BIT: Bayesian Identification of Transcriptional Regulators

Zeyu Lu[1], 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. 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:

BIT is a novel Bayesian hierarchical model capable of predicting transcriptional regulators (TRs) from the input of user-provided epigenomic regions. TRs are critical molecules in transcriptional regulation. Many diseases and cancers are linked to the dysfunction of TRs. Knowing TRs in certain biological process can help find new biomarkers or therapeutic targets. Thus, BIT formulates a novel Bayesian hierarchical model with the Pólya-gamma data augmentation strategy. Based on collected ChIP-seq datasets, BIT can identify TRs responsible for the genome-wide binding pattern within the user-provided epigenomic regions. BIT has been validated by using a simulation study and three applications.

Key Words: transcriptional regulator, Bayesian hierarchical modelling, Pólya-gamma data augmentation, epigenomics regions,

Introduction

Transcriptional regulators (TRs), comprising transcription factors, cofactors, chromatin modulators, and various regulatory molecules, play a pivotal role in regulating the gene transcription[1-3]. TRs are essential in multiple cellular processes, including growth, differentiation, morphogenesis, and apoptosis[4-6]. Dysfunction of TRs can lead to various diseases, including cancers[7-9], diabetes[10,11], and heart diseases[12,13]. Thus, gaining insight into the identity and function of TRs can be helpful in a broad range of biological and medical research, such as finding new biomarkers and therapeutic targets[14-16].

Beyond traditional labor-intensive wet lab experiments, a growing interest in leveraging statistics has emerged to discern TRs based on the NGS data and the intricate interplay between epigenomic features. However, it is noted that the current available state-of-art computational methods to identify TRs based on NGS data suffer from various significant limitations. (i) Most of these methods rely on a gene list as input[17-21], requiring the conversion of peak regions derived from epigenomic data into associated genes. However, there is a lack of clear standards on how to perform this conversion, resulting in potentially inconsistent outcomes when applying different cutoffs to the same peaks. (ii) Some methods (e.g., MEME-ChIP[22], HOMER[23], and WhichTF[24]) rely on DNA sequence motifs. However, highly similar motifs shared among many TRs make accurate identification challenging[25], motifs alone cannot characterize any in-vivo binding event, and the lack of well-defined motifs for some TRs further reduces the reliability of using motifs for identification[2,26]. (iii) Each TR's binding pattern can vary significantly in different cell types or environmental conditions. However, published methods have no rigorous treatment of the within-TR heterogeneity and homogeneity when dealing with multiple TR ChIP-seq datasets from the same TR, resulting in biased results. (iv) All the existing methods employ frequentist approaches for TR statistical significance, lacking coherent probabilistic interpretation or

uncertainty estimation. Given the dynamic nature of TR binding events, the ability to quantify uncertainty is not just beneficial but essential.

To overcome the proposed limitations, we developed a novel Bayesian hierarchical approach for the identification of TRs responsible for genome-wide binding patterns, named BIT (Bayesian identification of Transcriptional regulators). Bayesian hierarchical modeling is one of the most powerful approaches to model complicated biological phenomena and integrate diverse information. First, BIT eliminates the need for an input gene list, taking peak regions derived from an epigenomic dataset as input. This removes the need for peaks-to-genes conversion, reducing false positives and broadening its application compared to the existing methods. Moreover, BIT does not rely on DNA sequence motifs, and so avoids the use of computationally predicted motifs that might cause unreliable TR predictions in motif-based approaches[22,23,27]. Second, via a unique hierarchical model setup, BIT integrates information across multiple TRs and multiple ChIP-seq datasets of the same TRs, appropriately accounting for both between- and within-TR homogeneity and heterogeneity. This formal model-based treatment leads to more reliable and robust prediction, as shown in our results. By contrast, many existing methods need to conduct thousands of separate ad-hoc statistical tests (e.g., BART[18], Lisa[17], and ChIP-Atlas[28]) and therefore tend to rank TRs with many ChIP-seq datasets higher[29]. Third, BIT employs a Bayesian framework to handle its modeling and computational needs. A Bayesian approach is inherently attractive in that (i) it pools information from various sources for better estimation of model parameters, which is essential for effective prediction; (ii) it offers the advantage of readily quantifying the estimation uncertainty such as providing a standard error or 95% credible interval of any quantity of interest (e.g., a parameter estimate or predicted importance score); (iii) in situations when meaningful prior knowledge is available, it enables researchers to formally incorporate such knowledge through prior specification for improved results while in all other situations, with the default non-informative prior setting, it becomes a

fully automated procedure that does not require any tuning parameter. By contrast, none of the existing methods can offer any of these advantages.

Results

The integration of omics data has been proved as a feasible approach to identify TRs in biomedical research[30]. It is well recognized that the binding sites of TRs critical to a specific cellular process are primarily enriched in the regulatory epigenomic regions of multiple downstream genes that drive the process[31]. Such epigenomic regions can be directly measured by epigenomic technologies[32,33] (e.g., H3K27ac ChIP-seq/ATAC-seq) or converted from genes identified by transcriptomic technologies and serve as an input “fingerprint”[34]. On the other hand, the binding sites of an individual TR can be measured by TR ChIP-seq[35], representing the most accurate available in-vivo data type for TR binding, with each ChIP-seq experiment unveiling unique binding profile for a specific TR. The pre-collected TR ChIP-seq datasets act as a reference database, using stored binding profiles as distinctive "fingerprints" for individual TRs. Users can compare the "fingerprint" of a newly studied cellular process with those in this reference database to pinpoint functional TRs that have similar epigenomic chromatin profiles.

BIT relies on two biological assumptions: (i) If a TR governs the user-provided epigenomic regions, the genome-wide binding pattern of this TR should be more consistent with the input regions. (ii) The distribution formulated by the similarity of TR ChIP-seq datasets from the same TR should be able to differentiate the TR from the population. The two assumptions are commonly used in other bioinformatics tools[18]. The first assumption implies that when a functional TR is actively involved in a specific cellular process that derives the user-provided epigenomic regions, this leads to a more frequent interaction between this functional TR and the DNA. Consequently, a significant number of the captured epigenomic regions are likely to encompass or be adjacent to the binding sites of this functional TR. The second assumption recognizes that certain variations exist within ChIP-seq datasets for the same TR, stemming from factors such as measurement errors, dataset quality, and cell type specificity. Despite these variations, it is expected that the general consistency observed across these ChIP-seq

datasets — a result of the TR's preference for binding to specific sites on chromatin — will create a distinctive binding pattern. When comparing with the input epigenomic regions, this pattern should be substantial enough to differentiate the functional TRs from other non-functional TRs.

BIT only requires user-provided epigenomic regions as input and consists of two main steps: Data pre-processing and Bayesian posterior inference with Gibbs sampler. First, BIT transforms the coordinates of user-provided epigenomic regions and ChIP-seq peaks to fixed-length binary vectors with 0s and 1s. Second, BIT formulates a Bayesian hierarchical model to integrate information and uses the MCMC algorithm to perform the posterior estimation of the importance of TRs. Finally, BIT assigns each TR a “BIT score” (**Fig. 1**). Based on the assumptions stated above, the BIT score can be used as the overall significance indicator of the TR being functional. Alternatively, it can be interpreted as the Jaccard similarity coefficient that quantifies the distance between the binary vector representation of the TR unique binding pattern and the binary vector representation of the input epigenomic regions.

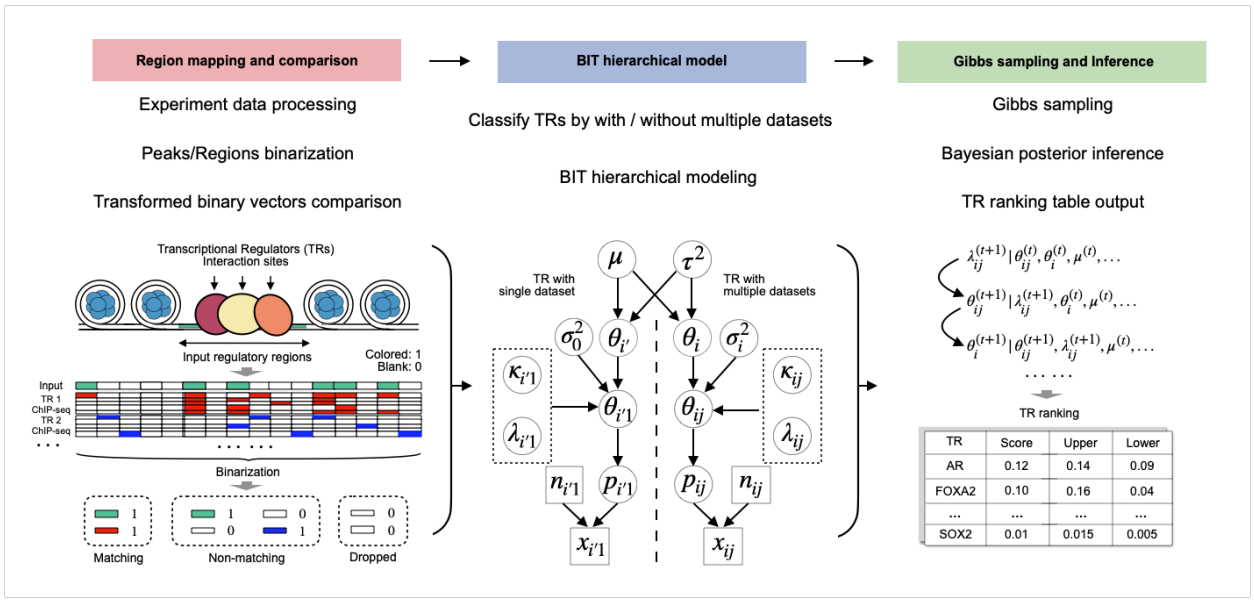


Figure 1 A. Transcription regulators such as transcription factors, cofactors, and chromatin modulators mostly bind to accessible chromatin regions to regulate gene transcription. NGS

techniques such as TR ChIP-seq and ATAC-seq can acquire the information of binding sites and accessible regions, respectively. Input regions and collected TR ChIP-seq peaks are transformed to binary vectors, the ratio of the “Good” cases over all “informative" cases serve as the raw score. B. BIT is a Bayesian hierarchical framework integrated with Pólya-Gamma data augmentation strategy to pool information from multiple TR ChIP-seq datasets. C. Gibbs sampler is used to derive the posterior estimation of TR level BIT score and give the final ranking table. The identified TRs can then be used for future research.

To verify BIT's performance, we first used a realistic and comprehensive simulation study to examine the accuracy of the posterior estimation. As BIT is the first Bayesian model-based approach that directly models over the accumulated TR ChIP-seq datasets, we used the naïve method as the baseline. Second, we examined BIT’s feasibility with three different applications: (i) the identification of TRs by using differential accessible regions (DARs) that are derived from TR perturbation experiments. Perturbed TRs should be responsible for the DARs, and therefore these DARs can serve as ground truth datasets. (ii) we tested BIT in the identification of TRs by using cancer-type specific accessible epigenomic regions. The identified TRs are cross-validated by findings in previous publications and results derived from CRISPR/Cas9 loss-of-function analysis. We also benchmarked results from BIT against results from several other methods to show its efficiency. (iii) given the fast development of single-cell techniques, cell-type level accessible regions can now be easily derived from scATAC-seq data. We further applied BIT to two sets of cell type specific accessible regions and validated the results.

Simulation shows BIT can estimate parameters with high accuracy.

It is essential to use simulation study to show the performance of BIT in the posterior estimation, as we can generate data from known distributions. The simulation study evaluated BIT in three steps. (i). The mean squared error (MSE) of estimating the hyperparameters μ, τ^2, σ_0^2 and TR level parameters θ_i and $\theta_{i'}$. (ii). The ability to recover the original ranks of generated θ_i and $\theta_{i'}$. The metric used is spearman rho. (iii). The robustness of BIT that whether it heavily depends on

148 the normality assumption. We used the naïve method as the baseline and generated 100
149 replicates in each simulation setup.

150 In the simulation, we aimed to mirror realistic scenarios in which approximately 1,000
151 transcriptional regulators are associated with single or multiple datasets. We adopted three
152 conditions ($I = 500, 1000, 1500$) for the total number of θ_i and $\theta_{i'}$, each θ_i or $\theta_{i'}$ represents one
153 TR (with or without multiple datasets). It is reasonable to examine the performance of BIT over
154 different total numbers of θ_i and $\theta_{i'}$. By now, we collected 10770 TR ChIP-seq datasets that can
155 be associated with 987 TRs. However, these numbers can further increase as more TR ChIP-
156 seq experiments could be conducted in the laboratory. In contrast, certain TRs could be
157 excluded from the analysis if they have been validated before for the cellular process that
158 generates the input epigenomic regions.

159 Based on these considerations, we varied the number of θ_i and $\theta_{i'}$ generated in the simulation
160 to cover more scenarios. Similarly, the true values of hyperparameters are varied to examine
161 their effects on the model performance. We expect BIT to be effective regardless of the number
162 of TRs generated and the actual value of hyperparameters used. Furthermore, given that the
163 distribution of the number of collected datasets for each TR is highly right skewed and has a
164 long tail, we therefore used truncated non-central t distribution to generate the number of
165 datasets (θ_{ij} and $\theta_{i'1}$.) in the simulation.

166 In general, BIT can successfully capture the hyperparameters with high accuracy (**Fig. 2A**).
167 First, it is noted that the MSE for the BIT estimation of hyperparameters μ, τ^2, and σ_0^2 are
168 consistently lower than the baseline method and are close to zero in all cases. Second, using a
169 low number of TRs can have relatively higher MSE, but the difference is negligible in BIT. Third,
170 the MSE for the estimation of σ_0^2 is positively correlated with the actual value, which is
171 reasonable: as variability increases, the spread of data and the likelihood of outliers also

increases, which leads to a greater uncertainty in the estimation. Nevertheless, BIT still significantly outperforms the baseline method.

The min-max range of MSE for estimating the θ_i and $\theta_{i'}$ over 100 replicates also show BIT have better performance than baseline method (**Fig. 2B**). Unlike hyperparameters μ, τ^2, and σ_0^2, which are fixed prior to the simulation, θ_i and $\theta_{i'}$ are randomly generated in each replicate. The MSE in this context indicates the aggregate deviation between the estimation and the true values generated within one replicate. The result shows that BIT consistently achieves higher precision in estimating θ_i and $\theta_{i'}$ surpassing the baseline method. Besides the aggregated deviation, we also examine the ability of BIT to recover the original rank of randomly generated θ_i and $\theta_{i'}$. It is noted that τ^2 and σ_0^2 can affect the performance of methods (**Fig. 2C**), which is reasonable. τ^2 control the overall dispersion of θ_i and $\theta_{i'}$. Thus, a larger τ^2 will increase the distances among the true value of parameters, making them more easily distinguished. σ_0^2 is the noise used to generate dataset level parameters $\theta_{i'1}$. Increasing σ_0^2 will lead to a larger error in the estimation of $\theta_{i'}$, reducing the performance.

We also checked the robustness of BIT as multiple Gaussian distributions are assumed in the BIT model. We used to different non-Gaussian distributions, student t distribution and gamma distribution, to generate the θ_i and $\theta_{i'}$. The dispersion of simulated θ_i and $\theta_{i'}$ by these two distributions are different from those generated from Gaussian distribution (**Fig. 2E**). student t distribution is known for heavy-tailedness, as well as gamma distribution is known for skewness. The simulated data are shifted and scaled to recover the same mean and variability for comparison while keeping the pattern of heavy-tailedness and skewness. We noted the performance of BIT is not sensitive to the heavy-tailedness or skewness of the distribution of simulated data (**Fig. 2F**).

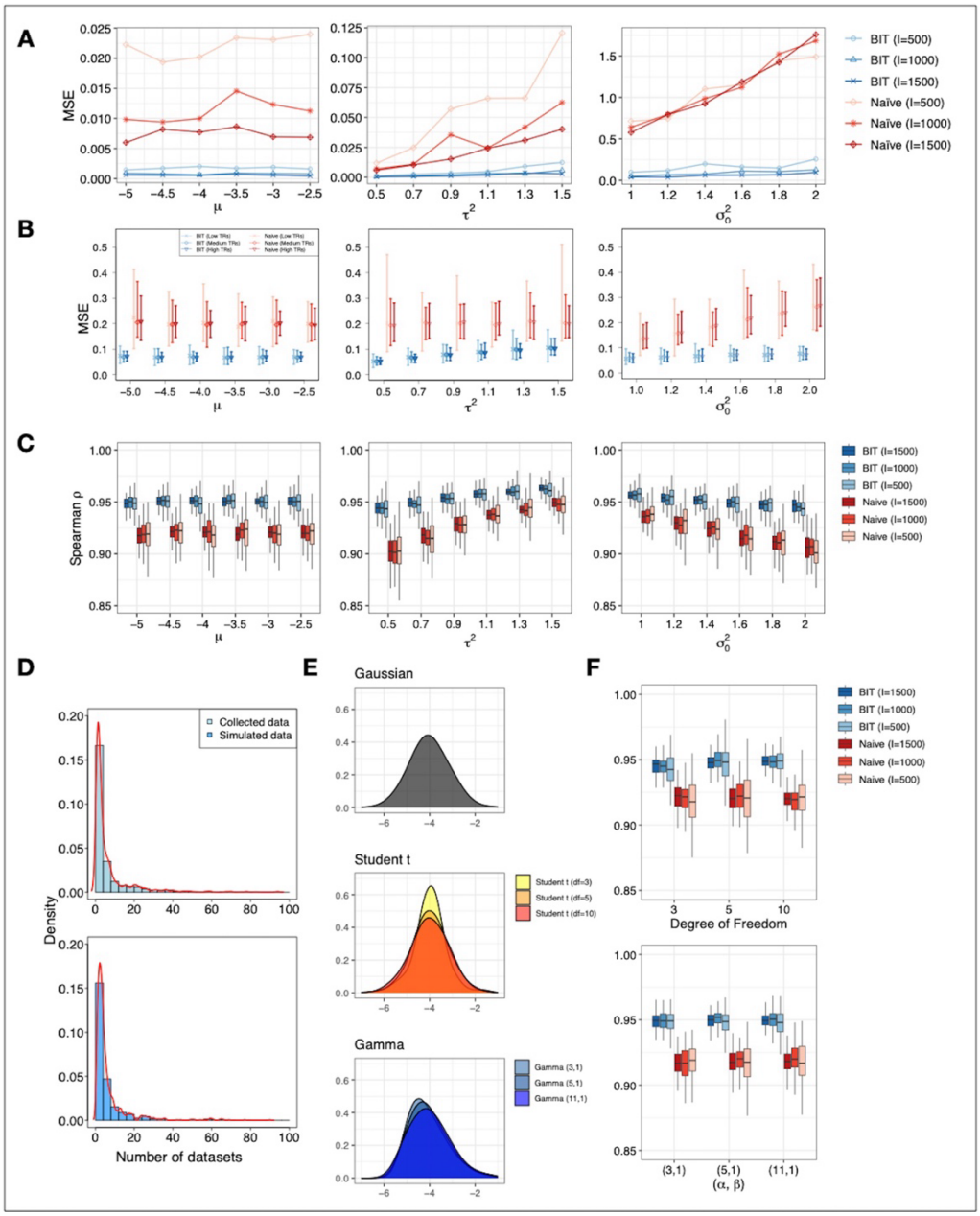


Figure 2. A. The MSE of the estimation for hyperparameters by using 100 replicates, BIT shows very high accuracy with MSE close to 0. B. The Min-Max MSE range bar of simulated θ_i and $\theta_{i'}$, still BIT show better performance in the estimation. C. The spearman rho correlation of BIT or Naive ranked simulated TRs versus true rank of simulated TRs. It is noted the BIT method outperform naive method and stay stable regardless of the change of hyperparameters. D. The simulated number of datasets under each TR mimic the collected number of datasets to make sure the simulation is realistic. E. The normality check, by using simulated data generated from t and gamma distribution, we examine whether the heavy-tailedness or skewness will significantly affect the BIT ranked simulated TRs F. The boxplot shows that BIT will not be significantly affected by normality problem.

206 In summary, we conducted several simulation studies to check whether the proposed BIT model
207 can lead to a reasonable estimation of parameters and recover the original ranks of the TRs as
208 expected. In all the simulation studies, BIT performs better than the baseline method, it can
209 estimate the parameter with high accuracy and recover the original ranks of randomly generated
210 θ_i and $\theta_{i'}$. In addition, it is noted that BIT is not sensitive to the normality assumption, and the
211 performance stays stable when the distributions used to generate simulated data have certain
212 heavy-tailedness or skewness.

213 **BIT can identify perturbed TRs from differential accessible regions (DARs)**

214 Differentially accessible regions (DARs) derived from differential accessibility analysis between
215 two biological conditions are critical in understanding the activity of TRs (**Fig. 3A**). We
216 demonstrate the feasibility of BIT in identifying TRs by using sets of DARs derived from three
217 separate TR perturbation experiments. Since the perturbation experiments are designed to
218 target specific TRs, these regulators are expected to have a direct connection to the derived
219 DARs. Therefore, the derived DARs can serve as ground truth data, and BIT is expected to
220 identify the perturbed TRs by using them as input. The first experiment is an acute depletion of
221 CTCF in the MLL-rearranged human B cell lymphoblastic leukemia (B-ALL) cell line SEM[36]. The
222 second one is from a ZBTB7A-knockout experiment in the HUDEP-2 cell line[37]. The third
223 experiment involves a FOXA2-knockout in the pancreatic progenitors 1 (PP1) differential stage
224 of human pluripotent stem cells (hPSCs)[38]. BIT has successfully identified the perturbed TRs
225 from the three experiments, and it has also identified several other TRs that can be reasonably
226 explained (**Fig. 3D**).

227 CTCF was ranked top 1 in the first experiment. Other top-ranked TRs include, SMC3, and
228 SMC1A, which are all components of cohesion complex. SMC proteins and RAD21 are core
229 proteins that formulate the ring-shaped cohesion complex (cohesion ring)[39,40]. CTCF and
230 cohesion ring can be crucial in forming chromatin loops and topologically associating domains

231 (TADs). The two structures are critical in the regulation of gene transcription[41,42]. Evidence has
232 shown that both CTCF and cohesion ring are essential for the maintenance of the two structures
233 when using acute depletion assays to handle proteins[43,44], which can explain why BIT also
234 identified the other three TRs after the acute depletion of CTCF.

235 In the second experiment, except for the ZBTB7A, GATA1 and KLF1 were also ranked in the top
236 10. Evidence has shown that erythroid differentiation and mutation can be controlled by GATA1
237 dependent transcription[45,46]. A previous study found ZBTB7A is a cofactor of GATA1 and plays
238 an essential antiapoptotic role during terminal erythroid differentiation[47], which explains why
239 GATA1 was also ranked at the top after the knockout of ZBTB7A. KLF1 can have a
240 multifunctional role in erythropoiesis and act as one of the master regulators[48].

241 BIT ranked FOXA2 in the top 3 in the third experiment. Though the ranking of perturbed TR is
242 not as high as the previous two experiments, we observed that other TRs, including PDX1,
243 GATA6, FOXA1, and GATA4 were also ranked in top 10. The observation is justifiable. These
244 TRs are both known for regulating the development and differentiation of pancreas[49-52].
245 Mutations and dysregulation of these regulators have been associated with various pancreatic
246 disorders. FOXA1 and FOXA2 are commonly recognized as pioneer factors that can change the
247 accessibility of chromatin[53]. PDX1, GATA6, and GATA4 might be recruited later to the accessible
248 regions to co-regulate the transcription process.

249 We further categorized the DARs as gain or loss accessible regions based on whether these
250 regions show increased (gain) or decreased (loss) accessibility after the perturbation of TRs
251 (**Fig. 3C and 3E**). It is noted that CTCF and FOXA2 can be well identified by BIT from increased
252 accessibility regions. Given that ZBTB7A is widely acknowledged as a repressor, it is not
253 surprising to observe that ZBTB7A is more significant with reduced accessibility regions[54]. Thus,
254 BIT can also help understand the functional role of the identified TRs in transcription.

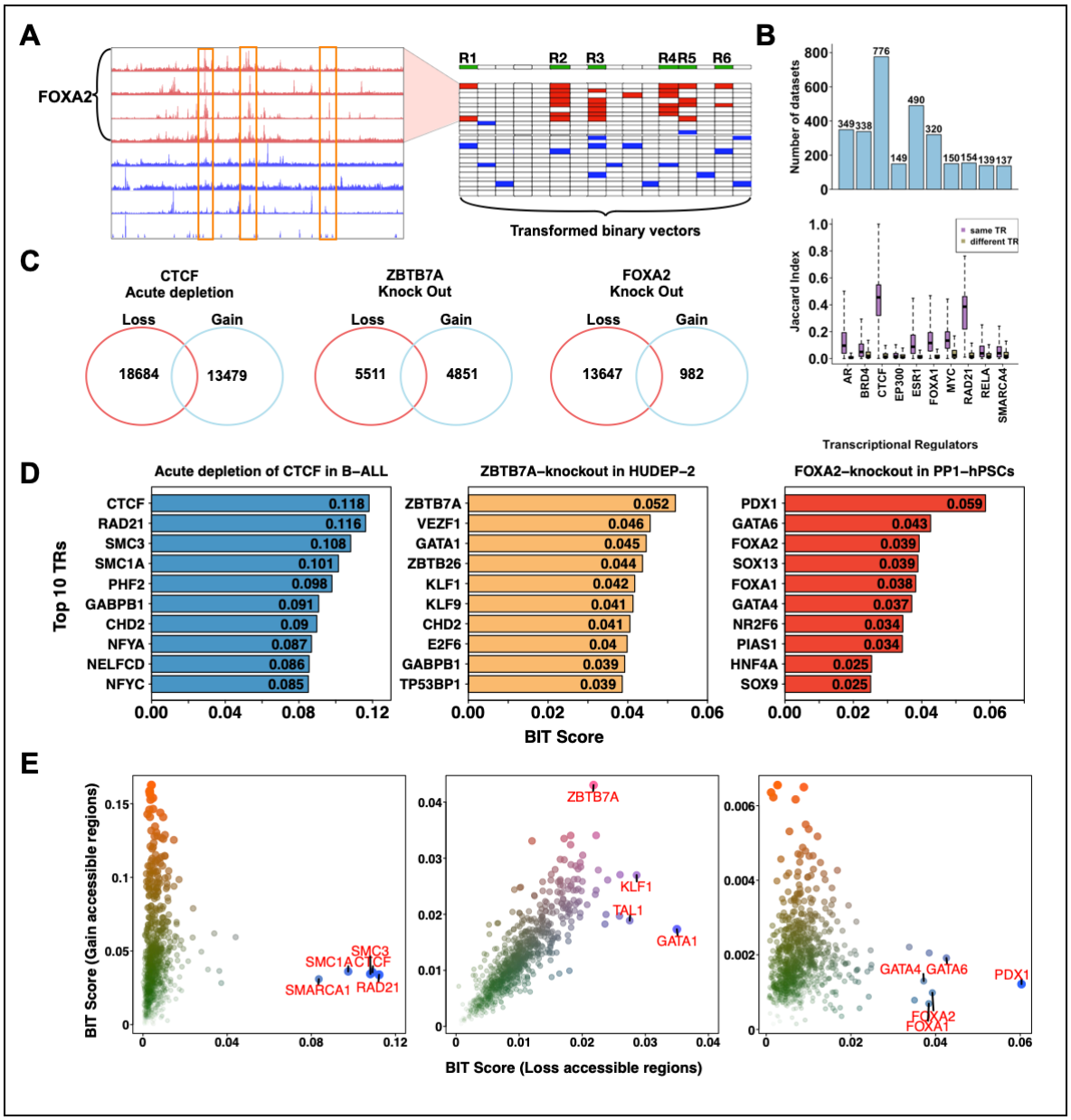


Figure 3. A. BIT first transforms DARs and all TR ChIP-seq datasets into binary vectors. If perturbation of specific TR is the main reason the derivation of DARs, these DARs should significantly match the binding sites of this TR. R1-R6 represent DARs, while red regions are peaks from FOXA2 TR ChIP-seq datasets and blue regions are from other non-FOXA2 TR ChIP-seq datasets. B. The Jaccard index of two transformed binary vectors randomly selected from same or different TR, which indicates regardless of factors such as cell line measured, the binding sites of same TR are much similar than different TR. C. The number of gain or loss accessible regions in each experiment. D. The top 10 TRs identified by BIT in each experiment. E. BIT score calculated by using only gain or loss accessible regions can be associated with TRs function as activator or repressor.

266 The results of the three TR perturbation experiments provide strong evidence for BIT's feasibility
267 in identifying functional TRs from DARs. BIT not only ranks the perturbed TRs in the top
268 positions but also demonstrates that it can find other experimentally validated TRs strongly
269 related to the perturbed TR or the cellular process.

270 **BIT can identify cancer-type specific TRs**

271 To further validate BIT's efficacy, we leveraged cancer-type specific accessible regions obtained
272 from The Cancer Genome Atlas (TCGA) database (details in **Table. S1**). TRs are well-
273 established drivers of crucial cellular processes in various cancers. However, a significant
274 portion of accessible regions recur across multiple cancer types. Including these non-specific
275 regions can hinder the identification of TRs critical to specific cancers. Therefore, employing
276 cancer-type specific accessible regions offered a more focused approach. We utilized regions
277 specific to nine cancer types: breast, bladder, colon, lung, liver, mesothelium, prostate,
278 squamous cell, and testis. More precisely, the ATAC-seq samples used to generate these
279 regions encompassed a broad range of carcinomas, including breast invasive carcinoma
280 (BRCA), bladder urothelial carcinoma (BLCA), and so on. The identified TRs were cross-
281 validated through existing publications and CRISPR/Cas9 screening data from The Cancer
282 Dependency Map Consortium (DepMap)[55], which evaluates cancer cell viability after gene
283 knockout.

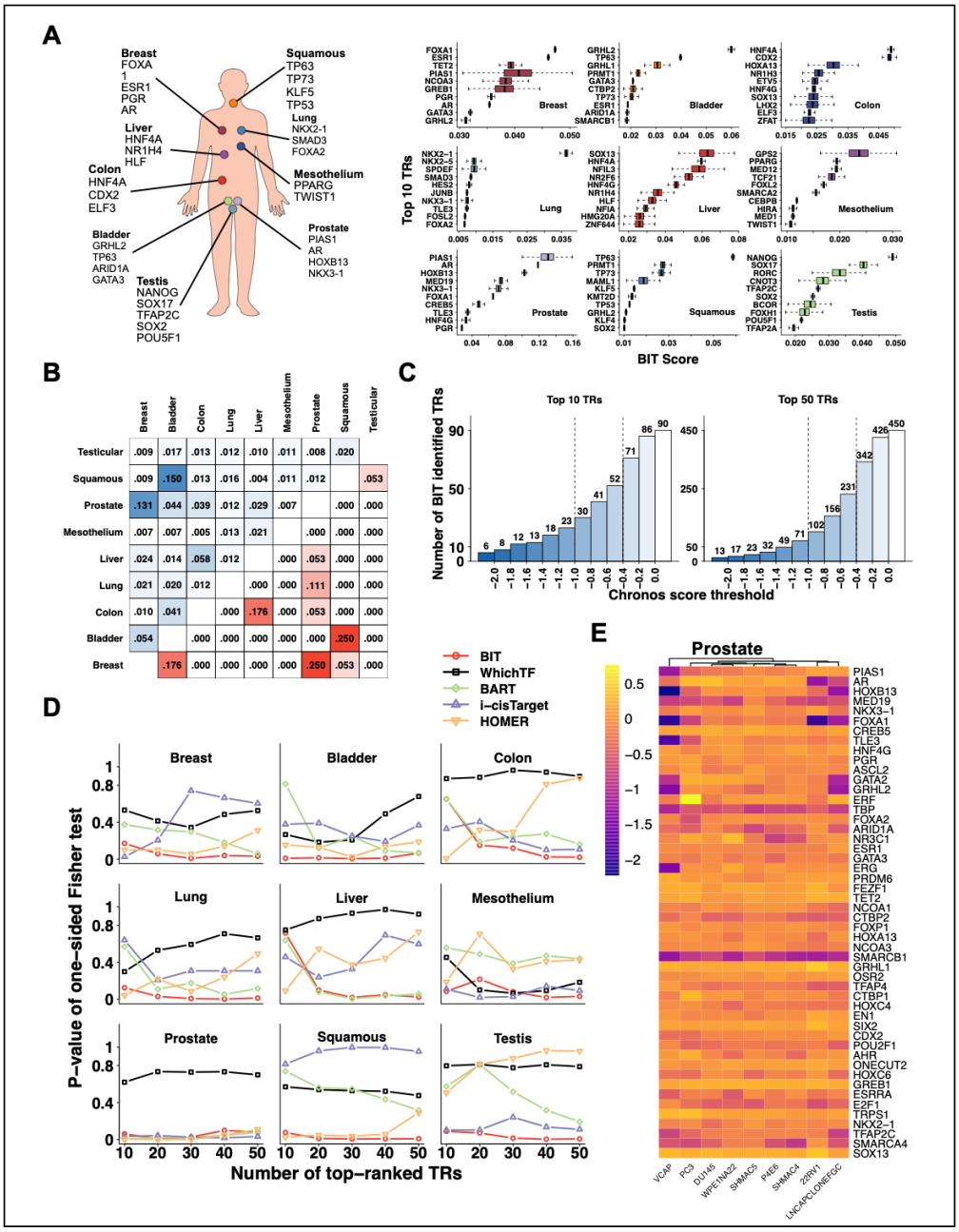
284 Firstly, Many BIT identified top-ranked TRs are well-known for their significant roles in regulating
285 normal cell behaviors such as differentiation and proliferation, and the mutation of these TRs
286 can be a major reason for the development of cancers. For instance, in breast cancer, top-
287 ranked TRs include FOXA1, ESR1, and PGR. PIAS1, AR, and HOXB13 were identified in
288 prostate cancer. NANOG, SOX17, and POU5F1 were identified by BIT in testicular cancer. The
289 detailed TR ranking list for each cancer type is in Table S2. A total of 33 TRs were cross-

290 validated by available publications (**Table. S3**), and some TRs have already been used as
291 biomarkers or therapeutic targets in cancer treatment[56,57].

292 Secondly, we examined whether the epigenomic regions retrieved from the TCGA database are
293 truly specific to the cancer types. The Jaccard similarity coefficients of the transformed binary
294 vectors ranged from 0.003 to 0.150, which revealed minimal overlap between different cancer
295 types (**Fig. 4B** and **Table. S4**), affirming the specificity of the retrieved accessible regions to
296 their corresponding cancer types. We further examined the overlap of top-ranked TRs identified
297 by BIT across different cancer types. The top 10 TRs identified by BIT vary significantly among
298 cancer types, consistent with the Jaccard similarity coefficients of transformed binary vectors,
299 confirming that BIT can identify different TRs for different cancer types.

300 Thirdly, it is noted that several top-ranked TRs, such as NCOA3 and TET2 in breast cancer, lack
301 known motifs[2]. This absence means these TRs cannot be detected using motif-based methods,
302 despite solid experimental evidence supporting their significant roles in these cancers
303 respectively[58,59]. In contrast, BIT uses TR ChIP-seq data to rank these TRs to top positions
304 regardless of such concerns. This finding further highlights the BIT's advantage over traditional
305 motif-based methods, especially in cases where known motifs are ambiguous or absent.

306 Next, given the critical role of many TRs in cellular viability[7,60,61], we anticipated a correlation
307 between BIT identified top-ranked TRs and high-dependency TRs in corresponding cancer
308 types. CRISPR/Cas9 knockout screenings highlighted the dependency of TRs in various cell
309 lines. Our analysis revealed that by counting the number of high-dependency TRs in BIT
310 identified top 10 and top 50 TRs in each cancer type, about half of the BIT identified TRs are
311 highly correlated with cancer cell viability (**Fig. 4C**, 52/90 for top 10 and 231/450 for top 50, high
312 dependency TRs are determined as TRs with minimum Chronos score < -0.4). The dependency
313 scores of the top 50 TRs identified by BIT in prostate cancer cell lines show that TRs can have
314 different level of influence across different cell lines (**Fig. 4E**).



315

316 **Figure 4.** A. The top 10 TRs identified by BIT in each cancer type, some of the TRs have wider
317 confidence interval than other TRs, many TRs have been proved to be significant by available
318 publications. B. the upper triangle shows the Jaccard index of cancer-specific accessible
319 regions between any two cancer types, the bottom triangle shows the similarity between the BIT
320 identified top TRs between any two cancer types. C. The number of BIT identified top 10 and
321 top 50 TRs that are also high dependency TRs in each cancer type. D. The comparison of
322 enrichment of top identified TRs with high dependency TRs from CRISPR/Cas9 loss-of-function
323 analysis between BIT and other methods (HOMER, WhichTF, BART, i-cisTarget). E. the
324 dependency score of BIT identified top 50 TRs in different prostate cancer cell lines.

325 Lastly, we compared the enrichment of BIT-identified top-ranked TRs in high-dependency TRs
326 from CRISPR/Cas9 data with results from other state-of-art methods, including BART, ChIP-
327 Atlas, i-cisTarget, WhichTF, and HOMER. Using a one-sided Fisher exact test, the results show
328 BIT is the best or close to the best method in all cancer types (**Fig. 4D**). However, it is important
329 to remember that cellular viability is not an exclusive indicator of a TR's significance in cancer
330 cells. Therefore, multiple factors can contribute to the variability of the results. Still, BIT proves
331 its efficacy in identifying TRs from cancer-type specific accessible regions.

332 **BIT can identify cell type specific TRs**

333 The rapid advancement towards single-cell and spatial omics highlights the growing need for
334 tools capable of analyzing gene transcription within cell clusters generated by common single-
335 cell data processing methods. To acquire an accurate epigenomic landscape of chromatin
336 accessibility, the analysis of scATAC-seq data frequently categorizes individual cells into
337 clusters to mitigate the scarcity of fragments per cell. One major task for the downstream
338 analysis is to investigate the activity of TRs based on the derived cell cluster level pseudo-bulk
339 accessible regions. Therefore, we also examined the applicability of BIT in identifying critical
340 TRs using cell type-specific accessible regions derived from scATAC-seq data.

341 Two datasets, including 10K Peripheral blood mononuclear cells (PBMCs) and primary liver
342 cancer samples, were used to study BIT's performance. The cells in two datasets were
343 previously annotated based on the dimension reduction results and marker genes. The UMAP
344 projection of the datasets shows a clear separation of different cell types in each dataset (**Fig.
345 5A**). SnapATAC2 was applied to generate cell type marker accessible regions. Specifically, we
346 applied BIT to marker accessible regions derived from B cells in PBMCs dataset and malignant
347 cells in liver cancer datasets. The results were thoroughly validated by different approaches.

348 We examined the credibility of top 20 TRs identified by BIT in the two investigated cell types
349 (**Fig. 5B**). The credibility of each TR in the list was evaluated based on existing publications and
350 by examining whether the TR was designated as a significantly elevated gene in the Human
351 Protein Atlas for the respective cell type (**Table. S4**). Notably, BIT successfully captured several
352 TRs critical for B cell development. These include PAX5 and EBF1, which were previously
353 reported for their role in maintaining B cell identity. Additionally, POU2AF1 and POU2F2 can act
354 as pivotal regulators in B cell proliferation, differentiation, the germinal center transcriptional
355 program, and immune response. The gene activity plot also shows a high specificity of identified
356 TRs in the two cell types (**Fig. 5D**).

357 For liver cancer malignant cells, consistency is observed between the top TRs identified from
358 the scATAC-seq data and from the aggregated bulk ATAC-seq data as shown in the previous
359 section. Specifically, SOX13, HNF4A, and HLF are again ranked highly by BIT. It is noted that
360 the accessible regions used in the previous section were generated from bulk ATAC-seq data,
361 while those employed here are derived from scATAC-seq data specific to malignant cells. This
362 observed consistency strengthens the argument for BIT's robustness in identifying critical TRs.

363 We further assessed the top 20 TRs identified through motif enrichment analysis implemented in
364 SnapATAC2. Notably, as anticipated and mentioned in the introduction, the similarity of binding
365 motifs can hinder the differentiation between TRs from the same protein family. This is evident in
366 the top-ranked TRs identified by using motif methods implemented in the SnapATAC2, where B
367 cell results include the POU, IRF, and PAX families, and the liver cancer malignant cells exhibit
368 enrichment for the ZNF and FOX families (**Fig. 5C**). It is crucial to recognize that not all
369 members within these protein families hold equal importance in the biological processes of their
370 respective cell types. Their inclusion in the top rankings solely reflects the inability of motif
371 enrichment analysis to identify which specific TR within the family directly interacts with the motif
372 site in the accessible chromatin regions. Next, we performed Gene Ontology (GO) enrichment

373 analysis for the top 20 TRs identified by BIT and motif enrichment in B cells. The results
374 revealed a strong enrichment for biological processes relevant to B cells or mononuclear cells in
375 the BIT-identified TRs. Conversely, the motif enrichment-identified TRs displayed enrichment for
376 the miRNA metabolic process (**Fig. 5E**).

377 Next, for BIT identified TRs from liver malignant cells, we investigated the potential of these TR
378 being as prognostic biomarkers for patients with liver cancers. Kaplan-Meier survival plots
379 revealed a statistically significant correlation between the differential expression of multiple TRs
380 and patient survival time (in months). This suggests that the BIT identified TRs could be
381 promising candidates for prognostic biomarkers in liver cancer (**Fig. 5F**). In addition, we similarly
382 examined the CRISPR/Cas9 post-knockout viability of top-ranked TRs in liver cancer cell lines.
383 The results show that BIT can identify more high dependency TRs than the motif enrichment
384 methods implemented in SnapATAC2 (**Fig. 5G** and **Fig. 5H**).

385 Collectively, our findings demonstrate the feasibility of using BIT for downstream analysis of
386 cluster-level accessible regions in scATAC-seq data. Beyond feasibility, BIT offers valuable
387 insights into potential regulatory mechanisms. This makes it a compelling alternative to the
388 current, dominantly motif-based enrichment methods used in scATAC-seq analysis.

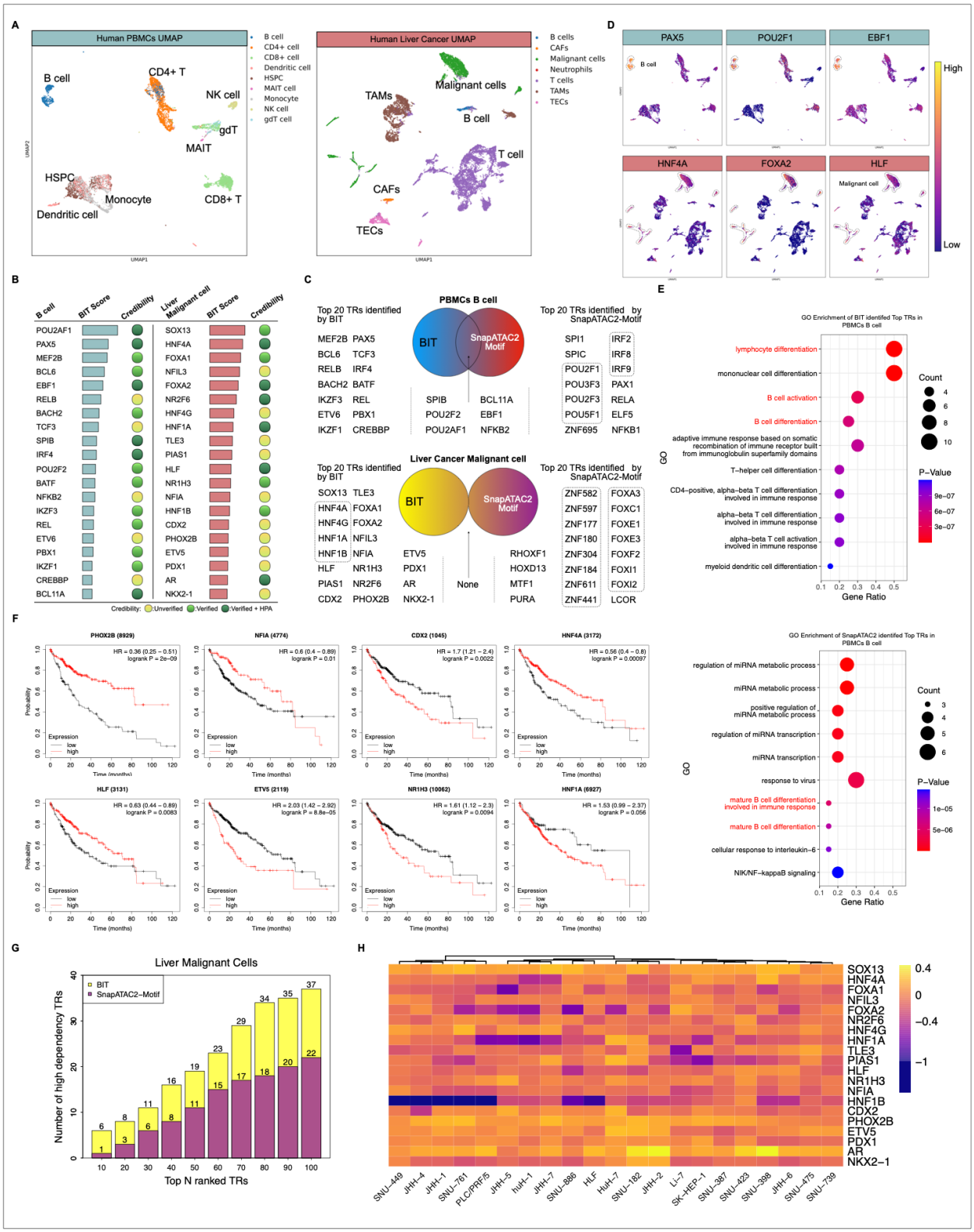


Figure 5. A. Sub-clustering UMAP projection of human Peripheral blood mononuclear cells and primary liver cancer cells. B. We manually verified top 20 TRs identified by BIT by using B cell and liver malignant cell type specific ATAC-seq peaks. C. Compared to TRs identified by BIT,

393 motif enrichment analysis cannot accurately differentiate TRs share similar binding motifs, such
394 as POU family in B cell and ZNF / FOX family in malignant cells. D. UMAP visualization of TR
395 activity. Top ranked cell type specific TRs identified by BIT show much higher gene activity in the
396 respective cell types. E. Gene ontology enrichment of top 20 TRs identified by BIT given
397 mononuclear cell or b cell related biological process, while the enrichment of top 20 TRs
398 identified by motif enrichment analysis shows miRNA metabolic process. F. Kaplan-Meier Plots
399 show the expression level of BIT identified TRs have significant correlation with liver cancer
400 patient survival rate, which suggests these TRs could be potential survival biomarkers.

401

402

Discussion

In this study we proposed a Bayesian hierarchical model that uses publicly available NGS data to identify transcriptional regulators with the input of user-provided epigenomic regions. Using simulation and application studies, we demonstrate the effectiveness of BIT and compare it with baseline or other existing methods. BIT has certain advantages compared to current available methods. It is the first rigorous Bayesian hierarchical model to identify TRs from user-provided epigenomic regions by leveraging accumulated TR ChIP-seq data. For example, BIT can give the confidence interval of the estimation based on the MCMC algorithm, which has never been explored before by other methods and can increase the Interpretability of the results.

We also observe several patterns that are worth to be noted. From the analysis based on differential accessible regions (DARs), our results show perturbation of TRs that have different functions can lead to different directionality of the change of chromatin accessibility. For example, ZBTB7A is well known to be a major repressor in HUDEP-2 cells, and the reason why BIT identifies it from loss accessible regions may have a direct relation to its functionality. Thus, distinguishing the directionality of the change to chromatin accessibility is important in identifying TRs from DARs.

The accuracy of BIT can be affected by multiple factors: (i) The quality of the data that generates user-provided epigenomic regions. (ii) The quality and availability of the TR ChIP-seq data. (iii) The consistency of the binding sites of TRs across different biological conditions. (iv) The dependency of user-provided epigenomic regions to the TRs. The regulation mechanism of gene transcription is complicated. Although BIT has been validated using multiple applications, many factors can affect the validation results. For example, removing certain TRs might not directly impact cell viability since other regulators can compensate for their function. The lack of data for certain TRs due to the limited available research resources also constitutes a challenge (e.g., lack of suitable antibodies). It also should be noted though BIT take the cellular

428 heterogeneity between TR ChIP-seq datasets into consideration, the current available
429 combination of TRs and cell lines may only represent a few of all possible combinations. The
430 heterogeneity between different cell types, cell states, or due to the response to the
431 environmental factors remains the biggest challenge so far for all TR prediction method.

432 In summary, BIT provides valuable insights into the regulation of gene transcription. With the
433 growing popularity of chromatin accessibility profiling techniques, especially the bulk and single-
434 cell ATAC-seq, BIT fits the emergency need for tools to properly handle omics data and provides
435 more rigorous identification from a Bayesian perspective by using the accumulated TR ChIP-
436 seq data that represent the most accurate in-vivo genome-wide binding patterns.

437

Conclusion

In high-throughput data analysis, TR binding profiles can provide critical information for identifying functional TRs from the user-provided epigenomic regions. However, currently available methods either primarily use genes as input or heavily rely on motif enrichment analysis to identify TRs from epigenomic regions, which suffer from various limitations. These considerations motivate us to propose BIT, a Bayesian hierarchical model to identify TRs from user-provided epigenomic regions using accumulated TR ChIP-seq data. We have proved the feasibility and effectiveness of BIT in simulation and multiple application studies. Our results demonstrate that BIT is a powerful tool for identifying TRs and provides a novel Bayesian perspective in high-throughput data analysis.

Methods

Data processing

In the TR perturbation derived DARs analysis, raw ATAC-seq sequence reads were first cleaned for adaptors using NGmerge (v.0.3). Read quality was checked by FastQC (v.0.11.9). Paired-end reads were mapped to the human genome (hg38) using bowtie2 (v.2.5.2) applying parameters “bowtie2 --very-sensitive -X 2000”. SAMtools (v.1.16.1) 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 (v.2.2.9.1). BAM files, index files, and peak files were used for DAR analysis using DiffBind (v.3.10.1). Reference TR ChIP-seq datasets were retrieved from GTRD, datasets with too low number peaks (<100) and individual peaks with fold enrichment <5 were removed. Cancer type specific accessible regions were retrieved from TCGA. 10K PBMCs and primary liver cancer scATAC-seq fragment files were retrieved from the GEO archive. The cell type marker accessible regions were generated from each dataset using SnapATAC2. When applicable, the cell type annotations were also retrieved and applied. Codes for the detailed data processing steps are available online.

Region mapping

BIT only requires the user-provided epigenomic regions as input, while TR ChIP-seq peaks serve as the reference. All binding sites (peaks) or epigenomic regions from a single dataset are mapped to nonoverlapping, pre-defined regions (or bins) of 1000 bps over the entire genome. Each bin is assigned either “0” or “1” according to whether the summit of any of the peaks 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 datasets (input and references) are transformed into binary vectors of the same length. We then compare the binary vector from the input with that from each of the TR ChIP-seq datasets, where we define three conditions for any given bin: (i) “matching” if (1,1); (2) “mismatching” if

(1,0) or (0,1); and (3) “non-informative” if (0,0). In all input vs. reference comparisons, the first two cases (“informative” bins) are much fewer than the third case (“non-informative” bins) and we drop the third case to mitigate the sparsity while not losing much useful information.

BIT model

BIT is the first model-based method that formally distinguishes TRs with multiple datasets from those with only a single dataset in the reference database, to account for both within-TR and between-TR heterogeneity and allow for information borrowing across different TRs and datasets via a hierarchical setup. We index TRs with multiple reference datasets by i, $i = 1,2,\ldots,\mathcal{M}$, each with $j = 1,2,\ldots,J_i$ datasets. TRs with a single reference dataset are indexed by $i' = 1,2,\ldots,\mathcal{M}^c$, all with $j \equiv 1$ dataset. Let x_{ij} $(x_{i'1})$ be the number of “matching” bins out of n_{ij} $(n_{i'1})$ “informative” bins by comparing the vectors between the user input and the jth reference dataset of the ith TR (or the only dataset of the i' th TR). The probability of a bin being “matching” given it is “informative” is 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 $\theta_{ij} \equiv logit(p_{ij})$ $(\theta_{i'1} \equiv logit(p_{i'1}))$ be the logit-transformed p_{ij} $(p_{i'1})$, which can be interpreted as the importance score at the dataset level. 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 and τ^2 represent the global mean and variability of TR-level importance scores. In this way, for each TR with multiple datasets, we model its importance scores by a normal distribution with distinctive mean θ_i and variance σ_i^2, reflecting its overall importance and within-TR heterogeneity, respectively; all TRs with only one dataset share a common variance σ_0^2 for information pooling. Unlike many existing algorithms[17,18,21], 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.

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

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

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

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

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

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

Where:

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

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

logit transformed importance θ_{ij} $(\theta_{i'1})$

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

And TR level score θ_i $(\theta_{i'})$

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

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. 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., $a = b = 0.01$).

$$p(\mu) = \frac{1}{U_\mu - L_\mu}, L_\mu \leq \mu \leq U_\mu$$

$$p(\tau^2) = (\tau^2)^{-a-1}\, exp\left(-\frac{b}{\tau^2}\right)$$

$$p(\sigma_0^2) = (\sigma_0^2)^{-a_0-1}\, exp\left(-\frac{b_0}{\sigma_0^2}\right)$$

$$p(\sigma_i^2) = (\sigma_i^2)^{-a_i-1} exp\,(-\frac{b_i}{\sigma_i^2})$$

Pólya-gamma data augmentation and Gibbs sampler

The BIT framework uses Gibbs sampler that is a specific Monte Carlo Markov chain algorithm to generate posterior samples from complex joint probability distribution. However, to compute the MCMC, the full conditionals should be obtained in advance. As the joint posterior distribution can be written as $p(\Theta|X) \propto p(\Theta, X)$, the full conditional posterior distribution of each parameter can be written as $p(\theta|X, \Theta \setminus \theta)$, where $\Theta \setminus \theta$ denote the remaining parameters after removing the parameter θ from the parameter set Θ. For μ, τ^2, σ_0^2, and σ_i^2, these parameters full conditional distributions are easily to be computed as

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

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

$$p(\sigma_0^2|X,\Theta \setminus \sigma_0^2) \propto IG\left(a_0 + \frac{M^c}{2}, b_0 + \frac{(\theta_{i'1} - \theta_{i'})^2}{2}\right)$$

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

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

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

However, it is difficult to sample from the full conditional of $p\left(\theta_{ij}|X,\Theta \setminus \theta_{ij}\right)$ and $p(\theta_{i'1}|X,\Theta \setminus \theta_{i'1})$ as these are intractable. Therefore, we instead implement a data augmentation strategy that is based on Pólya-Gamma (PG) latent variables[62]. We further define auxiliary variables λ_{ij} and $\lambda_{i'1}$, which follow the Pólya-Gamma distribution.

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

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

The full conditionals based on the defined auxiliary parameters can be easily written as the form of known distribution.

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

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 full conditional distribution of $\theta_{i'1}$. Given that all posterior 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.

Simulation naïve baseline method

To compare the performance of BIT in the estimation of parameters, we used naïve method as baseline.

Where hyperparameter μ is:

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

Hyperparameter τ^2 is:

570 $$\hat{\tau}^2 = \frac{\sum_{i=1}^{M}\sum_{j=1}^{J_i}\left(logit\left(\frac{x_{ij}}{n_{ij}}\right)-\hat{\mu}\right)^2 + \sum_{i'=1}^{M^c}\left(logit\left(\frac{x_{i'1}}{n_{i'1}}\right)-\hat{\mu}\right)^2}{\sum_{i=1}^{M} J_i + M^c - 1}$$

571 Parameter σ_0^2 is:

572 $$\sigma_0^2 = \frac{1}{M^c - 1}\sum_{i'=1}^{M^c}\left(logit\left(\frac{x_{i'1}}{n_{i'1}}\right)-\hat{\mu}\right)^2 - \hat{\tau}^2$$

573 And θ_i and $\theta_{i'}$ are:

574 $$\hat{\theta}_i = \frac{\sum_{j=1}^{J_i} logit\left(\frac{x_{ij}}{n_{ij}}\right)}{J_i},\ \hat{\theta}_{i'} = logit\left(\frac{x_{i'1}}{n_{i'1}}\right)$$

575 Though the probability is extremely low, in the case if any simulated data point has $\frac{x_{ij}}{n_{ij}} = 0$ or

576 $\frac{x_{i'1}}{n_{i'1}} = 0$, to allow the estimation to be done properly for naïve method, we add a tiny positive

577 value to these extreme cases.

578 **Cross-validation with CRISPR-Cas9 screening data**

579 Data of TR knock out effects on viability of cell lines were retrieved from DepMap. We filtered

580 out TRs that are not included in the BIT database or the DepMap data. Cell lines under the

581 same lineage are grouped to find TRs with a Chronos score lower than -0.4, which are classified

582 as high dependency TRs. 2-by-2 contingency tables were created based on whether top ranked

583 TRs identified by each method are also high dependency TRs. Finally, one-sided Fisher exact

584 tests were applied to each table to compare the capability of methods to identify TRs that are

585 also high dependency TRs in each cancer type.

586

587 **Data Availability**

588 TR perturbation experiments data were retrieved from GEO (GSE153237[36], GSE173416[37], and
589 GSE114102[38]). scATAC-seq data are retrieved from 10XGenomics and GEOXX, The reference
590 ChIP-seq datasets were retrieved from GTRD, and transformed binary vectors are available on
591 SMU Box. Cancer-type specific accessible regions were retrieved from TCGA. CRISPR-Cas9
592 knockout screening data were retrieved from DepMap. BIT software is available on GitHub, we
593 also have a reference manual available on readthedocs.

594

Funding

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

(R01CA258584)(to X.W.)

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

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

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

Reference

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

10 Rottner, A. K. *et al.* A genome-wide CRISPR screen identifies CALCOCO2 as a regulator of beta cell function influencing type 2 diabetes risk. *Nat Genet* **55**, 54-65 (2023). https://doi.org/10.1038/s41588-022-01261-2

11 Balakrishnan, S., Dhavamani, S. & Prahalathan, C. beta-Cell specific transcription factors in the context of diabetes mellitus and beta-cell regeneration. *Mech Dev* **163**, 103634 (2020). https://doi.org/10.1016/j.mod.2020.103634

12 Hong, J. H. & Zhang, H. G. Transcription Factors Involved in the Development and Prognosis of Cardiac Remodeling. *Front Pharmacol* **13**, 828549 (2022). https://doi.org/10.3389/fphar.2022.828549

13 Bauer, A. J. & Martin, K. A. Coordinating Regulation of Gene Expression in Cardiovascular Disease: Interactions between Chromatin Modifiers and Transcription Factors. *Front Cardiovasc Med* **4**, 19 (2017). https://doi.org/10.3389/fcvm.2017.00019

14 Bhagwat, A. S. & Vakoc, C. R. Targeting Transcription Factors in Cancer. *Trends Cancer* **1**, 53-65 (2015). https://doi.org/10.1016/j.trecan.2015.07.001

15 Bushweller, J. H. Targeting transcription factors in cancer - from undruggable to reality. *Nat Rev Cancer* **19**, 611-624 (2019). https://doi.org/10.1038/s41568-019-0196-7

16 Darnell, J. E., Jr. Transcription factors as targets for cancer therapy. *Nat Rev Cancer* **2**, 740-749 (2002). https://doi.org/10.1038/nrc906

17 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

18 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

19 Keenan, A. B. *et al.* ChEA3: transcription factor enrichment analysis by orthogonal omics integration. *Nucleic Acids Res* **47**, W212-W224 (2019). https://doi.org/10.1093/nar/gkz446

20 Puente-Santamaria, L., Wasserman, W. W. & Del Peso, L. TFEA.ChIP: a tool kit for transcription factor binding site enrichment analysis capitalizing on ChIP-seq datasets. *Bioinformatics* **35**, 5339-5340 (2019). https://doi.org/10.1093/bioinformatics/btz573

21 Roopra, A. MAGIC: A tool for predicting transcription factors and cofactors driving gene sets using ENCODE data. *PLoS Comput Biol* **16**, e1007800 (2020). https://doi.org/10.1371/journal.pcbi.1007800

22 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

23 Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-589 (2010). https://doi.org/10.1016/j.molcel.2010.05.004

24 Tanigawa, Y., Dyer, E. S. & Bejerano, G. WhichTF is functionally important in your open chromatin data? *PLoS Comput Biol* **18**, e1010378 (2022). https://doi.org/10.1371/journal.pcbi.1010378

25 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

26 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

27 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

28 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

29 Lu, Z., Xiao, X., Zheng, Q., Wang, X. & Xu, L. Assessing NGS-based computational methods for predicting transcriptional regulators with query gene sets. *bioRxiv*, 2024.2002.2001.578316 (2024). https://doi.org/10.1101/2024.02.01.578316

30 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

31 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

32 Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 21931-21936 (2010). https://doi.org/10.1073/pnas.1016071107

33 Yan, F., Powell, D. R., Curtis, D. J. & Wong, N. C. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biology* **21**, 22 (2020). https://doi.org/10.1186/s13059-020-1929-3

34 Wang, S. *et al.* Modeling cis-regulation with a compendium of genome-wide histone H3K27ac profiles. *Genome Research* **26**, 1417-1429 (2016). https://doi.org/10.1101/gr.201574.115

35 Kharchenko, P. V., Tolstorukov, M. Y. & Park, P. J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature Biotechnology* **26**, 1351-1359 (2008). https://doi.org/10.1038/nbt.1508

36 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

37 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

38 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

39 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

40 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

41 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

42 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

43 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

44 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

45 Leonard, M., Brice, M., Engel, J. D. & Papayannopoulou, T. Dynamics of GATA transcription factor expression during erythroid differentiation. *Blood* **82**, 1071-1079 (1993).

46 Cheng, Y. *et al.* Erythroid GATA1 function revealed by genome-wide analysis of transcription factor occupancy, histone modifications, and mRNA expression. *Genome Res* **19**, 2172-2184 (2009). https://doi.org/10.1101/gr.098921.109

47 Maeda, T. *et al.* LRF is an essential downstream target of GATA1 in erythroid development and regulates BIM-dependent apoptosis. *Dev Cell* **17**, 527-540 (2009). https://doi.org/10.1016/j.devcel.2009.09.005

48 Siatecka, M. & Bieker, J. J. The multifunctional role of EKLF/KLF1 during erythropoiesis. *Blood* **118**, 2044-2054 (2011). https://doi.org/10.1182/blood-2011-03-331371

49 Wang, H. *et al.* Pdx1 level defines pancreatic gene expression pattern and cell lineage differentiation. *J Biol Chem* **276**, 25279-25286 (2001). https://doi.org/10.1074/jbc.M101233200

50 Watt, A. J., Zhao, R., Li, J. & Duncan, S. A. Development of the mammalian liver and ventral pancreas is dependent on GATA4. *BMC Dev Biol* **7**, 37 (2007). https://doi.org/10.1186/1471-213x-7-37

51 Fujimoto, K. & Polonsky, K. S. Pdx1 and other factors that regulate pancreatic beta-cell survival. *Diabetes Obes Metab* **11 Suppl 4**, 30-37 (2009). https://doi.org/10.1111/j.1463-1326.2009.01121.x

52 Allen, H. L. *et al.* GATA6 haploinsufficiency causes pancreatic agenesis in humans. *Nat Genet* **44**, 20-22 (2011). https://doi.org/10.1038/ng.1035

53 Iwafuchi-Doi, M. & Zaret, K. S. Pioneer transcription factors in cell reprogramming. *Genes Dev* **28**, 2679-2692 (2014). https://doi.org/10.1101/gad.253443.114

54 Norton, L. J. *et al.* KLF1 directly activates expression of the novel fetal globin repressor ZBTB7A/LRF in erythroid cells. *Blood Adv* **1**, 685-692 (2017). https://doi.org/10.1182/bloodadvances.2016002303

55 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

56 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

57 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

58 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

59 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

60 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

61 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

62 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