On the identification of differentially-active transcription factors from ATAC-seq data

Felix Ezequiel Gerbaldo[1☯], Emanuel Sonder[1,2,3,5☯], Vincent Fischer[4], Selina Frei[4], Jiayi Wang[3], Katharina Gapp[4], Mark D. Robinson[3,5], Pierre-Luc Germain[1,3,5*]

1 Computational Neurogenomics, D-HEST Institute for Neurosciences, Zürich, Switzerland
2 Systems Neuroscience, D-HEST Institute for Neurosciences, Zürich, Switzerland
3 Department of Molecular Life Sciences, University of Zürich, Zürich, Switzerland
4 Epigenetics and Neuroendocrinology, D-HEST Institute for Neurosciences, Zürich, Switzerland
5 SIB Swiss Institute of Bioinformatics, University of Zurich, Switzerland

☯These authors contributed equally to this work. * pierre-luc.germain@hest.ethz.ch

Abstract

ATAC-seq has emerged as a rich epigenome profiling technique, and is commonly used to identify Transcription Factors (TFs) underlying given phenomena. A number of methods can be used to identify differentially-active TFs through the accessibility of their DNA-binding motif, however little is known on the best approaches for doing so. Here we benchmark several such methods using a combination of curated datasets with various forms of short-term perturbations on known TFs, as well as semi-simulations. We include both methods specifically designed for this type of data as well as some that can be repurposed for it. We also investigate variations to these methods, and identify three particularly promising approaches (a chromVAR-limma workflow with critical adjustments, monaLisa and a combination of GC smooth quantile normalization and multivariate modeling). We further investigate the specific use of nucleosome-free fragments, the combination of top methods, and the impact of technical variation. Finally, we illustrate the use of the top methods on a novel dataset to characterize the impact on DNA accessibility of TRAnscription Factor TArgeting Chimeras (TRAFTAC), which can deplete TFs – in our case NFkB – at the protein level.

Author summary

Transcription factors regulate gene expression by binding sites in the genome that often harbor a specific DNA motif. The collective accessibility of these motif-matching regions, measured by technologies such as ATAC-seq, can be used to infer the activity of the corresponding transcription factors. Here we use curated datasets of 11 TF-specific perturbations as well as 116 semi-simulated datasets to benchmark various methods for identifying factors that differ in activity between experimental conditions. We investigate important variations in the analysis and make recommendations pertaining to such analysis. Finally, we illustrate the application of the top methods to characterize the effects of a novel method for perturbing transcription factors at the protein level.

1 Introduction

Transcription factors (TFs) regulate transcription by binding specific sites in the genome, which often harbor a corresponding DNA motif. Identifying which TFs regulate a given transcriptional signature is critical to a mechanistic understanding of gene expression. The expression of the TFs themselves can sometimes be a useful indicator of their involvement, especially between cell types and developmental stages. However, it is often a poor proxy for their activity in dynamic settings, where their activity is often regulated post-transcriptionally. For example, the TF ATF4 is critical to the integrated stress response, and its mRNA is already present in cells but its translation is blocked by an upstream uORF until a translation initiation factor is modified [1, 2]. Another example is the cAMP-response element binding protein (CREB), which is fundamental to a number of processes: CREB is already translated but its phosphorylation strongly enhances its DNA binding ability and allows for the recruitment of co-factors ([3]). Further common examples are nuclear receptors, which in general are outside the nucleus until activated by their ligand, upon which they translocate to the nucleus to affect transcription [4]. It is therefore critical to have means of measuring TF activity independently of their expression.

An obvious approach is to profile TF binding through ChIP-seq or derivative technologies, which is however time-consuming, can only be done one TF at a time, and depends on the availability of high-quality antibodies. DNA accessibility instead promises a less biased approach for the discovery of TFs underlying a change in cellular phenotype. In particular, the Assay for Transposase-Accessible Chromatin with sequencing (ATAC-seq) has emerged as a simple, fast accessibility assay that requires little starting material [5, 6], being even amenable to single-cell analysis [7]. Since TF binding is associated with DNA accessibility, such assays offer genome-wide evidence for TF binding, though without TF specificity. In conjunction with a compendium of DNA binding motif models, however, the activity of different TFs can in principle be resolved in terms of global accessibility of the motif matches (i.e. regions whose sequence matches the motif to a reasonable degree). A number of methods from transcriptomics can be repurposed for this task [8, 9, 10], and a number of others have been developed specifically for it [11, 12, 13, 14]. In general, they can be divided into two families: 1) 'fold-change-based' methods do not operate on individual ATAC-seq samples, but on a vector of per-peak fold-changes between conditions (e.g. obtained from differential accessibility analysis), while 2) 'sample-based' methods assign each motif model a score in each sample, which can then, if desired, be compared between experimental groups. To our knowledge, these methods were never thoroughly compared and evaluated.

Here we benchmark methods for differential TF activity inference from ATAC-seq. We curated a set of ATAC-seq datasets with replicates, profiling genome-wide accessibility upon perturbation (activation or repression) of a specific TF with an annotated DNA binding motif model. To limit the effect of indirect downstream effects, which would make the true perturbation (and hence the benchmark) ambiguous, we included only datasets in which the samples were profiled relatively early after the intervention. For ligand-based activation, which is very fast [15], we included only datasets up to 4h after intervention. For CRISPRi knockdowns, which take much longer to take effect, we used a cutoff of 72h. These cutoffs are obviously not perfect and were largely set in a compromise to gain a sufficient number of datasets while mitigating the impact of indirect effects. The included datasets are shown in Table 1, and the magnitude of peak-level changes in accessibility upon treatment are illustrated in Supplementary Fig S1. We further complemented these real datasets with semi-simulations incorporating perturbations of only one specific TF at a time using ChIP-seq derived peak coordinates.

target	perturbation	system	n/group	delay	ref
NR3C1	Agonist (Dexa.)	A549	3	1h	[15]
ESR1/2	Agonist (Estradiol)	MCF-7	3	45min	[16]
NR1H3	Agonist (GW3965)	Mouse liver	2	4h	[17]
NR1H4	Agonist (GW4064)	Mouse liver	3	4h	[17]
GATA1	CRISPRi KD	K562	3	72h	[18]
GATA2	CRISPRi KD	K562	3	72h	[18]
RUNX1	CRISPRi KD	K562	3	72h	[18]
RUNX2	CRISPRi KD	K562	3	72h	[18]
KLF1	CRISPRi KD	K562	3	72h	[18]
MYC	CRISPRi KD	K562	3	72h	[18]
BANP	dTAG KD	mouse ESC	2	2h	[19]

Table 1.

Description of the benchmark datasets.

2 Materials and methods

2.1 Datasets preprocessing

For the baseline data for simulation, as well as for the NR3C1 benchmark dataset, filtered alignments and peaks were downloaded from ENCODE (NR3C1: ENCSR385LRX; for the datasets related to the simulations, see Table 2).

For the SpearATAC-seq datasets (GSE168851: GATA1, GATA2, RUNX1, RUNX2, KLF1, MYC), the fragment file and processed SummarizedExperiment object were downloaded from GEO. We used the processed object to identify the 3-day cell barcodes assigned to each guide RNA. Then, we merged the fragments for each guide RNA, creating 21 pseudobulk samples (3 guides per 6 target genes + 3 control guide RNAs). For methods that specifically require BAM files, the fragment files were converted using bedToBam.

For other datasets (ESR1/2: GSE117940, NR1H3/4: GSE149075, BANP: GSE155604), and the in-house TRAFTAC dataset, we downloaded the raw reads (if published) from SRA, removed adapters using Trimmomatic 0.38 [20] with ILLUMINACLIP:$adapters:2:15:4:4:true LEADING:20 TRAILING:20 SLIDINGWINDOW:4:15 MINLEN:25, aligned using Bowtie2 2.3.4.3 [21] with –dovetail –no-mixed –no-discordant -I 15 -X 2000 against either GRCh38.p10 or GRCm38. Duplicated fragments were marked by using Picard 2.18.27, and peak calling was done using MACS2 2.2.5 using BAMPE (or BEDPE) mode. Peaks overlapping ENCODE blacklisted regions [22] were removed. Consensus peaks were obtained by merging the peak calls of the individual samples (i.e. GenomicRanges::reduce). Since some methods recommend resizing peaks to a common size, we resized peaks (to 300bp, which is recommended by chromVAR and was close to the median peak size) for all methods, and then generated peak count tables using 'chromVAR::getCounts'.

For nucleosome-free filtering, we used fragments between 30 and 120bp. Given the reduction in effective number of reads, we called peaks with MACS2 on the merged fragments across samples, using a q-value threshold of 0.01. We then resized peaks to 150bp (again because this was very close to the median peak size).

Peak-level differential analysis was performed with edgeR 3.32.1 [23].

2.2 Motif models and motif matching

For all methods, we used the HOCOMOCO 10-11 motif models [24] available in the MotifDb package version 1.32.0 [25]. Specifically, we queried for HOCOMOCO motif models of the respective species in the database and kept the newest and highest-quality motif for each gene symbol (see the 'getNonRedundantMotifs()' function of the repository). The BANP motif model was not available in any database and therefore, we obtained the published one [19] from the authors and added it manually to the collection. Similarly, NR1H3 had a human motif model, but no mouse model; we therefore complemented the mouse motif collection with the human one.

To avoid this being a source of variation between methods, whenever possible we tried to standardize the motif matching across methods, using a common fimo 4.11.2 [26] motif scan with default parameters. monaLisa, however, did not allow us to provide the motif matches and performed its own motif matching. While it is a little more stringent, its global number of motif matches was also within 10-20% different of the fimo matches, indicating that it is nevertheless fairly similar.

2.2.1 Clustering of motif models

For each species, the motif models were compared with the universalmotif package, using the pearson correlation of the Position Probability Matrices with positions weighted by total information content. Hierarchical clustering with complete linkage was performed on the distances (1 - correlation), and the dendrogram was cut at a height (0.35 for mouse and 0.45 for human, which has more motifs) established so that the merge of the clustered motifs did not have a correlation above 0.7. Motif merging was done with the ALLR method of the universalmotif package. The clustered and merged motifs are henceforth referred to as 'archetypes' (similar to [27]).

2.3 Differential motif activity methods and their application

2.3.1 chromVAR

While originally designed to analyze sparse chromatin-accessibility data from scATAC-seq data, chromVAR [11] is also commonly applied to bulk ATAC-seq data. The algorithm computes deviations between the summed fragment counts of peaks containing a match for a certain motif model and the expected number of fragments based on the average of all samples/cells, and then compares these deviations to that of random sets of (background) peaks with similar GC content and average accessibility as those containing motif matches, to adjust for technical biases. Specifically, the method produces two matrices of sample/cell- and motif-specific scores: bias-adjusted deviations, and z-scores compared to a background. These scores can then be used to perform differential analysis across groups of cells/samples. The package includes a function to this effect, 'chromVAR::differentialDeviations'. In addition, we tested a variation using 'limma::eBayes' [28] to test for differences in chromVAR activity scores between groups, either based on the adjusted deviations (denoted 'chromVAR(deviations)>limma') or on the z-scores (denoted 'chromVAR(z)>limma'), as well as different normalizations of these scores before differential analysis. Importantly, we observed that the results of the differential analysis were not reproducible across different random seeds unless we considerably increased the default number of background sets (Supplementary Fig S2A). All comparisons were therefore run with 2000 background iterations. Otherwise, the package's guidelines were followed.

2.3.2 monaLisa: Binned motif enrichment

MonaLisa [12] comes with two distinct algorithms, and the first follows a binned enrichment analysis approach. First, ATAC-seq peaks are grouped into bins according to their change in accessibility (i.e. log-fold-change, or logFC) across conditions. Over-representation of motif matches is then assessed for each bin, either in comparison to the bin centered around a logFC of zero, i.e. the peaks not changing (this variant is here denoted as 'monaLisa.vsZero'), or against all other bins (denoted 'monaLisa.vsOthers'). The method can also correct for differences in sequence composition including GC bias in the same fashion as HOMER [29]. In some rare cases, the binning procedure triggered an error due to ties in the logFCs; to avoid this, we therefore added a very small amount of noise to the logFC vector (jitter with a factor of 0.001). monaLisa produces a p-value for each bin and motif model, which is not easily comparable to alternative methods that produce a single p-value per motif. Since the bins' p-values cannot be deemed independent, we tested different p-value aggregation methods that did not assume so, including Simes, Cauchy [30], mean and geometric mean p-values and settled for Simes' as offering the best results in terms of both the rank of the true TF and the network score in the benchmark datasets (see metrics and ranking sections below).

2.3.3 monaLisa: Randomized Lasso stability selection

The second method included in monaLisa follows a randomized lasso stability selection. The goal here is to select TFs as predictors for the observed changes in accessibility. Due to the nature of the robust regression approach, similar motifs compete to be selected. To correct for GC bias, the GC content of the peak sequences can be included as an additional predictor. The model returns a selection probability for each predictor, which we used as a ranking of the motifs for benchmarking. The selection probability cannot be interpreted as a p-value, and since the authors recommend a threshold of 0.9 for stringent selection, we used this threshold for computing precision and recall.

2.3.4 diffTF

diffTF [13] is another method based on peaks logFC between conditions, but which bins peaks by GC content to be more robust to technical variations. Within each bin, the logFCs of peaks with matches for the given motif are compared to those without. It can be run in two modes: in the sample permutation mode, the differences are compared to differences between permuted groups of samples, while the analytic mode establishes p-values analytically. The latter is more appropriate for small sample sizes, as is the case in our benchmark datasets. We nevertheless tested both modes, and since (as expected) the analytic mode systematically performed better or equal to the permutation mode, the latter was omitted from the main figures. diffTF was used via the pre-configured Singularity container provided by the authors, following the quick start guide and using a maximum of 8 cores. To make the method comparable with alternatives, we provided the motif matches using the same scans as for other methods. Everything else was kept to the default settings.

2.3.5 VIPER and msVIPER

The viper package was especially developed to infer differentially-active TFs from transcriptomic data, using either co-expression networks [10] or curated regulons [31]. While it was not used in the context of ATAC-seq to our knowledge, given the similarity of the tasks it was a natural solution to try, simply by converting motif matches into regulons, which become the sets of peaks containing a match for each motif, along with

the motif matching scores (scaled by the maximum observed matching score for each motif model). The 'tfmode' (repression or activation) was set to 1 for all targets. Based on the same regulons, the package offers two main functions: the 'viper' method uses an analytical rank-enrichment analysis to compute per-sample TF activity score, while the 'msviper' method works on the vector of per-peak fold-changes between conditions. For the 'viper' method, we first transformed the expression matrix using DESeq2's variance-stabilizing transformation and then used viper's t-test method, and then used limma on the samples' activity scores. Otherwise, both methods were used with default parameters. We tested both approaches using either regulons weighted (likelihood of the links) by the motif scores (denoted e.g. 'VIPER(scores)'), or with all likelihoods set to 1 (denoted e.g. 'VIPER(binary)').

2.3.6 decoupleR

decoupleR is a generic package to extract, from -omics data (typically transcriptomics or proteomics), what it conceptualizes as 'biological activities' – of transcription factors, pathways, or enzymes, etc [9]. It includes a unified interface to use a variety of methods, as well as an ensemble ('consensus') method. We tested the top methods from the decoupleR paper: univariate linear models (ulm), multivariate linear models (mlm), weighted sum (wsum), univariate decision tree (udt), as well as the consensus method. Of note, decoupleR can in principle be applied to the peak logFCs, yielding direct significance estimates of differential activity, or (the typical workflow) to the normalized peak count matrix, yielding motif activity scores per sample. We tested both approaches, and for the second (denoted here 'decoupleR>limma') once again applied limma::eBayes to perform differential activity analysis on the motif activity scores.

2.3.7 fGSEA

Geneset Enrichment Analysis (GSEA) is a powerful way to look for enrichment of certain sets (in this case, sets of peaks containing a match for each motif model) in a vector of scores per element (in this case, peaks' logFC between conditions). To circumvent the long running time of traditional GSEA, we used the fast multi-level implementation from the fgsea package [8]. To avoid ties in the results, we set nPermSimple=10000 and eps=0.

2.3.8 ATACseqTFEA

ATACseqTFEA [32] works under the premise that TFs leave a footprint, and calculate, for each site, a weighted binding score that takes the insertion counts proximal to the motif match in relation to those at the matching sequence itself, weighted by the distal count. It then computes the difference in mean scores between the two experimental groups for each site and performs a GSEA-like enrichment score across the ranked peaks for all sets of peaks containing a match for each motif model. The method was run on motif matches inside peaks, as recommended.

2.3.9 BagFootLike

The original BagFoot method [14] does not involve proper statistical testing, is bound not to work for the many factors that do not leave a footprint [33], and furthermore we were unable to prepare the various annotation files needed by the software for the same genome used by other methods. For these reasons, we implemented a logic inspired by BagFoot, again using limma to test for differences in both the footprint depth and the flanking accessibility, and then combining those tests using Fisher's method. As in the original BagFoot, we used 20bp on each flank of the motif matches to calculate footprint

depth, while flanking accessibility was calculated on the 200bp on each side. We left out the sequence bias correction (reasoning that it is the same across samples, and should not impact relative analyses).

2.3.10 Insertion model

Inspired by various methods [14, 32] and the fact that, when hypothesizing a difference in motif activity, one commonly compares the aggregated insertion profiles around motif matches, we developed a method that counts the (shifted) insertion events around (+/-200bp) motif matches overlapping peaks for each sample, then for each motif model, weights the insertions based on their position to the motif match. The weights are based on the global insertion profile of the motif model across samples, smoothed and made symmetrical (see for example Supplementary Fig S2B). The weighted insertion counts are then summed to produce a score per motif and sample, and differential motif activity is done using limma.

2.3.11 MEIRLOP

Motif Enrichment In Ranked Lists of Peaks (MEIRLOP) performs a logistic regression of whether regions contain a motif on given region scores (in our case, accessibility log2FC across conditions) and sequence-based covariates [34].

2.3.12 Simple linear regression approaches

We also tested simple per-motif linear regression of the logFCs on either 1) whether or not the peak contains a match for the motif ('ulm(binary)'), 2) the motif matching scores ('ulm(scores)'), as well as 3) whether or not the peak contains a match plus the peak's GC content as a covariate ('ulm+GC').

Finally, we tested a lasso- or ridge-regularized linear regression of the logFCs on the matrix of (binary) motif matches.

As we tested multiple method variations, the short names referring to each and their description can be found in Supplementary Table S1.

2.4 Benchmark metrics

For rank-based metrics, the results of each method were sorted first according to uncorrected p-values, then by absolute effect size (e.g. absolute logFC, coefficient or enrichment score). This was especially important for permutation-based methods that often produce ties.

To compute the rank of the true motif, we considered the true motif model that of the manipulated TF, with some special cases: the estradiol treatment is expected to affect both estrogen receptors (ESR1 and ESR2), and therefore both were considered as true; as far as is known, MYC only binds to DNA in cells when in a heterodimer with MAX, and since the MYC motif model was not detected by any method but the MAX was by the top methods, we considered both as 'true motifs'; similarly, NR1H4 (also known as FXR) typically binds DNA as a heterodimer with RXR, and were therefore included also RXRA and RXRB as true motifs. For datasets that had more than one true motif, we used the best rank of any true motif per method as the 'rank of the true motif'.

For the 'network score', we first downloaded from BioGRID 4.4 all the known physical interactors of each true TF (in the species of the respective dataset), and any motif model that overlapped with these interactors (as well as the true motif itself) was considered a network member. For each method, and for k=1:100, we calculated the

proportion of top k motif models that are network members, producing a curve as shown in Supplementary Fig S2C. We then computed the area under this curve (AUC) and divided it by the maximum AUC theoretically obtainable for the dataset (which depends on the number of annotated interactors). The 'archetype score' was computed in exactly the same manner, but using (instead of interactors) motif models that clustered together with the true motif (see motif clustering above).

Precision and recall were calculated using an adjusted p-value <= 0.05, except for monaLisa's stability selection method (which does not provide p-values), for which we used a selection probability of 0.9 as the threshold. Otherwise, if the methods did not report an adjusted p-value, one was generated using the FDR method [35]. To calculate precision, all motifs belonging clustering with network members were counted as positives. However, since not all network members can be expected to be relevant in a given cellular context, only the aforementioned 'true motifs' were considered for calculating recall (i.e. sensitivity).

2.4.1 Ranking

For ranking the methods in the heatmaps (e.g. Fig 1), we transformed the ranks of the true motif using an inverse-logit-based transformation to put it on a similar scale as the other metrics, and so that it decays beyond a certain point. Specifically, we used:

$$2\sqrt{\frac{e^{-x}}{(1+e^{-x})}}$$

where x is the square-root of the rank. We then ranked according to the mean across all scores. For dataset/method combinations that had missing values, we imputed the values for the purpose of ranking by taking the median or mean for that dataset, whichever was the worst.

2.5 Semi-simulations

The semi-simulations were performed by introducing perturbations in the signal of one TF at a time in a baseline ATAC-seq dataset.

As a baseline, six (control) samples of a lymphoblastoid cell line (HG03442) dataset were downloaded from ENCODE [36] (see Table 2 for an overview of the data used for the semi-simulations). To obtain specific perturbations, the baseline ATAC-seq datasets were overlapped with ChIP-seq peaks of that TF from the same cell type. The baseline ATAC-seq samples were divided into two groups, and peak-overlapping regions were downsampled in one of the groups (see Supplementary Fig S10).

2.5.1 Fold-changes

The extent of downsampling followed one of two differential binding scenarios (TF activation and haploinsufficiency), each based on a real reference ChIP-seq dataset (Table 2) in which the distribution of log-fold-changes between conditions is associated to the degree of respective peaks' enrichment over the input (Fig 4B). Specifically, the enrichment of the peaks to be downsampled was quantile normalized to those of the reference peaks in order to select the fold-change of the closest matching reference peak. The matched per-peak fold-changes (FC_i) were introduced in the baseline ATAC-seq samples by downsampling in one of the groups ATAC-seq fragments overlapping the corresponding ChIP-seq peaks (Fig 4A). To test sensitivity of the methods to the strength of the perturbations, the fold-changes were further multiplied by a factor

$p \in \{0, 0.25, 0.5, 1, 3\}$, that we refer to as 'perturbation strength'. For a sample j of the group with a perturbation to be introduced for peak i, the number of fragments to sample n^s_{ij} was finally obtained by:

$$n^s_{ij} = \min\{\lceil \frac{n^0_{ij}}{FC_i * p} \rceil, n^0_{ij}\}$$

where n^0_{ij} denotes the number of fragments overlapping peak i in sample j in the original dataset. The group in which the fragments overlapping peak i were downsampled, was determined by the sign of the $\log_2(FC_i)$. A perturbation strength of 0 corresponds to the original dataset without any downsampling of ATAC-seq fragments in ChIP-seq peak regions.

Of note, the semi-simulations rest on the assumption that changes in binding are associated to corresponding changes in accessibility, which is not always the case.

2.5.2 Technical variation

Further, technical variations were inserted in some semi-simulations by varying GC content and fragment length (FL) distributions (see Supplementary Fig S10C for an overview of the semi-simulated datasets). Fragment GC content distributions were varied sample-wise by sampling fragments according to their GC content, mimicking reference GC content distributions. FL distributions of the baseline ATAC-seq samples were varied by sampling nucleosome-free-, mono-, di- and multi-nucleosome (>2) containing fragments with differing sampling probabilities. Fragments were classified by applying fixed thresholds on their lengths ($(0, 120]$, $(120, 300]$, $(300, 500]$, $(500, \infty)$).

After obtaining the semi-simulated ATAC-seq profiles by (under-)sampling fragments in ChIP-seq peak regions or according to their GC content or length, peaks were called for each of the groups separately using MACS2 and merging the peaks of both groups using the bedtools [37] merge command.

2.6 TRAFTAC experiment

2.6.1 Cell culture

HEK293 cells were cultured in Dulbecco's Modified Eagles Medium (DMEM) with 10% Fetal Bovine serum (FBS) (Gibco,10500064) and 1% Penicillin/Streptomycin (P/S) (Gibco,15140122). All of the incubation steps were performed at 37 °C and CO2 content of 5%. For the transfection cells at a confluency of 70-90% were incubated for 6 h with 1 μg/mL of a plasmid encoding N-terminal HA-HaloTag7-containing dCas9 fusion protein and 25 nM TRAFTAC using Lipofectamine 3000 following the suppliers' instructions. The plasmid used was published and kindly provided by the Crews lab [39]. Cells were then incubated for 19h with 5 ng/mL of TNF-α (Sigma-Aldrich, H8916) and 25 μM HP14 (dissolved in DMSO) (WuXiAppTec) respectively DMSO. 30 min before harvest, cells were treated with 200 units/ml DNase (Sigma, DN25). HEK293 cells were then washed twice with PBS (Gibco, 10010023) and detached with TrypLE (Gibco, 12605010). Enzymatic digestion was quenched with growth medium and cells were washed once again with PBS.

2.6.2 Protein detection

Western blot was performed to qualitatively and quantitatively assess NFkB levels. Overall protein abundance was determined via BCA assay following the ThermoFisher scientific protocol. The obtained values were normalized and utilized to calculate the sample volumes to be loaded. Samples were mixed with Laemmli buffer (BioRad,

Database	ID	Purpose	Cell line	ref
ENCODE	ENCFF495DQP ENCFF130DND ENCFF447ZRG ENCFF966ELR ENCFF358GWK ENCFF963YZH	Baseline ATAC-seq samples	Human lymphoblastoid cell lines	[36]
GEO	GSE98477	Log fold-change distribution (haploinsufficiency)	Lymphoblastoid cell lines	[38]
ENCODE	ENCFF395FIP	Peaks (activation)	A549	[15]
ENCODE	ENCFF416YUN ENCFF860QUX ENCFF456CMP ENCFF181HLP ENCFF870WJP	Reads for log fold-change distribution (activation)	A549	[15]
ENCODE	ENCFF156OCY	CEBPB peaks	GM12878	[36]
ENCODE	ENCFF592UDD	CTCF peaks	GM12878	[36]
ENCODE	ENCFF250FJC	MAZ peaks	GM12878	[36]
ENCODE	ENCFF500EWB	ZNF143 peaks	GM12878	[36]
GEO	GSM3315605	alternative GC content reference distribution	MCF-7	[36]

Table 2.

Data used for the semi-simulations.

#1610747) and β-Mercaptoethanol (Sigma, M6250) (1:10) and boiled for 10 min at 70 °C. SDS-PAGE was performed using 10% Mini-PROTEAN TGX precast gels and the proteins were transferred from the gel to a Trans-Blot Turbo Mini nitrocellulose membrane via electroblotting. Membranes were then blocked in 5% milk in TBST (1 x TBS, 0.1% Tween 20, pH=7.6) for 1h and incubated overnight in primary antibody against NFkB (SantaCruz, sc-8414) (1:200 diluted in 5% milk in TBST), HA (Cell signalling, 3724) (1:1000 diluted in 5% milk in TBST) and GAPDH (Millipore, ABS16) (1:1000 diluted in 5% milk in TBST) at 5 °C. The next day the membranes were incubated for 1 h in HRP-coupled secondary antibody (Sigma Aldrich, AP308P) (1:20 000 diluted in 5% milk in TBST). Proteins were then visualized via chemiluminescent detection (Clarity Western ECL Substrate) at a BioRad imager.

2.6.3 Library preparation for ATAC-sequencing

ATAC libraries were prepared from 50 000 cells per sample according to the published OmniATAC protocol [6] with minor adaptations. In short, samples were washed with cold 50 μl of PBS, afterwards cells were incubated in 50 μl lysis buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM MgCl2, 0.1% NP-40, 0.1% Tween-20, 0.01% Digitonin) for 3 min on ice. Cells were washed with 1 ml of lysis buffer without Digitonin and NP-40. Transposition was done in 50 μl reaction buffer (1x tagmentation buffer (Diagenode, C01019043), 0.1% Tween-20, 0.01% and 2.5 μl Tn5 (Diagenode, C01070012)) for 30 min at 37 °C and 1000 RPM shaking. Subsequent DNA purification was performed with MinElute Reaction Cleanup Kit (Qiagen, 28204). PCR amplification was done with NEBNext High Fidelity 2X PCR Master mix (NEB, M0541) and UDI primer (Diagenode, #C01011034) with concentrations according to manufacturer's

recommendation. Total number of needed amplification cycles (8x) was controlled with qPCR. Library purification and fragment size selection was done with AMPure XP beads (Beckmann, A63880). 20 μl of 10 nM was sent for sequencing externally with Novogene (Cambridge, UK). A minimum of 20 Mio PE reads were sequenced per sample. The data is available under the GEO accession GSE260504.

2.7 Packages versions and code

We used the following package versions: chromVAR 1.12.0, monaLisa 1.5.0, ATACseqTFEA 1.5.0, viper 1.24.0, decoupleR 2.3.3, fgsea 1.16.0, diffTF 1.8.

The code used to run the methods and simulations, evaluate the results, and create the figures is available on https://github.com/ETHZ-INS/DTFAB.

The semi-simulated datasets can be retrieved from Zenodo: 10.5281/zenodo.10732704, 10.5281/zenodo.10781849, 10.5281/zenodo.10781759, 10.5281/zenodo.10781109.

3 Results

3.1 The top methods were specifically designed for chromatin data

Fig 1 shows the results of the methods on the benchmark datasets. A good method for identifying differentially accessible TF should rank the motif model of the perturbed TF (henceforth 'true motif') top, i.e. with the smallest p-value. We therefore first monitored the rank that each method attributed to the true motif (Fig 1, left). Of note, the MYC and NR1H4 motifs were not accurately detected by any method in their corresponding datasets, but their obligatory binding partners (respectively MAX and RXRA/B) were. These were therefore included in the set of true motifs for the benchmark. The top performing methods were all methods specifically designed for chromatin data, in particular, chromVAR (adapted, see below) and monaLisa. Proven methods for geneset enrichment analysis (GSEA) as well as methods with demonstrated performance for transcriptome-based master regulator analysis, such as decoupleR- and viper-based methods [9, 31], underperformed.

Of note, the two datasets in which none of the methods ranked the true motif in the top 10, RUNX2 and MYC, both showed no peak-level change in accessibility with genome-wide significance (Supplementary Fig S1A), suggesting the lack of a clear treatment effect in the first place. However, for some other datasets that showed no significant peak-level change (e.g., RUNX1 and GATA2), some methods ranked the true motif towards the top (though not with an adjusted p-value $<= 0.05$), thus indicating that global differences in motif accessibility can be detected even in the absence of significant peak-level changes. In the case of RUNX2, it should additionally be noted that the RNA expression of the gene in that cell type is very low, and hence perhaps the factor had a low activity to begin with.

TFs do not act in isolation, but most often form heterodimers or complexes that cooperatively regulate gene expression. Therefore, perturbations on one TF will likely disrupt whole complexes and the binding of its co-factors as well. We therefore devised a 'network score', based on the ranking of known physical interactors of the perturbed TF (see Methods and Supplementary Fig S2C), shown in the middle part of Fig 1. This second metric confirms the findings of the first: except at very low (1-3) or high (>100) ranks (i.e. where ranks cease to be discriminatory), the two metrics show a good agreement, although the network score is not comparable across datasets

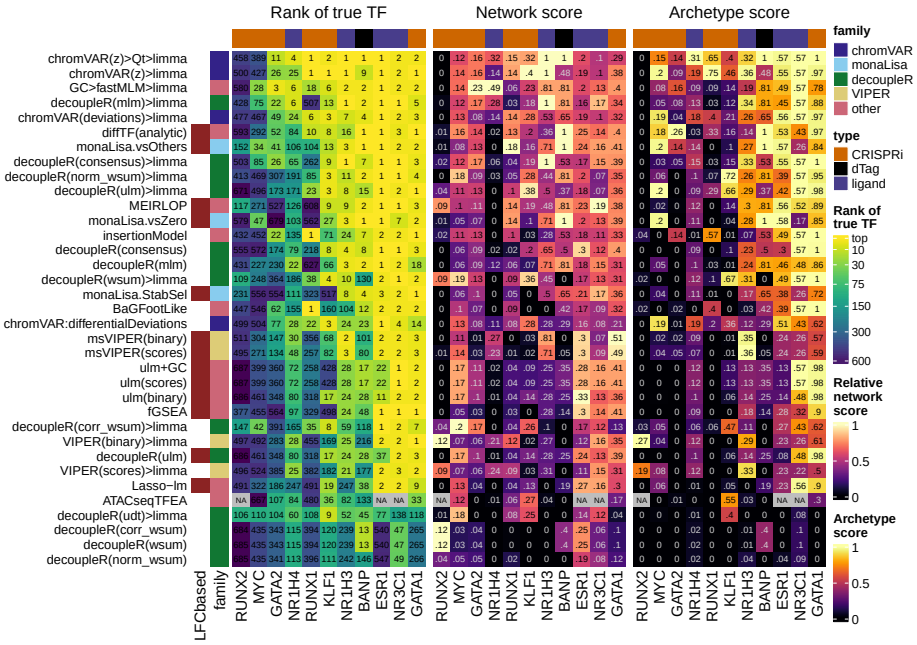

Fig 1. Benchmark of differential TF activity methods. Ranking-based metrics for each method across each dataset. The left part of the heatmap indicates the rank of the true TF (top one if the dataset has more than one true target), the numbers are the actual ranks and the colors are mapped to the squared root of the ranks (because differences in very high ranks are irrelevant in practice). The central part shows the network score, i.e. the AUC of the proportion of top X TFs that are in the network of interactors of the true TF, relative to the maximal possible AUC for the dataset (see Methods and Supplementary Fig S2C for details). The labels indicate the actual scores, while for the color mapping, the values are relative to the maximum achieved value for each dataset. Finally, the right part indicates a similar AUC-based score using motif models that are similar to (i.e. cluster together with) the true motif (see Methods). The methods are ranked based on the average of transformations of the three metrics across datasets (see methods). For ATACseqTFEA, NA values indicate that the method repeatedly crashed on those datasets, presumably due to unsustainably high (>70GB) memory usage. For methods for which variations were tested, we show here only the original and top variant(s). All variants are available in Supplementary Fig S5. Supplementary Table S1 contains full descriptions for the short names displayed here.

(Supplementary Fig S2D). Finally, some motif models in the collection are highly similar to the true motif, and these should also tend to be ranked towards the top. To assess this, we clustered similar motif models into archetypes (see Methods), and then computed a score similar to the network score, but based on motif models clustering together. This is shown in the rightmost heatmap of Fig 1, and also largely recapitulates the rank-based relative performances.

Ranking the true motif (and that of its interactors) top is critical, but it does not assess whether the significance of the differential activity reported by a method is well calibrated. We therefore measured the sensitivity in detecting the true TF (recall) in relation to precision. Fig 2A reports those values for the top variant of each method. To calculate precision, all motif models clustering with the true TF or its known interactors were considered as true positives (considering only interactors yielded very similar results). As expected, we observed a trade-off between precision and recall, with fGSEA offering the highest sensitivity but at the cost of a very high FDR, while VIPER, for example, controlled FDR but had a very low sensitivity (Fig 2A). As with the rank-based metrics, our adjusted chromVAR method (chromVAR>Qt>limma) showed the best trade-off between precision and recall.

While monaLisa (without stability selection) performed well in ranking the true motifs (Fig 1), it had a low precision, indicating that it should be used for discovery rather than confirmatory purposes, or in conjunction with the stability selection strategy. That said, it should be noted that monaLisa reports a p-value for each motif model and fold-change bin, and that while our use of Simes' aggregation to obtain a single p-value per motif (needed for benchmarking purposes) is not unreasonable, it does not come from the method's authors.

The estimated mean runtime per dataset for each method is shown in Fig 2B. Despite multi-threading (see Methods), diffTF had very long runtimes; fGSEA, ATACseqTFEA and decoupleR also had elevated runtimes, taking several hours per dataset, while all other methods performed within a reasonable time.

3.2 Improvements to chromVAR-based approaches

The chromVAR package includes a function to perform differential analysis ('chromVAR::differentialDeviations'), which performs t-tests on the adjusted deviation scores of each motif. However, this method did not perform well, and we found that using limma [28] on chromVAR's motif z-scores ('chromVAR(z)>limma') considerably outperformed this original approach (Fig 1 and Fig 2). Presumably, this is because of limma's moderation of the variance. In addition, we observed that in some datasets, the per-sample distributions of z-scores could exhibit major shifts and differences in width that were not associated with experimental groups (Supplementary Fig S3A-B). In particular, not resizing peaks to a common size (as is recommended) tended to result in large shifts in the distributions, but some variations in the distributions – especially in their width – persisted despite resizing (Supplementary Fig S3A).

Since these distributional differences were in our experience sometimes correlated to technical sources of variation such as sequencing depth, and are unlikely to have a biological meaning, we hypothesized that something in the chromVAR method did not perfectly adjust for technical differences between the samples. Internally, chromVAR uses a simple library size normalization to compare each sample to expectations, which is known often not to be adequate for NGS data [40]. We therefore hypothesized that these shifts could be due to inadequate normalization, and adapted chromVAR to use more robust normalization procedures, including the well-established TMM normalization [40]. However, these changes had little impact on the distributional differences, nor the differential TF identification (Supplementary Fig S3C-D).

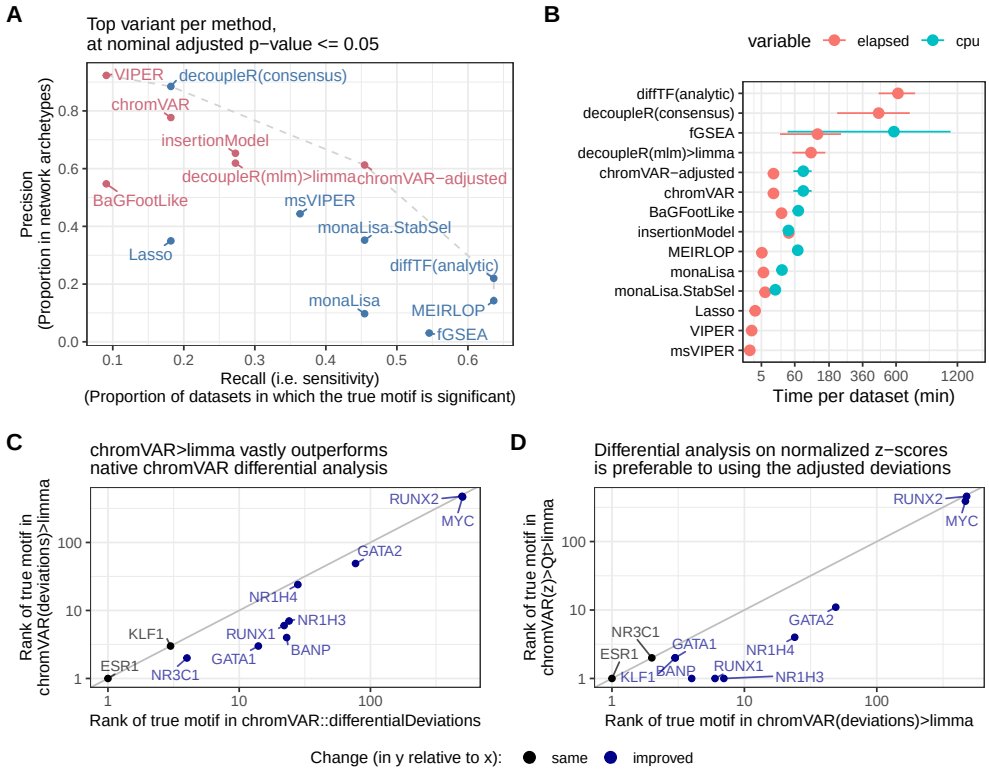

Fig 2. Further benchmark results, and improvements over native chromVAR. **A.** Sensitivity and specificity (considering interactors and members of the same archetypes as positives) of the top alternative per approach. The methods are coloured by family (sample-wise methods in pink, logFC-based methods in blue). **B.** Mean (and standard error of the mean) running time (elapsed, as well as total CPU time when multithreading) across datasets of the top alternative per approach (the x axis is squared-root-transformed for visibility). Note that because it was done separately for reasons of standardization, the running time does not include the generation of the peak-count matrix, nor, except for monaLisa, the motif scanning. monaLisa is therefore disadvantaged in this comparison, and these times should be interpreted as rough estimates. **C.** Comparison of the rank of the true motifs obtained by a *limma* analysis on the chromVAR deviations, versus using chromVAR's native *differentialDeviations*. **D.** Comparison of the rank of the true motifs obtained by a *limma* analysis on the normalized chromVAR z-scores, versus on the chromVAR deviations.

We therefore tested whether a further, downstream normalization of the motifs z-scores would improve the detection accuracy. Quantile normalization appears to perform best (Fig 1A and Supplementary Fig S4A). This was also confirmed by comparing the t-value attributed to the motif of the true TF, which is more sensitive (Supplementary Fig S4B).

In light of these results, we took advantage of the fact that some of the benchmark datasets are in fact (aggregated) single-cell ATAC-seq datasets [18] to see whether such normalization also had a beneficial impact at the single-cell level. As a proxy for the signal-to-noise ratio, we computed the standardized mean difference of the respective motif's z-scores upon each knockdown. Centering the z-score distributions was always beneficial, quantile normalization also tended to be beneficial most of the time, while unit-variance scaling showed mixed results (Supplementary Fig S4C).

Finally, we also tried to optimize another of the top methods, monaLisa. Changing the number of bins or the logFC threshold did not lead to improvements (Supplementary Fig S5). Since observing a dose-response across the monaLisa bins is a further evidence of a pattern, we also tried to capture this computationally by dividing, for the purpose of ranking, the aggregated p-values by the absolute Spearman correlation between the fold-enrichment and the bins order (denoted 'monaLisa.vsOthers+spearman' in Supplementary Fig S5), however it did not yield a clear improvement.

3.3 Using GC smooth quantile normalization improves non-specialized methods

Sample-specific GC bias is a known source of variation in ATAC-seq, and methods have recently been developed to address it, in particular, smooth-quantile normalization within GC bins [41]. We therefore tested whether using this normalization, in particular for calculating the peak-level log(fold-change) between conditions, could improve in particular logFC-based and non-specialized methods. As shown in Fig 3A, which compares key methods with and without GC smooth quantile normalization, most methods benefited from this specialized normalization method, vindicating the procedure (see Supplementary Fig S5 for the full rank-based metrics on all variants). Indeed, even monaLisa stability selection, which already includes some form of control for GC bias, seemed improved. With this normalization, the sample-based decoupleR(mlm)>limma approach even became even more competitive with the top chromVAR variant (Fig 3B-C). We therefore re-implemented the method in a more performant fashion, and included it (in combination with GC smooth quantile normalization) in the following analyses as 'GC>fastMLM>limma'.

3.4 Impact of stringency in motif matching and using fewer, more distinct motif models

The previous analyses used default thresholds for motif matching. To test the impact of variations in this step, we decreased the motif matching p-value threshold by a factor of ten, i.e. from 1e-4 to 1e-5, leading to an average 82% decrease in global motif matches (for monaLisa, where only a score threshold can be set, we changed the threshold from 10 to 13, leading to a 74% decrease in matches). The impact of these changes on a selection of the top methods was relatively mild, but overall a more stringent motif matching had negative effects (Supplementary Fig S6). As previously mentioned, many motif models in the collection are highly similar to each other, and as such will get very similar scores from the majority of activity inference methods that test each motif model separately. Multivariate methods, instead, can be sidetracked by very similar motifs. Given these considerations, we assessed whether the relative performance of a selection of the top methods was altered when using the motif archetypes (i.e. a smaller collections of motif models that are less similar to each other). Overall, using motif archetypes yielded a similar or lower performance, in particular a slightly lower power (Supplementary Fig S7).

3.5 Focusing on nucleosome-free peaks/fragments yields mixed results

ATAC-seq data has a characteristic, roughly periodic fragment length distribution, where smaller fragments (e.g. <100-125bp) are too small to be wrapped around a nucleosome and are therefore deemed 'nucleosome-free' (NF), while larger peaks

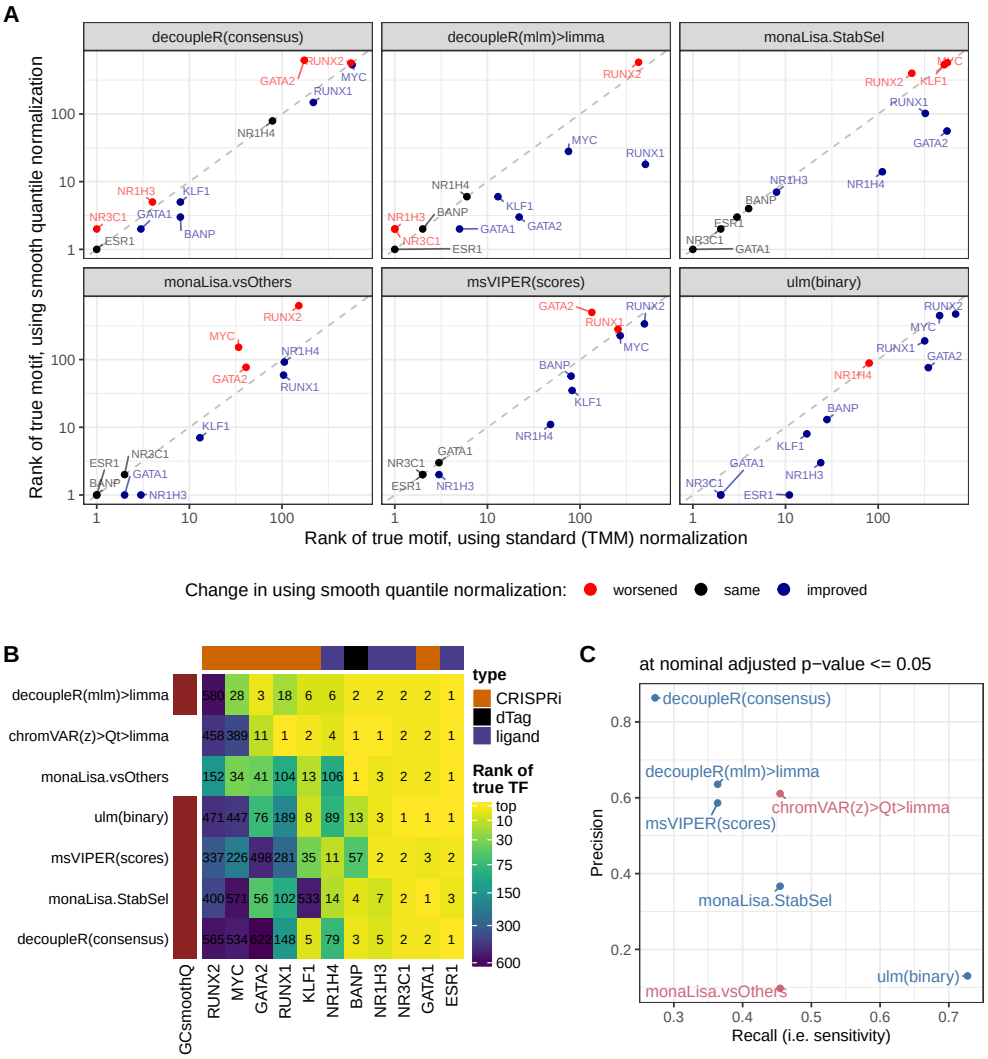

Fig 3. Impact of using GC smooth quantile normalization. A. Comparison, across key methods, of the ranks of the true motif when using GC smooth quantile normalization instead of standard (TMM) normalization to calculate per-peak log(fold-change). **B.** Comparison of the ranks of the true motif with GC smooth quantile normalization, compared to the two best-performing methods from the earlier benchmark. **C.** Comparison of the precision and recall (at adjusted p-value <= 0.05) when using GC smooth quantile normalization (in blue), compared to the two best-performing methods (in pink).

correspond to mono-nucleosome-, di-nucleosome-, and eventually tri-nucleosome-containing fragments. Since most TF compete with nucleosomes for binding, and NF fragments are enriched at the core promoter, NF fragments are often thought to be particularly characteristic of TF binding. Therefore, we tested whether using filtering based on fragment size could improve the benchmark results, focusing on the top method for each of the two families of approaches. Specifically, in addition to the previous results based on all fragments, we tested i) counting only NF fragments, ii) counting all fragments but on peaks called only on NF fragments, as well as iii) counting only NF fragments on NF peaks (i.e. using only NF fragments throughout).

While these alternatives did appear beneficial on some datasets, their impact on discriminatory power was not consistent across datasets (Supplementary Fig S8). In particular, chromVAR hardly ever benefited from these variants, whereas it sometimes made major differences to monaLisa, for instance with GATA2 being much better ranked when counting NF fragments (Supplementary Fig S8). Overall, however, these differences did not generalize into clear recommendations.

3.6 Limited gain from simple method aggregation

Reasoning that different approaches could in principle offer complementary advantages, we next investigated whether aggregating the results of the three top-performing methods could yield superior results. We tested common p-value aggregation methods (despite the assumption of independence underlying some of them being violated), as well as a non-parametric, rank-based alternative (Supplementary Fig S9). The latter proved best but did not show consistent improvement upon the best single methods. We next tested combining the per-sample motif activity scores produced by chromVAR and by the multivariate model applied on GC smooth quantile normalized counts, before computing differential analysis (Supplementary Fig S9). This was always better or equivalent to chromVAR, although the improvements were mild. While it is possible that more sophisticated ensemble methods could yield an improvement, these results suggest a low degree of complementarity between methods.

3.7 Semi-simulations confirm the results from real benchmark datasets

Even when focusing on early time-points, the real benchmark datasets have a problem of ambiguity with respect to the truth, as interfering with a factor is likely to impact other interacting factors, as well as indirectly interacting factors competing for binding. To confirm the robustness of our results, we therefore developed a semi-simulation procedure. We generated, from baseline ATAC-seq profiles of lymphoblastoid cell lines, cohorts of two groups of samples with a difference in accessibility at binding sites of a single given TF at a time (Fig 4A; see also Supplementary Fig S10 and methods). Briefly, fragments overlapping the ChIP-seq peaks of the given TF were subsampled in one of the two groups following one of two realistic differential binding scenario (Fig 4B): an activation (where the high-occupancy peaks are the most affected), modeled on the ligand-based activation of the glucocorticoid receptor [15], and a haploinsufficiency (where the TF amount is halved, and the remaining TF preferentially goes to high-occupancy peaks, leaving the low-occupancy peaks most affected), modeled on the heterozygous loss of function of YY1 [38]. In the activation paradigm, the difference between conditions affects virtually all sites (and many by a lot), whereas in the haploinsufficiency many peaks are very mildly affected (Fig 4B). Specifically, we simulated various perturbation strengths (i.e. multiplicators of the original differential binding fold-changes) for four factors that had 1) high quality ChIP-seq data for the simulated celltype (lymphoblasts), 2) a clear associated motif model, and 3) represented distinct factor families and properties (Supplementary Fig S11).

We first confirmed that the simulated data was distributed similarly to the real data (Supplementary Fig S12), and that the effects introduced were as expected (Supplementary Fig S13-14), and next ran the methods on each dataset. The results, a subset of which is presented in Fig 4C (full results in Supplementary Fig S15), afford the following observations. As expected, stronger perturbation strengths made the task easier, and the haploinsufficiency scenario was systematically harder to detect, owing potentially both to the lower general fold-changes and comparative sparing of high-occupancy sites (Fig 4B). Contrarily to the real datasets, msVIPER appears to

perform very well in the simulations (Supplementary Fig S15), however this is chiefly because for two hard datasets, it gives by chance low ranks to the TF already in the absence of any simulated effect, which remains low across the simulations. Aside from this, as with the real datasets limma on the chromVAR z-scores gave the best overall results, followed by monaLisa (vsOthers) and MEIRLOP, which provided a good performance so long as the perturbation strength was sufficient. Furthermore, monaLisa (vsOthers) offered greater sensitivity than other methods, being the only approach marking the true factor for all simulated TFs and paradigms as significant and also in some cases at lower perturbation strengths, however at the cost of a generally lower precision (Fig 6, Supplementary Fig S18). Once more the original chromVAR differentialDeviations did not perform well, and the multivariate model had mixed, somewhat unstable performance.

Given the controlled setting of the simulations, any difference between the factors necessarily hinges on the relationship between its motif model and its experimental binding sites. The easiest factor was CTCF, which had the highest number of experimental binding sites, but especially the highest proportion of those harboring matches for its motif, as well as the highest enrichment for motif matches in bound (i.e. overlapping a ChIP-seq peak) vs all ATAC-seq peaks (Supplementary Fig S11A). On the opposite spectrum, CEBPB was the hardest, with few binding sites, few of which with a motif match. Interestingly, ZNF143 was easier to detect than MAZ, despite having a similar number of sites and a motif model that is less specific to the bound peaks (Supplementary Fig S11A). Closer investigation revealed that a large number of motifs co-occur with MAZ, and in fact show a stronger enrichment in its bound peaks (Supplementary Fig S11B). While it is also the case with ZNF143 that many other motifs are more enriched in its bound peaks, this is chiefly because a part of the ZNF143 peaks show a higher GC content (Supplementary Fig S11C-D), and the fact that the top methods correct for GC presumably enables an accurate detection nonetheless. To confirm the importance of the relationship between motif models and binding site, we created simulations in which only the ChIP-seq peaks containing a motif match for the respective TF were perturbed. As expected, for all TFs but in particular for CEBPB and MAZ performance of all methods improved considerably (Supplementary Fig S14).

Finally, the semi-simulation procedure further enabled us to introduce sample-wise variations in two known factors of technical variation: fragment length (FL) distribution (Supplementary Fig S16) and GC content (Supplementary Fig S17). The top methods were generally robust to both kinds of variation, despite not correcting for FL biases. However, for some of the biased settings (both GC and FL), monaLisa (vsOthers) detected false positives in datasets were no perturbation was simulated (perturbation strength 0).

In summary, the simulations confirm the main results of the benchmark on real datasets, reveal the top methods to be robust to technical biases, and indicate the relevance of how distinct and informative the motif model is for actual binding.

3.8 Application of the methods to characterize TRAFTAC-mediated TF degradation

PROTACs (PROteolysis TArgeting Chimera) are a particularly successful technology that hijacks the cell's endogenous protein degradation machinery (the ubiquitin-proteasome system) and targets it to a protein of interest [42]. However, each PROTAC requires careful chemical design and validation. A recent, attractive derivative technology harnesses the versatility of the CRISPR technology to target and degrade TFs in a more flexible fashion. With TRAFTACs, or TRAnscription Factor TArgeting Chimeras [39], the targeting specificity is not encoded in the design of the

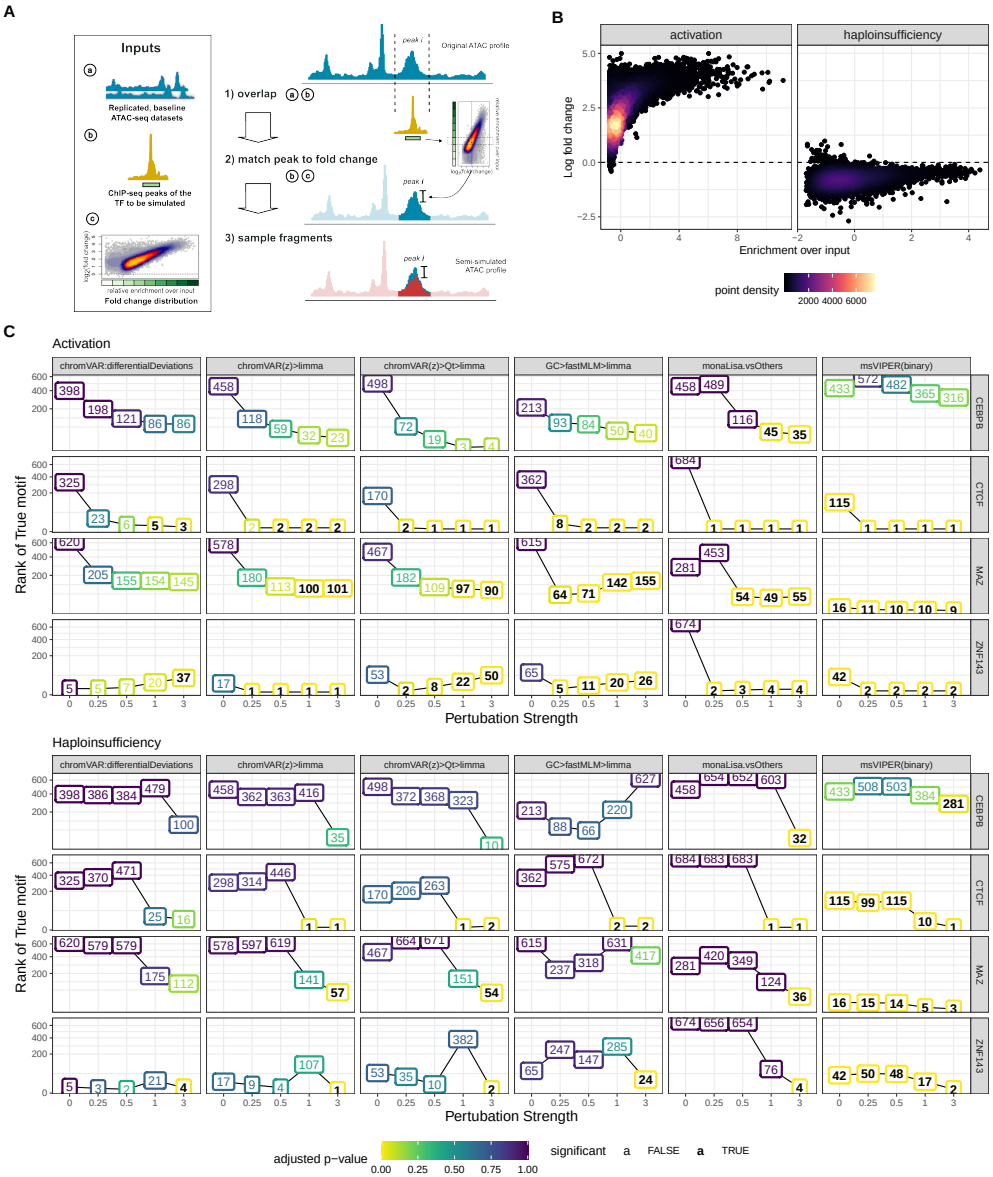

Fig 4. Semi-simulations and impact of perturbation strength. A. Strategy for sampling ATAC-seq fragments: ChIP-seq peaks of the TF to be perturbed are overlapped with the baseline ATAC-seq profiles (1); a fold-change is determined by per ChIP-seq peak fold changes from a reference distribution (2) and used to sub-sample the overlapping fragments of the ATAC-seq profiles of one of the two groups (3). **B.** Two group designs were simulated by downsampling ATAC-seq fragments in one group according to the obtained fold changes of the overlapping peaks. Perturbation strength was varied by multiplying fold changes with different factors (0, 0.25, 0.5, 1, 3), i.e. the perturbation strength. Further technical variations were introduced in some datasets by varying GC contents and fragment length distributions. **C.** Performance of a subset of the top performing methods on the semi-simulated datasets with varying perturbation strength. Numbers in the line plot represent the rank of the perturbed TF as detected by the respective method. Colors of the boxes signify the corresponding adjusted p-value obtained in the differential activity analysis. Bold vs plain text indicates if the adjusted p-value was found to be significant (adjusted p-value <= 0.05).

PROTAC molecule, but by a double-stranded DNA containing the target TF's binding motif. This oligo is covalently linked to a single-stranded RNA that functions as a guide to recruit a deactivated Cas9 fused with a HaloTag7, which in turn can be bound by a generic HP14 PROTAC. As a result, binding of the oligo was shown to lead to ubiquitination and subsequent cleavage of the target TF [39].

Using the top differential motif accessibility methods identified above, we sought to characterize the impacts of a validated TRAFTAC against NFKB [39]. Replicating the authors' experiment, we transfected HEK cells with the dCas9-HT7 plasmid and DNA-RNA hybrids (NFkB or scrambled TRAFTAC) for 6h, before treating the cells TNF-α (to trigger NFkB activation) and the PROTAC against the Halo tag (HP14) for 19h. We confirmed the reduction of NFKB1 protein levels using western blot (37% reduction with respect to scrambled control, Supplementary Fig S19), and generated ATAC-seq profiles. We then ran differential motif accessibility analyses using chromVAR and monaLisa (Fig 5). Interestingly, although the NFkB motif matches have a weak trend of decreased accessibility, they are far from significance (Fig 5A). Instead, the strongest signal is in motifs related to the AP-1 complex, i.e. JUN/FOS family members (which all have very similar motif models), showing a decreased accessibility in the knockdown according to both the chromVAR and monaLisa analyses. Of note, despite the key role of the AP-1 complex in regulation response to stimulus, JUN/FOS-related motifs were significant in only one of the benchmark dataset (BANP), arguing against a widespread sensitivity to perturbation, and for a role specifically related to NFkB/TNF-α. Beside activating NFkB through IKK, TNF-α is expected, through MKK7, to trigger the MAPK/JNK pathway, which in turn activates AP-1 [43]. However, the reduction in AP-1 activity that we observe in the NFkB knockdown suggests a cross-talk between these two downstream signaling pathways.

Fig 5C shows some example significant motifs across the top three methods. Some are consistent: for example, the GC-rich GLIS2 motif appears enriched in a dose-dependent fashion in monaLisa bins that increase in accessibility, and depleted in those that decrease; it is moreover increased in activity in the knockdown according to both chromVAR and GC>MLM>limma. NFE2 and FOSL1 show the same consistency, with the opposite pattern. In contrast, some inferred motifs activities are inconsistent across methods: while both chromVAR and MLM suggest a reduction in NFYA accessibility in the knockdown, which would be consistent with reported links between NFkB and CCAAT elements [44], the monaLisa results shows that the motifs are depleted in both upregulated and downregulated peaks, and instead enriched in the bin of peaks that do not change. JUNB and FOS, instead, show the same pattern as FOSL1 according to both monaLisa and chromVAR, as expected given that the motifs are virtually the same, while MLM shows no clear pattern. This is precisely because the three motifs are largely equivalent: because the multivariate approach regresses accessibility on the combined effect of the different motifs, highly-similar motifs will nevertheless have different coefficients. This is a good illustration of the fact that, whereas most methods benefited from the extensive collection of motif models, multivariate methods should be used with a reduced collection of more distinct motif models (Supplementary Fig S7).

4 Discussion

A major issue in assessing methods for differential TF activity is the lack of a ground truth, even in datasets with known experimental perturbations. Proteins interact in complex ways that modulate both their stability and activity, so that knocking down or altering the activity of a TF is highly likely to affect the activity of other, interacting TFs. Moreover, TFs typically function in cascades, so that intervening on a TF will

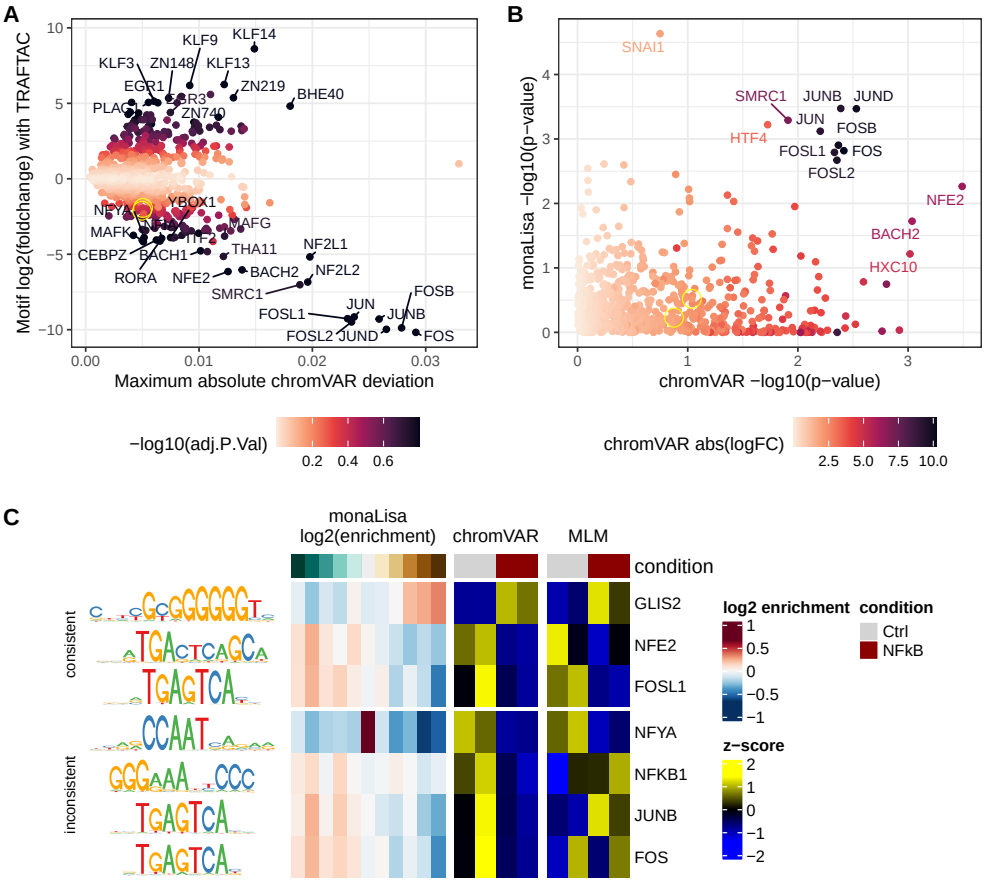

Fig 5. Characterizing the impacts of a TRAFTAC against NFkB after TNF-α activation in HEK cells. A. Results of the 'chromVAR(z)>limma' analysis, highlighting the motifs with strongest changes between conditions and high absolute deviations. The two overlapping yellow circles highlight NFKB1 and NFKB2. **B.** Comparison of the per-motif p-values assigned by monaLisa (Simes p-value) and chromVAR, showing an agreement only on the strongest differences. **C.** Representative examples of motifs called as differentially-accessible by some method, showing side by side monaLisa's enrichment scores across logFC bins, as well as the per-sample chromVAR z-scores and z-scores of the MLM t-values.

affect the transcription, and eventually activity, of other downstream TFs, leading to indirect, yet no less genuine changes in TF activity. We adopted a number of complementary strategies to mitigate these issues. First, we focused on datasets profiled shortly after intervention to minimize indirect effects. Second, developed a network score and computed precision in a way that considers known interactors of the perturbed TFs. Finally, we developed a semi-simulation procedure that avoids both issues. While none of these solutions are perfect, the general agreement between simulations and real datasets, as well as between evaluation metrics, provide us with the confidence that our results genuinely reflect the relative performances of the methods.

4.1 Summary of the results and general recommendations

Fig 6 summarizes the main benchmark results on the union of the top 5 methods from real and semi-simulated datasets. Overall, our adjusted chromVAR workflow performed

best on rank-based metrics, and provided the best tradeoff between sensitivity and specificity. Specifically, we recommend increasing the number of backgrounds iterations (to 1000-2000) to improve reproducibility (not easily done with single-cell data), and using limma on the z-scores to perform differential analysis (as opposed to the native 'differentialDeviations', which did not perform well). Moreover, its results were often (though not always) improved by post-processing of the scores, suggesting that some technical variation is not entirely accounted for by chromVAR's background strategy. Furthermore, we found the sample-wise methods useful in relaying to the analyst the variability across samples. On the other hand, those methods are limited to detecting global increases or decreases in TF activity, and in principle cannot detect the displacement of a TF's binding at constant global activity (although it is unclear how likely a scenario this is). MEIRLOP, diffTF and monaLisa also gave good results, although with a low precision. monaLisa's bin-based strategy and plotting functions have the advantage of providing the user with a more fine-grained view of where the enrichment is observed, relative to where the changes in accessibility are happening. Observing a dose-response (i.e. an increased enrichment in bins of greater fold-change) bolsters our confidence in the enrichment.

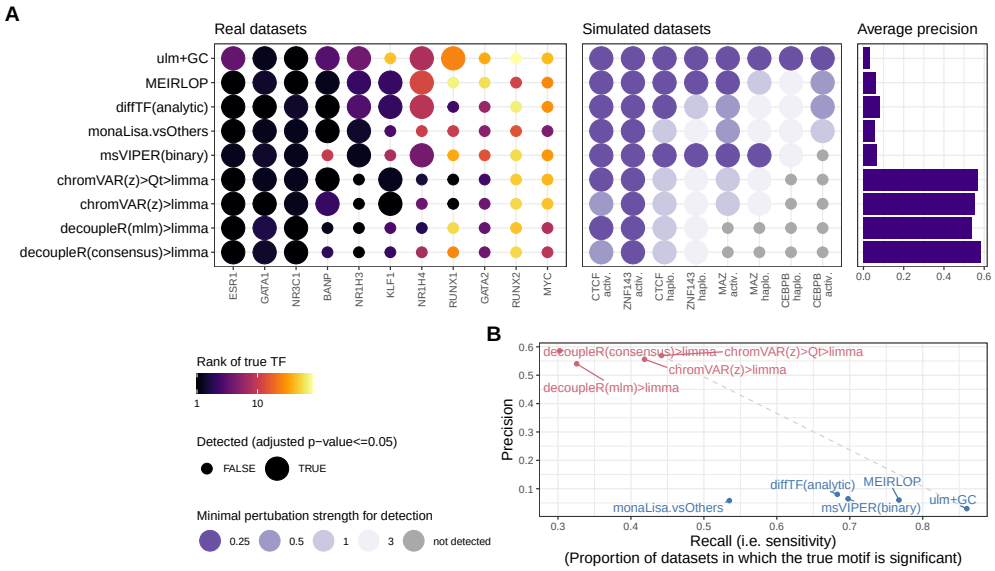

Fig 6. Summary of the benchmark on variants of the top methods. A. Sensitivity (left and center) and precision (right) of each method across real (left) and semi-simulated (datasets). The size of the points indicate whether or not the method found the true motif significant (adjusted p-value <= 0.05). For real datasets, the color indicates the rank of the true motif, while for semi-simulated datasets it indicates the perturbation strength at which the true motif was found significant. **B.** Sensitivity and specificity (considering interactors as positives) of the top methods, aggregating both real and semi-simulated datasets. The methods are coloured by family (sample-wise methods in blue, logFC-based methods in pink).

While methods not specifically designed for chromatin data performed poorly in the main benchmark, they were improved when using smooth quantile normalization in GC bins, confirming the value of the normalization method. In particular, the GC>decoupleR(mlm)>limma approach performed as good as, and in some datasets better than, the top specialized method. This method fits a multivariate linear model on the peak accessibility of each sample, using all motifs as covariates. Of note, it is one

of the few methods (along with monaLisa stability selection and the similar lasso-linear model, which are however logFC-based rather than sample-based) that handles motifs together, rather than separately. This suggests that this general strategy should be further investigated. However, as the TRAFTAC example illustrates (Fig 5C), a high motif similarity within the collection causes major problems for multivariate approaches, whether regularized or not. When using such methods, we rather recommend using a reduced collection of more distinct motif models. Most likely, harnessing the potential of the multivariate approach will involve a more refined handling of the similarity between motif models.

Simple results aggregation of these three methods did not yield clear improvements. Nevertheless, we recommend using more than one of the top methods to increase confidence, given their different strategies and benefits. Furthermore, we observed that the p-values were not generally well-calibrated, i.e. the observed false discovery rate was nearly always higher than the nominal threshold, despite not counting similar motifs and interactors as negatives (admittedly, the knowledge of interactors taken into account for calculating precision is neither exhaustive nor specific to the cell types used). Precision estimates show strong differences across datasets, and high precision typically only occurred when no effect was detected (Supplementary Fig S18). Hence, we recommend caution in interpreting the obtained precision estimates, and for this reason we especially focused on ranks of the true motif for reporting performance. However, for similar reasons also the ranks should be interpreted on a rather coarse scale, and might not be comparable across TFs given differences in the number of factors co-binding/co-occurring with the true factor.

During the revision of this manuscript, a related study was published evaluating methods for TF prioritization ATAC-seq and, especially, a large number of H3K27ac datasets with known perturbations [45]. Consistent with the present study, they also identified monaLisa as one of the top methods, indicating its robustness to different types of epigenomic signals. On the other hand, they also reported MEIRLOP as a top method, which performed decently but not as well in our study. Of note, the workflow from Santana *et al.* focused on methods that take peak-level fold-changes across condition as input, and therefore did not consider methods that compute per-sample activity scores, such as chromVAR. This, along with differences in the methods included and datasets used (in particular, a large number of H3K27ac datasets versus ATAC-seq simulations in our case), make the two studies complementary, although largely agreeing on some common observations.

4.2 Open questions

An issue that was barely touched in our study is the definition of the consensus peaks on which to work. How peak calling should best be done in ATAC-seq, especially across samples, is still an open debate. Our attempt to restrict to peaks called on nucleosome-free fragments, in particular the fact that counting all or only nucleosome-free fragments on nucleosome-free peaks both give similar results, suggests that the peaks, rather than the counting, is the decisive factor. However, further study will be needed to establish what type of peak calls, and whether leniency or stringency in the thresholding, is best for TF activity inference.

We have shown that using motif archetypes rather than an extensive set of similar motif does not generally improve results except for the multivariate approach, but the impact of the motif collection (or binding sites or other proxy to it) used will require further research. Our simulations indicate that, as expected, with constant perturbation strength the difficulty in detecting a difference in TF activity is strongly related to how good a proxy for binding the motif is. Unfortunately, it is well known that motifs are in general poorly informative of TF binding. Further work should be put in comparing

motif collections against other forms of binding site predictions, or collections of experimental binding sites [46].

Finally, our benchmark is based on datasets with a small number of samples (2-3 per experimental group). While such a design is very common, it is possible that some methods will benefit more from larger sample sizes. In particular, it is plausible that the sample-permutation mode of diffTF, and more generally methods that compute per-sample scores, would benefit more from increased sample sizes.

5 Supporting information

Table S1 Name and description of each method variation tested.

Fig S1 Overview of the benchmark datasets and TFs. A. Volcano plot of the peak-level differential accessibility analyses, illustrating the extent and significance of changes upon treatment in each dataset. The dashed line represents a 0.05 FDR threshold, and indicated are the number of ATAC-seq peaks passing this threshold, as well as the proportion of peaks with an uncorrected p-value lower than 0.05. **B.** Enrichment for experimental binding sites of the factor (main TF after which the datasets are named) in ATAC-seq peaks containing matches for the respective motifs, versus all ATAC-seq peaks. **C.** Proportion of ChIP-seq peaks (overlapping ATAC-seq peaks) that contain matches for the respective motif.

Fig S2 Reproducibility of chromVAR-based differential accessibility analysis. A. Each color represents a dataset. The line and shaded area respectively represent the mean and standard error of the pairwise correlations across 7 independent runs of the chromVAR>limma pipeline for each dataset and setting, of the ranks of the (union of) top 20 motifs. **B. Illustration of the motif-specific position weights for the insertionModel method. C. Illustration of the computation of the network score.** The motifs are ranked for each method, and for each k=1:100 the proportion of top k motifs that are part of the known network of the true TF is computed. The area under the curve is then reported, relative to the AUC of the best possible motif ordering for that factor (gray area). The example plotted here is from the GATA1 dataset. **D. Agreement between the two rank-based metrics.** Each point represents the results of a method (colour) in a dataset (shape). The network score is not entirely comparable across datasets. Beside this effect, and except at very low (1-3) or high (>100) ranks (i.e. where the ranks stop being discriminatory), there is a good agreement between the two metrics. Capping ranks to 100, the overall correlation between between ranks and network score is -0.61, and the median correlation across datasets is -0.75.

Fig S3 Differences in distributions of chromVAR z-scores across samples A-B. Example of a subset of ATAC-seq samples from Caradonna, Paul and Marrocco (2022) (GSE200670), where different samples show different shapes of z-scores distributions across motifs/TFs. The colors indicate different experimental conditions. When peaks are not resized to all have the same width, as recommended, the distributions are often globally shifted (**A**, left). Even when peaks are resized (**A**, right), differences in the width of the z-scores distributions that are not related to experimental groups can persist in some datasets. This is not specific to the z-scores, but also present in the bias-corrected deviations (**B**). **C-D.** Effect on replacing the internal library size normalization performed by chromVAR with a TMM normalization. TMM normalization does not reduce the differences in z-score distributions (**C**), nor does it improve the rank of the true TF (**D**).

Fig S4 Impact of normalizing the chromVAR z-score distributions. A-B: Effect on the rank of the true TF (**A**) and its t-value (**B**) of normalizing the chromVAR activity scores (i.e. z-scores or deviations) distributions across samples before running limma. In **B**, for datasets where a downregulation was expected, the t-values were inverted. The heatmap colors show the methods deviation relative to the dataset average, while the numbers indicate the actual (rounded) t-value. **C:** Impact on the single-cell z-scores. As a proxy to signal-to-noise ratio, we computed the standardized mean difference of the respective motif for each guide RNA (individual lines) to the mean of the control guide RNAs. Centering is always beneficial (Wilcoxon p-value $<$ 8e-06), unit-variance scaling is often beneficial (Wilcoxon p-value $\sim$ 0.07) but sometimes detrimental, and quantile normalization also tends to be beneficial (Wilcoxon p-value $\sim$ 0.041).

Fig S5 Rank-based metrics of all variants, including smooth quantile normalization in GC bins. Same as Fig 1, but showing all method variants tested.

Fig S6 Impact of stringency in motif matching. Rank of the true TF (**A**) and Precision-Recall (**B**) of selected methods, when using a stringent motif matching vs the default.

Fig S7 Impact of using motif archetypes. Rank of the true motif/archetype (**A**) and Precision-Recall (**B**) of selected methods, when using motifs archetypes instead of the full collection of motifs. Given the difference in the size of the two sets of motifs, ranks were capped at 144 for the purpose of creating the heatmap.

Fig S8 Impact of using nucleosome-free fragments. Comparison of the rank of the true motif obtained by the top method of each family of approaches, when using alternative peaks or overlap counts based on nucleosome-free fragments. 'All frags on NF peaks' stands for counting overlaps with fragments of any size, but on peaks called on nucleosome-free (NF) fragments only. 'NF frags on NF peaks' stands for using only NF-fragments, i.e. counting overlaps of NF fragments on peaks called only on NF fragments. Finally, 'NF frags on full peaks' stands for counting overlaps of only NF fragments, but on the standard peaks (called using all fragments).

Fig S9 Simple aggregation of top methods' results does not improve inferences. A. The results of the top method from each of the two families of approach (on the left) are compared to those of simple aggregation methods (on the right). Beside the established Fisher's, Stouffer's and Simes' methods, a rank-based permutation approach was tested, establishing the probability of having a sum of ranks across methods lower or equal to the observed one. We also tested using limma on merged per-sample activity scores from chromVAR and GCsmooth+MLM. **B.** Same comparison, in terms of Precision and Recall at adjusted p-value $<=$ 0.05.

Fig S10 Overview of the semi-simulations. A. 6 baseline ATAC-seq lymphoblastoid cell lines (LCLs) are divided into two groups. Differences will be introduced through downsampling of fragments in one of the groups, and eventual per-sample GC and fragment length biases are introduced. **B.** The same as shown in main Fig 4A. The extent of downsampling is based on the enrichment over the input of the ChIP-seq peak, following one of two differential activity scenarios: an activation and a haploinsufficiency (see Fig 4B). The magnitude of the effect is further scaled up and down (perturbation strength). **C.** Settings used for the semi-simulated datasets. The resulting datasets can be divided into three groups, a first set where only the strength of

the perturbation was varied by a factor (perturbation strength), a second and third one were additionally differences in respectively fragment length distribution and GC were introduced across samples.

Fig S11 Overview of the simulated TFs. Data shown corresponds to the semi-simulated datasets with a perturbation strength of 1 in the activation paradigm. **A.** Number of peaks, proportion of ChIP-seq and/or ATAC-seq peaks with a motif, and fold-enrichment of the ATAC-seq peaks overlapping ChIP-seq peaks (i.e. bound peaks) relative to all ATAC-seq peaks for the motif across the four TFs. **B-C.** Fold-enrichment of all tested motifs among the bound peaks (relative to all ATAC-seq peaks), with the true motif labeled in black. The dots are colored either by co-occurence with the true motif (**B**) or by GC content of the motif, i.e. the summed probabilities of G and C across positions in the position probability matrix divided by the total sum (**C**). **D.** GC content distributions of the ChIP-seq peaks of each TF. With dotted lines the GC content of the true motif is indicated, with dashed lines the median GC content of the motifs showing stronger enrichment in the ChIP-seq peaks than the true motif.

Fig S12 Distribution of fragment counts and variance of the semi-simulations. A. Per peak fragment count distributions of the original (dataset: "0") and simulated samples (datasets: "1", "0") with 1-fold and 3-fold perturbation strengths, for an example TF (CEBPB) across both paradigms. **B.** Mean-variance relationship of the original and simulated samples.

Fig S13 MA plots of the simulated effects. For each factor, simulated paradigm and perturbation strength, the distribution of peaks $\log_2$FC of raw fragment counts between conditions as a function of mean accessibility is plotted as 2d kernel density, separating either (**A**) ATAC-seq peaks overlapping a ChIP-seq peak or not, or (**B**) ATAC-seq peaks containing a motif or not for the TF. For visibility, the peaks were subsampled to a more comparable size across the semi-simulated datasets (1.2×10^5 peaks per dataset). Colored Points indicate the means of the respective distributions.

Fig S14 Positive control of the semi-simulations. Shown are ranks and significance assigned by each method to the true motif across simulation parameters (TF, paradigm and perturbation strength). In light gray are the normal simulations (as shown in Fig 4), while dark line and colored boxes correspond to the positive controls where a difference between the groups was only introduced in the peaks overlapping a motif of the true TF.

Fig S15 Heatmap of performance of variants of top methods on the semi-simulated datasets. As in main Fig 1, the heatmap shows the ranks the methods achieved across different simulated perturbation strengths and fold change distributions (activation and haploinsufficiency). The annotation bar to the left shows the mean network and archetype scores on perturbation strengths greater than zero. Ranking of the methods was performed as in Fig 1 and described in the methods section using all the datasets with a perturbation strength greater than zero. Where needed imputation of missing values for the ranking was performed as previously described using only datasets with the same perturbation strength and fold change distribution. ATACseqTFEA was excluded for runtime reasons and diffTF could not be run on two of the datasets.

Fig S16 Robustness to simulated variations in fragment length. In addition to the simulated TF effect, two different fragment length distributions (**B**) were simulated by differential sampling of the fragments according to their length for the different samples as in scheme **A**. The distribution in light green corresponds to sampling mononucleosome containing fragments with higher probability, whereas for the distribution in dark green nucleosome-free fragments were sampled preferentially. The results of the methods are shown in **C**. Colors of the points correspond to the rank of the true TF obtained, point size if the corresponding adjusted p-value was found to be significant (adjusted p-value $<= 0.05$).

Fig S17 Robustness to simulated variations in GC content. In addition to the simulated TF effect, two different GC content distributions (**B**) were simulated by differential sampling of the fragments according to their GC content for the different samples as in scheme **A**. The results of the methods are shown in **C**. Colors of the points correspond to the rank of the true TF obtained, point size if the corresponding adjusted p-value was found to be significant (adjusted p-value $<= 0.05$).

Fig S18 Estimated precisions per dataset. A. Heatmap of precision estimates obtained per dataset and method on the 11 real datasets (**A**), as well as averaged across each perturbation strength of the semi-simulations (**B**).

Fig S19 Successful downregulation of NFkB1 at the protein-level using the TRAFTAC. Western Blot (**A**) and derived densitometry (**B**) of NFkB1 (aka p50) in HEK cells transfected with the dCas9-HT7 plasmid and a TRAFTAC (either against NFkB or a scrambled control) for 6h, before treating the cells TNF-α (to trigger NFkB activation) and the PROTAC against the Halo tag (HP14) for 19h. The DMSO control sample was the same as a NFkB1, except that it received DMSO instead of the PROTAC.

6 Financial Disclosure Statement

This work was supported by research grants (ETH-25 02-2 to PLG and 23-2 ETH-015 to KG) from the Swiss Federal Institute of Technology (ETH Zurich). The salary of ES is paid by the ETH-25 02-2 grant. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

7 Acknowledgments

We thank Lukas Burger for providing the BANP motif.

References

[1] H. Ait Ghezala et al. "Translation termination efficiency modulates ATF4 response by regulating ATF4 mRNA translation at 5' short ORFs". In: *Nucleic Acids Res* 40.19 (2012). Ait Ghezala, Hayet Jolles, Beatrice Salhi, Samia Castrillo, Katia Carpentier, Wassila Cagnard, Nicolas Bruhat, Alain Fafournoux, Pierre Jean-Jean, Olivier eng Research Support, Non-U.S. Gov't England 2012/08/21 Nucleic Acids Res. 2012 Oct;40(19):9557-70. doi: 10.1093/nar/gks762. Epub 2012 Aug 16., pp. 9557–70. ISSN: 1362-4962 (Electronic) 0305-1048 (Print) 0305-1048 (Linking). DOI: 10.1093/nar/gks762. URL: https://www.ncbi.nlm.nih.gov/pubmed/22904092.

[2] K. M. Vattem and R. C. Wek. "Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells". In: *Proc Natl Acad Sci U S A* 101.31 (2004). Vattem, Krishna M Wek, Ronald C eng R01 GM049164/GM/NIGMS NIH HHS/ GM49164/GM/NIGMS NIH HHS/ GM643540/GM/NIGMS NIH HHS/ Research Support, U.S. Gov't, P.H.S. 2004/07/28 Proc Natl Acad Sci U S A. 2004 Aug 3;101(31):11269-74. doi: 10.1073/pnas.0400541101. Epub 2004 Jul 26., pp. 11269–74. ISSN: 0027-8424 (Print) 1091-6490 (Electronic) 0027-8424 (Linking). DOI: 10.1073/pnas.0400541101. URL: https://www.ncbi.nlm.nih.gov/pubmed/15277680.

[3] M. Kobayashi et al. "Phosphorylation of ATF-1 enhances its DNA binding and transcription of the Na,K-ATPase alpha 1 subunit gene promoter". In: *Nucleic Acids Res* 25.4 (1997). Kobayashi, M Shimomura, A Hagiwara, M Kawakami, K eng Research Support, Non-U.S. Gov't England 1997/02/15 Nucleic Acids Res. 1997 Feb 15;25(4):877-82. doi: 10.1093/nar/25.4.877., pp. 877–82. ISSN: 0305-1048 (Print) 1362-4962 (Electronic) 0305-1048 (Linking). DOI: 10.1093/nar/25.4.877. URL: https://www.ncbi.nlm.nih.gov/pubmed/9016641.

[4] Richard Sever and Christopher K. Glass. "Signaling by Nuclear Receptors". In: *Cold Spring Harbor Perspectives in Biology* 5.3 (Mar. 1, 2013). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, a016709. ISSN: , 1943-0264. DOI: 10.1101/cshperspect.a016709. URL: http://cshperspectives.cshlp.org/content/5/3/a016709 (visited on 12/12/2023).

[5] Jason D. Buenrostro et al. "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position". In: *Nature Methods* 10.12 (Dec. 2013). Number: 12 Publisher: Nature Publishing Group, pp. 1213–1218. ISSN: 1548-7105. DOI: 10.1038/nmeth.2688. URL: https://www.nature.com/articles/nmeth.2688 (visited on 02/10/2022).

[6] M. Ryan Corces et al. "An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues". In: *Nature Methods* 14.10 (Oct. 2017), pp. 959–962. ISSN: 1548-7105. DOI: 10.1038/nmeth.4396. URL: https://www.nature.com/articles/nmeth.4396 (visited on 02/26/2019).

[7] Jason D. Buenrostro et al. "Single-cell chromatin accessibility reveals principles of regulatory variation". In: *Nature* 523.7561 (July 2015). Number: 7561 Publisher: Nature Publishing Group, pp. 486–490. ISSN: 1476-4687. DOI: 10.1038/nature14590. URL: https://www.nature.com/articles/nature14590 (visited on 12/11/2023).

[8] Gennady Korotkevich et al. "Fast gene set enrichment analysis". In: *bioRxiv* (2021), p. 060012. DOI: 10.1101/060012. URL: https://www.biorxiv.org/content/biorxiv/early/2021/02/01/060012.full.pdf.

[9] I. Mompel P. Badia et al. "decoupleR: ensemble of computational methods to infer biological activities from omics data". In: *Bioinform Adv* 2.1 (2022). Badia-I-Mompel, Pau Velez Santiago, Jesus Braunger, Jana Geiss, Celina Dimitrov, Daniel Muller-Dott, Sophia Taus, Petr Dugourd, Aurelien Holland, Christian H Ramirez Flores, Ricardo O Saez-Rodriguez, Julio eng England 2023/01/27 Bioinform Adv. 2022 Mar 8;2(1):vbac016. doi: 10.1093/bioadv/vbac016. eCollection 2022., vbac016. ISSN: 2635-0041 (Electronic)

2635-0041 (Linking). DOI: 10.1093/bioadv/vbac016. URL: https://www.ncbi.nlm.nih.gov/pubmed/36699385.

[10] M. J. Alvarez et al. "Functional characterization of somatic mutations in cancer using network-based inference of protein activity". In: *Nat Genet* 48.8 (2016), pp. 838–47. ISSN: 1546-1718 (Electronic) 1061-4036 (Print) 1061-4036 (Linking). DOI: 10.1038/ng.3593. URL: https://www.ncbi.nlm.nih.gov/pubmed/27322546.

[11] A. N. Schep et al. "chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data". In: *Nat Methods* 14.10 (2017). Schep, Alicia N Wu, Beijing Buenrostro, Jason D Greenleaf, William J eng U19 AI057266/AI/NIAID NIH HHS/ P50 HG007735/HG/NHGRI NIH HHS/ T32 HG000044/HG/NHGRI NIH HHS/ UM1 HG009436/HG/NHGRI NIH HHS/ R01 GM111990/GM/NIGMS NIH HHS/ 2017/08/22 Nat Methods. 2017 Oct;14(10):975-978. doi: 10.1038/nmeth.4401. Epub 2017 Aug 21., pp. 975–978. ISSN: 1548-7105 (Electronic) 1548-7091 (Print) 1548-7091 (Linking). DOI: 10.1038/nmeth.4401. URL: https://www.ncbi.nlm.nih.gov/pubmed/28825706.

[12] D. Machlab et al. "monaLisa: an R/Bioconductor package for identifying regulatory motifs". In: *Bioinformatics* 38.9 (2022). Machlab, Dania Burger, Lukas Soneson, Charlotte Rijli, Filippo M Schubeler, Dirk Stadler, Michael B eng Research Support, Non-U.S. Gov't England 2022/02/25 Bioinformatics. 2022 Apr 28;38(9):2624-2625. doi: 10.1093/bioinformatics/btac102., pp. 2624–2625. ISSN: 1367-4811 (Electronic) 1367-4803 (Print) 1367-4803 (Linking). DOI: 10.1093/bioinformatics/btac102. URL: https://www.ncbi.nlm.nih.gov/pubmed/35199152.

[13] I. Berest et al. "Quantification of Differential Transcription Factor Activity and Multiomics-Based Classification into Activators and Repressors: diffTF". In: *Cell Rep* 29.10 (2019), 3147–3159 e12. ISSN: 2211-1247 (Electronic). DOI: 10.1016/j.celrep.2019.10.106. URL: https://www.ncbi.nlm.nih.gov/pubmed/31801079.

[14] S. Baek, I. Goldstein, and G. L. Hager. "Bivariate Genomic Footprinting Detects Changes in Transcription Factor Activity". In: *Cell Rep* 19.8 (2017), pp. 1710–1722. ISSN: 2211-1247 (Electronic). DOI: 10.1016/j.celrep.2017.05.003. URL: https://www.ncbi.nlm.nih.gov/pubmed/28538187.

[15] Ian C. McDowell et al. "Glucocorticoid receptor recruits to enhancers and drives activation by motif-directed binding". In: *Genome Research* 28.9 (Sept. 1, 2018). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 1272–1284. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.233346.117. URL: https://genome.cshlp.org/content/28/9/1272 (visited on 04/27/2023).

[16] Jane Guan et al. "Therapeutic Ligands Antagonize Estrogen Receptor Function by Impairing Its Mobility". In: *Cell* 178.4 (Aug. 8, 2019), 949–963.e18. ISSN: 1097-4172. DOI: 10.1016/j.cell.2019.06.026.

[17] J. Kain et al. "Pioneer factor Foxa2 enables ligand-dependent activation of type II nuclear receptors FXR and LXRalpha". In: *Mol Metab* 53 (2021). Kain, Jessica Wei, Xiaolong Reddy, Nihal A Price, Andrew J Woods, Claire Bochkis, Irina M eng R01 DK121059/DK/NIDDK NIH HHS/ Research Support, N.I.H., Extramural Germany 2021/07/12 Mol Metab. 2021 Nov;53:101291. doi:

10.1016/j.molmet.2021.101291. Epub 2021 Jul 8., p. 101291. ISSN: 2212-8778 (Electronic) 2212-8778 (Linking). DOI: 10.1016/j.molmet.2021.101291. URL: https://www.ncbi.nlm.nih.gov/pubmed/34246806.

[18] S. E. Pierce, J. M. Granja, and W. J. Greenleaf. "High-throughput single-cell chromatin accessibility CRISPR screens enable unbiased identification of regulatory networks in cancer". In: *Nat Commun* 12.1 (2021). Pierce, Sarah E Granja, Jeffrey M Greenleaf, William J eng RM1 HG007735/HG/NHGRI NIH HHS/ T32 CA009302/CA/NCI NIH HHS/ U19 AI057266/AI/NIAID NIH HHS/ P50 HG007735/HG/NHGRI NIH HHS/ UM1 HG009436/HG/NHGRI NIH HHS/ UM1 HG009442/HG/NHGRI NIH HHS/ Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. England 2021/05/22 Nat Commun. 2021 May 20;12(1):2969. doi: 10.1038/s41467-021-23213-w., p. 2969. ISSN: 2041-1723 (Electronic) 2041-1723 (Linking). DOI: 10.1038/s41467-021-23213-w. URL: https://www.ncbi.nlm.nih.gov/pubmed/34016988.

[19] R. S. Grand et al. "BANP opens chromatin and activates CpG-island-regulated genes". In: *Nature* 596.7870 (2021). Grand, Ralph S Burger, Lukas Grawe, Cathrin Michael, Alicia K Isbel, Luke Hess, Daniel Hoerner, Leslie Iesmantavicius, Vytautas Durdu, Sevi Pregnolato, Marco Krebs, Arnaud R Smallwood, Sebastien A Thoma, Nicolas Vermeulen, Michiel Schubeler, Dirk eng Research Support, Non-U.S. Gov't England 2021/07/09 Nature. 2021 Aug;596(7870):133-137. doi: 10.1038/s41586-021-03689-8. Epub 2021 Jul 7., pp. 133–137. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking). DOI: 10.1038/s41586-021-03689-8. URL: https://www.ncbi.nlm.nih.gov/pubmed/34234345.

[20] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. "Trimmomatic: a flexible trimmer for Illumina sequence data". In: *Bioinformatics* 30.15 (2014), pp. 2114–2120.

[21] Ben Langmead and Steven L. Salzberg. "Fast gapped-read alignment with Bowtie 2". In: *Nature Methods* 9.4 (Apr. 2012), pp. 357–359. ISSN: 1548-7105. DOI: 10.1038/nmeth.1923. URL: https://www.nature.com/articles/nmeth.1923 (visited on 12/12/2023).

[22] Haley M. Amemiya, Anshul Kundaje, and Alan P. Boyle. "The ENCODE Blacklist: Identification of Problematic Regions of the Genome". In: *Scientific Reports* 9.1 (June 27, 2019). Number: 1 Publisher: Nature Publishing Group, p. 9354. ISSN: 2045-2322. DOI: 10.1038/s41598-019-45839-z. URL: https://www.nature.com/articles/s41598-019-45839-z (visited on 12/11/2023).

[23] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: *Bioinformatics* 26.1 (Jan. 1, 2010), pp. 139–140. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp616. URL: https://doi.org/10.1093/bioinformatics/btp616 (visited on 04/24/2023).

[24] Ivan V. Kulakovskiy et al. "HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis". In: *Nucleic Acids Research* 46 (D1 Jan. 4, 2018). Publisher: Oxford Academic, pp. D252–D259. ISSN: 0305-1048. DOI: 10.1093/nar/gkx1106. URL: https://academic.oup.com/nar/article/46/D1/D252/4616875 (visited on 03/09/2020).

[25] Paul Shannon and Matt Richards. "MotifDb: An Annotated Collection of Protein-DNA Binding Sequence Motifs". In: *Bioconductor* (2020). R package version 1.32.0.

[26] Charles E. Grant, Timothy L. Bailey, and William Stafford Noble. "FIMO: scanning for occurrences of a given motif". In: *Bioinformatics* 27.7 (Apr. 1, 2011), pp. 1017–1018. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr064. URL: https://doi.org/10.1093/bioinformatics/btr064 (visited on 12/12/2023).

[27] J. Vierstra et al. "Global reference mapping of human transcription factor footprints". In: *Nature* 583.7818 (2020). Vierstra, Jeff Lazar, John Sandstrom, Richard Halow, Jessica Lee, Kristen Bates, Daniel Diegel, Morgan Dunn, Douglas Neri, Fidencio Haugen, Eric Rynes, Eric Reynolds, Alex Nelson, Jemma Johnson, Audra Frerker, Mark Buckley, Michael Kaul, Rajinder Meuleman, Wouter Stamatoyannopoulos, John A eng R35 HG011317/HG/NHGRI NIH HHS/ U54 HG007010/HG/NHGRI NIH HHS/ UM1 HG009444/HG/NHGRI NIH HHS/ Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't England 2020/07/31 Nature. 2020 Jul;583(7818):729-736. doi: 10.1038/s41586-020-2528-x. Epub 2020 Jul 29., pp. 729–736. ISSN: 1476-4687 (Electronic) 0028-0836 (Print) 0028-0836 (Linking). DOI: 10.1038/s41586-020-2528-x. URL: https://www.ncbi.nlm.nih.gov/pubmed/32728250.

[28] Matthew E Ritchie et al. "limma powers differential expression analyses for RNA-sequencing and microarray studies". In: *Nucleic Acids Research* 43.7 (2015), e47. DOI: 10.1093/nar/gkv007.

[29] Sven Heinz et al. "Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities". In: *Molecular Cell* 38.4 (May 28, 2010). Publisher: Elsevier, pp. 576–589. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2010.05.004. URL: https://www.cell.com/molecular-cell/abstract/S1097-2765(10)00366-7 (visited on 12/12/2023).

[30] Yaowu Liu et al. "ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies". In: *The American Journal of Human Genetics* 104.3 (Mar. 7, 2019), pp. 410–421. ISSN: 0002-9297. DOI: 10.1016/j.ajhg.2019.01.002. URL: https://www.sciencedirect.com/science/article/pii/S0002929719300023 (visited on 12/11/2023).

[31] Luz M. Garcia-Alonso et al. "Transcription factor activities enhance markers of drug sensitivity in cancer". In: *Cancer Research* (Jan. 1, 2017), canres.1679.2017. ISSN: 0008-5472, 1538-7445. DOI: 10.1158/0008-5472.CAN-17-1679. URL: http://cancerres.aacrjournals.org/content/early/2017/12/09/0008-5472.CAN-17-1679 (visited on 05/01/2019).

[32] Jianhong Ou. "ATACseqTFEA: Transcription Factor Enrichment Analysis for ATAC-seq". In: *Bioconductor* (2023). R package version 1.5.0. URL: https://github.com/jianhong/ATACseqTFEA.

[33] Myong-Hee Sung et al. "DNase Footprint Signatures Are Dictated by Factor Dynamics and DNA Sequence". In: *Molecular Cell* 56.2 (Oct. 2014), pp. 275–285. ISSN: 10972765. DOI: 10.1016/j.molcel.2014.08.016. URL: https://linkinghub.elsevier.com/retrieve/pii/S1097276514006716 (visited on 12/11/2023).

[34] Nathaniel P. Delos Santos, Lorane Texari, and Christopher Benner. "MEIRLOP: improving score-based motif enrichment by incorporating sequence bias covariates". In: *BMC Bioinformatics* 21.1 (Sept. 2020), p. 410. ISSN: 1471-2105. DOI: 10.1186/s12859-020-03739-4. URL: https://doi.org/10.1186/s12859-020-03739-4 (visited on 05/05/2024).

[35] Yoav Benjamini and Yosef Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". en. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57.1 (1995). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1995.tb02031.x, pp. 289–300. ISSN: 2517-6161. DOI: 10.1111/j.2517-6161.1995.tb02031.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1995.tb02031.x (visited on 06/07/2024).

[36] The ENCODE Consortium. "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489.7414 (Sept. 2012), pp. 57–74. ISSN: 1476-4687. DOI: 10.1038/nature11247. URL: https://www.nature.com/articles/nature11247 (visited on 01/23/2024).

[37] Aaron R. Quinlan and Ira M. Hall. "BEDTools: a flexible suite of utilities for comparing genomic features". In: *Bioinformatics* 26.6 (Mar. 15, 2010), pp. 841–842. ISSN: 1367-4811, 1367-4803. DOI: 10.1093/bioinformatics/btq033. URL: https://academic.oup.com/bioinformatics/article/26/6/841/244688 (visited on 02/13/2024).

[38] Michele Gabriele et al. "YY1 Haploinsufficiency Causes an Intellectual Disability Syndrome Featuring Transcriptional and Chromatin Dysfunction". In: *The American Journal of Human Genetics* 100.6 (June 1, 2017), pp. 907–925. ISSN: 0002-9297. DOI: 10.1016/j.ajhg.2017.05.006. URL: http://www.sciencedirect.com/science/article/pii/S0002929717301933 (visited on 08/21/2019).

[39] Kusal T. G. Samarasinghe et al. "Targeted degradation of transcription factors by TRAFTACs: TRAnscription Factor TArgeting Chimeras". In: *Cell Chemical Biology* 28.5 (May 20, 2021). Publisher: Elsevier, 648–661.e5. ISSN: 2451-9456, 2451-9448. DOI: 10.1016/j.chembiol.2021.03.011. URL: https://www.cell.com/cell-chemical-biology/abstract/S2451-9456(21)00150-1 (visited on 12/12/2023).

[40] Mark D. Robinson and Alicia Oshlack. "A scaling normalization method for differential expression analysis of RNA-seq data". In: *Genome Biology* 11.3 (Mar. 2, 2010), R25. ISSN: 1474-760X. DOI: 10.1186/gb-2010-11-3-r25. URL: https://doi.org/10.1186/gb-2010-11-3-r25 (visited on 12/12/2023).

[41] Koen Van Den Berge et al. "Normalization benchmark of ATAC-seq datasets shows the importance of accounting for GC-content effects". In: *Cell Reports Methods* 2.11 (Nov. 2022), p. 100321. ISSN: 26672375. DOI: 10.1016/j.crmeth.2022.100321. URL: https://linkinghub.elsevier.com/retrieve/pii/S2667237522002119 (visited on 08/27/2023).

[42] Dhanusha A. Nalawansha and Craig M. Crews. "PROTACs: An Emerging Therapeutic Modality in Precision Medicine". In: *Cell Chemical Biology* 27.8 (Aug. 20, 2020). Publisher: Elsevier, pp. 998–1014. ISSN: 2451-9456, 2451-9448. DOI: 10.1016/j.chembiol.2020.07.020. URL: https://www.cell.com/cell-chemical-biology/abstract/S2451-9456(20)30293-2 (visited on 04/06/2021).

[43] Guadalupe Sabio and Roger J. Davis. "TNF and MAP kinase signalling pathways". In: *Seminars in Immunology* 26.3 (June 2014), pp. 237–245. ISSN: 10445323. DOI: 10.1016/j.smim.2014.02.009. URL: https://linkinghub.elsevier.com/retrieve/pii/S1044532314000220 (visited on 02/22/2024).

[44] Alexey M. Chumakov et al. "Modulation of DNA binding properties of CCAAT/enhancer binding protein epsilon by heterodimer formation and interactions with NFkappaB pathway". In: *Blood* 109.10 (Jan. 2007), pp. 4209–4219. ISSN: 0006-4971. DOI: 10.1182/blood-2005-09-031963. eprint: https://ashpublications.org/blood/article-pdf/109/10/4209/1477394/zh801007004209.pdf. URL: https://doi.org/10.1182/blood-2005-09-031963.

[45] Leonor Schubert Santana et al. "Benchmarking tools for transcription factor prioritization". English. In: *Computational and Structural Biotechnology Journal* 0.0 (May 2024). Publisher: Elsevier. ISSN: 2001-0370. DOI: 10.1016/j.csbj.2024.05.016. URL: https://www.csbj.org/article/S2001-0370(24)00164-8/fulltext (visited on 05/14/2024).

[46] Alexandra B Keenan et al. "ChEA3: transcription factor enrichment analysis by orthogonal omics integration". In: *Nucleic Acids Research* 47 (W1 July 2, 2019), W212–W224. ISSN: 0305-1048. DOI: 10.1093/nar/gkz446. URL: https://doi.org/10.1093/nar/gkz446 (visited on 12/12/2023).