Epiregulon: Inference of single-cell transcription factor activity to dissect mechanisms of lineage plasticity and drug response

Tomasz Włodarczyk[1], Aaron Lun[1], Diana Wu[2], Shreya Menon[3,4,5], Shushan Toneyan[1], Kerstin Seidel[2], Liang Wang[2], Jenille Tan[2], Shang-Yang Chen[1], Timothy Keyes[1], Aleksander Chlebowski[1], Yu Guo[1], Ciara Metcalfe[2], Marc Hafner[1,2], Christian W. Siebel[2], M. Ryan Corces[3,4,5], Robert Yauch[2]*, Shiqi Xie[2]*, Xiaosai Yao[1,2]*

[1]Department of Computational Sciences, Genentech Inc, South San Francisco, CA, USA
[2]Department of Discovery Oncology, Genentech Inc, South San Francisco, CA, USA
[3]Gladstone Institute of Neurological Disease, Gladstone Institutes, San Francisco CA, USA
[4]Gladstone Institute of Data Science and Biotechnology, Gladstone Institutes, San Francisco CA, USA
[5]Department of Neurology, University of California San Francisco, San Francisco CA, USA

*correspondence

yauch.bob@gene.com; xie.shiqi@gene.com; yao.xiaosai@gene.com

Abstract

Transcription factors (TFs) and transcriptional coregulators represent an emerging and exciting class of targets. By quantifying target gene modulation, gene regulatory networks (GRNs) delineate disease biology and evaluate pharmacological agents targeting these regulators. However, none of the existing methods are explicitly designed to measure the effects of perturbations in which TF expression is decoupled from its activity. We present Epiregulon, a method that constructs GRNs from single-cell ATAC-seq and RNA-seq data for accurate prediction of TF activity. Our weight estimation, based on co-occurrence of TF expression and chromatin accessibility, avoids erroneous inflation of TF activity as seen with TF expression only approaches. Furthermore, our utilization of ChIP-seq data expands inference to transcriptional coregulators lacking defined motifs. Our extensive network of regulators facilitates identification of cell-state specific interaction partners. Using Epiregulon, we uncover divergent cell fate transitions of prostate cancer cells driven by NKX2-1 and GATA6 overexpression. We accurately predicted the effects of AR inhibition across various drug modalities. Finally, Epiregulon was able to infer the context-dependent activity of a chromatin remodeler lacking a defined motif sequence and recapitulate the unique etiologies of prostate cancer. By mapping out the network of key regulators across a multitude of perturbations, Epiregulon can accelerate the discovery of new therapeutics targeting transcription factors.

Introduction

Transcription factors and transcriptional coregulators shape cell fates and lineage commitment, and their dysregulation drives congenital diseases and tumor growth. Transcription factors bind to specific DNA sequences whereas coregulators interact with transcription factors in a context-specific manner and lack defined motifs. These transcriptional modulators represent an important class of therapeutic targets in oncology and beyond. Therapeutics targeting transcriptional regulators inhibit their activity by either blocking the ligand binding domains[1], degrading the protein[2,3] or disrupting protein-protein interactions[4]. Delineating the functions of transcriptional regulators can accelerate our understanding of disease biology and drug discovery.

Gene regulatory networks (GRNs) model the underlying circuitry of gene regulation and have been applied to understand lineage commitment and plasticity and drug response[5]. Early methods to construct gene regulatory networks relied exclusively on gene expression data and sought to identify association between expression of target genes and transcription factors[6,7]. Now, single-cell multiomics technologies provide chromatin accessibility information in addition to gene expression and recent GRN inference methods leverage both modalities to improve performance (Table 1). Most of these methods, including CellOracle[8], FigR[9], Pando[8] and GRaNIE[10], rely on linear relationships between transcription factors and their target genes to model gene regulation. In addition, SCENIC+ utilizes random forest regression to model the regulatory relationships between TF, regulatory elements and target genes[11].

Several challenges still impede the broader utilization of these GRN inference methods in basic biology and drug discovery. First, none of these methods are specifically designed to predict changes in which the TF activity is decoupled from its gene expression. These include drug perturbations in which inhibitors disrupt protein complex formation or localization, and post-translational modifications that can impact TF activity and also CRISPR genome editing in which genes can be silenced by insertions or deletions at the DNA level but without causing appreciable changes in gene expression. Second, the use of motif sequences in GRN construction precludes prediction of transcriptional coregulators, an important class of regulators without defined motifs. Lastly, current studies rely on ChIP-seq and HiC data to benchmark GRNs but TF occupancy and chromatin conformation do not directly measure the true regulatory activity of TFs. Therefore, erroneous estimation of TF activity misled by gene expression, sole reliance on TF motifs and lack of appropriate benchmarking datasets are three limitations in current GRN inference methods.

Here we present Epiregulon, a method that constructs a GRN from single-cell multiomics data and TF occupancy data to infer activity of transcriptional regulators. Unlike most other GRN tools, Epiregulon utilizes accompanying ChIP-seq data, thus allowing inference of both transcription factors and transcriptional coregulators lacking defined motifs. Epiregulon utilizes weight estimation methods that account for co-occurrence of both TF gene expression and chromatin accessibility to handle decoupling of TF expression and activity. Functional interpretation of GRNs through Jaccard similarity of target genes and pathway annotations uncovers drivers and interaction partners of lineage transition and drug response in a cell-state specific and context-aware manner. We performed simulations and generated ground truth data to robustly benchmark

its performance in a multitude of datasets including lineage development and drug treatment. Epiregulon is implemented as a R package and interfaces with ArchR[12] or any other preprocessing tools.

Table 1: Comparison of different GRN inference methods

Feature/package	**Epiregulon**	**GRaNIE**	**SCENIC+**	**FigR**	**Pando**	**cellOracle**
Prior information to predict TF occupancy	Granges list of ChIP-seq peaks or motif coordinates	BED file with motif coordinates	Pre-computed database with motif scores	Motif database compiled from cisBP	Motif database compiled from cisBP and JASPAR	Motif data from cisBC v. 2.0
TF-target gene model	Wilcoxon-based co-occurrence, correlation and mutual information	Expression correlation	Gradient Boosting Machine regression	TF-DORC correlation and TF enrichment	Regularized GLM, Bayesian regression models, gradient boosting regression, bagging and ridge Bayesian models	Regularized machine - learning linear model
Pruning method	TF-RE-TG co-occurrence (chi-sq)	Based on permuted data	Filtering TF-TG pairs using GSEApy	No pruning	Linear regression	Incorporated in the main model
Addressing sparsity	pseudobulks	no	imputation	KNN smoothing	no	KNN imputation, pseudocells
Data pre-processing compatibility	ArchR or SingleCellExperiment	Peak and gene expression matrices	scanpy, pycisTopic, pycisTarget	Summarized Experiment (peak), gene expression matrix	Signac	cicero, scanpy

Results

Epiregulon constructs GRNs to infer TF activity at the single cell level

Epiregulon aims to map out target genes of transcriptional regulators and track changes in their regulatory potential in each cell state, thereby identifying master regulators of cell fate changes. It assumes that a transcription factor or a transcriptional coregulator (collectively referred to as TF for simplicity) binds to regulatory elements (REs) and modulates the expression of nearby target genes (TGs). Epiregulon assigns regulatory elements (REs) to their putative target genes based on correlation between chromatin accessibility and gene expression (Figure 1). Each regulatory element is then interrogated for TF occupancy based on user-provided or/and a compilation of public TF ChIP-seq data from ChIP-Atlas and ENCODE. The use of ChIP-seq data enables us to extend activity inference to chromatin modifiers and co-factors which do not have well-defined binding motifs. Furthermore, accumulating evidence suggests that TF occupancy is determined not just by motif sequences but also by other factors including interactions with other TFs and co-factors, local DNA structure and genomic features[13]. A substantial proportion of TF binding sites do not harbor motifs[14]. Therefore, ChIP-seq data is likely to yield a more relevant set of target genes than motif annotations are.

We combine the RE-TG links with TF occupancy at REs to establish an initial unweighted tripartite graph with TFs, REs and TGs being the nodes and their connections being edges. An optional pruning step of the graph removes potentially spurious regulatory relationships and TFs with undetectable gene expression. We then estimate the weights of the TF-RE-TG edges using the Wilcoxon test statistic, computed by comparing the ranks of the target gene expression between the group of cells that express the corresponding TF and have chromatin accessibility at that particular RE versus the remaining cells. This graph becomes the final gene regulatory network returned by Epiregulon. Finally, the activity of a TF is defined as the weighted sum of its target gene expression normalized by the number of target genes. The exact algorithm is described in the Methods; see also Supplementary Note Section 1 for comparison of parameter choices.

It is of great interest to understand the functions of a TF through its differential activity between cell states or treatment conditions. Differential TF activity can result from at least two different scenarios; 1) TFs can be turned on or off by changes in gene expression, or 2) they can redistribute to different genomic regions due to mutations or hijacking by other factors. The first scenario can be adequately captured by differences in total TF activity whereas the second scenario requires the consideration of regulon membership. In addition to quantifying differential total TF activity between cell states, Epiregulon employs a second method for differential network analysis. We use the Wilcoxon method to derive cluster-specific weights, subtract the edge weights between two networks and compute the degree centrality of the resulting edge-subtracted network. See Supplementary Note Section 2 for simulation of differential networks.

Finally, it is desirable to interpret the functions of a TF through the pathway enrichment of its regulons and through its interaction partners. Epiregulon provides pathway enrichment of the regulons and further utilizes Jaccard similarity to estimate the overlap between regulons to identify

potential interaction partners. These functional annotations of GRNs serve to provide biological interpretability.

Epiregulon identifies cell fate changes

To evaluate Epiregulon's performance, we generated our own data using Reprogram-Seq[15] to model lineage transition of prostate adenocarcinomas by overexpressing defined transcription factors. We separately introduced lentivirus encoding NKX2-1 and GATA6 into a prostate adenocarcinoma cell line LNCaP and tracked changes in transcriptomic and epigenomic profiles at the single cell level. NKX2-1 is a known driver of AR-independence and neuroendocrine transition in prostate cancer[16,17] whereas GATA6 has unknown function in the prostate. These factors are not expressed in parental LNCaP and were used for subsequent benchmarking analysis. Reprogram-Seq data could serve as the ground truth to benchmark the inference of TF activity because we expect to see increased TF activity in cells with exogenous expression of NKX2-1 and GATA6.

Expression of these transcription factors altered the cell states of LNCaP, with cluster 1 enriched in cells expressing GATA6 and cluster 3 enriched in cells expressing NKX2-1 (Figure 2a). Gene expression of these factors could not accurately map the identity of TFs being introduced, largely because transcription factors are lowly expressed and 3' bias of scRNAseq makes detection of exogenous transcripts difficult (Figure 2b). In contrast, TF activity computed from Epiregulon recapitulated the correct cell identity despite the low expression of these TFs (Figure 2b).

We performed a systematic benchmarking exercise to evaluate the performance of Epiregulon against other multiomic GRN inference methods. Compared to FigR, Pando, cellOracle, GRaNIE and SCENIC+, Epiregulon was the only method that was able to predict NKX2-1 activity (Figure 2c). This is because the use of ChIP-seq data allowed Epiregulon to predict the largest collection of regulators (Figure 2d). We evaluated the GRN inference methods on whether TF activity is able to classify cells expressing the matching TF vs the remaining cells. For GATA6, Epiregulon had the second highest area under the receiving operating characteristic (AUROC) value (Figure 2c). We next examined Epiregulon's performance across its different weight methods. Because TF activity was driven by increased TF expression in reprogram-seq, weight estimation based on the correlation between TF-TG expression was the best performing method (Figure S1a). This implies that Epiregulon gives consistent predictions as other top-performing GRN inference methods and offers improved sensitivity in predicting activity of those factors with low gene expression.

We further investigated the accuracy of the target genes inferred. We showed that both unpruned and pruned GATA6 and NKX2-1 targets inferred by Epiregulon ranked highly by differential expression (Figure S1b). Furthermore, cluster-specific GRN showed that NKX2-1 targets were exclusively expressed in cluster 3 while GATA6 targets were exclusively expressed in cluster 1 (Figure S1c). As another independent validation, we engineered a dox-inducible cell line expressing NKX2-1 or eGFP as a control. The log changes of our inferred NXK2-1 targets upon dox treatment correlated with the NKX2-1 target gene weights inferred by Epiregulon (Figure 2e, S1d), suggesting that the GRN weights inferred by Epiregulon are informative of regulatory strength.

Global analysis of TF activity identified other TFs whose activities were concomitantly altered by introduction of these reprogramming factors. Jaccard similarity found FOSL2 (a paralog of FOSL1), YAP1 and TEAD1 to be amongst the top interaction partners of GATA6 (Figure 2f, S1e). These factors are known to reprogram prostate adenocarcinomas towards a stem cell-like (SCL) state[18]. This led us to examine whether GATA6 itself could also reprogram LNCaP cells towards the SCL subtype. We examined regions that are uniquely accessible in the SCL tumors[18]. Strikingly, GATA6 overexpression led to increased chromatin accessibility at these SCL-specific regions and GATA6 activity was highly correlated with the combined chromatin accessibility score (Figure 2g, 2h). Lineage trajectory resulting from GATA6 expression differed from the neuroendocrine cell state induced by NKX2-1 (Figure 2g, 2h, S1f). Taken together, Epiregulon uncovered a previously unknown causal relationship between GATA6 expression and lineage transition to SCL subtype of prostate cancer.

Epiregulon predicts responses towards TF targeting drugs

In the previous example, Epiregulon was able to predict changes in TF activity when they are largely driven by changes in their expression. A more challenging task is to predict changes when TF activity is decoupled from its gene expression. We generated a second dataset to evaluate changes in AR activity following drug treatment. In total, 4 prostate cancer cell lines (2 AR-dependent and 2 AR-independent) were treated with 3 AR-targeting agents (Figure 3a). The first agent is the clinically approved AR antagonist, enzalutamide, which interferes with AR protein function by blocking its ligand binding domain[1]. The second agent is ARV-110, a degrader of AR protein which acts by bringing an E3 ubiquitin ligase in close proximity to an AR protein[3]. We also synthesized a third agent, SMARCA2.4_1, a degrader of the SWI/SNF chromatin remodeling complex (Figure S2a-b); inhibition of the same protein complex is known to decrease AR activity by evicting it from the chromatin[19]. These pharmacological agents do not directly or consistently suppress the transcription of AR target genes, thus serving as an example of the decoupling of the TF expression from its activity.

Epiregulon predicted decreased AR activity following enzalutamide and ARV110 treatment in the known sensitive cell lines, LNCaP and VCaP and minimal changes in AR activity in the AR-independent cell lines, DU145 and H660 (Figure 3b). Other GRN inference methods failed to predict decreased AR activity in either or both cell lines, with CellOracle even predicting an increased AR activity in VCaP (Figure 3c). Gene expression of AR was not informative of the effect on AR activity because enzalutamide increased AR gene expression in VCaP cells despite its antagonistic function (Figure 3d). Like the Reprogram-Seq data, Epiregulon recovered the greatest number of regulators (Figure 3e). By estimating weights based on both TF expression and RE chromatin accessibility, Epiregulon outperformed methods utilizing TF expression alone, and robustly predicted decreased AR activity despite discordant AR gene expression (Figure 3c, Figure S2c).

We next asked whether ChIP-seq data generated from merging ChIP-atlas and ENCODE (thereafter referred as merged ChIP-seq data) could yield comparable performance as cell line-

matched ChIP-seq data. TF activity inferred from AR and FOXA1 ChIP-seq data of LNCaP and VCaP cells correlated well with that produced by merged ChIP-seq (Figure S2c, S2d). Taken together, GRN constructed using merged ChIP-seq data could accurately recapitulate TF activity.

GRN constructed from ChIP-seq data is a good predictor of TF activity

Epiregulon uses ChIP-seq data as the default method to predict TF occupancy but can also incorporate motif annotations. Given that the majority of the other GRN inference methods start with motif sequences, we would like to systematically compare the accuracy of TF activity prediction starting from ChIP-seq data or motif annotations.

Using the Reprogram-Seq dataset, we showed that ChIP-seq (obtained by merging public ChIP-seq datasets from ENCODE and ChIP-Atlas) is superior to motif annotations in predicting TF activity for both GATA6 and NKX2-1 (Figure 4a, 4b). We further asked whether selecting a set of high-confidence motifs can improve TF activity prediction. We leveraged deep learning sequence-to-function models or deep neural networks (DNNs) to prioritize motif patterns of importance within the regulatory elements. DNNs are capable of learning complex patterns in the input data associated with accessible regions to make predictions of ATAC-seq coverage[20-22]. Here in addition to comparing ChIP-seq and motif annotations, we also applied thresholds to motif importance scores assigned by two independent sequence deep learning models, Basenji and ChromBPNet (see Methods). However, thresholded motifs did not improve the performance of TF activity; in fact, overly stringent thresholding resulted in a reduced number of target genes and worse performance (Figure 4a, 4b, Figure S3a, S3b). The weights of the target genes contributed to the accuracy of TF prediction since shuffling of weights reduced AUROCs (Figure S3c, S3d).

Next, we leveraged the prostate dataset to evaluate the performance of GRN. We found that in contrast to the Reprogram-Seq data, AR motifs outperformed merged ChIP-seq in terms of differentiating drug treated vs DMSO treated cells (Figure 4c). We could however improve the accuracy of ChIP-seq by utilizing cell line-matched ChIP-seq. In fact, combining motif annotation with cell line-matched ChIP-seq binding was the best predictor of AR activity (Figure 4c). DNN-based models did not further improve the performance of motif annotation (Figure 4c). Taken together, we showed that target genes inferred by ChIP-seq data outperforms motif annotations in the estimation of TF activity.

Epiregulon predicts response of drug targeting transcriptional coregulator

Transcriptional coregulators cooperate with transcription factors to regulate gene expression and are attractive targets in oncology. The challenge with predicting responses of transcriptional coregulators is that these factors do not contain defined motifs as they can interact with different transcription factors depending on the cellular context. As such, GRN inference methods which rely on TF motifs cannot predict the activity of these factors. SMARCA2 and SMARCA4, two mutually exclusive paralog proteins encoding the ATPase subunit of the SWI/SNF chromatin remodeler are examples of transcriptional coregulators. We evaluated the performance of Epiregulon based on its ability to predict decreased SMARCA2/4 activity upon treatment with a degrader of SMARCA2/4 protein. We utilized the same AR dataset as in Figure 3 but focused on the third agent, SMARCA2.4_1. We performed most of our analysis on SMARCA4 since it is the

263 predominant paralog that affects the proliferation of prostate cancer cell lines[23]. Pharmacological
264 degradation of SMARCA4 allows benchmarking of activity inference for a highly context-
265 dependent transcriptional coregulator lacking a defined motif.
266
267 To circumvent the lack of motifs, Epiregulon utilizes public SMARCA4 ChIP-seq to infer the most
268 likely target genes in a cell-line aware manner. Epiregulon predicted decreased SMARCA4
269 activity across all 4 prostate cancer cell lines (Figure 5a). None of the other GRN inference
270 methods we benchmarked were able to predict the activity of SMARCA4. Within Epiregulon,
271 weight estimation methods that take into account both TF gene expression and RE chromatin
272 accessibility outperformed methods that only take into account TF expression (Figure S4a).
273 Furthermore, SMARCA4 activities calculated from merged ChIP-seq and cell line-matched ChIP-
274 seq were consistent (Figure S4b). The SMARCA4 regulons recapitulated the unique etiologies of
275 prostate cancer, with AR-dependent cell lines LNCaP and VCaP enriched for androgen response,
276 stem-cell like cell line DU145 enriched for activation of YAP activity[18] and the neuroendocrine cell
277 line NCI-H660 enriched for the loss of TP53, activation of EZH2[24,25] and MAPK signaling[26] (Figure
278 5b).
279
280 Epiregulon can inform mechanisms of action of the SMARCA4 degrader by discovering TFs co-
281 perturbed with SMARCA4 (Figure 5c). We identified differential TFs by subtracting DMSO treated
282 from SMARCA4 degrader treated networks. TFs were ranked by Jaccard similarity with
283 SMARCA4 after background correction using edge permuted networks. Common to all 4 cell
284 lines, highly ranked TFs included components of general transcriptional machinery including
285 BRD4 and EP300 and also regulators of chromatin structure including components of the cohesin
286 complex (RAD21, SMC3, SMC1A) and chromosome looping (YY1, CTCF and MED1). These
287 proteins have been previously implicated to associate with SMARCA4 (Table 2). Unique to the
288 individual cell lines, SMARCA4 interacts with AR in the AR-dependent cell lines (LNCaP and
289 VCaP). In the neuroendocrine cell line NCI-H660, SMARCA4 interacts with TCF12, a factor
290 known to be important for neuronal development[27] and EZH2, a factor known to mediate
291 neuroendocrine transition in prostate cancer[25]. Taken together, Epiregulon expands the inference
292 of activity to include transcriptional coregulators lacking defined motifs and is able to capture their
293 context specificity and infer the most relevant biological processes.
294
295 Table 2: Associated partners of SMARCA4 identified by differential network analysis

Function	**Gene**	**References**
Transcription	BRD4	25,28, 29, 30, 31
Transcription	EP300	32, 33, 34
Chromosome structure	Cohesin (RAD21, SMC1A, SMC3)	35, 36
Chromatin looping	YY1 and CTCF	37, 38

296
297
298 **Epiregulon identifies lineage factors and predicts effects of perturbations**
299 Finally, we ask whether regulons generated from single-cell multiomic data can be applied to
300 scRNA-seq dataset lacking chromatin information. We first used Epiregulon to compute the
301 activity of lineage factors in a scMultiomic dataset of human peripheral blood mononuclear cells
302 (PBMCs)[39] (Figure 6a). Epiregulon recapitulated the lineage-specific activity of known lineage
303 factors including GATA3 in T and NK cells, EBF1 and PAX5 in B cells and SPI1 in myeloid cells
304 (Figure 6b), and gene set enrichment of their target genes is consistent with their known biological
305 functions (Figure S5a). We used the Wilcoxon method to obtain cell-type specific weights for all
306 TF-TG pairs. As an example to illustrate the specificity of the GRN, the B cell lineage factor EBF1
307 had its target genes largely confined to preB cells but not to CD4 memory T cells (Figure 6c). We
308 further utilized Epiregulon's similarity function to identify known interacting partners (Figure 6d).
309 The most similar TF to EBF1 was PAX5; EBF1 and PAX5 regulate B cell development in a highly
310 coordinated manner[40]. We further compared the global network topology of monocytes and CD4
311 memory T cells; degree centrality of the differential graph revealed known lineage factors of
312 monocytes (SPI1, CEBPA, NR4A1)[41,42] and CD4 T cells (TBX21, GATA3, LEF1, FOXO1)[43,44],
313 suggesting that Epiregulon is able to prioritize lineage factors from differential network analysis.
314
315 Next, we applied the GRN of the above PBMC dataset to a new scRNAseq dataset containing
316 K562, a lymphoblast cell line, engineered to express CRISPRa and guides targeting defined
317 factors[45]. The barcode information in each cell encodes the guide RNAs corresponding to the TFs
318 they activate and serves as the ground truth. GRN from Epiregulon better predicted the effect of
319 CRISPRa perturbations than gene expression, lineage-agnostic regulons (Dorothea[46] and
320 CollecTRI[47]) and lineage-specific regulons (Humanbase[48]) did, as reflected in the AUROC values
321 (Figure 6f, Figure S5b). Epiregulon also captured the highest number of TFs (Figure 6f). The
322 improvement in performance is largely due to the reduction of data sparsity by aggregating the
323 appropriate target genes, as shown by PRDM1 (Figure 6g). Taken together, based on this ground
324 truth dataset, GRNs reconstructed from Epiregulon could accurately identify known lineage
325 factors and predict effects of gene perturbation in datasets containing only scRNA-seq.
326
327
328 **Discussion**
329 We develop Epiregulon, a method to construct gene regulatory networks and infer activity of both
330 transcription factors and transcriptional coregulators lacking defined motifs, an important class of
331 regulators largely overlooked by previous GRN inference methods. When benchmarked against
332 other GRN inference methods, Epiregulon performs most robustly across a multitude of datasets,
333 especially when TF activity is decoupled from its gene expression such as in drug perturbations.
334 Epiregulon predicts the largest repertoire of regulators, constructing an extensive regulatory
335 network that is functionally annotated with pathway enrichment analysis and interaction partners.
336
337 Unlike most other GRN inference methods, Epiregulon allows construction of GRNs from ChIP-
338 seq data but still preserves the option to filter ChIP-seq peaks based on the presence of motifs.
339 ChIP-seq data allows the prediction of chromatin co-activators and modifiers lacking defined

motifs. Importantly, we were able to infer the activity of the chromatin remodeler SMARCA4 in a context-aware manner. We systematically compared the performance of TF prediction based on GRNs constructed from ChIP-seq vs motif annotations, fine tuning motif annotations using deep-learning neural networks models. We found that TF occupancy obtained from public or user-provided ChIP-seq data can accurately predict TF activity.

Weight estimation by Epiregulon enhances the accuracy of TF activity by assigning the importance scores to target genes. The Wilcoxon weight estimation method, based on co-occurrence of TF gene expression and chromatin accessibility, is especially well suited for applications in which TF activity is discordant with its gene expression such as drug perturbations. The Wilcoxon method also enables estimation of cell-cluster specific weights and subsequent differential network topology analysis. Epiregulon correctly recapitulates the response of multiple AR-targeting agents in various prostate cancer cell line models, and is well positioned to predict TF activity in a wide range of biological applications and systems.

Epiregulon enables us to make new discoveries from the ground truth data we generated. We found an unexpected connection between GATA6, YAP1 and TEAD1 which led to the discovery that GATA6 overexpression leads to reprogramming of the prostate adenocarcinoma cell line LNCaP towards a stem-cell like subtype. This was possible because compared to other methods, Epiregulon covers the most extensive regulators including coregulators lacking motifs (in this case YAP1), enabling the identification of interaction partners based on shared target genes.

Several limitations of our study remain. While Epiregulon is robust in predicting the overall TF activity, we have not formally evaluated the accuracy of individual target genes. Second, Epiregulon does not explicitly model cooperativity between transcription factors and partners, so direct interactions between functionally related factors may have been overlooked. Lastly, given the use of ground truth data, our evaluation of performance is limited to a few transcription factors and it remains unclear whether Epiregulon uniformly performs well across multiple TFs. We speculate that the quality of available ChIP-seq data greatly influences the prediction of TF activity, and wherever possible, ChIP-seq data of TF of interest should be generated in the matching cell lines to ensure robustness of the biological findings.

We envision that Epiregulon can enable drug discovery by evaluating response to therapeutic agents not limited to oncology and identifying new drivers of resistance. Epiregulon is available at https://github.com/xiaosaiyao/epiregulon along with the data package scMultiome (https://bioconductor.org/packages/release/data/experiment/html/scMultiome.html).

Methods

Epiregulon

The Epiregulon workflow consists of several components: GRN construction, pruning of networks (optional), estimation of weights, calculation of activity, functional annotation of regulons, differential network analysis and identification of interaction partners. Epiregulon is designed to work seamlessly with ArchR but can accept any input as long as they are formatted into

SingleCellExperiment objects. It is available as three related R packages. Epiregulon is the core package that performs GRN construction and activity inference (https://github.com/xiaosaiyao/epiregulon/). Epiregulon.extra contains differential analysis and plotting functions (https://github.com/xiaosaiyao/epiregulon.extra). Epiregulon.archr contains functions interfacing with ArchR (https://github.com/xiaosaiyao/epiregulon.archr).

Data preprocessing

Epiregulon assumes that prior preprocessing of the data has been performed by users' methods of choice, requiring paired gene expression matrix (scRNA) and peak matrix (scATAC) along with dimensionality reduction matrix as the input. If gene expression and peak matrix are not paired, integration of scATAC-seq and scRNA-seq data must be performed prior to Epiregulon using methods such as ArchR's addGeneIntegrationMatrix function.

For the datasets included in this manuscript, reads were mapped by the Cell Ranger ARC 2.0 and were further processed with the use of ArchR[12]. Briefly, cells were filtered based on TSS enrichment (>6) and number of ATAC-seq reads mapped to the nuclear genome (>2500 and $<10^5$). Moreover, doublets were removed using dedicated ArchR functions. ATAC-seq data was represented as a 500bp tile matrix on which iterative LSI was performed. Similarly, normalized gene expression data was subject to iterative LSI. The dimensionality reductions from both modalities were then combined into a single one that was used for cell clustering. The ATAC-seq peaks were called separately for each cluster and after normalization were merged to produce one peak set for all cells (peak x cell matrix). The cells were assigned to their sample barcodes using demuxEM[49].

Network construction

Epiregulon provides two similar methods to establish peak to gene links using the calculateP2G function. If there is an existing ArchR project, Epiregulon retrieves the peak to gene links that have been previously assigned by ArchR. Briefly, ArchR creates 500 cell aggregates, resampling cells if needed. ArchR computes the correlation between chromatin accessibility and target genes within a size window (default ±250 kb) and retains peak-gene pairs that exceed a correlation threshold (default Epiregulon cutoff is 0.5). In the absence of an ArchR project, Epiregulon defines cell aggregates using k-means clustering based on the reduced dimensionality matrix and performs correlation in the same manner as ArchR does. If cluster labels are provided, cell aggregates are formed within each cluster and cluster-specific correlations are reported in addition to overall correlations.

TF occupancy data

Each regulatory element is then interrogated for TF occupancy based on a compilation of public TF ChIP-seq binding sites. ChIP-seq were downloaded from ChIP-Atlas (chip-atlas.org) and ENCODE (encodeproject.org). TF ChIP-seq peaks with FDR $< 1 \times 10^{-5}$ were retained. ChIP-seq datasets of the same TFs were merged, yielding 1557 unique factors (human) and 768 unique factors (mouse). Data is provided in the scMultiome package as GRanges list objects and can be accessed using the getTFMotifInfo function from the Epiregulon package or directly from the scMultiome package using the tfBinding function.

If desired, ChIP-seq peaks can be further annotated for the presence of motifs using the addMotifScore function. If motif annotation has been performed previously using ArchR's addMotifAnnotations function, motif annotations can be easily retrieved and appended to the peak matrix. In the ArchR independent workflow, Epiregulon can annotate peak matrix using motifmatchr's motif matching function[50] and cisbp as the reference motif database[51]. Alternatively, users can start entirely from motif annotations and leverage deep learning motifs to select motifs based on motif importance scores (see section on DNN models).

Network pruning (optional)

Epiregulon prunes the network by performing tests of independence on the observed number of cells jointly expressing transcription factor (TF), regulatory element (RE) and target gene (TG) vs the expected number of cells if TF/RE and TG are independently expressed using the pruneRegulon function.

We define n as the total number of cells, k as the number of cells jointly expressing TF, TG and RE above a set threshold, g as the number of cells jointly expressing TF and RE above a threshold and h as the number of cells expressing TG above a threshold. p, the expected probability of cells jointly expressing TF, TF and RE above a threshold is defined as follows:

$$P(cells\ expected\ to\ jointly\ express\ TF\ t, RE\ r\ and\ TG\ g) = p\ = \frac{g}{n} \times \frac{h}{n}$$

Two tests of independence are implemented, the binomial test and the chi-square test. In the binomial test, the expected probability is p, the number of trials is the total number of cells n, and the observed number of successes is k, the number of cells jointly expressing all three elements.

In the chi-square test, the expected probability for having all 3 elements active is also p. The observed cell count for the active category is k, and the cell count for the inactive category is $n - k$. P-values are calculated from chi-squared distribution with degree of freedom equal to 1. Cluster-specific p-values are calculated if users supply cluster labels.

Estimation of weights

While network pruning provides statistics on the joint occurrence of TF-RE-TG, we would like to further estimate the strength of regulation using the addWeights function. Biologically, this can be interpreted as the magnitude of gene expression changes induced by changes in transcription factor activity. Epiregulon provides 3 different methods to estimate weights. Two measures (correlation and Wilcoxon statistics) give both the magnitude and directionality of changes whereas weights computed by mutational information (MI) are always positive. Within 2 of the methods (correlation and MI), there is an option of modeling the TG expression based only on TF expression, or on the product of TF expression and RE chromatin accessibility. Consideration of both TF expression and RE chromatin accessibility is highly recommended, especially for scenarios in which TF activity and TF gene expression are decoupled (as in the case of drug perturbations or CRISPR genome editing). The correlation and mutual information statistics are

computed on pseudobulks by user-provided cluster labels and yield a single weight across all clusters per each TF-RE-TG triplet. In contrast, the Wilcoxon method groups cells based on the joint expression of TF, RE and TG in each single cell or in cell aggregates. Cell aggregation uses a default value of 10 cells and can help overcome sparsity and speed up computation. If cluster labels are provided, we can obtain cluster-specific weights using the Wilcoxon method.

1) Wilcoxon test statistics

Cells were divided into two groups, with the first group jointly expressing TF gene expression and chromatin accessibility at RE, and the remaining cells in the second group. The default cutoff is 1 for normalized gene expression and 0 for normalized chromatin accessibility. There is also an option to use the median of each feature as an adaptive cutoff.

2) Correlation

If we consider only TF expression (tf_re.merge set to FALSE), the weight is the correlation coefficient between the TF gene expression and TG gene expression. If we consider both TF expression and RE chromatin accessibility (tf_re.merge set to TRUE), the weight is the correlation coefficient between the product of TF gene expression and RE chromatin accessibility vs the TG gene expression.

3) Mutual information between the TF and target gene expression,

If we consider only TF expression (tf.re_merge is set to FALSE), the weight is the mutual information between the TF gene expression and the TG gene expression. If we consider both TF expression and RE chromatin accessibility (tf.re_merge is set to TRUE), the weight is the mutual information between the product of TF gene expression and RE chromatin accessibility vs the TG gene expression.

Calculation of TF activity

The activities for a specific TF in each cell are computed by averaging the weighted expressions of target genes linked to the TF:

$$\boldsymbol{K_{t,c}} = \frac{\mathbf{1}}{\boldsymbol{q_t}} \sum_{\boldsymbol{g \in R_t}} \boldsymbol{\beta_{t,g,r} Y_{g,c}}$$

where $K_{t,c}$ is the activity of a TF t for a cell c, q is the total number of target genes for a TF t, Y_g is the normalized count of target gene g where g is regulated by TF t and $\beta_{t,g,r}$ is the regulatory weight of TF t on target gene g at RE r. R_t is the regulon of TF t.

If cluster labels are provided, cluster-specific weights are used.

Gene set enrichment of regulons

Gene set enrichment of a regulon is performed by testing whether the target genes of a TF are over-represented in known gene signatures such as those provided by MSigDB using a

hypergeometric test. Target genes can be refined by filtering the regulons on user-defined weights.

Differential TF activity by total activity

This differential analysis compares the differences in the total activity of each transcription factor between conditions. This analysis is well suited for identifying factors that have contrasting levels of activities, for instance, lineage factors that are turned on or off during certain developmental stages or cell types. TF activities are compared between groups using any standard statistical methods. Here we use scran's findMarkers function to find differential activity between user provided groups of cells (https://rdrr.io/bioc/scran/man/findMarkers.html).

Differential TF activity by network topology

A second approach to investigate differential TF activity is to compare and contrast target genes or network topology. This is useful when a transcription factor differs in the target genes it regulates, while maintaining a similar level of total activity. This can happen when a transcription factor redistributes to a different set of genomic regions, due to mutations in the transcription factors or changes in the interaction partners. Differences in network topology are calculated by taking the degree centrality of the edge subtracted graphs between two conditions, with cluster-specific regulon weights representing edge weights in each condition.

$G^{(1)}$ and $G^{(2)}$ are networks with identical node sets N and respective edge sets $E^{(1)}$ and $E^{(2)}$:

- $E^{(1)} = \{e^{(1)}_1, e^{(1)}_2, e^{(1)}_3, \ldots e^{(1)}_m\}$
- $E^{(2)} = \{e^{(2)}_1, e^{(2)}_2, e^{(2)}_3, \ldots e^{(2)}_m\}$

where $e^{(j)}_i$ represents the edge weight of the *i*th edge in network $G^{(j)}$.

The edge-subtracted network G' induced by $G^{(1)}$ and $G^{(2)}$ such that each edge e'_i in G' is given by $e'_i = e^{(1)}_i - e^{(2)}_i$. We simplify the tripartite TF-RE-TG graph to a bipartite TF-TG graph by taking the maximum of TF-RE-TG weights of the same TF-TG pairs. Degree centrality is the sum of the weights of the TF-TG e'_i. Degree centrality is further normalized to account for differences in the number of target genes of each transcription factor. The default normalization method is dividing degree centrality by the square root of the number of target genes. This strikes a balance between penalizing TFs with an abundance of target genes and prioritizing TFs with differential target genes. Transcription factors are ranked by normalized degree centrality.

Identification of interaction partners

To identify potentially interacting TF or co-factors, we calculate Jaccard similarity between all pairs of factors and rank factors by the degree of overlap in their regulons. The edges can be filtered on a weight cutoff to retain the most confident set of target genes.

Benchmarking

Benchmarking using Reprogram-Seq data

We introduced 4 transcription factors into LNCaP cells - FOXA1, FOXA2, NKX2-1 and GATA6 and obtained paired gene expression and chromatin accessibility information. Because FOXA1 was already highly expressed in LNCaP cells at the basal level, its introduction did not have any impact on lineage plasticity and therefore FOXA1 was excluded in subsequent analyses. We focused on NKX2-1 and GATA6 because their expression resulted in distinct cell clusters. Each cell is identified by the HTO tag corresponding to the well receiving virus encoding a particular TF and this information serves as the ground truth cell labels. For each TF, a cell is classified into either expressing this TF or not. AUROC was computed for all the packages being benchmarked as well as for gene expression of the TF being evaluated. Note that because each TF has different transduction efficacy and not all cells are able to take up the virus or express the TF, the theoretical maximum of AUROC will not reach 1.

Benchmarking using AR antagonist data

For the drug treatment dataset, we treated 4 prostate cancer cell lines with 3 different therapeutic agents and obtained paired gene expression and chromatin accessibility information. Only the 2 AR-dependent cell lines, LNCaP and VCaP, were included in the benchmark since they are known to respond to AR-targeting agents. Similarly, only AR antagonist, Enzalutamide[1] and AR degrader ARV-110[3] were used in the benchmark studies since they are known to specifically inhibit AR activity. Cells treated with Enzalutamide and ARV110 are supposed to show reduced AR activity compared to cells treated with DMSO control. Each cell line was analyzed separately by retaining only peaks found in each cell line and all the subsequent analyses were performed separately for both cell lines. Each cell is identified by the HTO tag corresponding to the well receiving DMSO or one of the two AR inhibitors and this information serves as the ground truth. For each of the two drugs (enzalutamide or ARV110), a cell is classified into either treated with DMSO control or an AR inhibitor. AUROC is computed for all the packages being benchmarked as well as for AR gene expression. Because TF activity is decoupled from TF expression, we used the Wilcoxon method to estimate weights with merged public AR ChIP-seq prefiltered for the presence of motifs as the TF occupancy prior.

We used 5 GRN inference tools (CellOracle, FigR, SCENIC+, *GRaNIE* and *Pando)* to benchmark the performance of Epiregulon. To ensure consistency, we applied the same gene expression and chromatin accessibility matrices across all tools. We first used ArchR to determine the peaks i.e. DNA regions with frequent Tn5 transposase insertion events from the fragment files output by CellRanger. Then the peak x cell matrix was produced by counting the number of insertions per peak per cell. From the same ArchR projects we also retrieved normalized gene expression to be used by Epiregulon, GRaNIE, FigR. For the remaining tools, we used a subset of cells in the gene expression data which matched the cells in the ArchR projects to make sure that all tools work on the same cells. Below we describe the workflows used in each tool. We followed the tutorial examples provided at official web sites with only minor changes to the default settings.

CellOracle

We performed benchmarking against CellOracle[52]. The peak count matrices retrieved from AchR projects were used to create a CellDataSet object which was subject to dimensionality reduction and cell aggregation. The ATAC-peaks were first annotated by HOMER to identify transcription

start sites (TSSs). Co-accessibility regions were further identified by the cicero R package using the run_cicero function. Pairs of peaks that show high co-accessibility (≥0.8) and overlap with a TSS were used to link peaks to their target genes. Peaks were scanned for the presence of TF binding motifs using TFinfo.scan. Cell neighborhoods were defined by diffusion maps on the dimensionality reduced matrix (PCA) and cell clusters were identified using the K nearest neighbor graph using the Louvain method.

CellOracle models the expression of the target gene as the weighted sum of gene expression of all the transcription factors it is regulated by.

$$\boldsymbol{Y_g} = \sum_{\boldsymbol{t}} \boldsymbol{\beta_{t,g} X_t} + \boldsymbol{c}$$

where Y_g is the expression of the target gene g, X_t is the expression of the transcription factor t, β is the fitted coefficient and c is the intersection.

The final activities were calculated using Epiregulon's calculateActivity function using $\beta_{t,g}$ as the weights. Weights were estimated for each cluster individually and the activities were assembled together.

FigR

We performed benchmarking against FigR [9]. In the first step of GRN construction the correlation between peak accessibility and target gene expression was determined using runGenePeakcorr. We used a non-default search range around TSS (250 kb) to be consistent across all the benchmarked tools. Correlation coefficients were also computed with background peaks and the significance of gene-peak association was determined with one-tailed z-test. Only gene-peak associations that show positive correlation and are statistically significant were retained ($p \leq 0.05$). DORaCs (High density domains of regulatory chromatin) correspond to genes with ≥7 associated peaks and were calculated by summing the ATAC-seq reads in peaks matched to each gene. Each TF is associated with a single motif by selecting the motifs most highly correlated with other motifs of the same TF. Each DORC is evaluated for enrichment of TFs by comparing the match frequency in its peakset vs the match frequency in a set of background peaks; the p-values of the enrichment were obtained by z-test ($P_{Enrichment}$). Smoothing of gene expression matrix and chromatin accessibility data summarized across target genes was performed after determination of k nearest neighbor cells based on LSI space retrieved from ArchR project and constructed using ATAC-seq data. The Spearman correlations between the smoothed DORC accessibility and smoothed TF gene expression were computed ($Correlation$) and their significance were obtained using z-test ($P_{Correlation}$). From the output of the main function (getFigRGRN) we retrieved scores indicating the strength of association between transcription factors and target genes which were computed as below:

$$Regulation\ score\ =\ sign(Correlation)\ *\ -log10[1-(1-P_{Enrichment})*(1-P_{Correlation})]$$

The regulation scores were used as weights when calculating activity with Epiregulon.

SCENIC+

We performed benchmarking against SCENIC+ [11]. Briefly, we used pycisTopic which uses Latent Dirichlet Allocation to group regulatory elements into the topics. The model evaluation indicated 20, 25 and 20 as the number of topics for VCaP cells from AR dataset, LNCaP cells from AR dataset, and Reprogram-Seq dataset respectively. The input peak matrix was retrieved from the AchR project. We then used pycistarget to identify TF-region links. Pycistarget identifies motif matches in the peak regions using HOMER, scores each region for motif importance and identifies differentially enriched motifs above background regions. Only regions showing NES > 3.0 and motifs with adjusted p-value <0.05 and log2FC >0.5 were retained. TF-gene importance scores were calculated using gradient-boosting machine regression by predicting TF expression from target gene expression. Region-gene importance scores were calculated using gradient-boosting machine regression by predicting target gene expression from region accessibility. Genes were ranked by TF-gene importance scores and only genes in the leading edge of the gene set enrichment were used for the eRegulon. Gene set enrichment was also performed for region-gene pairs. Peaks were ranked by imputed chromatin accessibility and genes were ranked by gene expression counts per cell. Enrichment score was defined as the AUC at 5% of the ranking and was calculated using AUCell. The enrichment score was used as the activity score in the benchmark assessment.

GRaNIE

We performed benchmarking against GRaNIE [10]. Briefly, GRaNIE overlapped TF binding sites obtained from HOCOMOCO-based TF motifs with ATAC-seq peaks. GRaNIE identified TF-peak connections using Pearson correlation between TF expression and the peak accessibility across samples. The cutoffs for the correlation were chosen based on an empirical FDR calculated from the ratio of TF-peaks in the background peaks over the total number of TF-peaks in both the background and foreground. Peak-gene connections were identified using correlation between the gene expression and chromatin accessibility. All the GRN edges have a weight of 1. The default threshold is FDR < 0.2 for TF-peak links and FDR < 0.1 for peak-gene links. Activities were computed using Epiregulon's calculateActivity.

Pando

We performed benchmarking against Pando[8]. Briefly, ATAC-seq peaks were intersected with PhastCons conserved elements and cCREs derived from ENCODE. TFs present in the 4000 most variable genes were included in the downstream analysis. Motifs for the TFs were obtained from JASPAR2020 and CISBP. Gene expression and chromatin accessibility counts were smoothened by averaging cells within a neighborhood. Pando then models target gene expression as the weighted sum of the product of TF expression and the chromatin accessibility of the region where the TF binds to.

$$\boldsymbol{Y_g} = \sum_{\boldsymbol{t}} \boldsymbol{\beta_{t,g,r} X_t A_r} + \boldsymbol{c}$$

where Y_g is the expression of the target gene g, X_t is the expression of the transcription factor t, A_r is the chromatin accessibility at region r, β is the fitted coefficient and c is the intersection.

Fitted coefficients were tested for significance using analysis of variance (ANOVA). Only edges with FDR <0.05 were retained in the final GRN. Activities were calculated using Epiregulon's calculateActivity function with the fitted coefficients $\beta_{t,g,r}$ as the weights.

Benchmarking using precomputed regulons

Gene expression and chromatin accessibility matrices were obtained from the unpaired PBMC data[39] and integrated using ArchR's addGeneIntegrationMatrix function. Regulons were constructed using Epiregulon and benchmarked against regulons obtained from Dorothea[46]) and Humanbase (hematopoiesis)[48].

These regulons were applied onto a single-cell RNA-seq dataset involving a CRISPRa screen that overexpressed transcription factors using gRNA in K562 cells[45]. The identity of the gRNA was used as the ground truth indicating which TF was activated in each cell. Regulon weights were computed in the K562 dataset, using the correlation method with gRNA as the cluster labels for all three regulons. ROC was computed using TF activity as the estimates for cluster labels indicating whether a particular TF is active or not. TF gene expression was also used as the estimates to generate a separate ROC.

DNN training

Steps for DNN-based TF-RE mapping involved dataset processing, model training, motif scoring and thresholded GRN construction.

Dataset processing

The reference genome hg38 and the pseudo-bulk ATAC (taken from ArchR) of individual clusters (C1 or C3) was collected. The whole genome excluding unmappable regions[53] was split into 3072 bp regions. Using the coordinates of each region the corresponding DNA sequence and bigwig coverage was obtained. The DNA sequence was converted to one-hot encoded form. The DNA and coverage information were saved in tf records files for faster I/O during training. Training, validation and test splits were done by chromosomes taking chr8 as test set and chr9 as validation set.

Model training

We used Basenji2 architecture with 32 base-pair resolution (binned coverage at 32 bp). The model was trained by randomly sampling 2048bp segment from the input DNA and taking the corresponding coverage (one cluster per model). We used Poisson NLL as loss, used reverse complement augmentation and trained for a maximum of 50 epochs with early stopping.

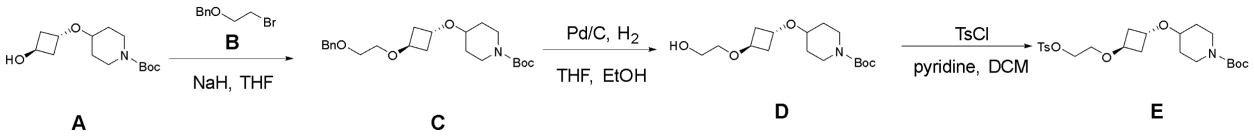
We additionally trained models with base-pair resolution using the ChromBPNet architecture. We first trained a custom Tn5 sequence bias model for this dataset on non-peak regions. We ensured that these models learned Tn5 bias by applying DeepSHAP to the model outputs. Using this bias model to regress out Tn5 sequence bias, we then trained ChromBPNet TF models. These models were trained on 2114bp sequences centered on ATAC-seq peaks.

729 *Motif scoring*
730 For each of the motif regions (taken from ArchR motif positions file) we obtained the DNA
731 sequence (denoted as wild type) by centering at the motif and extending to 2048bp. We occluded
732 the motif by replacing motif nucleotides with 'N' (or 0.25 in one-hot encoding) (denoted as mutant).
733 We computed the score - mean fraction change in predicted coverage by subtracting mutant from
734 wild type and normalizing by predicted wild type coverage.
735
736 *Thresholded GRN construction*
737 We computed several quartiles of scores as thresholds. We filtered for those motifs with
738 importance scores higher than the threshold corresponding to bigger differences in prediction.
739 Using the RE remaining we constructed the GRN as before.
740
741 **Reprogram-Seq**
742 *Construction of Lentivirus Plasmids for TF Over-expression*
743 All lentivirus plasmids were generated by GenScript. Briefly, the ORF of transcriptional factors
744 were codon optimized, synthesized and cloned after the hEF1a promoter. The puromycin
745 resistance gene is driven by a separate Cbh promoter to enable antibiotics selection. Maxi-prep
746 of each plasmid was performed to maximize the transfection efficiency.
747
748 *Cell Culture and Virus Packaging*
749 LNCaP Clone FGC cells were obtained from ATCC and cultured with RPMI-1640 media with 10%
750 FBS and 2mM L-Glutamine. The cells were split every 4-5 days to maintain the appropriate
751 density. 293T cells were cultured in DMEM with 10% FBS, 100μM NEAA, 2mM Glutamine. They
752 were split every 2-3 days. One day before transfection, the cell culture dish was treated with 5mL
753 1% gelatin in PBS, incubated for 10 minutes, then aspirated. $3.5X10^6$ 293T cells were seeded
754 into each 10 cm dish to reach ~80% confluence. On the day of transfection, 20 μl Lipofectamine
755 2000 was added to 480 μl OptiMEM. In a new tube, the plasmid was mixed in 500μl Opti-MEM at
756 the following ratio: carrier plasmid, 5 μg; delta8.9, 16 μg; VSVG, 1 μg. Both mixes were combined
757 and incubated for 20 minutes before adding to the dish. The dish was incubated at 37˚C for 6
758 hours and then 6 ml of complete media was added. Virus was then harvested 48 hours after
759 transfection. Briefly, all supernatants were harvested and filtered through a 0.45 μm filter bottle.
760 The virus was concentrated by using the Lenti-X concentrator (TAKARA Bio.) following the
761 manufacturer's instructions and resuspended in 1 ml 1% BSA in PBS per dish. The concentrated
762 virus was divided into 200μl aliquots and stored in -80˚C until infection.
763
764 *Infection of LNCaP Cells by Lentivirus*
765 Two days before infection, $4X10^5$ LNCaP cells were seeded into each well of a 6-well plate. On
766 the day of infection, aspirate the media and change to 0.5mL RPMI + 10% FBS + 1X Glutamax
767 with 8 μg/ml Polybrene. A total of 200 μl concentrated lentivirus was added to each well to achieve
768 high MOI, and then the plate was centrifuged at 1800 rpm for 45 minutes at room temperature.
769 After that the plate was put into the incubator for another 3 hours before 2 ml of the full media was
770 added to each well. Two days after infection, the media was refreshed with 1μg / ml puromycin.
771 The cells were then grown for another 7 days before harvesting for the single-cell analysis, during
772 which the cells were split accordingly when the confluence reached 100%.

773
774 *Engineering of dox-inducible NKX2-1 expressing cell lines*
775 LNCAP Clone FGC cells were passaged two days before electroporation. 200k cells per reaction
776 were transduced with 375 ng PBIND-NKX2-1-IRES-eGFP-PGK-puro and 125ng pBO (PiggyBac
777 transposase) (3:1 vector:transposase ratio) according to manufacturer's instructions for SF Cell
778 Line 4D-Nucleofector™ X Kit S, using the EN-120 electroporation protocol in 20 µl
779 Nucleocuvette™ Strip. After electroporation, 100ul of warm media (RPMI + 10% FBS + 1X
780 Glutamax) was added to the cuvette and the cuvette was placed in the incubator for 1 hour to
781 recover. Cells were then transferred into a 6 well plate. Puromycin selection (RPMI + 10% FBS +
782 1X Glutamax, 1ug/ml Puromycin) was performed after 4-5 days, depending on cell recovery.
783
784 **Single-cell RNA-seq and Single-cell RNA-ATAC Co-assay**
785 The single-cell RNA-seq was performed using the Chromium Single Cell 3' kit (V3.1) from 10X
786 Genomics with cell hashing. Briefly, the cells were trypsinized into single-cell suspension from the
787 6-well plates, and washed once with PBS with 1% BSA. The cells from different wells were stained
788 with human TotalSeq-A cell hashing antibodies (BioLegend) containing distinct barcodes
789 following the manufacturer's protocol. The cells were then washed twice with PBS with 1% BSA
790 before combining and the live cells were sorted using the SONY SH800S FACS machine. For
791 loading of the 10X chip G, we overloaded the channel with the aim of recovering 20K cells. The
792 library construction was following the standard 10X protocol with the following exceptions: 1) the
793 1µl of 10µM HTO additive primer was spiked in during the cDNA PCR step; 2) during the SPRI
794 beads cleanup step, the supernatant from the 0.6X cleanup was saved to recover the HTO
795 fragment. Along with the transcriptome library, the HTO library was amplified from the supernatant
796 using the HTO-specific primer. The detailed protocol can be found on the cell hashing website
797 (https://cite-seq.com/cell-hashing/).
798
799 The single-cell RNA-ATAC co-assay was performed using the 10X Genomics Single Cell
800 Multiome ATAC + Gene Expression kit. In a previous test, we found that the standard cell
801 permeabilization protocol used for nuclei preparation does not completely remove the plasma
802 membrane of the LNCap cells and robust B2M expression can still be detected with FACS (data
803 not shown). Therefore, we speculated that the same cell hashing strategy would work for the 10X
804 Multiome assay. Briefly, we followed the same hashing protocol as used in the scRNA-seq
805 experiment. After the antibody staining and pooling, the cells were permeabilized by following the
806 nuclei isolation protocol from 10X genomics (10X Nuclei Isolation Protocol). The cells were then
807 counted and loaded on the Chip J with a recovery aim of 12K cells per lane. For the library
808 construction, we followed the standard protocol with the following modifications: 1) 1µl of 10µM
809 HTO additive primer was spiked in during the pre-AMP and cDNA PCR steps; 2) during the 0.6X
810 cDNA cleanup, the supernatant was saved to amplify the HTO library, as described in the previous
811 section.
812
813 All the libraries were sequenced on Illumina NextSeq 2000 or NovaSeq 4000. We were aiming
814 for 20K, 40K and 2K raw reads per cell for the transcriptome, ATAC and HTO libraries,
815 respectively.
816

Synthesis of SMARCA2.4_1

Scheme 1. Preparation of intermediate **E**.



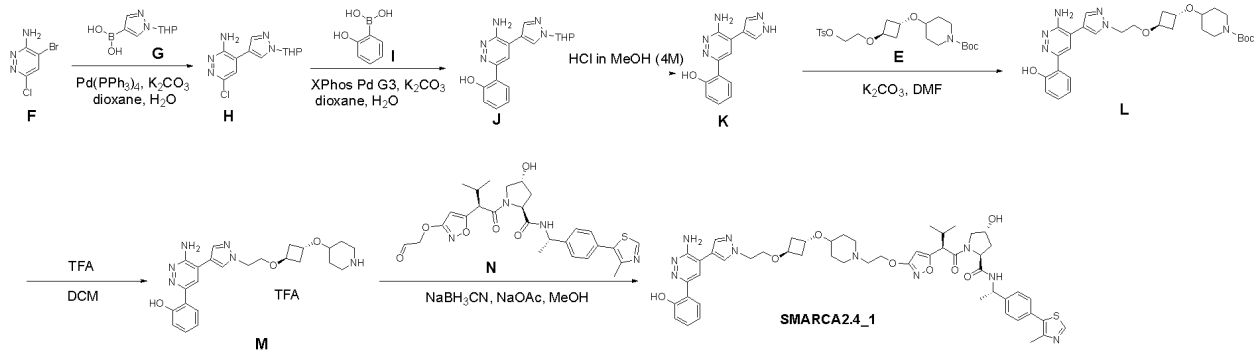
*tert-Butyl 4-((trans)-3-(2-(benzyloxy)ethoxy)cyclobutoxy)piperidine-1-carboxylate. (**C**)*. To a solution of *tert*-butyl 4-((trans)-3-hydroxycyclobutoxy)piperidine-1-carboxylate (**A**, 500 mg, 1.84 mmol, 1.0 eq) in THF (7 mL) was added NaH (60% in mineral oil, 111 mg, 2.77 mmol, 1.5 eq) at 0 °C. The resulting solution was stirred at 0 °C for 0.5 h afterwhich ((2-bromoethoxy)methyl)benzene (**B**, 592 mg, 2.76 mmol, 1.5 eq) was added. The mixture was warmed to room temperature and was stirred for an additional 16 h then was quenched with saturated aqueous NH_4Cl (5 mL). The resulting solution was extracted with EtOAc (10 mL x 3), and the combined organic layers were dried over sodium sulfate and concentrated under vacuum. The residue was purified by silica gel chromatography (EtOAc:petroleum ether, 0:100 to 24:76) to afford the title compound (330 mg, 44%) as a colorless oil. ^{1}H NMR (300 MHz, DMSO-d_6) δ 7.42-7.23 (m, 5H), 4.49 (s, 2H), 4.25-4.17 (m, 1H), 4.09-4.02 (m, 1H), 3.68-3.61 (m, 2H), 3.57-3.51 (m, 2H), 3.47-3.37 (m, 2H), 2.95 (t, *J* = 11.0 Hz, 2H), 2.23-1.98 (m, 4H), 1.79-1.67 (m, 2H), 1.39 (s, 9H), 1.35-1.17 (m, 3H). LC-MS: (ES+): m/z 406 $[M+H]^+$.

*tert-Butyl 4-((trans)-3-(2-hydroxyethoxy)cyclobutoxy)piperidine-1-carboxylate. (**D**)*. A solution of intermediate **C** (330 mg, 0.815 mmol) and 10% Pd/C (66.0 mg) in ethanol (10 mL) and THF (10 mL) was stirred at room temperature for 1 h under a hydrogen atmosphere (3 atm). The resulting mixture was filtered, and the filter cake was washed with DCM (10 mL x 2). The filtrate was concentrated under reduced pressure to afford the title compound (245 mg, 95% yield) as a white oil. ^{1}H NMR (300 MHz, DMSO-d_6) δ 4.58 (t, *J* = 5.5 Hz, 1H), 4.26-4.18 (m, 1H), 4.08-4.01 (m, 1H), 3.69-3.61 (m, 2H), 3.53-3.34 (m, 3H), 3.28 (t, *J* = 5.3 Hz, 2H), 2.96 (t, *J* = 11.4 Hz, 2H), 2.24-1.98 (m, 4H), 1.76-1.71 (m, 2H), 1.39 (s, 9H), 1.36-1.18 (m, 2H). LC-MS: (ES+): m/z 316 $[M+H]^+$.

*tert-Butyl 4-((1r,3r)-3-(2-(tosyloxy)ethoxy)cyclobutoxy)piperidine-1-carboxylate (**E**)*. To a solution of intermediate **D** (245 mg, 0.78 mmol, 1.0 eq) and pyridine (245 mg, 3.11 mmol, 4.0 eq) in DCM (2 mL) was added TsCl (223 mg, 1.16 mmol, 1.5 eq) at 0 °C. The mixture was then warmed to room temperature and stirred for 12 h. The solvent was subsequently removed under vacuum, and the residue was purified by silica gel chromatography (EtOAc:petroleum ether, 0:100 to 29:71) to afford the title compound (303 mg, 83%) as a colorless oil. ^{1}H NMR (300 MHz, DMSO-d_6) δ 7.82-7.72 (m, 2H), 7.52-7.41 (m, 2H), 4.22-4.05 (m, 3H), 4.01-3.88 (m, 1H), 3.67-3.59 (m, 2H), 3.45-3.34 (m, 3H), 2.94 (t, *J* = 11.4 Hz, 2H), 2.41 (s, 3H), 2.13-1.92 (m, 4H), 1.76-1.64 (m, 2H), 1.37 (s, 9H), 1.33-1.15 (m, 2H). LC-MS: (ES+): m/z 470 $[M+H]^+$.

Scheme 2. Preparation of SMARCA2.4_1 (compound **1)**.

855



856

857

858

859 *6-Chloro-4-(1-(tetrahydro-2H-pyran-2-yl)-1H-pyrazol-4-yl)pyridazin-3-amine (**H**)*. A
860 solution of 4-bromo-6-chloropyridazin-3-amine (**F**, 2.00 g, 9.60 mmol, 1.0 eq), (1-(tetrahydro-2H-
861 pyran-2-yl)-1H-pyrazol-4-yl)boronic acid (**G**, 2.46 g, 12.5 mmol, 1.3 eq), $Pd(PPh_3)_4$ (1.12 g, 0.969
862 mmol, 0.1 eq) and K_2CO_3 (2.67 g, 19.3 mmol, 2.0 eq) in 1,4-dioxane (20 mL) and water (2 mL)
863 was stirred at 100 °C for 2 h. The reaction was then cooled to room temperature and was
864 quenched with water (20 mL). The resulting solution was extracted with EtOAc (30 mL x 3), and
865 the combined organic layers were dried over sodium sulfate and concentrated under vacuum.
866 The residue was purified by silica gel chromatography (EtOAc:petroleum ether, 0:100 to 58:42)
867 to afford the title compound (2.40 g, 89%) as a red oil. ^{1}H NMR (300 MHz, DMSO-d_6) δ 8.52 (d,
868 J = 0.8 Hz, 1H), 8.11 (d, J = 0.8 Hz, 1H), 7.65 (s, 1H), 6.41 (s, 2H), 5.46 (dd, J = 9.9, 2.2 Hz, 1H),
869 4.07-3.90 (m, 1H), 3.76-3.57 (m, 1H), 2.24-2.05 (m, 1H), 2.00-1.91 (m, 2H), 1.79-1.64 (m, 1H),
870 1.60-1.53 (m, 2H). LC-MS: (ES+): m/z 280 $[M+H]^+$.

871 *2-(6-Amino-5-(1-(tetrahydro-2H-pyran-2-yl)-1H-pyrazol-4-yl)pyridazin-3-yl)phenol (**J**)*. A
872 solution of intermediate **H** (1.35 g, 4.83 mmol, 1.0 eq), (2-hydroxyphenyl)boronic acid (**I**, 870 mg,
873 6.31 mmol, 1.3 eq), XphosPdG3 (410 mg, 0.484 mmol, 0.1 eq) and K_2CO_3 (1.33 g, 9.64 mmol,
874 2.0 eq) in 1,4-dioxane (11.5 mL) and water (2.5 mL) was stirred at 95 °C for 5 h. The mixture was
875 cooled to room temperature, and the solvent was removed under vacuum. The residue was
876 purified by silica gel chromatography (MeOH:DCM, 0:100 to 5:95) to afford the title compound
877 (900 mg, 55%) as a yellow solid. ^{1}H NMR (300 MHz, DMSO-d_6) δ 13.77 (s, 1H), 8.60 (d, J = 0.8
878 Hz, 1H), 8.26-8.17 (m, 2H), 8.02 (dd, J = 8.3, 1.6 Hz, 1H), 7.28-7.23 (m, 1H), 6.97-6.85 (m, 2H),
879 6.51 (s, 2H), 5.48 (dd, J = 9.9, 2.2 Hz, 1H), 3.99-3.90 (m, 1H), 3.74-3.60 (m, 1H), 2.25-2.06 (m,
880 1H), 2.00-1.95 (m, 2H), 1.76-1.64 (m, 1H), 1.59-1.52 (m, 2H). LC-MS: (ES+): m/z 338 $[M+H]^+$.

881 *2-(6-Amino-5-(1H-pyrazol-4-yl)pyridazin-3-yl)phenol (**K**)*. A solution of intermediate **J**
882 (300 mg, 0.890 mmol) in HCl (4.0 M in MeOH, 1 mL) was stirred at room temperature for 2 h. The
883 solvent was then removed under vacuum, and the residue was purified by pre-packed C18
884 column (solvent gradient: 0-100% MeOH in water (0.1% TFA)) to afford the title compound (188
885 mg, 83%) as a brown solid. ^{1}H NMR (300 MHz, DMSO-d_6) δ 13.85 (s, 1H), 13.42 (s, 1H), 8.33
886 (s, 2H), 8.22 (s, 1H), 8.02 (dd, J = 8.3, 1.6 Hz, 1H), 7.30-7.24 (m, 1H), 6.95-6.90 (m, 2H), 6.50 (s,
887 2H). LC-MS: (ES+): m/z 254 $[M+H]^+$.

*tert-Butyl 4-((trans)-3-(2-(4-(3-amino-6-(2-hydroxyphenyl)pyridazin-4-yl)-1H-pyrazol-1-yl)ethoxy)cyclobutoxy)piperidine-1-carboxylate (**L**)*. A solution of intermediate **K** (100 mg, 0.395 mmol, 1 eq), intermediate **E** (223 mg, 0.475 mmol, 1.2 eq) and K_2CO_3 (136 mg, 0.985 mmol, 2.5 eq) in DMF (1.3 mL) was stirred at 80 °C for 2 h. The reaction mixture was cooled to room temperature, then was diluted with EtOAc (5 mL) and washed with brine (3 mL x 5). The organic layer was dried over sodium sulfate and was concentrated under vacuum. The residue was purified by silica gel chromatography (MeOH:DCM, 0:100 to 5:95) to afford the title compound (60 mg, 27%) as a brown solid. ^{1}H NMR (300 MHz, DMSO-d_6) δ 13.79 (s, 1H), 8.45 (d, J = 0.8 Hz, 1H), 8.24-8.17 (m, 2H), 8.03-7.97 (m, 1H), 7.25 (t, J = 7.7 Hz, 1H), 6.97-6.87 (m, 2H), 6.48 (s, 2H), 4.30 (t, J = 5.4 Hz, 2H), 4.15-4.06 (m, 1H), 4.06-4.00 (m, 1H), 3.69 (t, J = 5.4 Hz, 2H), 3.61-3.53 (m, 2H), 3.24-3.13 (m, 1H), 2.84 (t, J = 11.3 Hz, 2H), 2.05-2.01 (m, 4H), 1.66-1.60 (m, 2H), 1.35 (s, 9H), 1.24-1.17 (m, 2H). LC-MS: (ES+): m/z 551 $[M+H]^+$.

*2-(6-Amino-5-(1-(2-(trans)-3-(piperidin-4-yloxy)cyclobutoxy)ethyl)-1H-pyrazol-4-yl)pyridazin-3-yl)phenol, TFA salt (**M**)*. A solution of intermediate **L** (60 mg, 0.11 mmol) in TFA (0.3 mL) and DCM (0.3 mL) was stirred at room temperature for 0.5 h. The solvent was removed under vacuum to afford the title compound (66 mg, crude) as a yellow solid. ^{1}H NMR (300 MHz, DMSO-d_6) δ 8.46 (s, 1H), 8.26 (s, 1H), 8.14 (s, 1H), 7.71-7.68 (m, 1H), 7.39-7.33 (m, 1H), 7.03-6.91 (m, 2H), 4.32 (t, J = 5.4 Hz, 2H), 4.20-4.00 (m, 2H), 3.70 (t, J = 5.4 Hz, 2H), 3.53-3.45 (m, 1H), 3.12-3.08 (m, 2H), 2.90-2.84 (m, 2H), 2.09-2.05 (m, 4H), 1.91-1.78 (m, 2H), 1.59-1.47 (m, 2H). LC-MS: (ES+): m/z 451 $[M+H]^+$.

*(2S,4R)-1-((R)-2-(3-(2-(4-((trans)-3-(2-(4-(3-amino-6-(2-hydroxyphenyl)pyridazin-4-yl)-1H-pyrazol-1-yl)ethoxy)cyclobutoxy)piperidin-1-yl)ethoxy)isoxazol-5-yl)-3-methylbutanoyl)-4-hydroxy-N-((S)-1-(4-(4-methylthiazol-5-yl)phenyl)ethyl)pyrrolidine-2-carboxamide (**1**)*. A solution of intermediate **M** (66 mg, crude), intermediate **N** (prepared as described in Berlin, M.; et al. "PROTACs Targeting BRM (SMARCA2) Afford Selective In Vivo Degradation Over BRG1 (SMARCA4) and are Active in BRG1-Mutant Xenograft Tumor Models", *J. Med. Chem*. accepted for publication; 85 mg, 0.157 mmol) and NaOAc (54 mg, 0.658 mmol) in MeOH (0.8 mL) was stirred at room temperature for 1 h. Then $NaBH_3CN$ (21.0 mg, 0.33 mmol) was added, and the mixture was stirred at room temperature for an additional 0.5 h. The solvent was removed under vacuum, and the residue was purified by Prep-HPLC [column: Xselect CSH F-Phenyl OBD column, 30*150 mm, 5 μm; mobile Phase A: water (10 mmol/L NH_4HCO_3), mobile phase B: CH_3CN; flow rate: 60 mL/min; gradient: 43% to 63% B over 8 min; wavelength: 254/220nm; RT: 8.97 min] to afford the title compound (12.8 mg, 12% over two steps) as a white solid. ^{1}H NMR (300 MHz, DMSO-d_6) δ 13.83 (s, 1H), 9.00-8.99 (m, 1H), 8.51-8.46 (m, 1H), 8.43 (d, J = 7.6 Hz, 1H), 8.26-8.19 (m, 2H), 8.05-8.01 (m, 1H), 7.47-7.34 (m, 4H), 7.29-7.24 (m, 1H), 6.98-6.87 (m, 2H), 6.50 (s, 2H), 6.09-5.92 (m, 1H), 5.11-5.02 (m, 1H), 4.96-4.86 (m, 1H), 4.44-4.24 (m, 4H), 4.15 (t, J = 5.6 Hz, 2H), 4.11-4.00 (m, 2H), 3.74-3.51 (m, 4H), 3.46-3.43 (m, 1H), 3.14-3.08 (m, 1H), 2.66-2.63 (m, 2H), 2.57-2.53 (m, 2H), 2.46 (s, 3H), 2.31-2.16 (m, 1H), 2.07-1.94 (m, 7H), 1.82-1.74 (m, 1H), 1.67-1.63 (m, 2H), 1.46-1.37 (m, 3H), 1.34-1.20 (m, 2H), 0.98 -0.95 (m, 3H), 0.84-0.79 (m, 3H). LC-MS: (ES+): m/z 975 $[M+H]^+$.

Measurements of SMARCA2/4 degradation using In-cell Western

22Rv1 cells were seeded at 8000/well in 96-well black/clear-bottom plates with 180ul DMEM containing 1% pen-strep, 1% HEPES and 10% FBS and incubated overnight at 37 °C to allow adherence. The next morning cells were treated with 20 uL of 10× compound and incubated for an additional 20 h. Cells were then washed with ice cold DPBS, then 50 uL of ice cold 4% PFA (Electron Microscopy Sciences 15711)/DPBS was added, and plates were then incubated at RT for 20 min. PFA was then removed and 200 uL of TBS-T containing 0.5% Triton X 100 (Sigma T8787) was added. Plates were incubated at RT for 30 min. TBS-T/Triton X was then removed, 50 uL of Li-cor blocking solution (Li-Cor 927-50003) was added, and plates were incubated at RT for 1 h. Blocking solution was removed, 50 uL of Li-Cor blocking solution containing primary antibody to SMARCA2 (1:2000) and α-Tubulin (1:2000) or SMARCA4 (1:1000) and α-Tubulin (1:2000) was added, and plates were incubated at 4 °C overnight. The next day, plates were washed 3× with TBS-T and 50uL of secondary antibody cocktail in Li-Cor blocking solution was added (IRDye 800CW Goat anti-rabbit IgG, Li-Cor 926-32211 and IRDye 680RD Goat anti-mouse IgG, Li-Cor 926-68070; both secondary antibodies are diluted 1:5000). Plates were incubated at RT for 1 h protected from light. Then plates were washed twice with TBS-T and excess liquid removed. Plates were read on the Li-Cor Odyssey with default intensity setting of 5.0 for both channels. Li-Cor images were analyzed with the in-cell Western feature of Image Studio Lite. Assays were run with technical duplicates and multiple biological replicates (>2) with error calculated with GraphPad prism using 95% confidence interval.

Acknowledgments

We would like to thank Xinpeng Wang, Fang Zhang, Mingtao He, Yunxing Cheng and Michael Berlin (Arvinas) for the synthesis of SMARCA2.4_1 and Peter Dragovich for guiding the synthesis. We would like to thank Jayaram Kancherla for his help and advice.

Figure legends

Figure 1. Figure constructs GRNs to infer TF activity at the single cell level

Schematics of Epiregulon. RNA and ATAC counts of similar cells are first aggregated. Based on the correlation between ATAC and RNA, each target gene (TG) is assigned a regulatory element (RE) within a defined distance (default ±250kb). Each RE is interrogated for possible TF occupancy using ChIP-seq data from ENCODE and ChIP-Atlas. An optional network pruning step retains TF-RE-TG triples that co-occur within the same cells. Co-occurrence is defined by TF and TG expression and chromatin accessibility exceeding a threshold. The regulatory potential of TF for each TG is estimated using 1 of the 3 methods (correlation, mutual information and Wilcoxon) and this serves as the weight of each TF-TG pair. The weights are computed for all the cells and/or for cells in each cluster to yield cluster-specific weights. The TF activity is computed as the weighted sum of the target gene expression normalized by the number of genes with non-zero edges. The regulon of each TF can be annotated with gene set enrichment. Jaccard similarity between the target genes is used to identify transcription factors that potentially interact. Differential network is performed on edge-subtracted networks of two cell states.

Figure 2: Epiregulon constructs GRNs to infer TF activity at the single cell level

a) UMAP representation of LNCaP cells transduced with virus encoding GATA6 and NKX2-1. The cells were infected in individual wells, hashtagged with HTO and then pooled into a single run.
b) Shown is the level of TF gene expression or TF activity computed by Epiregulon for GATA6 and NKX2-1
c) Each cell was identified by the HTO tag corresponding to the well receiving virus encoding a particular TF and this information serves as the ground truth. For each TF, a cell was classified into either expressing this TF or not based on the TF activity. AUROC was computed for all the packages being benchmarked as well as for gene expression of the TF being evaluated. Note that because each TF had different transduction efficacy and not all cells were able to take up the virus or express the TF, the theoretical maximum of AUROC will not reach 1.
d) Number of TFs retrieved from the different packages being benchmarked
e) As a validation, LNCaP cells were transduced with dox-inducible constructs expressing either NKX2-1 or eGFP control. Correlation between weights of NKX2-1 targets (computed by the Wilcoxon method) against logFC of NKX2-1 targets between no dox and dox treatments.
f) Genes most similar to GATA6 were ranked by Jaccard similarity between the GATA6 regulon and regulons of all the other TFs.
g) Chromatin accessibility at neuroendocrine (NE) - and stem cell like (SCL)- specific regions defined by Tang et al was computed by ChromVAR
h) Shown are the correlation between GATA6 activity and chromatin accessibility at regions specific to stem cell like (SCL) subtype of prostate cancer, and the correlation between NKX2-1 activity and chromatin accessibility at neuroendocrine (NE) subtype of prostate cancer.

Figure 3: Epiregulon predicts responses towards TF-targeting drugs

a) UMAP representation of scMultiome data of 4 prostate cancer cell lines treated with 3 drug compounds. VCaP and LNCaP are AR-dependent whereas DU145 and H660 are AR-independent. The 3 drugs include an AR antagonist (enzalutamide), an AR degrader (ARV110) and a degrader of SMARCA2/4, the ATPase subunit of the SWI/SNF remodeling complex (SMARCA2.4_1). Enzalutamide and ARV110 are known and specific inhibitors of AR activity and were used for benchmarking.
b) AR activity computed by Epiregulon.
c) Each cell was identified by the HTO tag corresponding to the well receiving DMSO or one of the two AR inhibitors and this information served as the ground truth. For each of the two drugs (enzalutamide or ARV110), a cell was classified into either treated with DMSO control or an AR inhibitor based on AR activity. AUROC was computed for all the packages being benchmarked as well as for AR gene expression.
d) AR expression of LNCaP and VCaP cells treated with the 3 drugs
e) Number of TFs retrieved by each of the packages

Figure 4: Comparison of ChIP-seq and motif annotations

a) Each cell was identified by the HTO tag corresponding to the well receiving virus encoding a particular TF and this information served as the ground truth cell labels. For each TF, a cell was classified into either expressing this TF or not. GATA6 ChIP-seq was obtained by merging ChIP-seq from ChIP-atlas and ENCODE. Merged ChIP + motif refers to merged ChIP-seq peaks that contain GATA6 motifs. We trained a DNN model on the cluster of cells expressing GATA6 using Basenji2 and predicted an importance score for each motif based on the difference between the original sequence and motif occluded sequence. We filtered for those motifs with importance scores higher than quartiles of scores.
b) Same as a) but computed for NKX2-1 using Nkx2.1.Homeobox_182 as the motif.
c) Each cell was identified by the HTO tag corresponding to the well receiving DMSO or the AR degrader ARV110 and this information serves as the ground truth cell labels. Public ChIP-seq was derived from merged AR ChIP-seq from ChIP-atlas and ENCODE. Motif was derived from the AR motif from CisBP. Merged ChIP-seq + motif refers to merged public ChIP-seq peaks that contain AR motifs. We trained a DNN model on cells treated with DMSO using Basenji2 and predicted an importance score for each motif based on the difference between the original sequence and motif occluded sequence. We filtered for those motifs with importance scores higher than quartiles of scores.

Figure 5: Epiregulon predicts responses towards chromatin-modifying drugs

a) SMARCA4 activity computed by Epiregulon
b) Hypergeometric enrichment of gene sets (hallmark and oncogenic signature) of the target genes of SMARCA4 inferred by Epiregulon for each of the 4 prostate cancer cell lines. VCaP and LNCaP are AR-independent and DU145 and H660 are AR-independent.
c) Potential interaction partners of SMARCA4 were nominated using Jaccard similarity between its target genes and the target genes of all other transcription factors. The top interacting genes are reflective of the unique etiologies of the different prostate cancer subtypes.

Figure 6: Epiregulon-constructed GRNs predict effects of perturbation

a) UMAP representation of PBMC data with unpaired and integrated scATAC-seq and scRNA-seq for which Epiregulon was used to compute the GRN.
b) Representation of TF expression or activity inferred by Epiregulon of well-known lineage factors (EBF1, PAX5 - B cells, GATA3 - T cells and SPI1- myeloid cells). Weights were estimated using the Wilcoxon method.
c) Cell-type specific GRN weights were computed for PreB and CD4 memory T cells (CD4.M) respectively. The GRN weights are cell type aware since only PreB cells, not CD4 memory T cells, have positive weights in the target genes of the B cell lineage factor, EBF1. Weight cutoff was >0.
d) EBF1 was interrogated for potential interaction partners using Jaccard similarity between its target genes and the target genes of all other transcription factors. The top genes (PAX5, IRF4 and TCF3 (also known as E2A)) are known to interact with EBF1 in B cell development.

e) Differential networks between monocytes and CD4 memory T cells were computed by subtracting the edges of the cell type specific GRNs. Genes were ranked by degree centrality of the edge-subtracted graph. SPI1 and CEBPA are well known lineage factors of the myeloid lineage whereas TBX21 and GATA3 are well known lineage factors of the T cell lineage.

f) The GRN computed from the PBMC data was applied onto a CRISPRa dataset of K562 cells containing only gene expression data (no scATAC-seq). Each cell was identified by the guide RNA corresponding to a TF that was being activated and this information serves as the ground truth. For each TF, a cell was classified into containing either the TF activating guide or the non-targeting control based on its TF activity. Shown is the area under the receiver operating curves (AUROC) using either TF gene expression or TF activity computed from GRN constructed from the PBMC dataset as the classifier. AUROC is also calculated for activity inferred from other GRNs (Dorothea, CollecTRI or humanbase).

g) TF activity inferred by epiregulon could better classify PRDM1-activated cells from control cells than gene expression or TF activity estimated from the Dorothea regulons. TF activity avoids zero inflation as seen with single cell gene expression.

Figure S1

a) AUROC of the reprogram-seq dataset was computed with each of the 3 weight methods offered by Epiregulon. Correlation (corr) and mutual information (MI) weights are computed either using 1) transcription factor (TF) expression vs target gene (TG) expression or 2) TF expression multiplied with the chromatin accessibility at regulatory elements (RE) vs TG expression. Each cell is identified by the HTO tag corresponding to the well receiving virus encoding a particular TF and this information serves as the ground truth. For each TF, a cell is classified into either expressing this TF or not.

b) Differences in rank between target and non-target genes with respect to GATA6 and NKX2-1 transcription factors. Gene rank was created from the product of two ranks: one based on the differences in the mean expression between the treatment and control groups, and the second based on the proportion of the total gene expression across all the cells contributed by the treated cells.

c) Representation of target genes of NKX2-1 and GATA6. C1 cluster contains cells infected with GATA6, C3 contains cells infected with NKX2-1 and C5 contains uninfected cells. Edges with weight > 0 were retained and target genes of NKX2-1 and GATA6 were shown for each cluster.

d) As a validation, LNCaP cells were transduced with dox-inducible constructs expressing either NKX2-1 or eGFP control. Shown are the putative target genes of NKX2-1 ranked by weights.

e) Representation of target genes of YAP1 and GATA6 with C1-specific weights.

f) Chromatin accessibility at adenocarcinoma- and neuroendocrine-specific regions defined by Baca et al was computed by ChromVAR

Figure S2

a) Chemical structure of SMARCA2.4_1, a degrader of SMARCA2 and SMARCA4
b) Levels of SMARCA2 and SMARCA4 after SMARCA2.4_1 treatment. SMARCA2 DC50 = 0.22nM, SMARCA4 DC50 = 0.31 nM in 22Rv1 cells.
c) AUROC of the AR dataset was computed with each of the 3 weight methods offered by Epiregulon. Correlation (corr) and mutual information weights are computed using either 1) transcription factor (TF) expression vs target gene (TG) expression or 2) TF expression multiplied with the chromatin accessibility at regulatory elements (RE) vs TG expression.
d) TF activity was computed with either AR ChIP-seq merged across ENCODE and ChIP-atlas or with LNCaP or VCaP AR and FOXA1 ChIP-seq (cell line-matched). There is a slight discrepancy between the cell culture condition used to obtain scMultiome data (in this manuscript) vs the condition used to obtain cell line-matched ChIP-seq. The cell line-matched ChIP-seq was obtained in cell lines cultured in charcoal stripped medium supplemented with R1881 whereas the scMultiome data was performed in cell lines cultured in 10% fetal bovine serum (FBS) but the identities of the cell lines are matched.

Figure S3

a) Same as Figure 4a but performed with ChromBPNet
b) Same as Figure 4b but performed with ChromBPNet
c) To evaluate whether GRN weights improves performance of TF activity, weights of target genes were randomly permuted and AUROC is computed with GRNs with permuted weights
d) Same as S3c) but performed on NKX2-1

Figure S4

a) AUROC of the AR dataset was computed with each of the 3 weight methods offered by Epiregulon. Correlation (corr) and mutual information (MI) weights are computed either using transcription factor (TF) expression vs target gene (TG) expression or TF expression multiplied with the chromatin accessibility at regulatory elements (RE) vs TG expression. Each cell is identified by the HTO tag corresponding to the well receiving DMSO or the SMARCA2/4 degrader SMARCA2.4_1 and this information serves as the ground truth. A cell is classified into either treated with DMSO control or SMARCA2.4_1 based on the SMARCA4 activity inferred by Epiregulon.
b) TF activity was computed with either SMARCA4 ChIP-seq merged across ENCODE and ChIP-atlas or with VCaP SMARCA4 ChIP-seq (cell line-matched). There is a slight discrepancy between the cell culture condition used to obtain scMultiome data (in this manuscript) vs the condition used to obtain cell line-matched ChIP-seq. The cell line-matched ChIP-seq was obtained in cell lines cultured in charcoal stripped medium supplemented with R1881 whereas the scMultiome data was performed in cell lines cultured in 10% fetal bovine serum (FBS) but the identities of the cell lines are matched.

Figure S5

a) Gene set enrichment of regulons inferred by Epiregulon (Wilcoxon weight cutoff > 0). EBF1 and PAX5 are B cell lineage factors. GATA3 is a lymphoid lineage factor. SPI1 is a myeloid lineage factor.

b) Shown is the area under the receiver operating curves (AUROC) using either TF gene expression or TF activity computed from GRN obtained from humanase, collecTRI or Dorothea.

Figure S6

a) We used the scMultiome package to simulate paired scRNA-seq and scATAC-seq data (see Supplementary Note). Briefly, we provided a lineage tree that specified the identities of the individual cells and a ground truth GRN that specified the regulatory relationships between TF-TG and RE-TG. Gene expression and peak matrices were generated taking into account both cell identities and gene regulatory relationships. We obtained the ground truth activity of each TF by summing the differences in target gene expression as defined by the GRN. For all assessments from b-e, we use the fixed cutoff and tf_re.merge set to TRUE.
b) We varied the degree of sparsity by increasing the technical noise and evaluated the correlation of activity against ground truth TF activity across all genes for each of the weight methods. Each data point represents a TF. The ATAC effect was set at basic. GRN had 100% true connections.
c) We added noise to GRN by adding false connections and removing true connections, and correlated the activity against ground truth activity for each TF. The ATAC effect was set at basic and data sparsity was set at medium.
d) Cell clusters were first defined using k-means clustering with a median cluster size of 10 cells for the Wilcoxon and logFC methods. We correlated the activity against ground truth activity for each TF. The ATAC effect was set at basic and data sparsity was set at medium. GRN had 100% true connections.
e) We varied the effect of ATAC on the gene expression, with the strong ATAC effect indicating a greater importance of ATAC counts on shaping gene expression counts. GRN had 100% true connections. True counts were generated without data sparsity and observed counts were generated with medium data sparsity.
f) We spiked in 100% false connections to the true GRN. We examined the recall and precision of true GRN connections with varying degrees of z score cutoffs computed by chi-square statistics
g) We spiked in 100% false connections to the true GRN. We show that true connections have increased RE-TG correlations and this translated to higher weights in the true connections.
h) We evaluated the effects of pruning on the accuracy of TF activity predictions. We computed the ground truth TF activity from the predefined GRN with high data sparsity and strong ATAC effects. We spiked in 100% false connections and compared the correlation between the TF activity inferred from either the unpruned or pruned GRN vs the ground truth GRN. Shown is the difference between the two correlations. A positive difference indicates improved performance due to pruning.

Figure S7

a) We added noise to the network by increasing proportions of randomly permuted edges on top of true differential edges (15% of total edges). Centrality subtraction was computed by

subtracting the degree centralities of two networks whereas edge subtraction was computed by centrality of the edge subtracted network. See methods for details.

b) We varied the signal strength by increasing the proportions of altered edges (enrichment, depletion or permutations) with noise fixed at 10% of total edges.

Code availability

Epiregulon: https://github.com/xiaosaiyao/epiregulon
Epiregulon.extra: https://github.com/xiaosaiyao/epiregulon.extra
Epiregulon.archr: https://github.com/xiaosaiyao/epiregulon.archr
ScMultiome: https://github.com/xiaosaiyao/scMultiome
Epiregulon benchmarking: https://github.com/xiaosaiyao/epiregulon.benchmark
Deep neural network model: https://github.com/xiaosaiyao/epiregulon.sequence.modeling
Differential network simulation: https://github.com/xiaosaiyao/epiregulon.diffnetwork.simulation

1 Tran, C. *et al.* Development of a second-generation antiandrogen for treatment of advanced prostate cancer. *Science* **324**, 787-790, doi:10.1126/science.1168175 (2009).

2 Sakamoto, K. M. *et al.* Protacs: chimeric molecules that target proteins to the Skp1-Cullin-F box complex for ubiquitination and degradation. *Proc Natl Acad Sci U S A* **98**, 8554-8559, doi:10.1073/pnas.141230798 (2001).

3 Snyder, L. B. *et al.* Discovery of ARV-110, a first in class androgen receptor degrading PROTAC for the treatment of men with metastatic castration resistant prostate cancer. *Cancer Res* **81**, 43 (2021).

4 Hagenbeek, T. J. *et al.* An allosteric pan-TEAD inhibitor blocks oncogenic YAP/TAZ signaling and overcomes KRAS G12C inhibitor resistance. *Nat Cancer* **4**, 812-828, doi:10.1038/s43018-023-00577-0 (2023).

5 Badia, I. M. P. *et al.* Gene regulatory network inference in the era of single-cell multi-omics. *Nat Rev Genet*, doi:10.1038/s41576-023-00618-5 (2023).

6 Alvarez, M. J. *et al.* Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat Genet* **48**, 838-847, doi:10.1038/ng.3593 (2016).

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

8 Fleck, J. S. *et al.* Inferring and perturbing cell fate regulomes in human brain organoids. *Nature* **621**, 365-372, doi:10.1038/s41586-022-05279-8 (2023).

9 Kartha, V. K. *et al.* Functional inference of gene regulation using single-cell multi-omics. *Cell Genom* **2**, doi:10.1016/j.xgen.2022.100166 (2022).

10 Kamal, A. *et al.* GRaNIE and GRaNPA: inference and evaluation of enhancer-mediated gene regulatory networks. *Mol Syst Biol* **19**, e11627, doi:10.15252/msb.202311627 (2023).

11 Bravo Gonzalez-Blas, C. *et al.* SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nat Methods* **20**, 1355-1367, doi:10.1038/s41592-023-01938-4 (2023).

12 Granja, J. M. *et al.* ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat Genet* **53**, 403-411, doi:10.1038/s41588-021-00790-6 (2021).
13 Inukai, S., Kock, K. H. & Bulyk, M. L. Transcription factor-DNA binding: beyond binding site motifs. *Curr Opin Genet Dev* **43**, 110-119, doi:10.1016/j.gde.2017.02.007 (2017).
14 Worsley Hunt, R. & Wasserman, W. W. Non-targeted transcription factors motifs are a systemic component of ChIP-seq datasets. *Genome Biol* **15**, 412, doi:10.1186/s13059-014-0412-4 (2014).
15 Duan, J. *et al.* Rational Reprogramming of Cellular States by Combinatorial Perturbation. *Cell Rep* **27**, 3486-3499 e3486, doi:10.1016/j.celrep.2019.05.079 (2019).
16 Baca, S. C. *et al.* Reprogramming of the FOXA1 cistrome in treatment-emergent neuroendocrine prostate cancer. *Nat Commun* **12**, 1979, doi:10.1038/s41467-021-22139-7 (2021).
17 Han, M. *et al.* FOXA2 drives lineage plasticity and KIT pathway activation in neuroendocrine prostate cancer. *Cancer Cell* **40**, 1306-1323 e1308, doi:10.1016/j.ccell.2022.10.011 (2022).
18 Tang, F. *et al.* Chromatin profiles classify castration-resistant prostate cancers suggesting therapeutic targets. *Science* **376**, eabe1505, doi:10.1126/science.abe1505 (2022).
19 Xiao, L. *et al.* Targeting SWI/SNF ATPases in enhancer-addicted prostate cancer. *Nature* **601**, 434-439, doi:10.1038/s41586-021-04246-z (2022).
20 Kelley, D. R. Cross-species regulatory sequence activity prediction. *PLoS Comput Biol* **16**, e1008050, doi:10.1371/journal.pcbi.1008050 (2020).
21 Avsec, Z. *et al.* Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat Genet* **53**, 354-366, doi:10.1038/s41588-021-00782-6 (2021).
22 Avsec, Z. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods* **18**, 1196-1203, doi:10.1038/s41592-021-01252-x (2021).
23 Cyrta, J. *et al.* Role of specialized composition of SWI/SNF complexes in prostate cancer lineage plasticity. *Nat Commun* **11**, 5549, doi:10.1038/s41467-020-19328-1 (2020).
24 Beltran, H. *et al.* Divergent clonal evolution of castration-resistant neuroendocrine prostate cancer. *Nat Med* **22**, 298-305, doi:10.1038/nm.4045 (2016).
25 Ku, S. Y. *et al.* Rb1 and Trp53 cooperate to suppress prostate cancer lineage plasticity, metastasis, and antiandrogen resistance. *Science* **355**, 78-83, doi:10.1126/science.aah4199 (2017).
26 Bluemn, E. G. *et al.* Androgen Receptor Pathway-Independent Prostate Cancer Is Sustained through FGF Signaling. *Cancer Cell* **32**, 474-489 e476, doi:10.1016/j.ccell.2017.09.003 (2017).
27 Mesman, S. & Smidt, M. P. Tcf12 Is Involved in Early Cell-Fate Determination and Subset Specification of Midbrain Dopamine Neurons. *Front Mol Neurosci* **10**, 353, doi:10.3389/fnmol.2017.00353 (2017).
28 Liu, W. *et al.* BRD4 regulates Nanog expression in mouse embryonic stem cells and preimplantation embryos. *Cell Death Differ* **21**, 1950-1960, doi:10.1038/cdd.2014.124 (2014).
29 Choi, H. I. *et al.* BET inhibitor suppresses migration of human hepatocellular carcinoma by inhibiting SMARCA4. *Sci Rep* **11**, 11799, doi:10.1038/s41598-021-91284-2 (2021).
30 Rahman, S. *et al.* The Brd4 extraterminal domain confers transcription activation independent of pTEFb by recruiting multiple proteins, including NSD3. *Mol Cell Biol* **31**, 2641-2652, doi:10.1128/MCB.01341-10 (2011).
31 Xu, X. *et al.* Bromodomain Containing Protein 4 (BRD4) Regulates Expression of its Interacting Coactivators in the Innate Response to Respiratory Syncytial Virus. *Front Mol Biosci* **8**, 728661, doi:10.3389/fmolb.2021.728661 (2021).

32 Alver, B. H. *et al.* The SWI/SNF chromatin remodelling complex is required for maintenance of lineage specific enhancers. *Nat Commun* **8**, 14648, doi:10.1038/ncomms14648 (2017).
33 Blumli, S. *et al.* Acute depletion of the ARID1A subunit of SWI/SNF complexes reveals distinct pathways for activation and repression of transcription. *Cell Rep* **37**, 109943, doi:10.1016/j.celrep.2021.109943 (2021).
34 Sobczak, M., Pitt, A. R., Spickett, C. M. & Robaszkiewicz, A. PARP1 Co-Regulates EP300-BRG1-Dependent Transcription of Genes Involved in Breast Cancer Cell Proliferation and DNA Repair. *Cancers (Basel)* **11**, doi:10.3390/cancers11101539 (2019).
35 Shang, X. Y. *et al.* ARID1A deficiency weakens BRG1-RAD21 interaction that jeopardizes chromatin compactness and drives liver cancer cell metastasis. *Cell Death Dis* **12**, 990, doi:10.1038/s41419-021-04291-6 (2021).
36 Lazar, J. E. *et al.* Global Regulatory DNA Potentiation by SMARCA4 Propagates to Selective Gene Expression Programs via Domain-Level Remodeling. *Cell Rep* **31**, 107788, doi:10.1016/j.celrep.2020.107788 (2020).
37 Marino, M. M. *et al.* Interactome mapping defines BRG1, a component of the SWI/SNF chromatin remodeling complex, as a new partner of the transcriptional regulator CTCF. *J Biol Chem* **294**, 861-873, doi:10.1074/jbc.RA118.004882 (2019).
38 Liao, L. *et al.* Multiple tumor suppressors regulate a HIF-dependent negative feedback loop via ISGF3 in human clear cell renal cancer. *eLife* **7**, e37925, doi:10.7554/eLife.37925 (2018).
39 Granja, J. M. *et al.* Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nat Biotechnol* **37**, 1458-1465, doi:10.1038/s41587-019-0332-7 (2019).
40 Ungerback, J., Ahsberg, J., Strid, T., Somasundaram, R. & Sigvardsson, M. Combined heterozygous loss of Ebf1 and Pax5 allows for T-lineage conversion of B cell progenitors. *J Exp Med* **212**, 1109-1123, doi:10.1084/jem.20132100 (2015).
41 Yeamans, C. *et al.* C/EBPalpha binds and activates the PU.1 distal enhancer to induce monocyte lineage commitment. *Blood* **110**, 3136-3142, doi:10.1182/blood-2007-03-080291 (2007).
42 Thomas, G. D. *et al.* Deleting an Nr4a1 Super-Enhancer Subdomain Ablates Ly6C(low) Monocytes while Preserving Macrophage Gene Function. *Immunity* **45**, 975-987, doi:10.1016/j.immuni.2016.10.011 (2016).
43 Kanhere, A. *et al.* T-bet and GATA3 orchestrate Th1 and Th2 differentiation through lineage-specific targeting of distal regulatory elements. *Nat Commun* **3**, 1268, doi:10.1038/ncomms2260 (2012).
44 Newton, R. H. *et al.* Maintenance of CD4 T cell fitness through regulation of Foxo1. *Nat Immunol* **19**, 838-848, doi:10.1038/s41590-018-0157-4 (2018).
45 Gilbert, L. A. *et al.* Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell* **159**, 647-661, doi:10.1016/j.cell.2014.09.029 (2014).
46 Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D. & Saez-Rodriguez, J. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res* **29**, 1363-1375, doi:10.1101/gr.240663.118 (2019).
47 Müller-Dott, S. *et al.* Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities. *bioRxiv*, 2023.2003.2030.534849, doi:10.1101/2023.03.30.534849 (2023).
48 Greene, C. S. *et al.* Understanding multicellular function and disease with human tissue-specific networks. *Nat Genet* **47**, 569-576, doi:10.1038/ng.3259 (2015).
49 Gaublomme, J. T. *et al.* Nuclei multiplexing with barcoded antibodies for single-nucleus genomics. *Nat Commun* **10**, 2907, doi:10.1038/s41467-019-10756-2 (2019).

50 Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat Methods* **14**, 975-978, doi:10.1038/nmeth.4401 (2017).

51 Weirauch, M. T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431-1443, doi:10.1016/j.cell.2014.08.009 (2014).

52 Kamimoto, K. *et al.* Dissecting cell identity via network inference and in silico gene perturbation. *Nature* **614**, 742-751, doi:10.1038/s41586-022-05688-9 (2023).

53 Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep* **9**, 9354, doi:10.1038/s41598-019-45839-z (2019).

Figure 1: Epiregulon constructs GRNs to infer TF activity at the single cell level

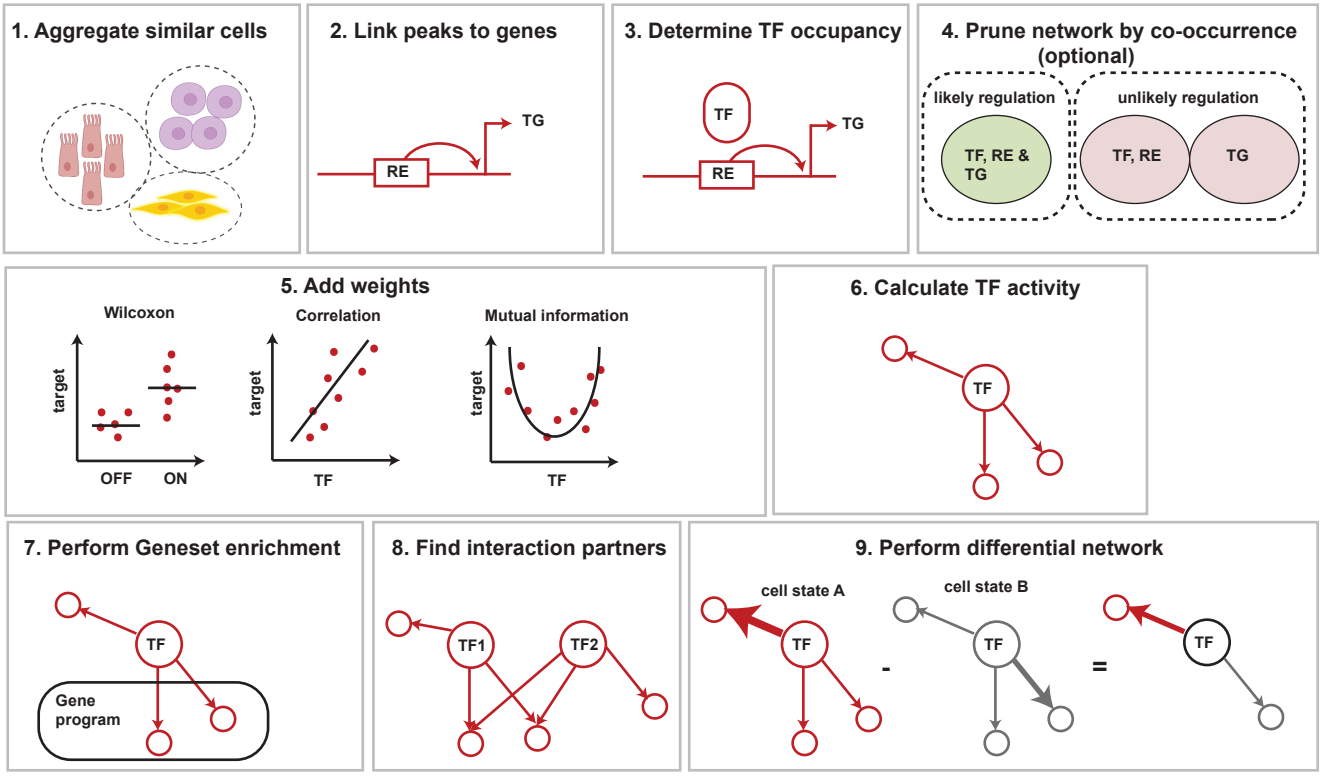


Figure 2: Epiregulon constructs GRNs to infer TF activity at the single cell level

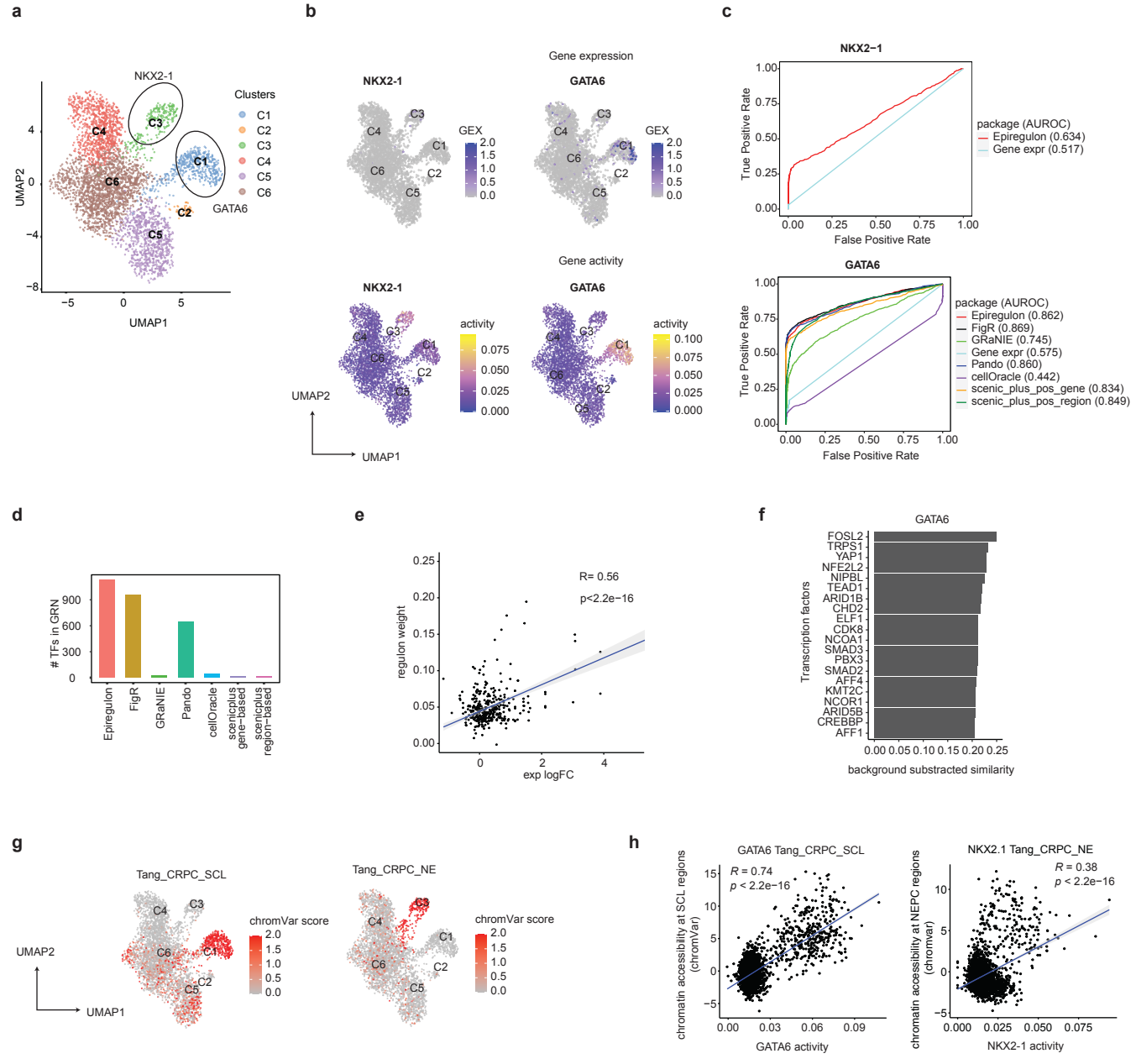


Figure 3: Epiregulon predicts responses towards TF-targeting drugs

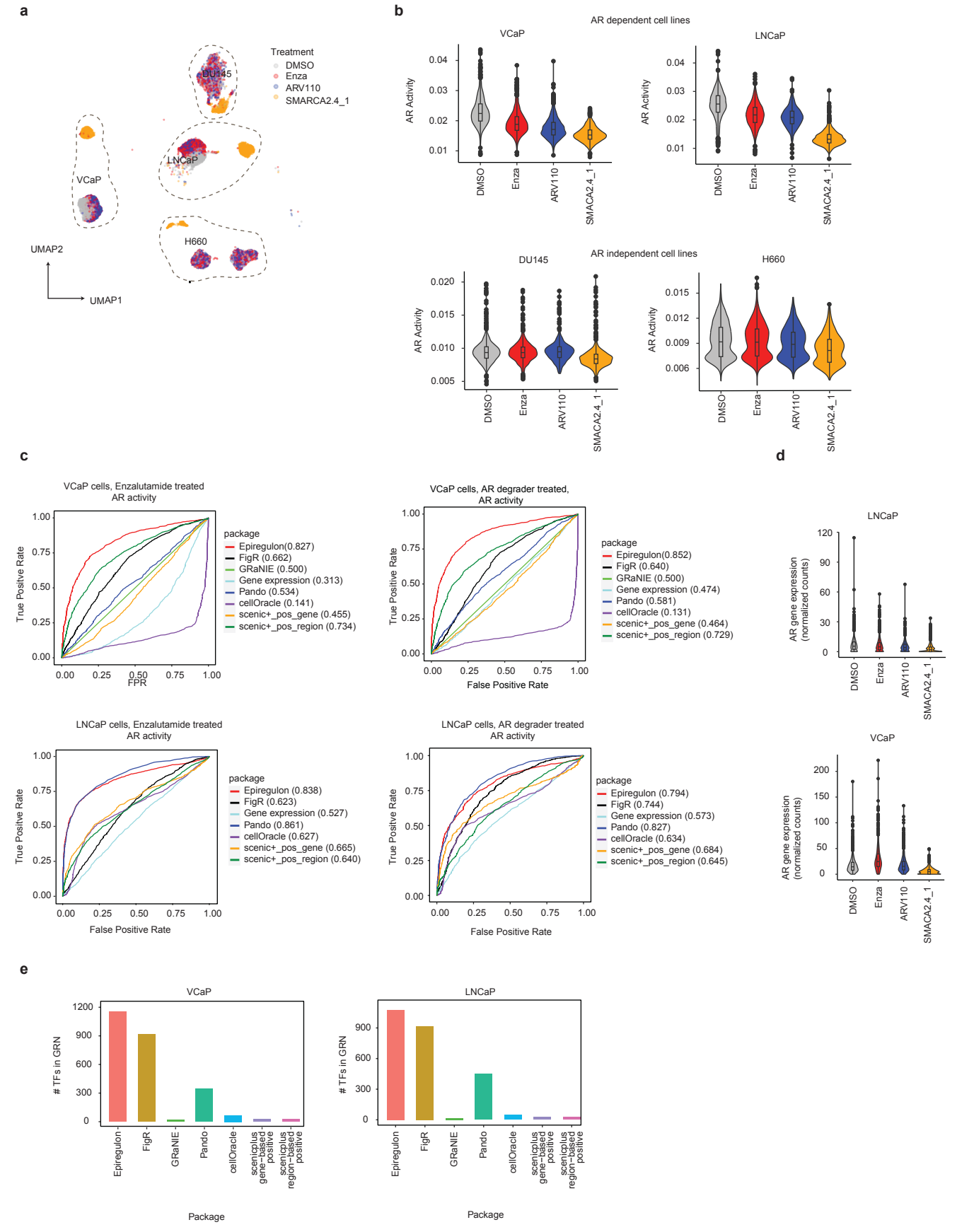


Figure 4: Comparison of ChIP-seq and motif annotations

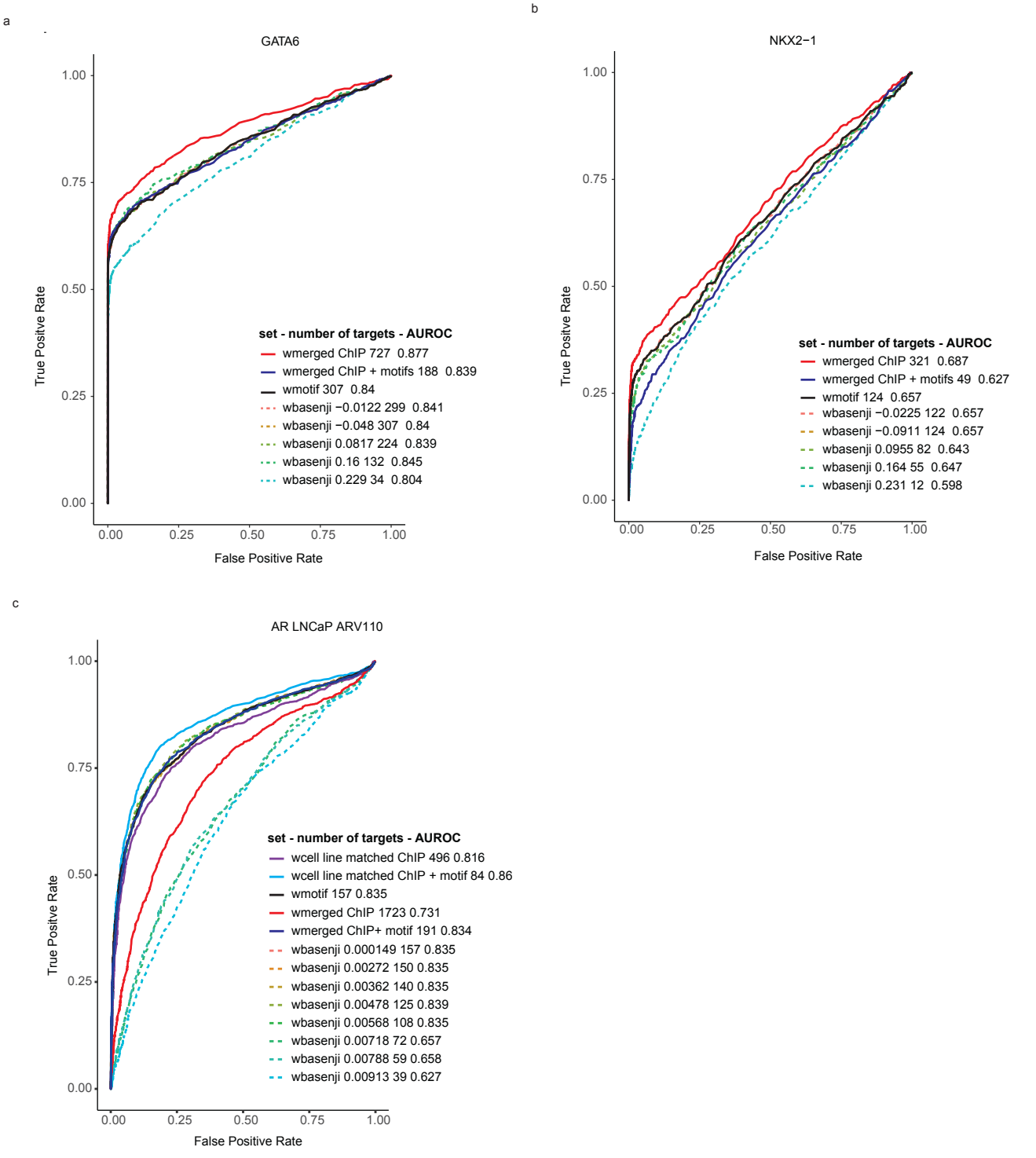


Figure 5: Epiregulon predicts responses towards chromatin-modifying drugs

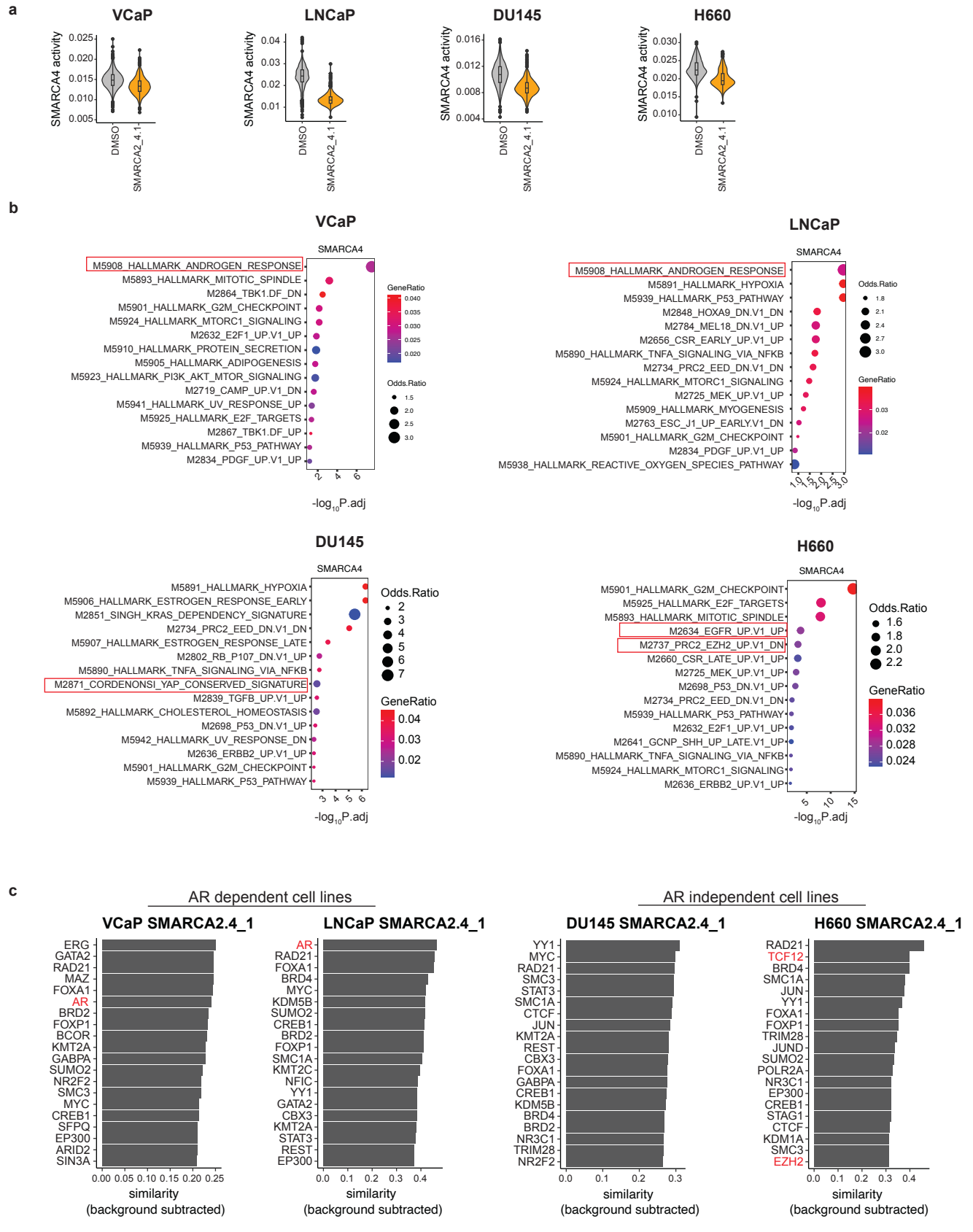
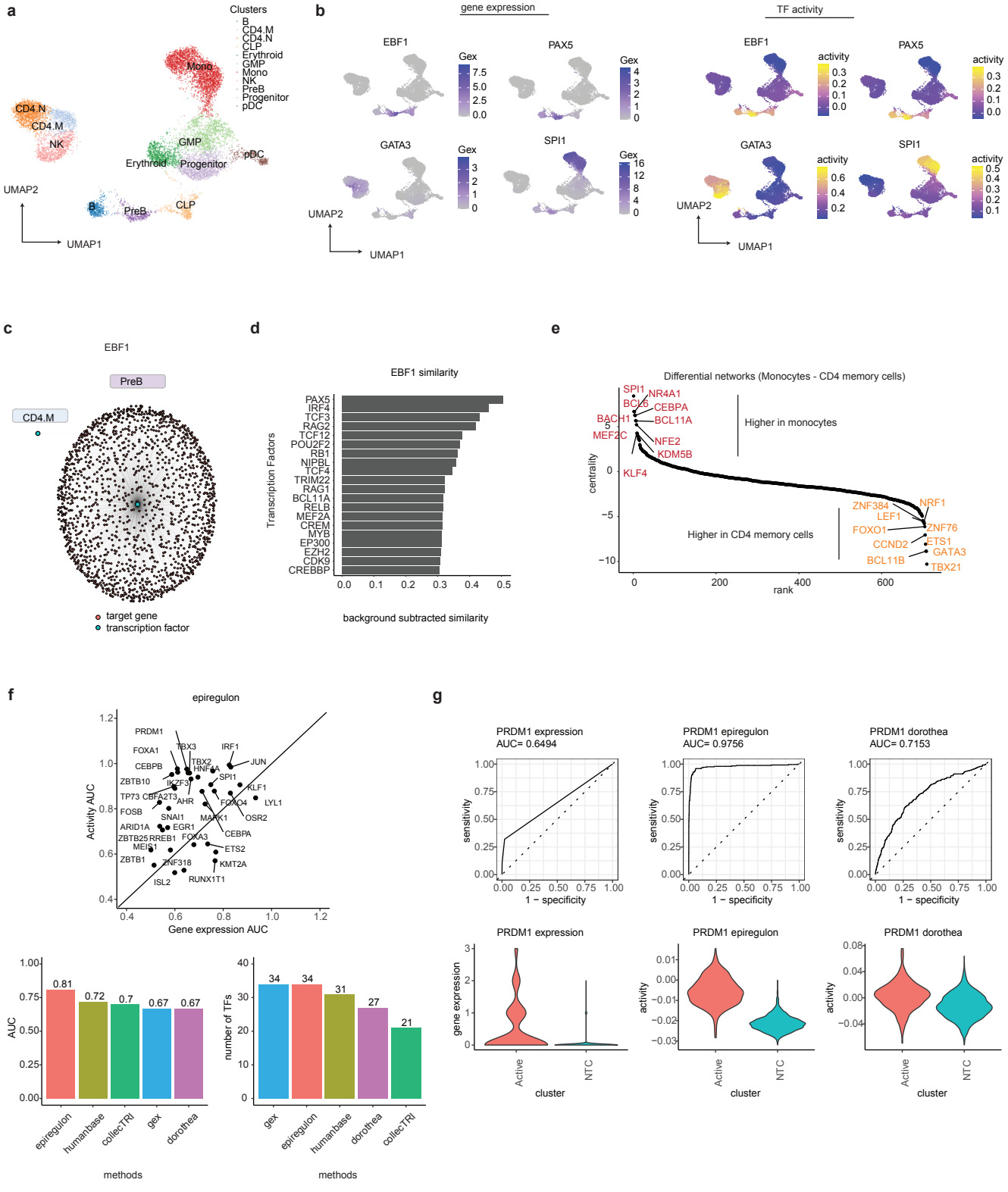



Figure 6: Epiregulon-constructed GRNs predict effects of perturbations