIT transforms the coordinates of processed peak files to fixed-length binary
159 vectors with 0s and 1s. Second, BIT formulates a Bayesian hierarchical model to integrate user-
160 provided information with TR binding information from the reference library and then uses Markov
161 Chain Monte Carlo (MCMC) to sample from the joint posterior distribution. Finally, BIT assigns
162 each TR a “BIT score” based on posterior draws (**Fig. 1**). Based on the assumptions stated above,
163 the BIT score can be used as the overall importance measure of the TR being functional.
164 Alternatively, it can be interpreted as the Jaccard index that quantifies the average similarity
165 between the binary vectors of the TR’s binding profiles and the binary vector of the input
166 epigenomic regions.

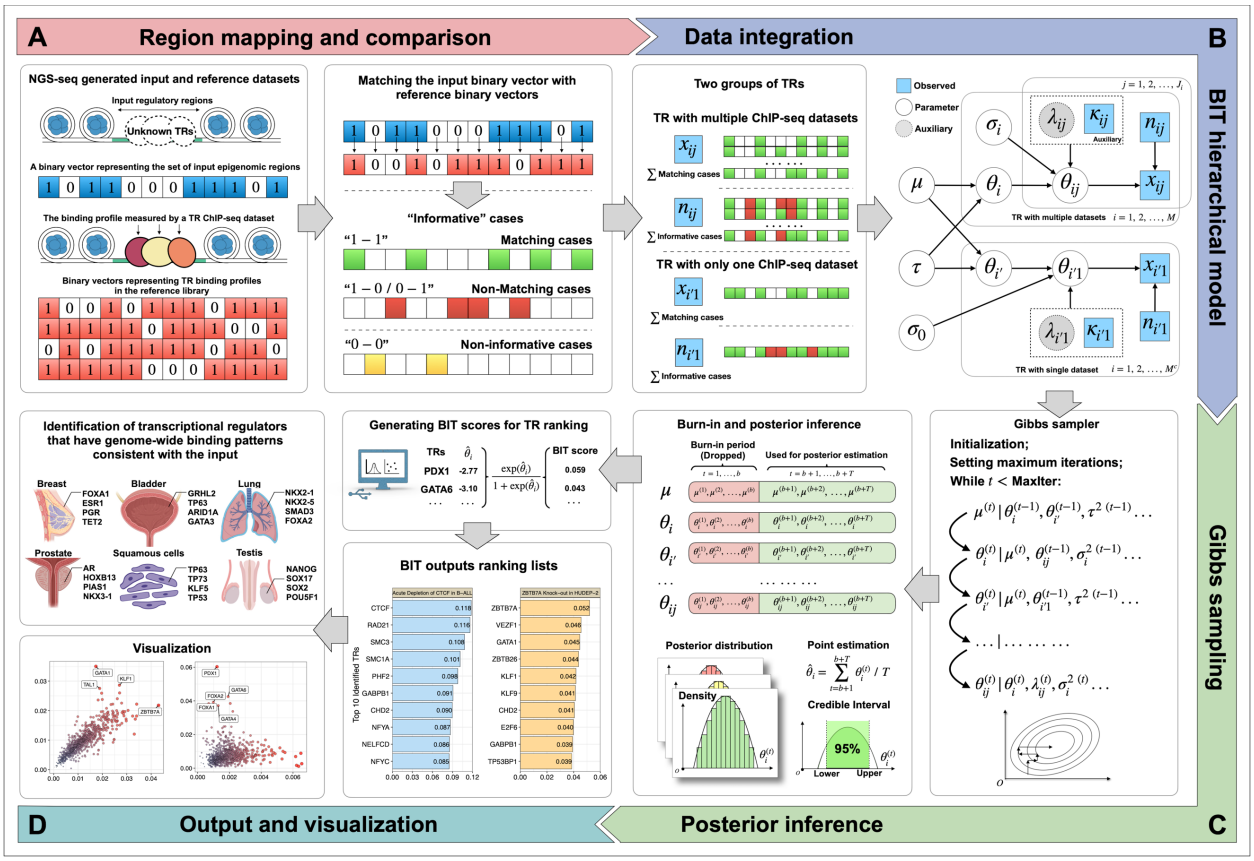


Figure 1. Overview of BIT's framework. A. *Transcription regulators (TRs) such as transcription factors, cofactors, and chromatin modulators mostly bind to accessible chromatin regions to regulate gene transcription. NGS techniques such as TR ChIP-seq and ATAC-seq can acquire information on TR binding sites and accessible regions, respectively. Input regions and collected TR ChIP-seq peaks are transformed into binary vectors, where the proportion of the "matching" cases among all "informative" cases serves as the response variable.* ***B.*** *BIT is a Bayesian hierarchical model, combined with Pólya-Gamma data augmentation, to integrate information from multiple TRs and multiple ChIP-seq datasets.* ***C.*** *A Gibbs sampler is used to draw posterior samples and estimate key model parameters.* ***D.*** *TR-level BIT scores are generated and ranked to output a TR ranking list.*

To verify BIT's performance, we first conducted simulations to examine the efficiency of its posterior estimation. As BIT is built on a unique hierarchical model with meaningful parameters, we used a "naïve" method as the baseline for comparison, which estimates the parameters based on straightforward summary statistics as described in **Methods**. Second, we examined BIT's performance with three different applications: (i) We identified TRs using differentially accessible regions (DARs) derived from TR perturbation experiments. Perturbed TRs should be responsible for the DARs, and therefore these DARs can serve as input with known ground truth. (ii) We tested BIT in TR identification using cancer-type specific accessible epigenomic regions. The identified

TRs were cross-validated by findings in previous publications and results derived from CRISPR/Cas9 loss-of-function analysis. (iii) Given the rapid development of single-cell techniques, cell-type-specific accessible regions can now be derived from scATAC-seq data. We further applied BIT to two sets of cell-type-specific accessible regions and validated the BIT predictions. We also benchmarked results from BIT against other state-of-the-art methods in the three applications.

BIT can estimate model parameters and TR rankings with high accuracy.

We explored the performance of BIT through extensive simulations. As will be detailed in the **Methods** section, our BIT model separates candidate TRs in the reference library into two distinct groups: (1) TRs with binding profiles generated from multiple TR ChIP-seq datasets and (2) TRs with only one binding profile generated from a single TR ChIP-seq dataset. We use i and i' to index TRs in the first and second groups, respectively. To cover potential scenarios, we first varied the global mean and variance parameters μ, τ^2 to control the distribution of θ_i. The θ_i denotes the mean (logit-scale) proportion of matching between the input binary vector representing the user-provided set of epigenomic regions and those representing binding profiles of the ith TR in the reference library, measuring its importance relevant to the biological process under study. We also varied σ_0^2, the common variance of $\theta_{i'1}$ within the second TR group, where $\theta_{i'1}$ is the logit-scale proportion of matching between the input binary vector and the reference vector representing the single binding profile of the i'th TR. In our simulation, we examined the effect of one such parameter at a time while keeping the other two fixed at their default values ($\mu = -4$, $\tau^2 = 0.75$, and $\sigma_0^2 = 1.5$), determined based on an exploratory data analysis of TR perturbation experiments in the **Results** Section. The range of the varied parameters was also decided based on the exploratory analysis to cover various real scenarios ($\mu \in \{-5, -4.5, \dots, -2.5\}$, $\tau^2 \in \{0.5, 0.7, \dots, 1.5\}$, and $\sigma_0^2 \in \{1, 1.2, \dots, 2\}$). Other parameters were then generated from Gaussian distributions with pre-determined global parameters according to our BIT model. Finally, the

number of input "matching" and "informative" bins were generated from a binomial distribution with logit^{-1} transformed θ_{ij} ($\theta_{i'1}$). We repeated the above steps and generated 100 replicate datasets for each combination of μ, τ^2, and σ_0^2.

We further investigated the effect of adjusting the number of candidate TRs (I) included in our analysis. Our human reference library covers 988 TRs, and our mouse reference library covers 607 TRs. Both libraries are expected to rapidly grow as more and more TR ChIP-seq datasets are being generated. Thus, we considered $I = 500, 1000, 1500$ in our simulation. In addition, the count of TR ChIP-seq datasets for the ith TR was simulated randomly from a distribution fitted using the actual counts within the first group in the human reference library.

We evaluated BIT in three aspects: (i) the mean squared errors (MSEs) for estimating global parameters μ, τ^2, σ_0^2 and TR-level importance parameters (θ_i's and $\theta_{i'}$'s); (ii) the ability to recover the original ranks of TRs according to their importance, measured by Spearman's rho. (iii) the robustness of BIT to the normality assumption of the BIT model. Comparison was made with a baseline method based on naïve estimation, as detailed in the **Methods** section.

Overall, BIT estimates all global parameters with high efficiency (**Fig. 2A**). First, it is noted that the MSEs of BIT for estimating μ, τ^2, and σ_0^2 are consistently lower than that of the baseline method and are close to zero in all cases. Second, a lower number of candidate TRs (i.e., smaller I) can result in relatively higher MSE, but the difference is much less obvious for BIT. Third, the MSE for the estimation of τ^2 tends to increase as the value of τ^2 goes up. This is reasonable as the increased variability would lead to greater estimation uncertainty. Nevertheless, BIT still significantly outperforms the baseline method in estimating τ^2.

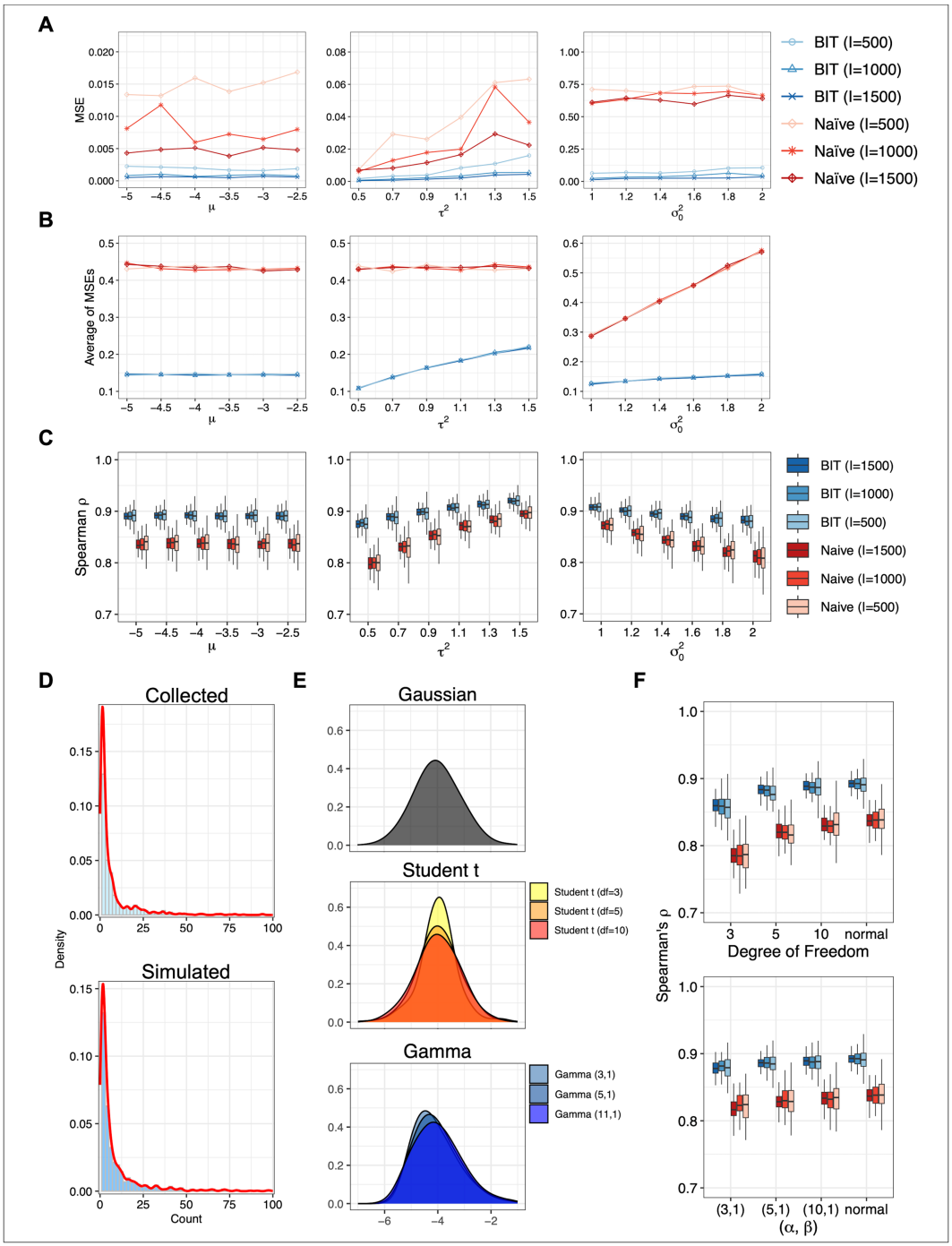


Figure 2. Simulation results under different parameter settings. A. *Mean squared errors for estimation of global parameters using 100 replicate datasets in each simulation setting.* ***B.*** *The average of MSEs for estimation of importance parameters (averaged over different TRs) over 100 replicates.* ***C.*** *Boxplots of Spearman's rho correlation over 100 replicates between ranks from BIT or the baseline method and the true ranks of simulated TRs.* ***D.*** *The distributions of simulated vs. real numbers of ChIP-seq datasets of individual TRs.* ***E.*** *Data generated in our sensitivity analysis from t and gamma distributions, to examine whether heavy-tailedness and skewness would significantly affect the performance of BIT.* ***F.*** *Boxplots of Spearman correlation over 100 replicates under t and gamma distributions, with baseline under normal distribution.*

BIT offers improved performance in estimating TR-level importance parameters as well, as evidenced by much lower average MSEs (averaged over the different TRs) in various settings (**Fig. 2B**). Accurate estimation of these parameters is crucial for determining precise TR rankings. We examined Spearman's rho correlation, which is the Pearson correlation between the rank values of the estimated and true importance parameters of the candidate TRs. We find that for both methods, larger τ^2 yields higher correlation while larger σ_0^2 does the opposite (**Fig. 2C**). As τ^2 controls the between-TR heterogeneity, larger τ^2 would make candidate TRs more easily distinguishable and so lead to better ranking performance (i.e., higher Spearman's ρ). On the contrary, σ_0^2 controls the noise level (i.e., how $\theta_{i'1}$ differs from $\theta_{i'}$) for TRs in the second group. Thus, increasing σ_0^2 would make the estimation of $\theta_{i'}$ more difficult, reducing the ranking performance.

The number of datasets associated with each TR in the first group was simulated using a lognormal distribution with right-skewness and heavy-tailedness, right truncated at 1,000, which closely mimics the actual count distribution observed in the human reference library (**Fig. 2D**). We also checked the robustness of BIT to the normality assumption in the BIT model. We used two types of non-Gaussian distributions, the Student's t-distributions (with degrees of freedom 3, 5, and 10) and gamma distributions (shape parameter α=3, 5, 10, scale parameter β=1), to generate θ_i's and $\theta_{i'}$'s. The overall shapes of these generated distributions are different from Gaussian distributions (**Fig. 2E**). The t-distributions with only a few degrees of freedom are known for heavy-tailedness, and gamma distributions for skewness. The simulated data are then shifted and scaled to recover the same mean and variability for comparison while keeping the heavy-tailedness or skewness. The ranking performance of both methods appears to decline as the deviation from the normality becomes larger. Nevertheless, BIT is much less sensitive and the advantages over the baseline method persist (**Fig. 2F**).

265 In summary, through simulations in settings mimicking various real scenarios, we showed BIT's
266 superior performance in accurately estimating parameters and recovering original TR ranks. It
267 consistently outperformed the baseline method, demonstrating high efficiency and robustness
268 even when normality is violated.

269 BIT can identify perturbed TRs from differentially accessible regions (DARs)

270 DARs derived from differential accessibility analysis between two biological conditions are critical
271 in understanding the activity of TRs. Here, we leverage DARs from TR perturbation experiments
272 to validate BIT's capacity to pinpoint key TRs. The targeted nature of these experiments provides
273 a ground truth, allowing us to assess BIT's ability to identify the perturbed TRs.

274 The first experiment is an acute depletion of CTCF in the MLL-rearranged human B cell
275 lymphoblastic leukemia (B-ALL) cell line SEM[40]. The second is a ZBTB7A-knockout experiment
276 in the HUDEP-2 cell line[41]. The third involves a FOXA2-knockout in the pancreatic progenitors 1
277 (PP1) differentiation stage of human pluripotent stem cells (hPSCs) [42]. BIT has successfully
278 identified each perturbed TR as one of the top-ranked TRs in all three experiments, and it has
279 also correctly identified several other TRs that contribute to the biological process involved (**Fig.
280 3A** and see **Tab. S1** for full lists).

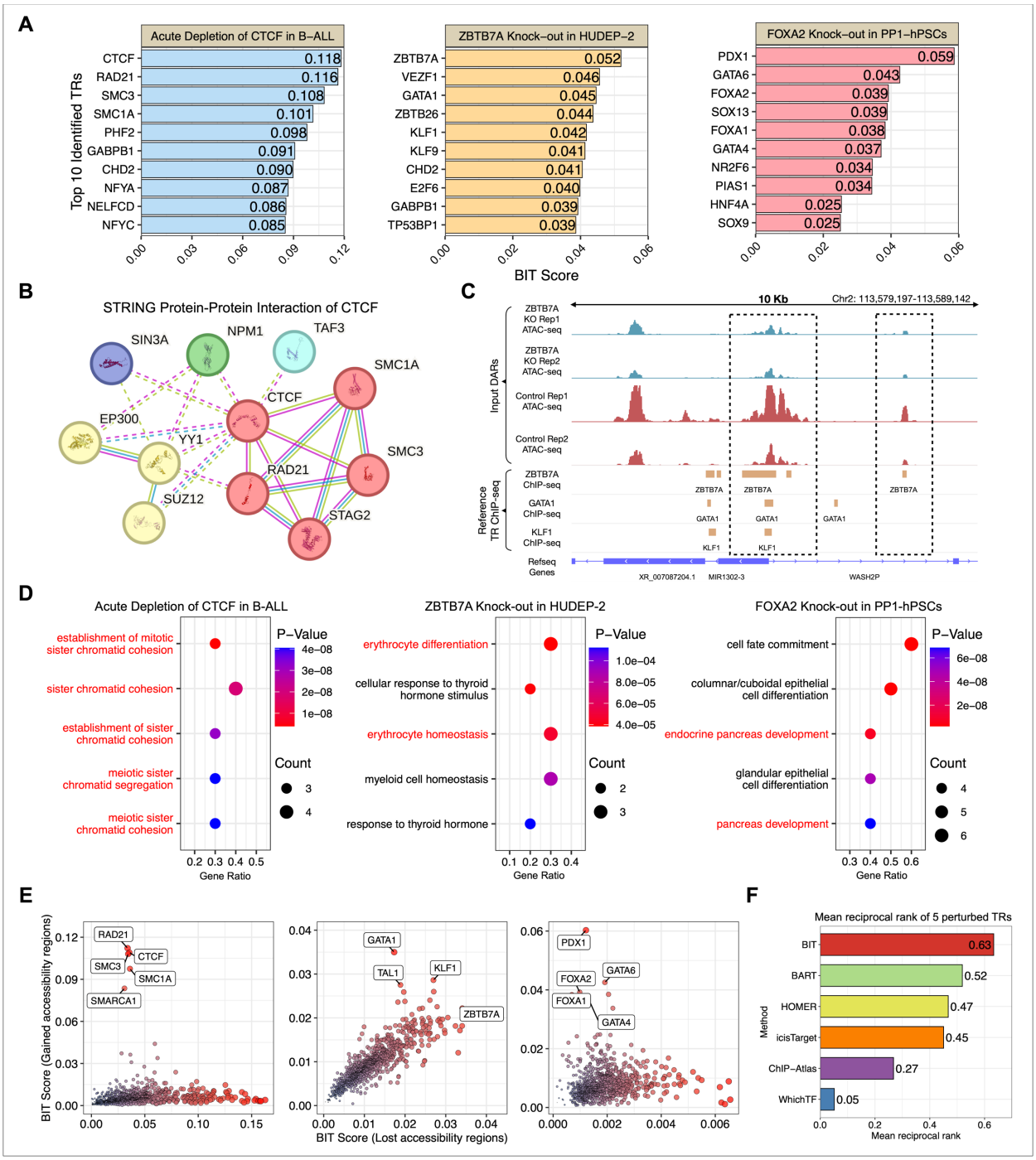


Figure 3. BIT can identify transcriptional regulators (TRs) from differentially accessible regions (DARs). A. *The top 10 transcriptional regulators (TRs) identified by BIT using DARs from each of the three TR perturbation experiments.* ***B.*** *Protein-protein interaction of CTCF derived from STRING.* ***C.*** *BIT-identified TRs have binding patterns consistent with the differentially accessible regions observed in TR perturbation experiments.* ***D.*** *The top 5 terms ranked by gene ontology enrichment analysis of top TRs in the output list of three cases.* ***E.*** *The scatterplot of TRs by using BIT score from gain accessible regions versus loss accessible regions.* ***F.*** *Mean reciprocal rank (MRR) of five perturbed TRs for BIT and five state-of-the-art methods (BART, ChIP-Atlas, i-cisTarget, WhichTF, HOMER).*

290 Using DARs from the first experiment as the input regions, the perturbed TR, CTCF, was ranked
291 first. Other top-ranked TRs include RAD21, SMC3, and SMC1A, which are all components of the
292 cohesin complex. SMC proteins and RAD21 are core proteins that form the ring-shaped cohesin
293 complex (cohesin ring)[43,44]. CTCF and the cohesin ring can be crucial in forming chromatin loops
294 and topologically associating domains (TADs). These two structures are critical in regulating gene
295 transcription[45,46]. Previous studies have confirmed that both CTCF and the cohesin ring are
296 essential for the maintenance of the two structures when using acute depletion assays to handle
297 proteins[47,48], which can explain why BIT also identified the other three TRs in this study. The result
298 was also validated by the protein-protein interactions of CTCF retrieved from the STRING
299 database[49], which shows strong associations among CTCF, SMC3, SMC1A, and RAD21 (**Fig.**
300 **3B**).

301 In the second experiment, in addition to the top-ranked ZBTB7A (the perturbed TR), GATA1 and
302 KLF1 were also ranked in the top 10, which have binding patterns consistent with the DARs
303 derived from the experiment (**Fig. 3C**). Early studies have revealed that erythroid differentiation
304 and mutation can be controlled by GATA1-dependent transcription[50,51], and that ZBTB7A is a
305 cofactor of GATA1 and plays an essential anti-apoptotic role during terminal erythroid
306 differentiation[52]. This explains why GATA1 was also ranked at the top after the knockout of
307 ZBTB7A. In addition, KLF1 has a multifunctional role in erythropoiesis and acts as one of the
308 master regulators[53].

309 In the third experiment, FOXA2 ranked third, but notably, other known pancreatic regulators
310 (PDX1, GATA6, FOXA1, and GATA4) were also identified. This is consistent with their established
311 roles in pancreatic development and disease[54-57]. For example, the pioneer factors FOXA1/2
312 could initiate chromatin accessibility[58], allowing subsequent binding and transcriptional regulation
313 by PDX1, GATA6, and GATA4.

314 The results of these three experiments were also validated using Gene Ontology (GO) enrichment
315 analysis, where we used the top TRs identified by BIT to examine enriched biological processes
316 (**Fig. 3D**). The count in this figure indicates the number of TRs associated with each process, and
317 the p-value was calculated using hypergeometric tests to evaluate whether the association of TRs
318 with a biological process is higher than expected by chance. We observed that the top five
319 enriched biological processes, ranked by p-values, include those involving sister chromatid
320 cohesion formation for the first experiment, erythrocyte differentiation and homeostasis for the
321 second experiment, and pancreas differentiation and development for the third experiment. This
322 further supports that BIT was able to identify TRs from DARs.

323 We further categorized the DARs into two groups based on whether these regions showed
324 increased (gained) or decreased (lost) accessibility after the perturbation of TRs. We separately
325 applied BIT to the two groups and noticed that the perturbed TRs were ranked differently
326 according to their BIT scores (**Fig. 3E**). For example, CTCF and FOXA2 were ranked higher in
327 the group of gained accessibility regions, while ZBTB7A was ranked higher in the group of lost
328 accessibility regions. This provides insights into how the perturbation of these TRs affects
329 chromatin accessibility and demonstrates that BIT can offer additional understanding of the
330 functional role of TRs from the DAR analysis.

331 We also applied BIT to two more datasets derived from TR knockout in mouse thymocytes to
332 study its performance on data from animal models and validated the results with GO enrichment
333 analysis (**Fig. S3**). BIT not only ranks the perturbed TRs in top positions (no. 3 and no. 2,
334 respectively) but also identifies other experimentally validated TRs strongly associated with
335 perturbed TRs or biological processes involved.

336 Finally, to compare the performance of BIT with other state-of-the-art methods (BART, ChIP-Atlas,
337 i-cisTarget, WhichTF, and HOMER), we calculated the mean reciprocal rank (MRR) of the five
338 perturbed TRs for each method (**Fig. 3F**). MRR is a commonly used metric to evaluate a method's

339 ability to rank relevant terms at the top positions[59]. Among the six methods compared, BIT
340 demonstrated the best overall performance.

341 BIT can identify cancer-type specific TRs

342 To further validate BIT's efficacy, we leveraged cancer-type-specific accessible regions obtained
343 from The Cancer Genome Atlas (TCGA) database[60]. TRs are well-established drivers of crucial
344 cellular processes in various cancers. However, a significant portion of accessible regions recur
345 across multiple cancer types. Including possible non-specific regions can hinder the identification
346 of critical TRs specific to one cancer type. Therefore, employing cancer-type-specific accessible
347 regions offered a more focused approach. We used regions specific to cancers originating from
348 nine tissue types to validate BIT's performance: breast, bladder, colon, lung, liver, mesothelium,
349 prostate, squamous cell, and testis. More specifically, the ATAC-seq samples used to generate
350 these regions encompassed a broad range of carcinomas, including breast invasive carcinoma
351 (BRCA), bladder urothelial carcinoma (BLCA), and others (see **Supplementary Note 2** for an
352 extended discussion). The identified TRs were cross-validated through existing publications and
353 CRISPR/Cas9 screening data from The Cancer Dependency Map Consortium (DepMap)[61], which
354 evaluates cancer cell viability after gene knockout.

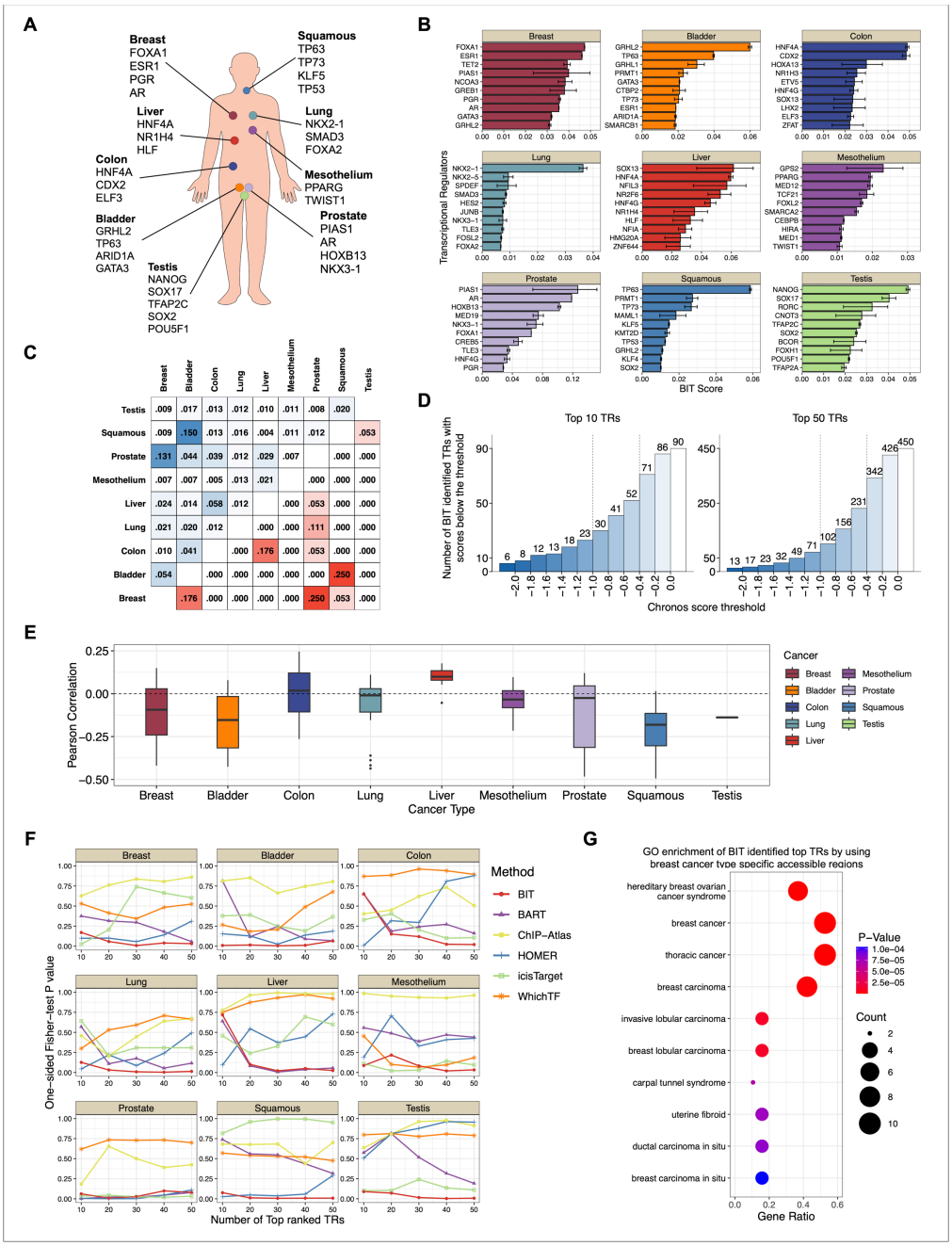


Figure 4. BIT can identify transcriptional regulators from cancer-type specific accessible regions. ***A.*** *32 TRs identified by BIT from TCGA cancer-type-specific accessible regions validated by existing literature.* ***B.*** *BIT scores of top 10 TRs with upper and lower bound of 95% credible intervals.* ***C.*** *Jaccard index compares cancer-specific accessible regions between any two cancer types (upper triangle) and compares the sets of top TRs identified by BIT between any two cancer types (lower triangle).* ***D.*** *Cumulative number of BIT identified top 10 and top 50 TRs with minimum Chronos scores of ≤−2 to >0.* ***E.*** *Boxplots showing the Pearson correlation coefficients between Chronos scores and BIT scores for the top 50 TRs identified by BIT across 209 cell lines from 9 cancer types.* ***F.*** *The results of the one-sided Fisher test evaluate whether the top TRs identified by BIT or other methods (BART, ChIP-Atlas, HOMER, i-cisTarget, WhichTF) significantly overlap with functionally essential TRs.* ***G.*** *The results of gene-ontology enrichment analysis of top 20 TRs identified by BIT from breast cancer.*

367 First of all, many of BIT's top-ranked TRs are well-known regulators of normal cell processes such
368 as differentiation and proliferation and have also been implicated in tumor development. A total of
369 32 TRs (listed by cancer type in **Fig. 4A**) were cross-validated with existing literature (**Tab. S2**),
370 with some already serving as biomarkers or therapeutic targets in cancer treatment[62,63]. For
371 instance, top-ranked TRs in breast cancer included FOXA1, ESR1, and PGR while prostate
372 cancer revealed PIAS1, AR, and HOXB13, and testicular cancer revealed NANOG, SOX17, and
373 POU5F1 (see **Tab. S3** for full lists). Beyond TR identification using estimated BIT scores, BIT also
374 gauges estimation uncertainty. For example, PIAS1 and SOX13 were identified in multiple cancer
375 types, but with higher uncertainty in their BIT score compared to other TRs in the same cancer
376 (**Fig. 4B**). This suggests their potential roles as pan-cancer regulators.

377 Second, we investigated the specificity of BIT-identified TRs across different cancer types. The
378 low Jaccard similarity coefficients (0.003 to 0.150) of the input binary vectors indicated minimal
379 overlap between cancer types (**Fig. 4C,** upper triangle), confirming that retrieved accessible
380 regions are generally specific to their corresponding cancer type. Furthermore, the top-ranked
381 TRs identified by BIT varied significantly across cancer types (**Fig. 4B**), consistent with the
382 generally low Jaccard similarity (**Fig. 4C,** lower triangle). This highlights BIT's ability to identify
383 key TRs specific to each cancer type.

384 Third, it is noted that several top-ranked TRs, such as NCOA3 and TET2 in breast cancer, lack
385 known motifs[2]. This absence means these TRs cannot be detected using motif-based methods,
386 despite solid experimental evidence supporting their significant roles in tumor development[64,65].
387 In contrast, BIT uses TR ChIP-seq data to prioritize these TRs, irrespective of motif information.
388 This highlights BIT's advantage over traditional motif-based methods, especially when known
389 motifs are ambiguous or absent.

390 Fourth, given the critical roles of many TRs in cellular viability[7,66,67], we hypothesized a significant
391 overlap between BIT's top-ranked TRs and functionally essential TRs (i.e., TRs strongly

392 correlated with cancer cell viability) identified in corresponding cancer types via CRISPR/Cas9
393 knockout screenings. Indeed, approximately half of the BIT-identified TRs were functionally
394 essential TRs (**Fig. 4D**): 52 out of 90 top-10 TRs and 231 out of 450 top-50 TRs across various
395 cancer types. Here, functionally essential TRs were defined as those with a minimum Chronos
396 (minC) score[61,68] (across all cell lines) less than -0.4; a more negative minC indicates greater
397 relevance. Furthermore, we computed Pearson correlation coefficients between Chronos and BIT
398 scores of the top 50 TRs identified by BIT in 209 cell lines across the nine cancer types. Boxplots
399 showed in seven out of nine cases, the two scores of most cell lines are negatively correlated,
400 which indicates that higher-ranked TRs generally exhibited lower Chronos scores, thus having a
401 greater impact on cancer cell viability after gene knockout (**Fig. 4E**).

402 Last but not least, we benchmarked BIT against state-of-the-art methods (e.g., BART, ChIP-Atlas,
403 i-cisTarget, WhichTF, and HOMER) by comparing the enrichment of their top-ranked TRs within
404 functionally essential TRs from CRISPR/Cas9 knockout data. Using a one-sided Fisher exact test,
405 we find that BIT consistently performed at or near the top across all the cancer types with lower
406 p-values (**Fig. 4F**), showing less variation as the number of top TRs increased. While cellular
407 viability is not the sole determinant of TR functionality in cancer, BIT's robust performance
408 highlights its efficacy in identifying crucial TRs from cancer-type-specific accessible regions, even
409 amidst the multifaceted nature of TR functionality.

410 In addition to the cross-validation with CRISPR/Cas9 knockout analysis, we conducted several
411 more analyses using various resources: (1) Gene ontology enrichment analysis using BIT's top
412 TRs identified from breast cancer-specific accessible regions revealed enriched GO terms
413 strongly associated with breast cancer (**Fig. 4G**). (2) We retrieved super-enhancer targeted genes
414 from Cistrome[69] for six available cancer types, which rank genes based on adjacent super-
415 enhancer activities. We noticed many BIT-identified TRs are also highly ranked by Cistrome (**Fig.**
416 **S4A**). (3) Kaplan-Meier plotter[70], a public analysis tool that has collected data from over 30,000

417 cancer samples, is used to investigate the correlation between gene expression levels and patient
418 survival time. It is observed that multiple BIT-identified TRs show a strong correlation with patient
419 survival time (**Fig. S4B**).

420 BIT can identify cell type-specific TRs

421 The rise of single-cell omics technologies has created a need for computational tools capable of
422 analyzing transcriptional regulation at the individual cell level. scATAC-seq, a powerful technique
423 for profiling chromatin accessibility in single cells, often involves clustering cells and annotating
424 them to specific cell types to overcome the challenge of sparse data. An important task in
425 downstream analyses of scATAC-seq data is to identify key TRs associated with accessible
426 regulatory regions in specific cell types.

427 In this study, we investigated the application of BIT for identifying crucial TRs using cell type-
428 specific accessible regions derived from scATAC-seq data. Two datasets, including 10K
429 peripheral blood mononuclear cells (PBMCs) and primary liver cancer samples, were used, where
430 the cell types were previously annotated based on the dimension reduction results and marker
431 genes. The UMAP projections of the datasets show a clear separation of different cell types in
432 each dataset (**Fig. 5A**). SnapATAC2[71] was applied to generate cell type-specific accessible
433 regions. We applied BIT to accessible regions derived from B cells in the PBMCs dataset and
434 malignant cells in the liver cancer dataset. The results were thoroughly validated by different
435 approaches, as described below.

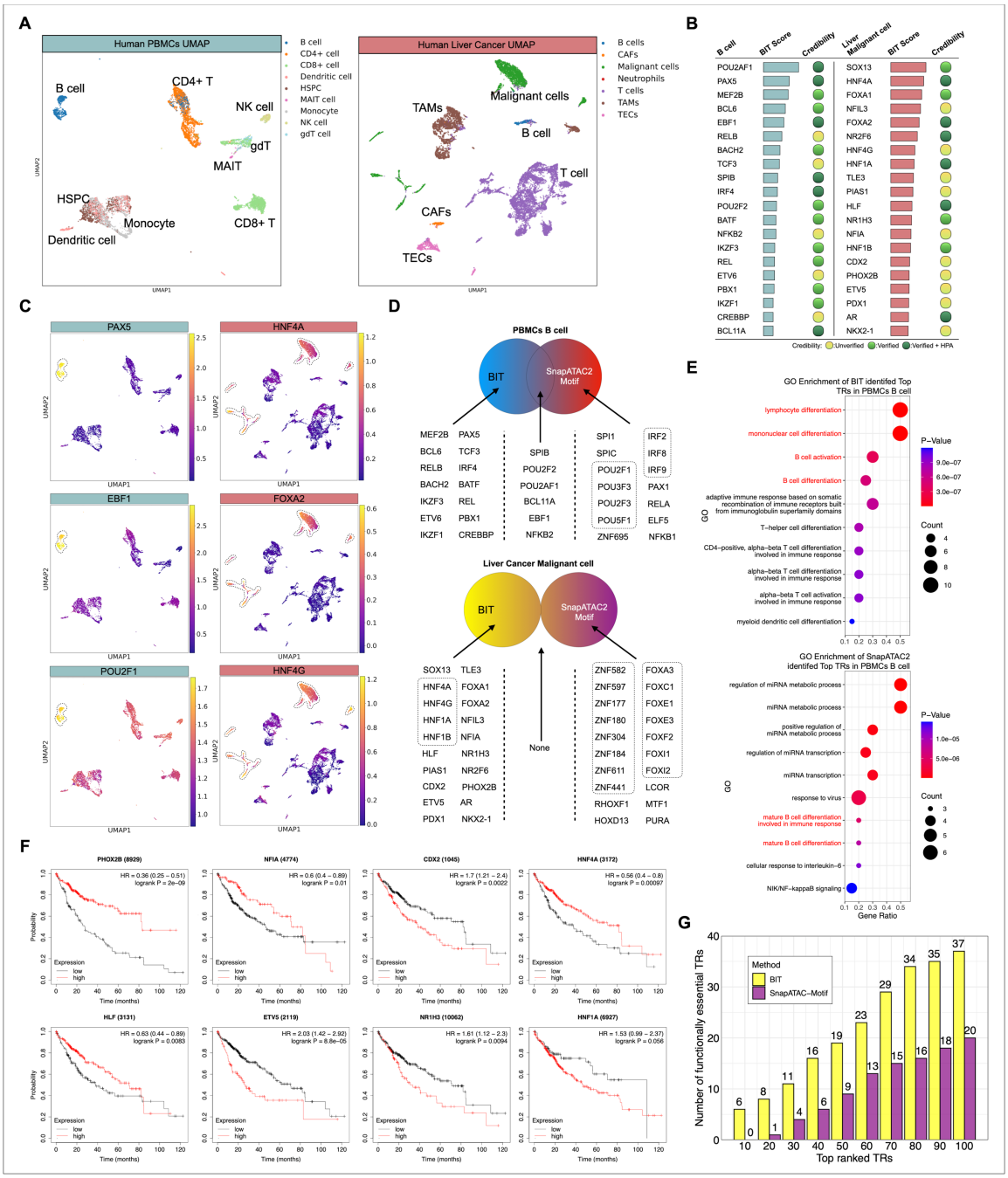


Figure 5. BIT can identify transcriptional regulators from cell-type-specific accessible regions. A. *Sub-clustering UMAP projection of human Peripheral blood mononuclear cells and primary liver cancer cells.* ***B.*** *Top 20 TRs identified by BIT from using B cell-specific and liver malignant cell-specific accessible regions, verified by existing literature and Human Protein Atlas.* ***C.*** *UMAP visualization of gene activity of PAX5, EBF1, POU2F1 in PBMCs and FOXA2, HNF4A, and HNF4G in liver cancer samples.* ***D.*** *The comparison of top 20 TRs identified by BIT and SnapATAC2.* ***E.*** *Gene ontology enrichment results of top 20 TRs identified by BIT and by SnapATAC2 motif enrichment analysis in PBMCs B cells.* ***F.*** *Kaplan-Meier Plots of PHOX2B, NFIA, CDX2, HNF4A, HLF, ETV5, NR1H3, HNF1A in liver cancer samples.* ***G.*** *Number of BIT and SnapATAC2 identified top 10 to 100 TRs from liver malignant cells with minimum Chronos score ≤−0.4.*

447 First, we evaluated the credibility of the top 20 TRs identified by BIT in the two investigated cell
448 types (**Fig. 5B**) based on existing publications and by examining whether each identified TR has
449 a significantly elevated protein expression in the Human Protein Atlas for the respective cell type
450 (**Tab. S4**). Notably, BIT successfully captured several TRs critical for B cell development. These
451 include PAX5 and EBF1, which were previously reported for their role in maintaining B cell identity.
452 In addition, POU2AF1 and POU2F2 can act as pivotal regulators in B cell proliferation,
453 differentiation, the germinal center transcriptional program, and immune response (see **Tab. S5**
454 for full lists).

455 For liver cancer malignant cells, we observed consistency between the top TRs identified from
456 scATAC-seq data and those previously found using bulk ATAC-seq data for liver cancer (**Fig. 4B**).
457 Specifically, SOX13, HNF4A, and HLF were again highly ranked by BIT. While the previous
458 analysis used accessible regions from aggregated bulk data, the current analysis leveraged
459 single-cell data specific to malignant cells. This consistency supports the robustness of BIT in
460 identifying critical TRs across different data types.

461 Second, gene activity plots were generated to demonstrate the high specificity of the identified
462 TRs in both B cells and malignant cells (**Fig. 5C**; left panel for B cells and right panel for malignant
463 cells). Gene activity was assessed by counting the number of Tn5 insertions within the gene's
464 regulatory regions, a standard approach in scATAC-seq analysis[71,72]. Multiple BIT-identified Top
465 TRs exhibited significantly higher activity in the studied cell types than other cell types, including
466 PAX5, EBF1, and POU2F1 in PBMCs B cells and HNF4A, HNF4G, and FOXA2 in Liver cancer
467 malignant cells.

468 Third, we also compared the top 20 TRs identified by BIT with those identified through motif
469 enrichment analysis implemented in SnapATAC2, where the two sets have a small or no overlap
470 at all. As mentioned in the introduction, the similarity of binding motifs can hinder the differentiation
471 between TRs from the same protein family. This is evident in the top-ranked TRs identified using

472 SnapATAC2, where results of B cells include the POU, IRF, and PAX families, and results of liver
473 cancer malignant cells exhibit enrichment exclusively for the ZNF and FOX families (**Fig. 5D**). It
474 is crucial to note that not all members within these families hold equal importance. Their inclusion
475 in the top rankings solely reflects the inability of motif enrichment analysis to identify which specific
476 TR within the family directly interacts with the motif site in the accessible chromatin regions. In
477 contrast, BIT's ability to pinpoint specific TRs, even within families with similar binding
478 motifs, offers a distinct advantage in understanding the precise regulatory mechanisms at play in
479 different cell types.

480 Next, we performed GO enrichment analysis for the top TRs in B cells identified by BIT and by
481 motif enrichment, separately. The results revealed a strong enrichment for biological processes
482 relevant to B cells or mononuclear cells in the BIT-identified TRs. Conversely, the motif
483 enrichment-identified TRs displayed enrichment for the miRNA metabolic process (**Fig. 5E**).

484 We further explored the potential of BIT-identified TRs from liver malignant cells as prognostic
485 biomarkers. Kaplan-Meier survival analysis demonstrated a statistically significant association
486 between the expression levels of multiple BIT-identified TRs and patient survival time (**Fig. 5F**),
487 suggesting their promise as prognostic biomarkers in liver cancer. In addition, CRISPR/Cas9
488 knockout experiments in liver cancer cell lines revealed that BIT identified a greater number of
489 functionally essential TRs compared to motif enrichment analysis in SnapATAC2 (**Fig. 5G**).

490 In addition to the above results for B cells and liver malignant cells, we generated results for
491 monocytes and dendritic cells in the PBMCs (**Fig. S5**) and for tumor-associated macrophages
492 and T cells in the primary liver cancer samples (**Fig. S6**). Collectively, our findings demonstrate
493 the feasibility of using BIT for the downstream analysis of cell type-specific accessible regions in
494 scATAC-seq data. Beyond feasibility, BIT offers valuable insights into potential regulatory
495 mechanisms. This makes it a compelling alternative to the current dominant motif-based
496 enrichment methods used in scATAC-seq analysis.

Discussion

We proposed a Bayesian hierarchical approach, called BIT, to identify TRs using user-provided epigenomic regions (peaks), often derived from genome-wide epigenomic profiling data. Using various simulation and application studies, we demonstrated the effectiveness of BIT and its advantages over existing state-of-the-art methods.

Biological Complexities of TR Regulation

Though BIT demonstrated superior performance consistently in various applications, BIT can suffer from the biological complexities of TR regulation, which can reduce the method's performance. TR binding is inherently dynamic, changing across cell types and time in response to diverse signals and environmental factors[73]. Chromatin structure and epigenetic modification can further complicate this by affecting the binding affinity and specificity[23,74]. In addition, cooperative binding can also change binding patterns as TRs frequently cooperate to regulate gene expression[75]. A large number of interactions between TRs, DNA, and other cellular components challenge accurate prediction and functional interpretation, while the currently gathered context-specific binding profiles from ChIP-seq experiments only represent a very tiny part of all possible experiments.

Data quality and sequencing biases

Similar to other NGS-based computational methods, data quality is a critical issue that can affect BIT's performance. Multiple types of sequencing biases, such as GC content bias or PCR amplification bias, can potentially reduce the reliability of the input epigenomic region set and binding profiles generated by NGS experiments[76]. For instance, the Illumina platform is known for a strong GC bias, wherein epigenomic regions with high or low GC content may be over- or under-represented in the sequencing data[77]. Preferential amplification of certain sequences can also

520 lead to the overrepresentation of specific genomic regions[78]. Therefore, the performance of BIT
521 can depend on the quality of the NGS data used.

522 Alternative input from epigenomic profiling data

523 In the manuscript, we primarily used ATAC-seq datasets that profile chromatin accessibility as
524 input data. However, other techniques such as DNase-seq can also profile the chromatin
525 accessibility and serve as input. In addition, alternative epigenomic profiling techniques are
526 available that also correlate with TR binding activity. For example, H3K27ac ChIP-seq data can
527 mark active promoters and enhancers[79], which are potential TR binding sites. Therefore, these
528 datasets can also be employed to identify TRs regulating the biological processes of interest.

529 Computational efficiency

530 BIT implements a Gibbs sampling scheme to draw posterior samples for thousands of parameters
531 defined within the model. The number of parameters is determined by the available TRs and
532 associated datasets in the reference library. With the growing availability of ChIP-seq
533 experiments, the number of parameters to infer is expected to increase significantly. In addition,
534 it is common for researchers to generate dozens or even hundreds of epigenomic profiling
535 datasets in a single project. Analyzing these datasets using the Gibbs sampler can become
536 increasingly time-consuming. Therefore, alternative approaches such as variational inference[80,81]
537 are currently being considered for improving computational efficiency.

538 In summary, BIT offers valuable insights into transcriptional regulation by providing accurate TR
539 identification through a rigorous Bayesian hierarchical model. This approach leverages the wealth
540 of accumulated TR ChIP-seq data, representing the most accurate *in vivo* context-specific binding
541 patterns. With the increasing popularity of epigenomic profiling techniques, particularly bulk and

542 single-cell ATAC-seq, BIT addresses a critical need for effective tools to analyze this rapidly
543 increasing omics data.

Methods

Preprocessing epigenomic NGS data

Bulk data: In the TR perturbation-derived DARs analysis, raw ATAC-seq sequence reads were first cleaned using Trim-galore. Read quality was checked by FastQC. Paired-end reads were mapped to the human genome (hg38) or mouse genome (mm10) using bowtie2 applying parameters “bowtie2 --very-sensitive -X 2000”. SAMtools was used for the following steps (i) convert SAM files to BAM files, (ii) remove unmapped reads and duplicates, (iii) generate index files for BAM files. Peaks were called using MACS2. BAM files, index files, and peak files were used for DAR analysis using DiffBind. Reference TR ChIP-seq datasets were retrieved from GTRD, datasets with too low number peaks (<500) and individual peaks with low q-value (<10) were removed. Cancer-type-specific accessible regions were directly retrieved from TCGA.

Single-cell data: 10K PBMCs and primary liver cancer scATAC-seq fragment files were retrieved from the 10XGenomics and GEO archive. The cell-type-specific accessible regions were generated from each dataset using SnapATAC2. When applicable, the cell type annotations were also retrieved and applied. We also used the SnapATAC2 implemented motif enrichment analysis method to contrast the results.

Region mapping: BIT only requires user-provided epigenomic profiling data as input, while leveraging a comprehensive reference library of TR binding profiles from a large collection of TR ChIP-seq datasets. All binding sites (peaks) from a single dataset are mapped to disjoint, consecutive, pre-defined bins of 1000 bps over the entire genome. Each bin is assigned either “0” or “1” according to whether the summit of the peak falls into the bin (in case that the summit information of a peak is missing, the middle point is used instead). Such mapping is also done to the user-provided epigenomic regions. Thus, all data (input and references) are transformed into binary vectors of the same length. We then compare the binary vector from the input with that

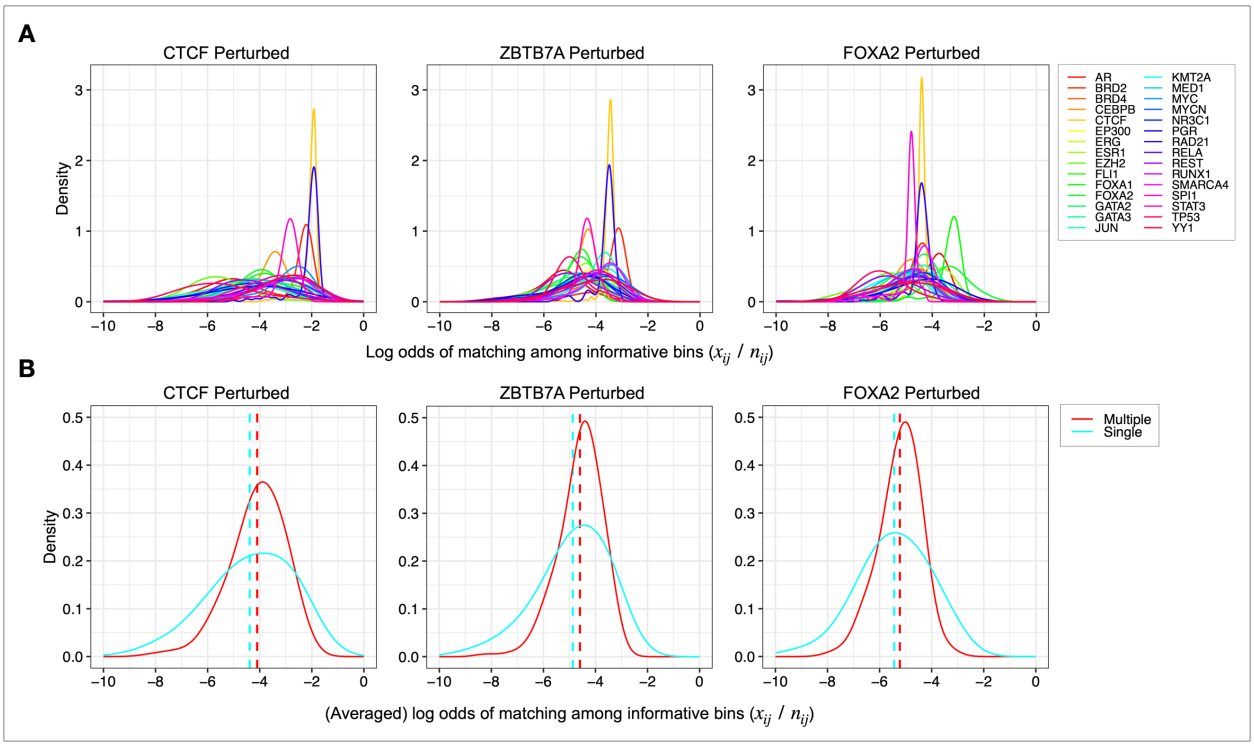
568 from each of the TR ChIP-seq datasets, where we define three conditions for any given bin: (i)
569 "matching" if (1,1); (2) "mismatching" if (1,0) or (0,1); and (3) "non-informative" if (0,0). In all input
570 vs. reference comparisons, the first two cases ("informative" bins) are much fewer than the third
571 case ("non-informative" bins) and we drop the third case to mitigate the sparsity while not losing
572 much useful information. Note that the proportion of matching bins among all informative bins is
573 the Jaccard Index, a measure of similarity between two binary vectors.

574 The data model of BIT

575 BIT is the first model-based method that formally distinguishes TRs with multiple datasets (or
576 binding profiles) from those with only a single dataset (or binding profile) in the reference database,
577 to account for both within-TR and between-TR heterogeneity and allow for information borrowing
578 across different TRs and datasets via a hierarchical setup. Our modeling choices are made
579 according to the principle of parsimony in statistics. Thus, BIT strikes a balance between clarity
580 and effectiveness in addressing complex biological questions. It aims to reduce the risk of
581 overfitting while preserving the key insights needed for accurate predictions. An exploratory
582 analysis using the three human TR perturbation datasets reveals the following: (1) When grouped
583 by the same TR, the distribution of the log odds (for matching among informative bins) often
584 exhibits a roughly bell shape, with different groups (TRs) showing different centers and spreads.
585 (2) When calculating the (average) log odds for each TR and categorizing these quantities based
586 on whether the TRs have single or multiple ChIP-seq datasets, the distributions of the two
587 categories exhibit similar centers but different variabilities (**Fig. 6**).

588 We index TRs with multiple reference datasets by i, $i = 1,2,\ldots,\mathcal{M}$, each with $j = 1,2,\ldots,J_i$
589 datasets. TRs with a single reference dataset are indexed by $i' = 1,2,\ldots,\mathcal{M}^c$, all with $j \equiv 1$
590 dataset. Let x_{ij} $(x_{i'1})$ be the number of "matching" bins out of n_{ij} $(n_{i'1})$ "informative" bins by
591 comparing the vectors between the user input and the jth reference dataset of the ith TR (or the

592 only dataset of the i' th TR). The probability of a bin being "matching" given it is "informative" is
593 denoted by p_{ij} $(p_{i'1})$ such that $x_{ij} \sim Binom(n_{ij}, p_{ij})$ and $x_{i'1} \sim Binom(n_{i'1}, p_{i'1})$. Let $\theta_{ij} \equiv$
594 $logit(p_{ij})$ $(\theta_{i'1} \equiv logit(p_{i'1}))$ be the log odds of matching among informative bins, which can be
595 interpreted as the Jaccard similarity index (or simply, the importance score) at the dataset level.
596 We further model θ_{ij} $(\theta_{i'1})$ by $\theta_{ij} \sim N(\theta_i, \sigma_i^2)$, $\theta_{i'1} \sim N(\theta_{i'}, \sigma_0^2)$ and all θ_i, $\theta_{i'} \sim N(\mu, \tau^2)$, where μ
597 and τ^2 represent the global mean and variability of TR-level importance scores. In this way, for
598 each TR with multiple datasets, we model its importance scores by a normal distribution with
599 distinctive mean θ_i and variance σ_i^2, reflecting its overall importance and within-TR heterogeneity,
600 respectively (**Fig. 6A**); all TRs with only one dataset share a common variance σ_0^2 for information
601 pooling; and all TRs, no matter which group they are in, share a common mean μ (**Fig. 6B**).



602

603 ***Figure 6. Exploratory analysis using ATAC-seq data from three TR perturbation experiments. A.***
604 *The distributions of log odds of "matching" bins among "informative" bins for the 30 TRs with the highest*
605 *numbers of TR ChIP-seq datasets available in the reference database.* ***B.*** *The distributions of the*
606 *(average) log odds for all TRs, where TRs are categorized into two groups based on whether each TR*
607 *has single or multiple TR ChIP-seq datasets.*

It turns out that the entire hierarchical model (**Fig. 1B**) is simple yet effective, as shown in our simulation and applications:

$$x_{ij} \sim Binom(n_{ij}, p_{ij}),\ x_{i'1} \sim Binom(n_{i'1}, p_{i'1})$$

$$\theta_{ij} = logit(p_{ij}),\ \theta_{i'1} = logit(p_{i'1})$$

$$\theta_{ij} \sim N(\theta_i, \sigma_i^2),\ \theta_{i'1} \sim N(\theta_{i'}, \sigma_0^2)$$

$$\theta_i,\ \theta_{i'} \sim N(\mu, \tau^2)$$

Unlike many existing algorithms[31,32,39], BIT inherently integrates information from reference datasets of different TRs to assign one final rank of each TR based on the TR-level importance score θ_i ($\theta_{i'}$) rather than conducting thousands of separate *ad-hoc* statistical tests.

The (hyper)parameters of our BIT model are collectively denoted by $\Theta = \{\mu, \tau^2, \sigma_0^2, \{\theta_i\}_{i=1}^{M}, \{\theta_{i'}\}_{i'=1}^{M_c}, \{\sigma_i^2\}_{i=1}^{M}, \{\theta_{ij}\}_{i=1,j=1}^{M,J_i}, \{\theta_{i'1}\}_{i'=1}^{M_c}\}$ and observed data are $X = \left\{ \{x_{ij}\}_{i=1,j=1}^{M,J_i}, \{x_{i'1}\}_{i'=1}^{M_c} \right\}$. Then the joint distribution of (X, Θ) can be written as

$$\begin{aligned} p(X,\Theta) &= \left(\prod_{i=1}^{M} \prod_{j=1}^{J_i} p\left(x_{ij}|\theta_{ij}, n_{ij}\right) p\left(\theta_{ij}|\theta_i, \sigma_i^2\right) \right) \\ &\times \left(\prod_{i'=1}^{M^c} p\left(x_{i'1}|\theta_{i'1}, n_{i'1}\right) p(\theta_{i'1}|\theta_{i'}, \sigma_0^2) \right) \\ &\times \left(\prod_{i=1}^{M} p\left(\theta_i|\mu, \tau^2\right) p\left(\sigma_i^2\right) \prod_{i'=1}^{M^c} p\left(\theta_{i'}|\mu, \tau^2\right) \right) p(\mu) p(\tau^2) p(\sigma_0^2) \end{aligned}$$

where

$$p\left(x_{ij}\middle|\theta_{ij}, n_{ij}\right) = \binom{x_{ij}}{n_{ij}} \frac{e^{x_{ij}\theta_{ij}}}{\left(1 + e^{\theta_{ij}}\right)^{n_{ij}}},$$

$$p(x_{i'1}|\theta_{i'1}, n_{i'1}) = \binom{x_{i'1}}{n_{i'1}} \frac{e^{x_{i'1}\theta_{i'1}}}{(1 + e^{\theta_{i'1}})^{n_{i'1}}}.$$

Prior specification

The global location parameter μ and all variance parameters including τ^2, σ_0^2, and $\{\sigma_i^2\}_{i=1}^{M}$ are assumed *a priori* independent. For μ, we consider the non-informative flat prior $\mu \sim uniform(L_\mu, U_\mu)$, which should provide a sufficiently wide coverage for all plausible values of μ suggested by data. Detailed discussion about how to set up the upper and lower bounds can be found in Jia et al. [82] For each variance parameter involved, we specify an inverse gamma prior distribution $IG(a, b)$, where a and b are tiny positive values to make prior very vague and diffuse (e.g., $a = b = 0.01$).

Pólya-gamma data augmentation and Gibbs sampler

BIT uses a Gibbs sampler, an efficient Monte Carlo Markov chain (MCMC) algorithm to generate posterior samples from the joint probability distribution $p(\Theta|X)$ that is proportional to $p(X, \Theta)$. We proceed to derive the full conditionals to set up the Gibbs sampler efficiently. The full conditional posterior distribution of each parameter can be written as $p(\theta|X, \Theta \setminus \theta)$, where $\Theta \setminus \theta$ denote the set of parameters after removing θ from Θ. For μ, τ^2, σ_0^2, and σ_i^2, the conditional posterior distributions are known distributions which can be directly sampled, given below:

$$p(\mu|X, \Theta \setminus \mu) \propto N\left(\frac{\sum_{i=1}^{M} \theta_i + \sum_{i'=1}^{M^c} \theta_{i'}}{M + M^c}, \frac{\tau^2}{M + M^c}\right) \times \mathbf{1}\{L_\mu \leq \mu \leq U_\mu\}$$

$$p(\tau^2|X, \Theta \setminus \tau^2) \propto IG\left(a + \frac{M + M^c}{2}, b + \frac{\sum_{i=1}^{M} (\theta_i - \mu)^2 + \sum_{i'=1}^{M^c} (\theta_{i'} - \mu)^2}{2}\right)$$

$$p(\sigma_0^2|X, \Theta \setminus \sigma_0^2) \propto IG\left(a + \frac{M^c}{2}, b + \frac{(\theta_{i'1} - \theta_{i'})^2}{2}\right)$$

$$p\left(\sigma_i^2 \middle| X, \Theta \setminus \sigma_i^2\right) \propto IG\left(a + \frac{J_i}{2}, b + \frac{\sum_{j=1}^{J_i} \left(\theta_{ij} - \theta_i\right)^2}{2}\right)$$

$$p(\theta_i|X, \Theta \setminus \theta_i) \propto N\left(\frac{\mu\sigma_i^2 + \sum_{j=1}^{J_i}\theta_{ij}\tau^2}{\sigma_i^2 + J_i\tau^2}, \frac{1}{\frac{1}{\tau^2} + \frac{J_i}{\sigma_i^2}}\right)$$

$$p(\theta_{i'}|X, \Theta \setminus \theta_{i'}) \propto N\left(\frac{\mu\sigma_0^2 + \theta_{i'1}\tau^2}{\sigma_0^2 + \tau^2}, \frac{1}{\frac{1}{\sigma_0^2} + \frac{1}{\tau^2}}\right)$$

However, it is difficult to sample from the conditional posteriors of $p(\theta_{ij}|X, \Theta \setminus \theta_{ij})$ and $p(\theta_{i'1}|X, \Theta \setminus \theta_{i'1})$ as these are intractable. We employ a data augmentation strategy based on Pólya-Gamma (PG) latent variables[83]. Specifically, we define auxiliary variables λ_{ij} and $\lambda_{i'1}$, which follow the Pólya-Gamma distribution.

$$p(\lambda_{ij}|\theta_{ij}, X) \sim PG(n_{ij}, \theta_{ij})$$

$$p(\lambda_{i'1}|\theta_{i'1}, X) \sim PG(n_{i'1}, \theta_{i'1})$$

Then the conditionals given the defined auxiliary parameters can be easily written as the form of known distribution.

$$p(\theta_{ij}|X, \Theta \setminus \theta_{ij}, \lambda_{ij}) \quad \propto N(V_{ij}, m_{ij})$$

where

$$V_{ij} = \left(\lambda_{ij} + \frac{1}{\sigma_i^2}\right)^{-1}$$

$$m_{ij} = V_{ij}\left(\kappa_{ij} + \frac{\theta_i}{\sigma_i^2}\right)$$

$$\kappa_{ij} = x_{ij} - \frac{n_{ij}}{2}$$

Similarly, we can derive the conditional posterior of $\theta_{i'1}$. Given that all conditionals are now known distributions, we can design an efficient Gibbs sampler in which all quantities are drawn sequentially and generated readily without using any built-in sampling algorithms (such as

Metropolis-Hastings or Acceptance/Rejection sampling) that can greatly slow down the computation.

Bayesian inference and TR rankings

Inference about TR-level importance parameters can be made by marginalizing over posterior samples generated from the Gibbs sampler. Let $\theta_i^{(t)}$ and $\theta_{i'}^{(t)}$ be the posterior draw of θ_i and $\theta_{i'}$ in the tth iteration of MCMC after the burn-in period, where $t = 1, \ldots, T$. Then θ_i and $\theta_{i'}$ are estimated by $\tilde{\theta}_i = \sum_{t=1}^{T} \theta_i^{(t)} / T$ and $\tilde{\theta}_{i'} = \sum_{t=1}^{T} \theta_{i'}^{(t)} / T$. Then, these logit-transformed probabilities $\{\tilde{\theta}_i\}_{i=1}^{M}$ and $\{\tilde{\theta}_{i'}\}_{i'=1}^{M_c}$ can be used to produce a final ranked TR list or to be further transformed back to the 0-1 scale. To gauge prediction uncertainty, credible intervals of any parameter of interest can be easily obtained from percentiles of draws or as highest posterior density intervals.

The "naïve" method as a baseline for comparison in simulation

To compare the performance of BIT in estimating model parameters, we used a naïve method as the baseline, where μ is estimate by

$$\hat{\mu} = \frac{\sum_{i=1}^{M} \sum_{j=1}^{J_i} logit\left(\frac{x_{ij}}{n_{ij}}\right) + \sum_{i'=1}^{M^c} logit\left(\frac{x_{i'1}}{n_{i'1}}\right)}{\sum_{i=1}^{M} J_i + M^c};$$

τ^2 is estimated by

$$\hat{\tau}^2 = \frac{\sum_{i=1}^{M} \sum_{j=1}^{J_i} \left(logit\left(\frac{x_{ij}}{n_{ij}}\right) - \hat{\mu}\right)^2 + \sum_{i'=1}^{M^c} \left(logit\left(\frac{x_{i'1}}{n_{i'1}}\right) - \hat{\mu}\right)^2}{\sum_{i=1}^{M} J_i + M^c - 1};$$

σ_0^2 is estimated by

$$\hat{\sigma}_0^2 = \frac{1}{M^c - 1} \sum_{i'=1}^{M^c} \left(logit\left(\frac{x_{i'1}}{n_{i'1}}\right) - \hat{\mu}\right)^2;$$

and θ_i and $\theta_{i'}$ are estimated by

$$\hat{\theta}_i = \frac{\sum_{j=1}^{J_i} logit\left(\frac{x_{ij}}{n_{ij}}\right)}{J_i},\ \hat{\theta}_{i'} = logit\left(\frac{x_{i'1}}{n_{i'1}}\right).$$

In the unlikely case that $\frac{x_{ij}}{n_{ij}} = 0$ or $\frac{x_{i'1}}{n_{i'1}} = 0$, we add a tiny positive value 0.0001 to the zeros.

BIT validation analyses

Computation of mean reciprocal rank: Let the rank of the perturbed TR in the ith output list generated by a method be denoted as R_i, the mean reciprocal rank (MRR) of the five perturbed TRs (CTCF, ZBTB7A, FOXA2, RUNX1, BCL11B) was computed as:

$$MRR = \frac{1}{5}\sum_{i=1}^{5}\frac{1}{R_i}$$

We computed the metric separately for the six methods (BIT, BART, ChIP-Atlas, i-cisTarget, WhichTF, and HOMER).

Cross-validation with CRISPR/Cas9 screening data: Data of TR knockout effects on viability of cell lines were retrieved from DepMap. We filtered out TRs that are not included in the BIT database or the DepMap data. Cell lines under the same lineage are grouped to find TRs with a Chronos score lower than -0.4, which are classified as functionally essential TRs. 2-by-2 contingency tables were created based on whether top-ranked TRs identified by each method are also functionally essential TRs. Finally, one-sided Fisher exact tests were applied to each table to compare the capability of methods to identify TRs that are also functionally essential TRs in each cancer type.

To compare with other state-of-the-art methods, the original implementation of BART, HOMER, and WhichTF were retrieved from https://github.com/zanglab/bart2, http://homer.ucsd.edu/homer/,

699 and https://bitbucket.org/bejerano/whichtf. For ChIP-Atlas and i-cisTarget, we used online portals.
700 The same epigenomic profiling datasets given to BIT were used as input, and output TR ranking
701 lists were used to perform Fisher-exact tests with the same steps as for BIT.

702 **Gene ontology enrichment analysis:** We used the R package “clusterProfiler” to conduct gene
703 ontology enrichment analysis. The top-ranked TRs in the output list of BIT were first converted to
704 their ensemble ids, next by contrasting them to the universe of all available genes. We computed
705 the significance of the enrichment of the top TRs in the GO terms. The top 10 TRs in TR
706 perturbation experiments, and top 20 TRs in the remaining cases were used. We used q-value
707 cutoff of 0.05 and listed the top 5 or 10 terms in the final outputs.

Data Availability

Only public datasets were used in this study. The reference ChIP-seq datasets were retrieved from GTRD. TR perturbation experiment data were retrieved from GEO (GSE153237[40], GSE173416[41], GSE114102[42], and GSE234331[84]). Cancer-type specific accessible regions were retrieved from TCGA[60]. CRISPR/Cas9 knockout screening data were retrieved from DepMap[61] (https://depmap.org/portal/). scATAC-seq data were retrieved from 10XGenomics and GSE227265[85]. We initialized a data repository on Zenodo (https://zenodo.org/records/13732877).

Code Availability

BIT software is available on GitHub (https://github.com/ZeyuL01/BIT), and through a web portal for online analysis (http://43.135.174.109:8080/).

Acknowledgements

This work was supported by the following funding: the Rally Foundation, Sam Day 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, R01HG011996, and R01HL144969) (to L.X.) (R01CA258584) (to X.W.)

Author Information

Authors and Affiliations

Department of Statistics and Data Science, Southern Methodist University, Dallas, TX, USA

Zeyu Lu

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

Lin Xu

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

Lin Xu

Division of Data Science, College of Science, University of Texas at Arlington, Arlington, TX, USA

Xinlei Wang

Department of Mathematics, University of Texas at Arlington, Arlington, TX, USA

Xinlei Wang, Zeyu Lu

Contributions

Lu, Z. conducted data-processing, model design, simulation, and validation.

Xu, L. and Wang, X. conceived the ideas, designed the study, and acquired the funding.

All members have read, revised, and approved the final manuscript.

Corresponding authors

Correspondence to Xinlei Wang or Lin Xu.

References

1 Casamassimi, A. & Ciccodicola, A. Transcriptional Regulation: Molecules, Involved Mechanisms, and Misregulation. *Int J Mol Sci* **20** (2019). https://doi.org/10.3390/ijms20061281

2 Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650-665 (2018). https://doi.org/10.1016/j.cell.2018.01.029

3 Andersson, R. & Sandelin, A. Determinants of enhancer and promoter activities of regulatory elements. *Nat Rev Genet* **21**, 71-87 (2020). https://doi.org/10.1038/s41576-019-0173-8

4 Mullen, A. C. *et al.* Master transcription factors determine cell-type-specific responses to TGF-beta signaling. *Cell* **147**, 565-576 (2011). https://doi.org/10.1016/j.cell.2011.08.050

5 Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307-319 (2013). https://doi.org/10.1016/j.cell.2013.03.035

6 Hwang, J. Y. & Zukin, R. S. REST, a master transcriptional regulator in neurodegenerative disease. *Curr Opin Neurobiol* **48**, 193-200 (2018). https://doi.org/10.1016/j.conb.2017.12.008

7 Bradner, J. E., Hnisz, D. & Young, R. A. Transcriptional Addiction in Cancer. *Cell* **168**, 629-643 (2017). https://doi.org/10.1016/j.cell.2016.12.013

8 Chen, Y., Xu, L., Lin, R. Y., Müschen, M. & Koeffler, H. P. Core transcriptional regulatory circuitries in cancer. *Oncogene* **39**, 6633-6646 (2020). https://doi.org/10.1038/s41388-020-01459-w

9 Pei, H., Guo, W., Peng, Y., Xiong, H. & Chen, Y. Targeting key proteins involved in transcriptional regulation for cancer therapy: Current strategies and future prospective. *Med Res Rev* **42**, 1607-1660 (2022). https://doi.org/10.1002/med.21886

773 10 Rottner, A. K. *et al.* A genome-wide CRISPR screen identifies CALCOCO2 as a regulator
774 of beta cell function influencing type 2 diabetes risk. *Nat Genet* **55**, 54-65 (2023).
775 https://doi.org/10.1038/s41588-022-01261-2
776 11 Balakrishnan, S., Dhavamani, S. & Prahalathan, C. beta-Cell specific transcription
777 factors in the context of diabetes mellitus and beta-cell regeneration. *Mech Dev* **163**,
778 103634 (2020). https://doi.org/10.1016/j.mod.2020.103634
779 12 Hong, J. H. & Zhang, H. G. Transcription Factors Involved in the Development and
780 Prognosis of Cardiac Remodeling. *Front Pharmacol* **13**, 828549 (2022).
781 https://doi.org/10.3389/fphar.2022.828549
782 13 Bauer, A. J. & Martin, K. A. Coordinating Regulation of Gene Expression in
783 Cardiovascular Disease: Interactions between Chromatin Modifiers and Transcription
784 Factors. *Front Cardiovasc Med* **4**, 19 (2017). https://doi.org/10.3389/fcvm.2017.00019
785 14 Bhagwat, A. S. & Vakoc, C. R. Targeting Transcription Factors in Cancer. *Trends Cancer*
786 **1**, 53-65 (2015). https://doi.org/10.1016/j.trecan.2015.07.001
787 15 Bushweller, J. H. Targeting transcription factors in cancer - from undruggable to reality.
788 *Nat Rev Cancer* **19**, 611-624 (2019). https://doi.org/10.1038/s41568-019-0196-7
789 16 Darnell, J. E., Jr. Transcription factors as targets for cancer therapy. *Nat Rev Cancer* **2**,
790 740-749 (2002). https://doi.org/10.1038/nrc906
791 17 Yan, F., Powell, D. R., Curtis, D. J. & Wong, N. C. From reads to insight: a hitchhiker's
792 guide to ATAC-seq data analysis. *Genome Biol* **21**, 22 (2020).
793 https://doi.org/10.1186/s13059-020-1929-3
794 18 Libermann, T. A. & Baltimore, D. Activation of interleukin-6 gene expression through the
795 NF-kappa B transcription factor. *Mol Cell Biol* **10**, 2327-2334 (1990).
796 https://doi.org/10.1128/mcb.10.5.2327-2334.1990

797 19 Gilley, J., Coffer, P. J. & Ham, J. FOXO transcription factors directly activate bim gene
798 expression and promote apoptosis in sympathetic neurons. *J Cell Biol* **162**, 613-622
799 (2003). https://doi.org/10.1083/jcb.200303026
800 20 Semenza, G. L. Targeting HIF-1 for cancer therapy. *Nat Rev Cancer* **3**, 721-732 (2003).
801 https://doi.org/10.1038/nrc1187
802 21 Lu, Z., Xiao, X., Zheng, Q., Wang, X. & Xu, L. Assessing next-generation sequencing-
803 based computational methods for predicting transcriptional regulators with query gene
804 sets. *Brief Bioinform* **25** (2024). https://doi.org/10.1093/bib/bbae366
805 22 Kulaeva, O. I., Nizovtseva, E. V., Polikanov, Y. S., Ulianov, S. V. & Studitsky, V. M.
806 Distant activation of transcription: mechanisms of enhancer action. *Mol Cell Biol* **32**,
807 4892-4897 (2012). https://doi.org/10.1128/mcb.01127-12
808 23 Tsompana, M. & Buck, M. J. Chromatin accessibility: a window into the genome.
809 *Epigenetics Chromatin* **7**, 33 (2014). https://doi.org/10.1186/1756-8935-7-33
810 24 Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory
811 epigenome. *Nat Rev Genet* **20**, 207-220 (2019). https://doi.org/10.1038/s41576-018-
812 0089-8
813 25 Minnoye, L. *et al.* Chromatin accessibility profiling methods. *Nature Reviews Methods*
814 *Primers* **1**, 10 (2021). https://doi.org/10.1038/s43586-020-00008-9
815 26 Park, P. J. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev*
816 *Genet* **10**, 669-680 (2009). https://doi.org/10.1038/nrg2641
817 27 Tanigawa, Y., Dyer, E. S. & Bejerano, G. WhichTF is functionally important in your open
818 chromatin data? *PLoS Comput Biol* **18**, e1010378 (2022).
819 https://doi.org/10.1371/journal.pcbi.1010378
820 28 Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime
821 cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-
822 589 (2010). https://doi.org/10.1016/j.molcel.2010.05.004

29 Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell* **152**, 327-339 (2013). https://doi.org/10.1016/j.cell.2012.12.009

30 Slattery, M. *et al.* Absence of a simple code: how transcription factors read the genome. *Trends Biochem Sci* **39**, 381-399 (2014). https://doi.org/10.1016/j.tibs.2014.07.002

31 Wang, Z. *et al.* BART: a transcription factor prediction tool with query gene sets or epigenomic profiles. *Bioinformatics* **34**, 2867-2869 (2018). https://doi.org/10.1093/bioinformatics/bty194

32 Oki, S. *et al.* ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. *EMBO Rep* **19** (2018). https://doi.org/10.15252/embr.201846255

33 Machanick, P. & Bailey, T. L. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* **27**, 1696-1697 (2011). https://doi.org/10.1093/bioinformatics/btr189

34 Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* **14**, 1083-1086 (2017). https://doi.org/10.1038/nmeth.4463

35 Tang, F. *et al.* Chromatin profiles classify castration-resistant prostate cancers suggesting therapeutic targets. *Science (New York, N.Y.)* **376**, eabe1505 (2022). https://doi.org/10.1126/science.abe1505

36 Wang, Z., Zhang, Y. & Zang, C. BART3D: inferring transcriptional regulators associated with differential chromatin interactions from Hi-C data. *Bioinformatics* **37**, 3075-3078 (2021). https://doi.org/10.1093/bioinformatics/btab173

37 McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**, 495-501 (2010). https://doi.org/10.1038/nbt.1630

38 Kolmykov, S. *et al.* GTRD: an integrated view of transcription regulation. *Nucleic Acids Res* **49**, D104-d111 (2021). https://doi.org/10.1093/nar/gkaa1057

39 Qin, Q. *et al.* Lisa: inferring transcriptional regulators through integrative modeling of public chromatin accessibility and ChIP-seq data. *Genome Biol* **21**, 32 (2020). https://doi.org/10.1186/s13059-020-1934-6

40 Xu, B. *et al.* Acute depletion of CTCF rewires genome-wide chromatin accessibility. *Genome Biol* **22**, 244 (2021). https://doi.org/10.1186/s13059-021-02466-0

41 Deng, W. *et al.* Reactivation of developmentally silenced globin genes by forced chromatin looping. *Cell* **158**, 849-860 (2014). https://doi.org/10.1016/j.cell.2014.05.050

42 Lee, K. *et al.* FOXA2 Is Required for Enhancer Priming during Pancreatic Differentiation. *Cell Rep* **28**, 382-393.e387 (2019). https://doi.org/10.1016/j.celrep.2019.06.034

43 Gruber, S., Haering, C. H. & Nasmyth, K. Chromosomal cohesin forms a ring. *Cell* **112**, 765-777 (2003). https://doi.org/10.1016/s0092-8674(03)00162-4

44 Haering, C. H., Löwe, J., Hochwagen, A. & Nasmyth, K. Molecular architecture of SMC proteins and the yeast cohesin complex. *Mol Cell* **9**, 773-788 (2002). https://doi.org/10.1016/s1097-2765(02)00515-4

45 Kadauke, S. & Blobel, G. A. Chromatin loops in gene regulation. *Biochim Biophys Acta* **1789**, 17-25 (2009). https://doi.org/10.1016/j.bbagrm.2008.07.002

46 Hansen, A. S., Cattoglio, C., Darzacq, X. & Tjian, R. Recent evidence that TADs and chromatin loops are dynamic structures. *Nucleus* **9**, 20-32 (2018). https://doi.org/10.1080/19491034.2017.1389365

47 Nora, E. P. *et al.* Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell* **169**, 930-944.e922 (2017). https://doi.org/10.1016/j.cell.2017.05.004

48 Rao, S. S. P. *et al.* Cohesin Loss Eliminates All Loop Domains. *Cell* **171**, 305-320.e324 (2017). https://doi.org/10.1016/j.cell.2017.09.026

49 Szklarczyk, D. *et al.* The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res* **51**, D638-d646 (2023). https://doi.org/10.1093/nar/gkac1000

873 50 Leonard, M., Brice, M., Engel, J. D. & Papayannopoulou, T. Dynamics of GATA
874 transcription factor expression during erythroid differentiation. *Blood* **82**, 1071-1079
875 (1993).
876 51 Cheng, Y. *et al.* Erythroid GATA1 function revealed by genome-wide analysis of
877 transcription factor occupancy, histone modifications, and mRNA expression. *Genome*
878 *Res* **19**, 2172-2184 (2009). https://doi.org/10.1101/gr.098921.109
879 52 Maeda, T. *et al.* LRF is an essential downstream target of GATA1 in erythroid
880 development and regulates BIM-dependent apoptosis. *Dev Cell* **17**, 527-540 (2009).
881 https://doi.org/10.1016/j.devcel.2009.09.005
882 53 Siatecka, M. & Bieker, J. J. The multifunctional role of EKLF/KLF1 during erythropoiesis.
883 *Blood* **118**, 2044-2054 (2011). https://doi.org/10.1182/blood-2011-03-331371
884 54 Wang, H. *et al.* Pdx1 level defines pancreatic gene expression pattern and cell lineage
885 differentiation. *J Biol Chem* **276**, 25279-25286 (2001).
886 https://doi.org/10.1074/jbc.M101233200
887 55 Watt, A. J., Zhao, R., Li, J. & Duncan, S. A. Development of the mammalian liver and
888 ventral pancreas is dependent on GATA4. *BMC Dev Biol* **7**, 37 (2007).
889 https://doi.org/10.1186/1471-213x-7-37
890 56 Fujimoto, K. & Polonsky, K. S. Pdx1 and other factors that regulate pancreatic beta-cell
891 survival. *Diabetes Obes Metab* **11 Suppl 4**, 30-37 (2009). https://doi.org/10.1111/j.1463-
892 1326.2009.01121.x
893 57 Allen, H. L. *et al.* GATA6 haploinsufficiency causes pancreatic agenesis in humans. *Nat*
894 *Genet* **44**, 20-22 (2011). https://doi.org/10.1038/ng.1035
895 58 Iwafuchi-Doi, M. & Zaret, K. S. Pioneer transcription factors in cell reprogramming.
896 *Genes Dev* **28**, 2679-2692 (2014). https://doi.org/10.1101/gad.253443.114
897 59 Craswell, N. in *Encyclopedia of Database Systems* (eds Ling Liu & M. Tamer ÖZsu)
898 1703-1703 (Springer US, 2009).

60 Corces, M. R. *et al.* The chromatin accessibility landscape of primary human cancers. *Science* **362** (2018). https://doi.org/10.1126/science.aav1898

61 Meyers, R. M. *et al.* Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat Genet* **49**, 1779-1784 (2017). https://doi.org/10.1038/ng.3984

62 Watson, P. A., Arora, V. K. & Sawyers, C. L. Emerging mechanisms of resistance to androgen receptor inhibitors in prostate cancer. *Nat Rev Cancer* **15**, 701-711 (2015). https://doi.org/10.1038/nrc4016

63 Patel, H. K. & Bihani, T. Selective estrogen receptor modulators (SERMs) and selective estrogen receptor degraders (SERDs) in cancer treatment. *Pharmacol Ther* **186**, 1-24 (2018). https://doi.org/10.1016/j.pharmthera.2017.12.012

64 Anzick, S. L. *et al.* AIB1, a steroid receptor coactivator amplified in breast and ovarian cancer. *Science* **277**, 965-968 (1997). https://doi.org/10.1126/science.277.5328.965

65 Kim, M. R., Wu, M. J., Zhang, Y., Yang, J. Y. & Chang, C. J. TET2 directs mammary luminal cell differentiation and endocrine response. *Nat Commun* **11**, 4642 (2020). https://doi.org/10.1038/s41467-020-18129-w

66 Durbin, A. D. *et al.* Selective gene dependencies in MYCN-amplified neuroblastoma include the core transcriptional regulatory circuitry. *Nat Genet* **50**, 1240-1246 (2018). https://doi.org/10.1038/s41588-018-0191-z

67 Gryder, B. E. *et al.* Histone hyperacetylation disrupts core gene regulatory architecture in rhabdomyosarcoma. *Nat Genet* **51**, 1714-1722 (2019). https://doi.org/10.1038/s41588-019-0534-4

68 Reddy, J. *et al.* Predicting master transcription factors from pan-cancer expression data. *Sci Adv* **7**, eabf6123 (2021). https://doi.org/10.1126/sciadv.abf6123

69 Liu, T. *et al.* Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol* **12**, R83 (2011). https://doi.org/10.1186/gb-2011-12-8-r83

70 Győrffy, B. Integrated analysis of public datasets for the discovery and validation of survival-associated genes in solid tumors. *Innovation (Camb)* **5**, 100625 (2024). https://doi.org/10.1016/j.xinn.2024.100625

71 Zhang, K., Zemke, N. R., Armand, E. J. & Ren, B. A fast, scalable and versatile tool for analysis of single-cell omics data. *Nat Methods* **21**, 217-227 (2024). https://doi.org/10.1038/s41592-023-02139-9

72 Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat Methods* **18**, 1333-1341 (2021). https://doi.org/10.1038/s41592-021-01282-5

73 Gertz, J. *et al.* Distinct properties of cell-type-specific and shared transcription factor binding sites. *Mol Cell* **52**, 25-36 (2013). https://doi.org/10.1016/j.molcel.2013.08.037

74 Zhang, Y. & Reinberg, D. Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails. *Genes Dev* **15**, 2343-2360 (2001). https://doi.org/10.1101/gad.927301

75 Jolma, A. *et al.* DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **527**, 384-388 (2015). https://doi.org/10.1038/nature15518

76 Meyer, C. A. & Liu, X. S. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat Rev Genet* **15**, 709-721 (2014). https://doi.org/10.1038/nrg3788

77 Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res* **40**, e72 (2012). https://doi.org/10.1093/nar/gks001

78 Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* **12**, R18 (2011). https://doi.org/10.1186/gb-2011-12-2-r18

79 Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**, 21931-21936 (2010). https://doi.org/10.1073/pnas.1016071107

80 Zhang, C., Bütepage, J., Kjellström, H. & Mandt, S. Advances in Variational Inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **41**, 2008-2026 (2019). https://doi.org/10.1109/TPAMI.2018.2889774

81 Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association* **112**, 859-877 (2017). https://doi.org/10.1080/01621459.2017.1285773

82 Jia, G. *et al.* RCRnorm: An integrated system of random-coefficient hierarchical regression models for normalizing NanoString nCounter data. *Ann Appl Stat* **13**, 1617-1647 (2019). https://doi.org/10.1214/19-aoas1249

83 Polson, N. G., Scott, J. G. & Windle, J. Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables. *Journal of the American Statistical Association* **108**, 1339-1349 (2013). https://doi.org/10.1080/01621459.2013.829001

84 Gamble, N. *et al.* PU.1 and BCL11B sequentially cooperate with RUNX1 to anchor mSWI/SNF to poise the T cell effector landscape. *Nat Immunol* **25**, 860-872 (2024). https://doi.org/10.1038/s41590-024-01807-y

85 Craig, A. J. *et al.* Genome-wide profiling of transcription factor activity in primary liver cancer using single-cell ATAC sequencing. *Cell Rep* **42**, 113446 (2023). https://doi.org/10.1016/j.celrep.2023.113446