Dissertation

zur Erlangung der Doktorwürde

an der

Gesamtfakultät für Mathematik,

Ingenieur- und Naturwissenschaften

der

Ruprecht-Karls-Universität Heidelberg

Thema:

Understanding gene regulation through the analysis of omics data

vorgelegt von:

Pau Badia i Mompel

10. Februar 2025

Gutachter:
Professor Dr. Carl Herrmann
Professor Dr. Julio Saez Rodriguez

Understanding gene regulation through the analysis of omics data

Pau Badia i Mompel
Oral examination: 10th February 2025

Abstract

The interactions between chromatin, transcription factors, and genes form intricate regulatory circuits, which can be modeled as gene regulatory networks (GRNs). Historically, GRNs have been inferred from bulk profiling omics data, as well as from literature sources. The emergence of single-cell multi-omics technologies has driven the creation of many novel computational methods that integrate genomic, transcriptomic, and chromatin accessibility data, allowing in principle to infer GRNs at better resolution. In the first chapter of this thesis, I describe the classic and new approaches to measure and model gene regulation through GRNs and their downstream applications. In the second chapter, I describe the development of decoupler, a computationally scalable framework for the inference of TF activities from omics data through the pairing of enrichment analysis with GRNs. There I also compare several enrichment methods and conclude that simple linear models outperform classic enrichment methods. Then, I showcase how decoupler together with transcription factor activity inference can be used to discover new biological insights in human diseases. In the third and last chapter, I showcase the design and implementation of Gene Regulatory nETwork Analysis (GRETA), a comprehensive cross-method benchmark of multimodal GRN inference, and compare their performance relative to several baselines. There I show that although the obtained GRNs have predictive properties and can moderately recover known biology, they do not exhibit causal properties, contrary to what is always assumed of them. Additionally, I show how they perform on par, or worse than literature-derived GRNs or GRNs inferred only from transcriptomics, suggesting that inferring de-novo regulatory programs might be an overly complex problem and that the incorporation of biological knowledge could aid in GRN inference.

Zusammenfassung

Die Interaktionen zwischen Chromatin, Transkriptionsfaktoren und Genen bilden komplexe regulatorische Netzwerke, die als Genregulationsnetzwerke (GRNs) modelliert werden können. Historisch gesehen wurden GRNs aus Omics-Daten von Bulk-Profiling sowie aus Literaturquellen abgeleitet. Das Aufkommen von Single-Cell-Multi-Omics-Technologien hat die Entwicklung vieler neuer computergestützter Methoden vorangetrieben, die genomische, transkriptomische und Chromatinzugänglichkeitsdaten integrieren und es in der Theorie ermöglichen, GRNs mit besserer Auflösung zu erschließen. Im ersten Kapitel dieser Dissertation beschreibe ich die klassischen und neuen Ansätze zur Messung und Modellierung der Genregulation durch GRNs sowie deren Anwendungen. Im zweiten Kapitel beschreibe ich die Entwicklung von decoupler, einem rechnerisch skalierbaren Framework zur Ableitung von TF-Aktivitäten aus Omics-Daten durch die Verknüpfung von Anreicherungsanalysen mit GRNs. Dort vergleiche ich auch verschiedene Anreicherungsmethoden und komme zu dem Schluss, dass einfache lineare Modelle klassische Anreicherungsmethoden übertreffen. Anschließend zeige ich, wie decoupler zusammen mit der Ableitung von Transkriptionsfaktoraktivitäten genutzt werden kann, um neue biologische Erkenntnisse bei menschlichen Krankheiten zu gewinnen. Im dritten und letzten Kapitel präsentiere ich das Design und die Implementierung von Gene Regulatory nETwork Analysis (GRETA), einem umfassenden Methodenvergleich für multimodale GRN-Ableitungen, und vergleiche ihre Leistung mit verschiedenen Baselines. Dort zeige ich, dass die gewonnenen GRNs zwar prädiktive Eigenschaften aufweisen und bekannte biologische Zusammenhänge moderat wiedergeben können, sie jedoch keine kausalen Eigenschaften zeigen, entgegen der üblichen Annahmen. Darüber hinaus zeige ich, dass sie gleichwertig oder schlechter abschneiden als GRNs, die aus Literaturquellen oder nur aus Transkriptomdaten abgeleitet wurden, was darauf hindeutet, dass die Ableitung von de-novo-Regulationsprogrammen ein übermäßig komplexes Problem sein könnte und die Einbeziehung biologischen Wissens die GRN-Ableitung unterstützen könnte.

Acknowledgements

The first time I heard about Heidelberg was when Celia, my girlfriend (now fiancée), found a PhD position here and without knowing anything about the city I decided to accompany her without a doubt, I just wanted to be with her. I was positively surprised to find out that it was such a hub for science but in particular for computational biology. I became even more positively surprised when Julio accepted me in his group to pursue my PhD in computational biology. This allowed Celia and I not only to pursue our professional careers, but also to finally start our lives living together for the first time, marking an important chapter in our lives.

Therefore, I want to start by immensely thanking Prof. Dr. Julio Saez Rodriguez for his unconditional trust, encouragement, support and freedom to pursue my academic goals during these years. His way to approach science, in a fun, friendly, collaborative and open way, is something that I will follow for the rest of my career. This philosophy has brewed a positive and friendly environment in the lab, as I couldn't be happier to have shared these years with so many amazing people. Be it to discuss science, celebrate triumphs, mourn bad news, laugh at memes or just plain gossip, you all have always made Heidelberg a second home (even if I couldn't stand the weather sometimes). For this, I thank all saezlab members, visitors and alumni that I had the chance to vibe with. On a personal note I'd like to thank Ricardo O. Ramirez Flores, Daniel Dimitrov and Sophia Müller-Dott for all their scientific input but most importantly for their friendship across these years.

I want to thank the funding provided by the Deutsche Forschungsgemeinschaft, Universitätsklinikum Heidelberg and GSK, which allowed me to pursue my scientific career (and pay rent at the same time).

No one would have guessed it, but I also have time to have friends outside of work. I'd like to thank my DKFZ gang, particularly Enrique Blanco Carmona for teaching me how to boulder and better understand my mental health, Sonia Jiménez Vázquez for always closing clubs together and Nicola Biondi for sharing all the juicy drama. Back home, I'd like to also thank all my friends from my master's and bachelor's degrees, and also my friends from Sant Cugat del Valles, my hometown.

And finally I'd like to thank my family and Celia, which I will do in my mother tongue, en català.

Mare i Pare, no puc estar més agraït de tot el suport i l'amor que m'heu donat d'ençà que vaig néixer, tot el vostre esforç es reflecteix en aquesta tesi, res de tot això no hauria estat possible sense vosaltres, us trobo a faltar. Joanet, a part de

discutir amb mi qui ha de desparar taula sempre has estat el meu germanet, moltes gràcies per aguantar-me i pel teu suport. Iber i Bru, moltes gràcies pel vostre amor incondicional, créixer amb vosaltres ha sigut una experiència meravellosa que guardaré sempre.

Celia, qui ho diria que estaríem aquí quan ens vam conèixer. Tot i que la gent al principi no donava ni un duro sobre nosaltres, així seguim, estimant-nos, aprenent i creixent junts. No em podria imaginar cap altra vida sense tu, t'estime, t'estimo, t'estim bonica.

Table of contents

Chapter 1: Gene regulation in systems biology

1.1. Quantification of gene regulation

Cells modulate gene transcription to orchestrate their activities in response to both internal and external cues[1]. Transcription is the process of copying genetic information from DNA to RNA, generating transcript molecules which are then translated into proteins to perform cellular functions[2]. Transcription is primarily controlled by transcription factors (TFs), proteins that attach to specific DNA sequences (DNA binding sites) and can either enhance or inhibit the transcriptional rate of their target genes[3]. Genomic DNA is compacted into structures called nucleosomes through interactions with structural proteins, forming the fundamental unit of chromatin, which makes many genes inaccessible to the transcriptional machinery. For transcription to proceed, the promoter region near a gene's transcription start site must be exposed by displacing the tightly packed nucleosomes. This shift in DNA accessibility can be initiated by the binding of pioneer TFs[4]. Additional TFs can bind to distal cis-regulatory elements (CREs) on the DNA, and together with cofactors and other proteins, interact with the promoter region through DNA loop extrusion[5]. This facilitates the recruitment and stabilization of the RNA polymerase complex, which synthesizes mRNA from the gene's DNA body (**Figure 1.1.**).

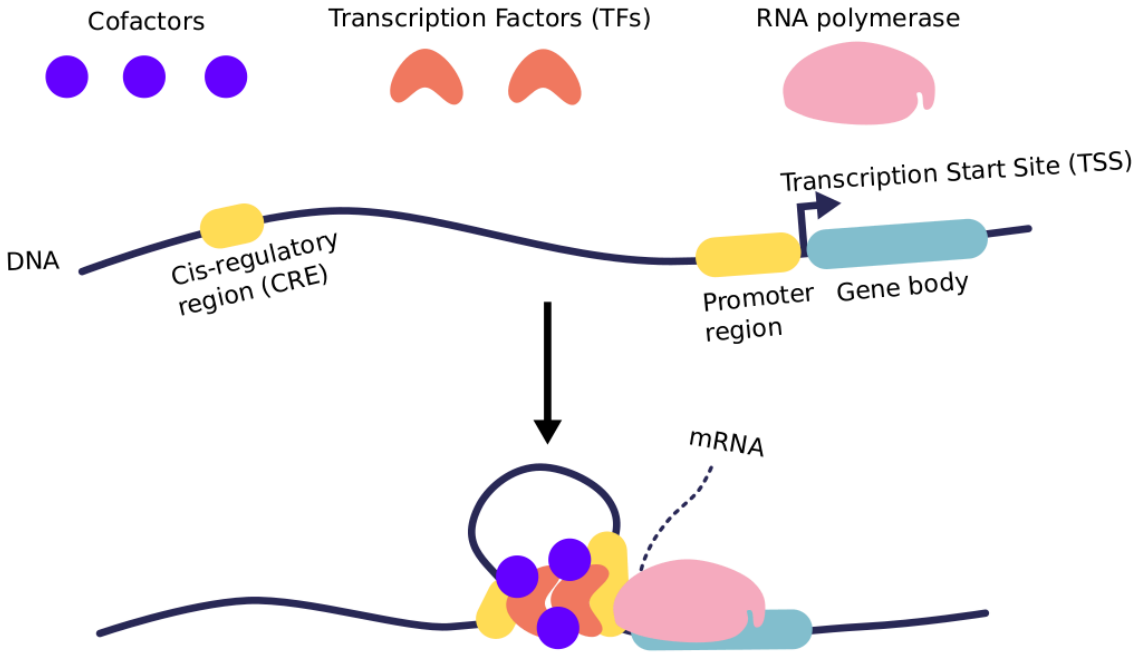

Figure 1.1. Key components of gene regulation. Transcription factors (TFs) bind to promoter regions and cis-regulatory elements (CREs), facilitating nucleosome displacement and exposing the transcription start site (TSS). The collaborative action of TFs, cofactors, and other proteins aids in the recruitment and stabilization of the RNA polymerase complex, which is responsible for synthesizing mRNA from the gene's DNA.

Omics technologies are laboratory techniques that enable high-throughput measurement of some of these biological molecules, providing a detailed snapshot of the underlying biology within a biological system of interest[6]. The most commonly used omics technologies for profiling gene regulation are RNA sequencing (RNA-seq)[7] and Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq)[8] (**Figure 1.2.**).

RNA-seq measures the quantity of transcript molecules in a biological sample[7]. First, RNA is extracted from the sample and converted into complementary DNA through reverse transcription. Next, the complementary DNA is fragmented and sequenced using high-throughput sequencing technologies. The resulting sequence reads are then aligned to a reference genome or transcriptome to quantify gene expression levels.

The most widely used technique to identify regions of open chromatin in the genome is ATAC-seq. This is due to its simplicity and relatively low cost, but other assays like DNase-seq[9] and NOME-seq[10] are also available (discussed elsewhere[11]). The process begins with the addition of a hyperactive mutant Tn5 transposase to the

sample, which inserts sequencing adapters into accessible DNA regions and fragments them from the rest of the genome. The DNA fragments are then sequenced and aligned to a reference genome. Regions with significantly higher read accumulation are annotated as open, a process known as peak calling[12]. Peaks are defined by their genomic coordinates in the following format: chromosome, starting base pair, and ending base pair. For instance, the peak "chr1-34593-34793" refers to a peak located on human chromosome 1, beginning at position 34,593 and ending at position 34,793, spanning a total of 200 base pairs. In this thesis I will interchangeably use peak or CRE to refer to defined regions of open chromatin in the genome.

Other omics technologies relevant to gene regulation exist but are less commonly used due to various limitations. Assays such as chromatin immunoprecipitation followed by sequencing (ChIP-seq)[13] and cleavage under targets and tagmentation (CUT&Tag)[14] provide genome-wide measurements of TF binding, also in the form of peaks, yet their low throughput and reliance on TF-specific antibodies limit their broader application. Additionally, chromosome conformation capture techniques such as Hi-C[15] measure the likelihood of physical interactions between genomic regions, revealing how distal CREs may be brought into proximity through three-dimensional DNA looping[5]. However, generating Hi-C data is costly and its results are hard to reproduce from run to run, which limits its widespread application[16].

Omics technologies can profile molecules at different resolutions (**Figure 1.2.**). The first iteration of omics technologies pooled millions of cells from a biological sample into a single molecular readout[6]. This resolution, referred to as “bulk”, contains a mixture of various cell-types in a single profile. For studying gene regulation, this resolution is problematic, as it cannot distinguish cell-type specific regulatory programs[17]. Current advances have allowed the profiling of individual nuclei, reaching single-cell resolution and allowing cell-type specific measurements for both transcriptomics (snRNA-seq)[18] and chromatin accessibility (snATAC-seq)[19]. However, these technologies disassociate the input samples into single nuclei, losing the information of their original cell organization and co-localization within the tissue.

State-of-the-art methods have recently been developed to simultaneously profile individual omic layers alongside the spatial location of reads within a two-dimensional tissue section[20]. Spatial technologies can be categorized into two classes: next-generation sequencing-based methods, which encode positional information onto transcripts before sequencing, and imaging-based methods such as

in situ sequencing or in situ hybridization[21]. Both approaches present unique challenges. Next-generation sequencing methods measure more features but sacrifice resolution and require deconvolution approaches to reach single-cell resolution[22]. In contrast, imaging methods provide higher resolution but measure fewer features and require segmentation techniques to define cell boundaries[23]. Additionally, both single-cell and spatial technologies suffer from large fractions of observed zeros in their readouts, an effect commonly termed as technical dropouts[24]. This sparsity complicates the analysis and interpretation of the data generated by these technologies.

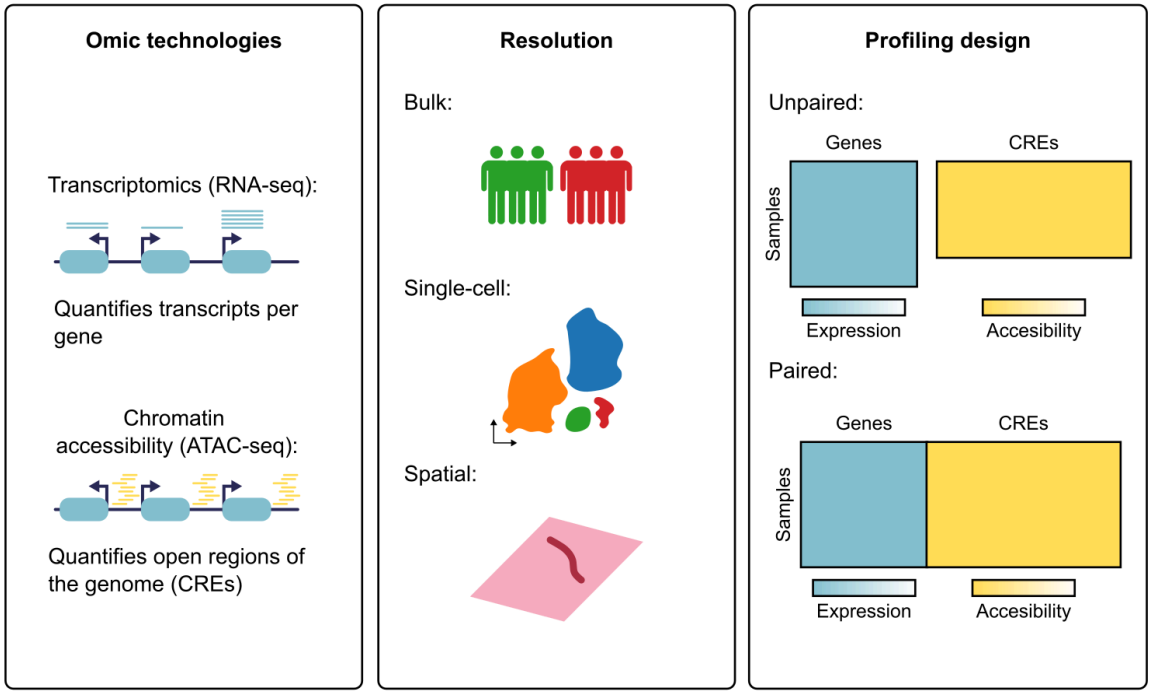

Figure 1.2. The most used omics technologies to quantify gene regulation are RNA-seq and ATAC-seq, which measure transcript abundance and genome-wide chromatin accessibility, respectively. Omics profiling can be done at the tissue (bulk), single-cell or spatial level. For single-cell or spatial, both omics can be profiled simultaneously from the same cells or spots. Depending on the profiling design, the generated datasets can be paired when the data was profiled from exactly the same cells or spots, or unpaired when they were profiled from different observations.

Individual omics layers provide a comprehensive view of biological processes but are largely descriptive, limiting their ability to enhance mechanistic understanding of gene regulation. Multi-omics techniques, which measure multiple omics simultaneously, offer the potential to better identify causal chains of events in gene regulation[25]. Multi-omics can be profiled from exactly the same cells, generating paired data, or from different cells coming from the same biological sample,

resulting in unpaired data (**Figure 1.2.**). In the case of unpaired data, integration approaches are needed to match cells between modalities[25]. Relevant for studying gene regulation, several technologies have been released and commercialized to jointly measure both RNA-seq and ATAC-seq at the single-cell level[26–28], with recent advancements also extending to spatial resolution[29].

Omics technologies are often used to profile molecules from both a control group and a group of interest, such as a disease or treatment cohort. Statistical or machine learning methods are then employed to associate features relevant to these groups. However, interpreting the biological significance of thousands of features that change between conditions is challenging. This issue is further exacerbated with multi-omics, as associations between omics layers also need to be considered. For example, the interplay between chromatin accessibility and gene expression can result in various combinations: a gene TSS may be open and expressed, open and repressed, closed and expressed, or closed and repressed. In systems biology, network-based approaches are used to summarize and interpret complex genome-wide relationships, which will be discussed in the following section.

1.2. Gene regulatory networks (GRNs)

Gene regulatory networks (GRNs) are interpretable computational models that represent gene expression regulation as networks[30,31]. GRNs can incorporate various elements of gene regulation, such as TFs, splicing factors, long non-coding RNAs, microRNAs, and metabolites. In my thesis, I focus on the simplest form of GRNs, which only captures the interactions between TFs and their target genes (**Figure 1.3.**). For alternative GRN representations, refer to the following reviews[32–35]. GRN interactions may be directed or undirected, indicating the presence or absence of causal relationships between genes. They can also be signed to represent positive or negative regulation and/or weighted to reflect their regulation strength. The fundamental unit of GRNs consists of a given TF and its associated target genes, which is referred to as a "regulon". The assembly of multiple connected regulons forms the overall structure of a GRN.

The study of GRNs has been a long-standing challenge in biology, exemplified by the seminal work from the 1960s that characterized the bacterial lactose (lac) operon[36], or their use in developmental biology at the beginning of this century[30]. Understanding their structure and dynamics is essential for learning how cellular identity is created and maintained[37], with significant implications for cell fate engineering[38] and disease prevention[39].

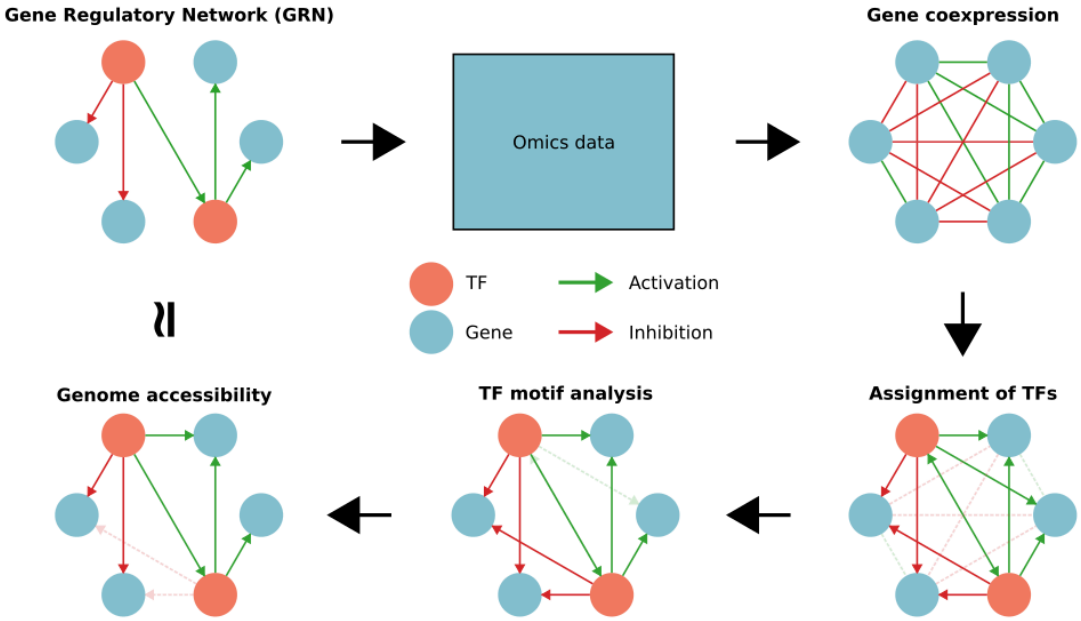

Figure 1.3. Gene regulatory networks (GRNs) can be constructed from measured omics data and further refined by incorporating additional information, such as TF binding predictions or chromatin accessibility, to potentially get closer to the real regulation system. In these networks, the nodes represent TFs and their regulated target genes, while the edges between nodes indicate the regulatory interaction, whether it is activation or inhibition.

Historically, GRNs have been assembled from experimentally validated regulatory interactions compiled in databases[40–42]. However, these interactions come from various cell-types, conditions and even sometimes organisms, making these GRNs generalistic. This poses a problem, as gene regulation can be very specific to the biological context at hand, with some TFs changing their targets and their mode of regulation depending on which other TFs are expressed in a particular cell-type or tissue[3].

GRNs can be inferred de novo from omics data using computational modeling approaches, which assume that the effects of a "hidden" underlying GRN are reflected in the measured data[43]. When enough data is available, this approach has the potential to generate networks that are better contextualized for the biological question at hand than generalistic networks. Different strategies for inferring GRNs are discussed in the next section.

1.3. Inference of GRNs

Over the years, various computational strategies have been developed to infer GRNs from omics data[31]. Early methods used individual omics at bulk resolution, specifically transcriptomics. Some later approaches attempted to combine different bulk readouts into a single network, but data generation remained challenging. GRN inference requires a large number of observations, and bulk resolution provides only a single molecular readout per sample, necessitating many tissue samples, which can be expensive and difficult to obtain. The advent of single-cell technologies, which generate thousands of observations for each sample, has made it easier to infer GRNs across different cell types, differentiation trajectories, and biological conditions. As a result, and with the introduction of single-cell multimodal profiling technologies, there has been a surge in novel GRN inference methods[31].

All GRN inference methods are based on variations of regression, a statistical method used to model the relationship between a dependent variable and one (univariate) or more (multivariate) independent variables[44]. The general formula of regression is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

Where:

- y is the dependent variable (the outcome being predicted),
- β_0 is the intercept (the value of y when all x's are zero),
- $\beta_1, \beta_2, \cdots, \beta_n$ are the coefficients for the independent variables $x_1, x_2, \cdots, x_n$, representing the effect of each variable on y,
- $x_1, x_2, \cdots, x_n$ are the independent variables (predictors),
- ϵ is the error term, accounting for the difference between the predicted and actual values of y.

For GRN inference, omics features measured across observations can serve as both dependent and independent variables. Based on which and how many different omics readouts are available, the modeling strategy will change. Most GRN methods use linear regression, but other alternatives exist. In the following sections several strategies are explained.

Regardless of the type of regression being used, it is essential to ensure sufficient variability between features. Extra caution is required when working with

single-cell data, as cell-type-specific GRNs are commonly inferred (**Figure 1.4.**). This is generally not an issue when cell-types show continuous variability, such as in developmental trajectories. However, it becomes problematic when using typical single-cell atlases, where most cells are already in a stable state and the observed variability is mainly between cell types. If not handled carefully, important cell-type-specific regulatory interactions could be missed during the regression process. For this reason, it is often preferable to infer GRNs at the atlas level rather than focusing solely on individual cell-types.

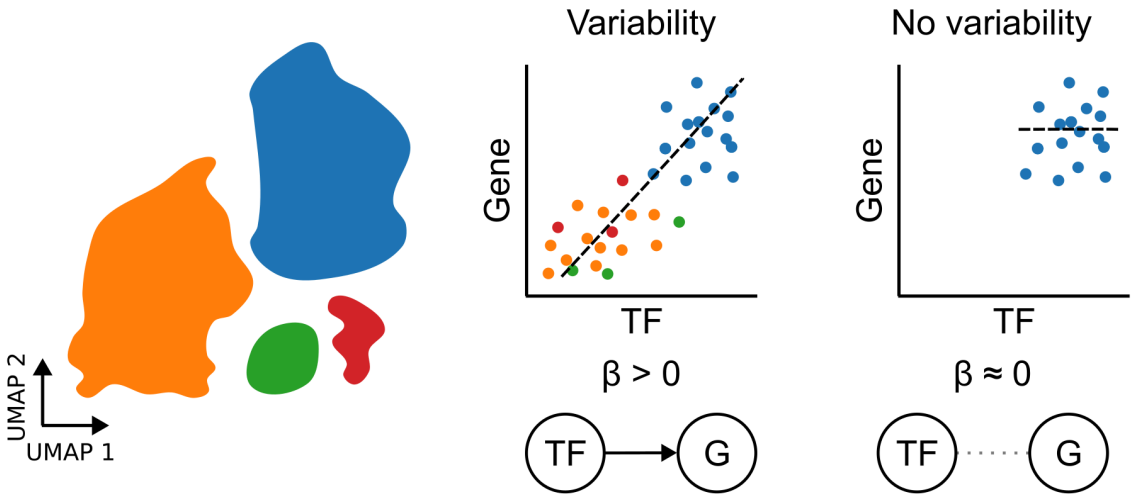

Figure 1.4. Example of GRN inference using univariate regression from snRNA-seq data. The observed gene expression of a target gene is explained by the expression levels of a TFs. When the whole atlas is used, there is enough variability for cell-type specific interactions to be recovered. When only the specific cell-type is used, these are missed due to low variability.

1.3.1. From single-omics

Methods in this category fit models that aim to explain the observed variability in gene expression based on the expression of other genes, or the accessibility of a CRE based on the accessibility of other CREs. Weighted gene co-expression network analysis (WGCNA)[45] has been one of the simplest and most widely used approaches. It performs pairwise correlations across the transcriptome to identify modules of co-expressed genes, resulting in a gene co-expression network formed by undirected interactions. While useful for unsupervised identification of gene modules, the absence of causal regulatory links limits its interpretability and often leads to a high number of spurious associations.

To overcome these challenges, methods like GENIE3[46] and its faster version GRNBoost2[47] distinguish TFs from target genes based on known regulatory activity[48] and then train models to predict target gene expression using only TF expression data. This strategy reduces the number of interactions considered and transforms undirected interactions into directed connections, introducing causal relationships.

Granger causality applied to temporally ordered RNA-seq data can also be used to obtain directed edges[49]. This approach tests whether past measurements, such as previous TF expression, can accurately predict future events, such as subsequent target gene expression. While generating temporal datasets with bulk resolution is challenging, single-nucleus snRNA-seq has made it easier to capture dynamic cell states in continuous processes like development, differentiation, or disease progression[50,51]. These processes create cells with gradually shifting transcriptomic profiles, forming cell trajectories. Trajectory inference, also called pseudotime analysis, orders these cells along a predefined trajectory based on gene expression similarities. LEAP[52] and SINCERITIES[53] are examples of GRN inference methods that use pseudotime ordering to determine the directionality between genes. However, pseudotime ordering requires the user to define the starting point, which can be arbitrary and non-trivial[54]. One potential solution is to compute "RNA velocity", which estimates the direction of cellular development by analyzing differences between unspliced and spliced mRNA reads[55]. Yet, technical limitations persist with RNA velocity, as it has been shown to fail even in developmental datasets where it should perform well according to the model assumptions[56,57].

Inferring GRNs from transcriptomics data alone can still lead to false positives, as key regulatory mechanisms, such as chromatin accessibility, are ignored. Furthermore, since many steps are required for a TF mRNA transcript to become a functional protein, transcript levels of TFs may not provide sufficient information[58]. These limitations can affect the accuracy of GRN inference, and as a result, these methods generally achieve only moderate success in accurately producing GRNs[59–61].

TF binding measurements from ChIP-seq or CUT&Tag can be used to construct GRNs by linking TF binding sites to potential target genes[62]. Despite the availability of some high-throughput alternatives[63,64], profiling TF binding remains expensive and is restricted to TFs with available antibodies. Moreover, using TF binding data alone often involves assigning bound TFs to target genes based on their nearest genomic proximity, overlooking possible distal interactions that play a key role in gene regulation[3].

Another approach is to solely use chromatin accessibility data to identify gene regulatory elements potentially targeted by TFs. These methods infer GRNs in two steps: first, TFs are assigned to CREs, and second, these elements are linked to neighboring target genes. In the first step, large-scale genome-wide binding data stored in databases is used to extract the most likely genomic sequences to which specific TFs bind, known as TF binding motifs[65]. Several databases have curated these assays and compiled collections of TF binding motifs for model organisms (**Table 1.1.**). Additionally, motif matcher algorithms have been developed to predict TF binding events by analyzing TF binding motifs and genomic sequence match (**Table 1.1.**). These algorithms calculate the probability of a TF binding event based on motif sequences and filter for significant matches. Given that different methods model TF binding in distinct ways, and that motif databases have different motif coverages, TF binding results may vary and should be carefully evaluated during GRN inference.

In the second step, regulatory elements are linked to genes based on their genomic proximity, as distal CREs such as enhancers or silencers generally interact with promoter regions within a defined genomic distance[3]. Methods like ATAC2GRN[66], LISA[67], and SPIDER[68] follow this two-step approach. However, these methods make the assumption that an accessible promoter region indicates active transcription, which is not always the case.

Table 1.1. Popular TF binding motif databases and motif matcher algorithms used across methods.

Name	Url	Refs
Binding motif databases		
CIS-BP	http://cisbp.ccbr.utoronto.ca/	69
cisTarget databases	https://resources.aertslab.org/cistarget/databases/	70
HOCOMOCO	https://hocomoco11.autosome.org/	71
JASPAR	https://jaspar.genereg.net/	72
Motif matcher algorithms		
GimmeMotifs	https://gimmemotifs.readthedocs.io/	73
HOMER	http://homer.ucsd.edu/homer/motif/	74
MOODs	https://github.com/jhkorhonen/MOODS	75,76
PWMScan	https://ccg.epfl.ch/pwmtools/pwmscan.php	77
pycisTarget	https://pycistarget.readthedocs.io/	70

1.3.2. From multi-omics

Incorporating multiomics into GRN inference is crucial because, much like the parable of the blind men and the elephant[78], relying on a single data type provides only a partial view of gene regulation, while integrating multiple layers of data offers a more comprehensive and accurate understanding of gene regulation. In the past, few methods combined multiple omics for network inference due to the difficulty in generating and accessing such datasets. Nevertheless, early efforts began to explore this approach. For instance, a pioneering study integrated ChIP-seq and transcriptomics data to refine the assignment of TFs to target genes, without relying solely on proximity to the nearest gene[79]. Another example is SCENIC[80], an extension of GRNBoost2, which introduced the concept of pruning edges inferred from co-expression patterns based on TF binding motif enrichment at gene promoter regions. This enabled it to incorporate cis-regulation information into the modeling without explicitly measuring any readout of it.

Advancements in simultaneously profiling snATAC-seq and snRNA-seq, commonly referred to as the commercial name “multiome” from the company 10X, have enabled the generation and accumulation of datasets ideal for this type of inference. Early studies used independently generated multi-omics data to infer GRNs in contexts such as human myeloid cell differentiation[81], mouse embryonic

development[82], and HIV infection in dendritic cells[83]. However, these studies did not make their methods available as tools for others to use.

These were followed by an explosion of novel methods for GRN inference that leverage both snRNA-seq and snATAC-seq (**Table 1.2.**). In my thesis I focus on a small collection of these based on their popularity and ease of use, but many more are available, which I reviewed elsewhere[31]. Some of these do not require paired chromatin accessibility and transcriptomics profiles for each cell, as they either summarize read-outs across groups of cells or build GRNs independently for each modality followed by a merging step. In contrast, other approaches model both modalities simultaneously within the same cell. When data are unpaired, these methods can still integrate both modalities through integration techniques. To enhance usability, methods such as figr[84] introduced their own integration strategies to obtain paired data.

Multimodal GRN inference methods extend and combine the steps used by single-modality methods to reconstruct GRNs (**Figure 1.5.**). Methods begin by preprocessing the omics matrices, creating a pool of candidate TFs, CREs, and genes. In this step, some methods try to reduce the observed sparsity of omics data. For example, celloracle[85] and figr[84], use k-Nearest Neighbors imputation to average readout values between similar cells, or granie[86], which performs in-silico bulks of the molecular profiles, also known as "pseudo-bulking". Another strategy, employed by pando[87], is to refine the measured CRE genomic coordinates by excluding exonic regions to ensure a better downstream TF binding prediction. Others, like dictys, employ simple quality control thresholds to ensure that sparse observations or features are removed before inference.

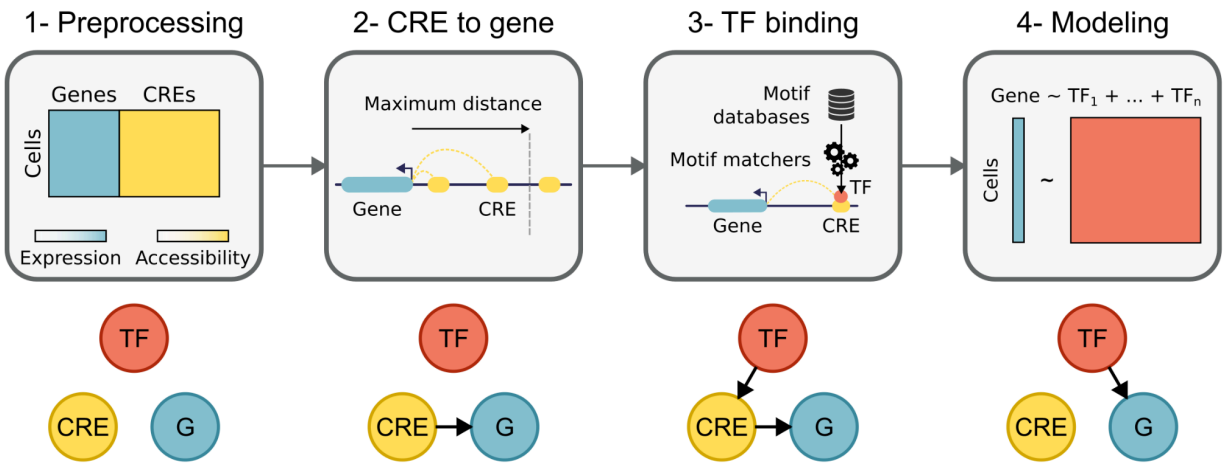

Figure 1.5. Flowchart of multimodal GRN methods. Methods start by preprocessing the omics matrices, generating a universe of candidate TFs, CREs and genes. This is followed by the assignment of CREs to neighboring genes based on genomic distances, generating CRE-Gene edges. Next, TF binding predictions are performed on the selected CREs using TF motif databases and motif matcher algorithms, generating TF-CRE-Gene triplets. Finally, methods use these triplets to build predictive models, generating simplified TF-Gene edges.

The next step involves linking CREs to nearby genes based on genomic distances, forming CRE-Gene connections. The goal of this filtering is to narrow the search space for each gene, minimizing computational demands and reducing the number of erroneous interactions. This step relies on the assumption that most genomic interactions tend to be proximal[88]. Methods vary in the genomic distance cutoffs that they consider (**Table 1.2.**). Some apply short-range cutoffs up to 10 kb, others use medium distances up to 100 kb, and some consider long-range interactions extending up to 1,000 kb. Since functionally validated interactions are common at shorter distances and decline significantly beyond 100 kb[3], the choice of cutoff can affect GRN inference results. However, there are known examples of CRE–gene interactions occurring over large distances, such as the enhancers of the MYC gene located nearly 2,000 kb downstream[89]. By restricting distance cutoffs, GRN methods might overlook critical long-range interactions, but extending it too much might introduce errors. Moreover, some interactions span across chromosomes, such as those observed in olfactory receptor selection[90], which current GRN methods are unable to account for.

Following the linking of CREs to genes, TF binding predictions are made on the identified CREs using TF motif databases and motif matcher algorithms, resulting in TF-CRE-Gene triplets, also known as enhancer regulons or "e-regulons". As mentioned in the previous section (**Section 1.3.1.**), methods use different, highly heterogeneous TF binding motif databases and prediction algorithms (**Table 1.2**).

Since TF binding motif databases vary in their coverage of TFs, and matching algorithms model binding differently, GRN inference methods may yield different results even when employing similar downstream modeling strategies. Most methods allow users to select alternative TF binding motif databases beyond the default option, but lock in the choice of motif matcher algorithm.

Table 1.2. Relevant tools for GRN inference from multimodal data.

Tool	Type of data	Modeling	Statistical framework	Motif database and matcher	Distance cutoffs	Lang.	Refs.
celloracle	Unpaired	Linear	Frequentist or Bayesian	CIS-BP, GimmeMotifs	500 kb, 500 kb	Python	[85]
dictys	Unpaired	Linear	Frequentist	HOCOMOCO, HOMER	500 kb, 500 kb	Python	[91]
figr	Paired	Linear	Frequentist	CIS-BP, MOODs	50 kb, 50 kb	R	[84]
granie	Paired	Linear	Frequentist	HOCOMOCO, PWMscan	250 kb, 250 kb	R	[86]
pando	Paired	Linear or non-linear	Frequentist or Bayesian	CIS-BP, MOODs	100 kb, gene body	R	[87]

Finally, the obtained TF-CRE-Gene triplets are used to build predictive models that simplify the network into classical TF-Gene connections using various mathematical strategies, generating a GRN. Some methods assume linear relationships between TFs, CREs, and genes, while others account for non-linear interactions (**Table 1.2.**). Linear modeling assumes that changes in one variable, such as gene transcripts, are directly proportional to changes in another variable, like TF transcripts or CRE accessibility. In contrast, non-linear modeling can capture more complex relationships, including synergistic effects between variables. However, these models tend to lose interpretability compared to linear ones, require more complex formulations, and often do not explicitly capture the direction of interactions. To enhance interpretability, some methods first infer regulatory interactions non-linearly and then determine interaction signs using correlation analysis between TF and gene expression. Although gene regulation is widely understood to be a non-linear process[92], many methods still opt for linear modeling for the mentioned reasons.

Regardless of the modeling strategy employed, the significance of regulatory interactions can be evaluated using either frequentist or Bayesian statistical frameworks (**Table 1.2.**). Frequentist approaches define the probability of an event as the frequency of its occurrence in a large number of repeated experiments, while

Bayesian methods interpret probability as the confidence in an event based on both observed data and prior knowledge. Although Bayesian methods can incorporate existing information, they generally require more computational resources than frequentist approaches, which can pose challenges when inferring genome-wide GRNs from large-scale single-cell data. Additionally, the effectiveness of Bayesian inference relies heavily on the quality of the prior knowledge used. Therefore, when no reliable prior data is available, or when it is considered inaccurate, frequentist inference may offer better accuracy.

Although all GRN methods use the same steps, they do not always adhere to the same order. For example, granie performs TF binding predictions before assigning CREs to genes. This choice is quite detrimental because it increases the number of tests, leading to more false positives and higher computational costs. Many CREs with predicted TF binding will ultimately not be linked to any target genes, rendering these predictions irrelevant. Therefore, the most efficient sequence is to begin with preprocessing, then link CREs to genes, perform TF binding predictions on the selected CREs, and finally model the scaffold triplets into a GRN.

1.4. Applications of GRNs

Once GRNs have been inferred, various analyses can be conducted to uncover novel biological insights into gene regulation (**Figure 1.6.**).

One fundamental analysis is topological analysis. While GRNs are straightforward and interpretable models of gene regulation, they can still contain many genes and an extensive amount of interactions. Network centrality measures can pinpoint key TFs or genes that are pivotal for network connectivity or information flow. Examples of network centrality measures include degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality. These metrics have been instrumental in identifying TFs that influence cell fate decisions across different biological contexts, such as direct lineage reprogramming[85], human myocardial infarction[93], and mouse development[94]. Another analysis to characterize the topology of GRNs involves spectral graph theory, which examines network properties through matrix representations. For example, non-negative matrix factorization of GRN adjacency matrices has revealed groups of TFs that collaboratively drive lineage transitions in mouse embryonic stem cells[95]. Similarly, clustering on the adjacency matrices of GRNs has identified key regulators in human hematopoietic cell differentiation[92] and in macrophage responses to interferon-γ[86]. The gene regulatory modules derived from these analyses can be further enriched with gene sets to elucidate their potential biological functions[96].

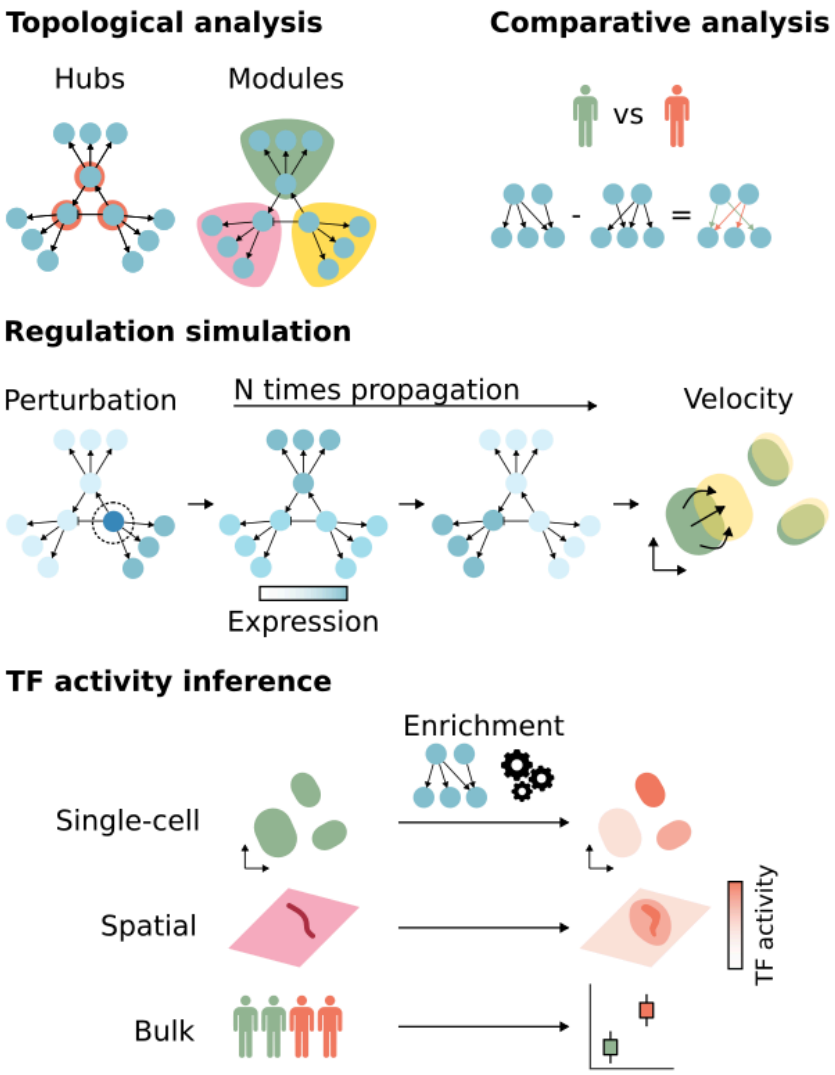

Figure 1.6. Available applications of GRNs. In topological analysis, network centrality measures can identify key TFs or genes that are highly connected. Clustering nodes based on their connectivity can reveal sub-network modules, which may be associated with specific biological functions. In comparative analysis, connectivities from different GRNs are contrasted via pairwise subtraction of TF-gene interactions, which can shed light on the rewiring of gene regulation across groups. In regulation simulation, propagation of changes in gene expression through the GRN over several iterations can be used to forecast cell fate decisions and the TFs driving it. In TF activity inference, enrichment methods identify active TFs from transcriptomics data at any resolution: bulk, single-cell or spatial.

Another approach for analyzing GRNs is comparative analysis, which can reveal rewiring events driving differences between cell types, states, disease conditions, treatment responses, and across species. The simplest method involves pairwise subtraction of TF-Gene interactions between GRNs, which has successfully

identified key regulators in various contexts, such as subpopulations of B cells in lymphocytic leukemia[97], TFs involved in fibroblast transdifferentiation to different human cell types[98], candidate Alzheimer's disease-specific trans-regulators[99], and cell state-specific regulators in human T cells[100]. This method has also been used to assess evolutionary conservation of TF-Gene interactions and transcriptional adaptation across species[101]. However, due to the sparse and noisy nature of GRNs, direct comparison of TF–gene interactions may lack robustness. Topic modeling techniques, such as latent Dirichlet allocation[102], which was originally developed for natural language processing, can generate dense, low-dimensional representations that reduce noise and potentially provide a more robust comparison of regulatory relationships. This approach has been effective in predicting cancer patient survival[103] and identifying rewiring events in human hematopoiesis[104].

One of the most interesting applications of GRNs is to simulate gene expression dynamics over time by iteratively propagating TF expression to target genes. This approach allows for in-silico perturbations, where the expression of a candidate TF is modified, and the resulting changes in the transcriptome are predicted after a specified number of iterations. The simulated gene expression values can then be compared with those from local neighboring cells to estimate cell identity transition probabilities, similar to RNA velocity analysis[55]. This strategy was first introduced by celloracle, which identified Zfp57 as crucial for generating and maintaining mouse-induced endoderm progenitors, an observation later confirmed through in vitro perturbation experiments[85]. This example highlights GRNs' ability to potentially model and capture complex regulatory events.

Finally, a straightforward yet valuable application of GRNs is inferring TF activities. By combining GRNs with enrichment methods, TF activities can be derived from transcriptomics data[105]. This approach integrates observed gene expression with GRN topology to identify TFs that may play significant roles in specific contexts. Common enrichment methods include GSEA[106], AUCell[80], and VIPER[107], among others[105]. In bulk studies, these methods have been used to identify druggable oncoproteins[107], stratify cell lines in response to drug treatments[108], and pinpoint a master regulator involved in metastasis promotion in breast carcinoma[109]. In single-cell studies, enrichment methods have revealed mechanisms of immunotherapy resistance in human T cells[110], regulators and inducers of oligodendroglioma[80], and potential druggable targets in pathological fibroblasts from COVID-19 patients[111]. Recently, these methods have also been applied to spatially resolved transcriptomics data, such as identifying regulators involved in the functional transition of cardiomyocytes across the border zone in

human myocardial infarction[93]. In the following chapter, I will elaborate on these methods and present some biological applications I have worked on.

Chapter 2: Inference of regulation activity and use cases

As mentioned in the last chapter, gene regulatory networks (GRNs) can be used to infer transcription factor (TF) activity from transcriptomics data through enrichment analysis. In this chapter, I will provide a more detailed explanation of what enrichment analysis entails and how it functions. Additionally, I will introduce the computational tool I have developed over the past few years, decoupler, a flexible and comprehensive framework for enrichment analysis[105]. Lastly, I will showcase real-world applications of TF activity inference using decoupler on human disease data.

2.1. Enrichment analysis

Omics technologies produce unbiased, high-dimensional molecular profiles. Their vast dimensionality, coupled with the intricate interconnectedness of the molecules they measure, makes their mechanistic interpretation challenging. For example, in transcriptomics identifying hundreds of relevant genes in a given biological context is common, but interpreting their functions individually is unfeasible. A key strategy to address this complexity is enrichment analysis, a statistical approach that reduces the dimensionality of omics data and generates more interpretable features[112,113].

Enrichment analysis involves testing whether a specific set of omics features is "overrepresented" or "coordinated" in the measured data compared to a background distribution (**Figure 2.1.**). Sets are predefined based on existing biological knowledge and may vary according to the specific omics technology used. For example, in RNA-seq, sets may consist of genes linked to particular pathways or TF regulons from a GRN. In ATAC-seq, sets can include CREs associated with GWAS traits or TF motifs. In phosphoproteomics, sets may encompass phosphosites that are known targets of specific kinases, among other examples.

These sets can be structurally represented as a bipartite graph. In this graph, the omics features act as child nodes, referred to as "targets", while the set to which they belong is the parent node, known as "sources". The edge weights indicate the strength of the association between the targets and the sources. Sets can consist of elements of a biological process, like the receptor and signaling kinases of a pathway, or the aftermath of it, such as which genes change after a perturbation. In

the latter case, these are referred to as "footprints" or signatures and are typically weighted[113].

Summary statistics f, commonly known as enrichment scores, are computed for a feature set S and a background B, which are then tested against the following hypotheses:

- $H_0: f(S) = f(B)$
- $H_1: f(S) \neq f(B)$

If the obtained $f(S)$ statistic shows a substantial difference from that of $f(B)$, the null hypothesis is rejected and the enrichment is considered significant.

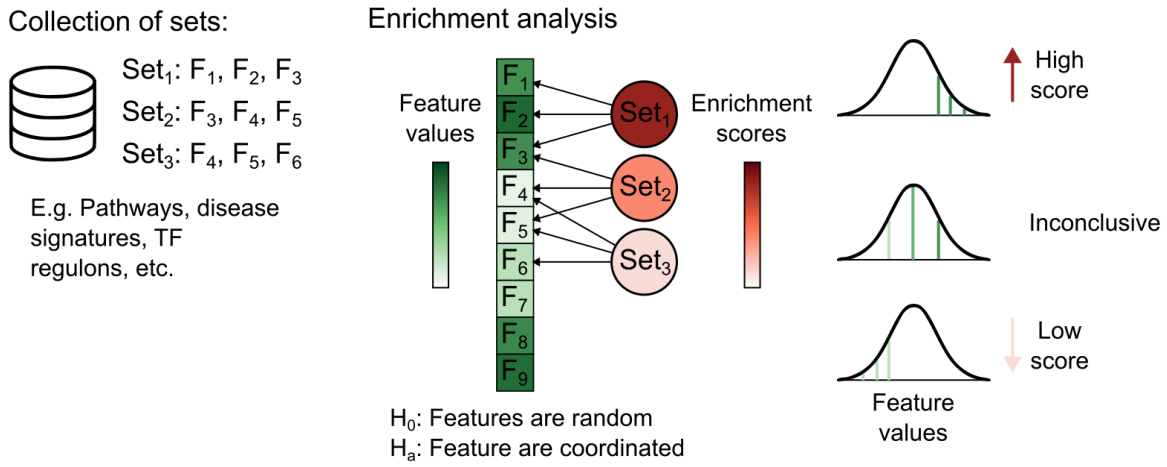

Figure 2.1. Enrichment analysis tests whether a collection of features is more coordinated than a background distribution. In this example of a competitive enrichment test across three different sets, the obtained enrichment scores are high when the set's features are consistently at the top of the distribution, and low when they are consistently at the bottom.

Depending on which background distribution is used, there are two categories of enrichment tests: self-contained and competitive[112]. In self-contained set testing, the sampling unit is the observation, meaning that multiple observations per group are necessary. It specifically evaluates whether the features within the set are differentially abundant between two groups, independent of the other measured features in the dataset. Alternatively, in competitive set testing, the focus is on whether the features in the predefined set are ranked higher relative to features not included in the set. Typically, this comparison is made with sets of similar sizes. The sampling unit in this case is the individual features, allowing the test to be conducted using a single observation. From now on, when I mention enrichment

analysis, I refer to competitive enrichment analysis, as the methods discussed in my thesis are all competitive and represent the most commonly used type of test.

Enrichment analysis can be conducted at the observation level, such as for a bulk profile, a single cell, or a spatial spot. It can also be performed after differential feature testing[114,115], utilizing a vector of contrast statistics. There is often a misconception that enrichment analysis is an activity score when done at the observation level, while it is considered classic enrichment when performed with contrast data, suggesting they are distinct concepts. However, both approaches represent the same process: a competitive enrichment analysis test against a background distribution. This confusion is further compounded by some methods using different terminology for functions depending on whether they use observational or contrast input data.

Enrichment statistics can be classified into two groups depending on the types of variables they use: categorical and numerical-based statistics.

2.1.1. Categorical-based statistics

Statistics in this category first perform feature selection based on the provided molecular readouts or feature contrast statistics, extracting a molecular-derived set X that is compared against a predefined biological set Y. Therefore, the two categories are, belongs to a molecular set ($\in X$ or $\in Y$) or not ($\notin X$, or $\notin Y$). In observational data, the top N features are chosen, whereas significance thresholds are applied in differential feature analysis results.

The most commonly used categorical-based method is overrepresentation, also called the one-tailed Fisher exact test[116]. It assesses the association between two binary variables in a contingency table, which shows how the variables are related by listing the counts of observations for each combination of categories (**Table 2.1.**). The rows represent the categories of one variable, while the columns represent the categories of the other variable. In the case of omics analysis, the binary categories are whether a feature belongs to one of the sets or not:

Table 2.1. Contingency table.

	$\in Y$	$\notin Y$
$\in X$	$X \cap Y$	$X - Y$
$\notin X$	$Y - X$	$X' \cap Y'$

The hypotheses of this non-parametric test are as follows:

- H_0: There is no association between the two variables.
- H_a: There is an association between the two variables in one direction.

A p-value can then be computed from the contingency table, which quantifies the significance of the overlap between the two sets. Normally, the resulting p-value is transformed using the $-\ log_{10}(p)$ to sort them in a descending direction.

Another popular categorical method is AUCell, introduced in the SCENIC publication[80]. It builds a recovery curve where the x-axis represents the filtered ranked molecular features X, and the y-axis tracks the number of features present in the biological Y set. Its f statistic is the "Area Under the Curve" (AUC) of this curve. If the features from set Y are highly ranked, the AUC will increase rapidly, reflecting strong enrichment. Conversely, if Y features are distributed more randomly or sparsely, the AUC grows more slowly. Unlike overrepresentation, AUCell only computes the AUC statistic without returning a p-value, as it does not perform any testing.

2.1.2. Numerical-based statistics

This group of statistics does not require feature selection, as seen in categorical-based methods. Instead, they use all available features and their numerical values. Many of these statistics treat the variables as ordinal and rank them accordingly.

The most famous numerical-based statistic is the weighted running sum, firstly introduced by GSEA[106]. It begins by ranking the observed molecular features or differential feature statistics in descending order, producing a vector r. The running sum for a feature set F is then calculated by moving down the ranked list, increasing the running sum when a feature from F is found, and decreasing it when a feature not in F is encountered. Mathematically, this is expressed as:

$$\delta(F,i) = \begin{cases} \frac{|r_i|}{\sum_{j \in F} |r_j|} & \text{if feature } i \in F \\ -\frac{1}{l} & \text{if feature } i \notin F \end{cases}$$

where:

- r_i is the value for feature i,
- r_j is the value for feature j in F,
- k is the number of features in F,
- N is the total number of features in r,
- and $l = N - k$ is the number of features not in F but present in r.

For each feature $\delta(F, i)$ is applied and stored as a sequence L:

$$L = \delta(F, i) \text{ for } i = 1, 2, \ldots, N$$

The enrichment score S is the signed maximum absolute deviation stored from the running sum:

$$S = L_{arg\,max\,|L|}$$

To evaluate the significance of the observed S, a permutation-based approach is used. Specifically, an empirical null distribution of enrichment scores is generated by repeatedly permuting the feature labels and recalculating the enrichment scores for each permutation.

$$p_{value} = \frac{S_{rand} \geq S}{P}$$

where:

- S_{rand} are the enrichment scores of the random permutations,
- and P is the total number of permutations.

This process yields a p-value that reflects the statistical significance of the observed enrichment score.

Finally, a normalized enrichment statistic NS is calculated by:

$$NS = \begin{cases} \frac{S}{\mu^+} & \text{if } S > 0 \\ \frac{S}{\mu^-} & \text{if } S < 0 \end{cases}$$

where:

- μ_+ is the mean of positive values in S_{rand},

- and μ_{-} is the mean of negative values in S_{rand}.

Another widely used numerical-based method is Virtual Inference of Protein-activity by Enriched Regulon analysis (VIPER)[107], which uses the Analytic Rank-based Enrichment Analysis (aREA) statistic. aREA calculates the enrichment score in a three-tailed approach. First, through a one-tail approach that ranks features based on their absolute values:

$$w = \frac{w}{max(|w|)}$$

$$l_{orig} = 1_{\{w \neq 0\}}$$

$$l = \frac{l_{orig}}{\sum_{i=1}^{k} \frac{l_i}{max(l_{orig})} max(l_{orig})}$$

$$q^{norm} = \Phi^{-1}(2|q - 0.5| + (1 + max(|q - 0.5|)))$$

$$S_1 = \sum_{i=1}^{k} q_i^{norm} l_i(1 - |w_i|)$$

where:

- $w \in [-1, +1]$ is a vector of interaction weights,
- $l \in [0, 1]$ is a vector of interaction likelihoods,
- $q \in [0, 1]$ is a vector of quantiles from the values of the omics input,
- Φ^{-1} is the inverse of the cumulative distribution function of the standard normal distribution,
- $q^{norm} \in [-\infty, +\infty]$ are the z-scores of the deviation of quantiles from 0.5,
- k is the number of features in q,
- and S_1 is the score of the one-tail approach.

This metric encodes for the magnitude of the score, irrespective of the interaction signs.

Then, a two-tail approach is employed, where the feature quantiles are z-transformed and weighted by their interaction weight and likelihood:

$$S_2 = \sum_{i=1}^{k} w_i l_i (\Phi^{-1}(q_i))$$

Unlike S_1, S_2 is taking the direction (sign) of interactions into consideration.

Afterwards, the three-tailed score is obtained:

$$S_3 = \begin{cases} (|S_2| + S_1) * sgn(S_2) & \text{if } S_1 > 0 \\ S_2 & \text{if } S_1 < 0 \end{cases}$$

The statistical significance of the obtained enrichment statistic is assessed by comparing it to a null model generated through an analytical approach that shuffles the features:

$$S_3^{norm} = S_3 \sqrt{\sum_{i=1}^{k} l_{o,i}^2}$$

$$p_{value} = \Phi(S_3^{norm})$$

When computing multiple sources simultaneously, a pleiotropic correction is also employed. In brief, all possible pairs of sources AB are generated under two conditions: (i) both A and B are significantly enriched ($p < 0.05$), and (ii) they share at least ten features. Subsequently, an enrichment score p-value is computed with aREA by both A (pA) and B (pB) based only on the shared features within the feature vector. Then the pleiotropy score (PS) is computed as:

$$PS = \begin{cases} \frac{1}{(1+|log_{10}(pB)-log_{10}(pA)|)^{\frac{20}{n_a}}} & \text{if } pA < pB \\ \frac{1}{(1+|log_{10}(pA)-log_{10}(pB)|)^{\frac{20}{n_b}}} & \text{if } pA > pB \end{cases}$$

where:

- n_a is the number of test pairs involving the source A,
- and n_b is the number of test pairs involving the source B.

This is then used to update l_{orig}:

$$\begin{cases} l_{\mathrm{i}}^{\mathrm{orig}} = PS * 1_{\{i \in A\}} & \text{if } pA < pB \\ l_{\mathrm{i}}^{\mathrm{orig}} = PS * 1_{\{i \in B\}} & \text{if } pA < pB \end{cases}$$

Finally, a new S_3^{norm} and p-value are calculated with aREA using the updated l_{orig}.

2.2. Decoupler, a flexible framework for enrichment analysis

In the past decade, numerous methods have been developed to infer enrichment scores from omics data, each incorporating unique assumptions and biases. Yet, the differences in scoring and predicting performance remain understudied. Additionally, running multiple methods is cumbersome due to their implementation in different software packages with numerous dependencies and varying formats for handling omics data and feature sets. Therefore, a unified framework would be beneficial to compare methods and make their access easier to the community. Previous efforts to compare or provide such frameworks were limited to older

methods, often required many dependencies, and were computationally inefficient[117–119]. To address these issues, I have developed decoupler[105], a flexible and efficient framework for enrichment analysis (code: https://github.com/saezlab/decoupler-py) (**Figure 2.2.**).

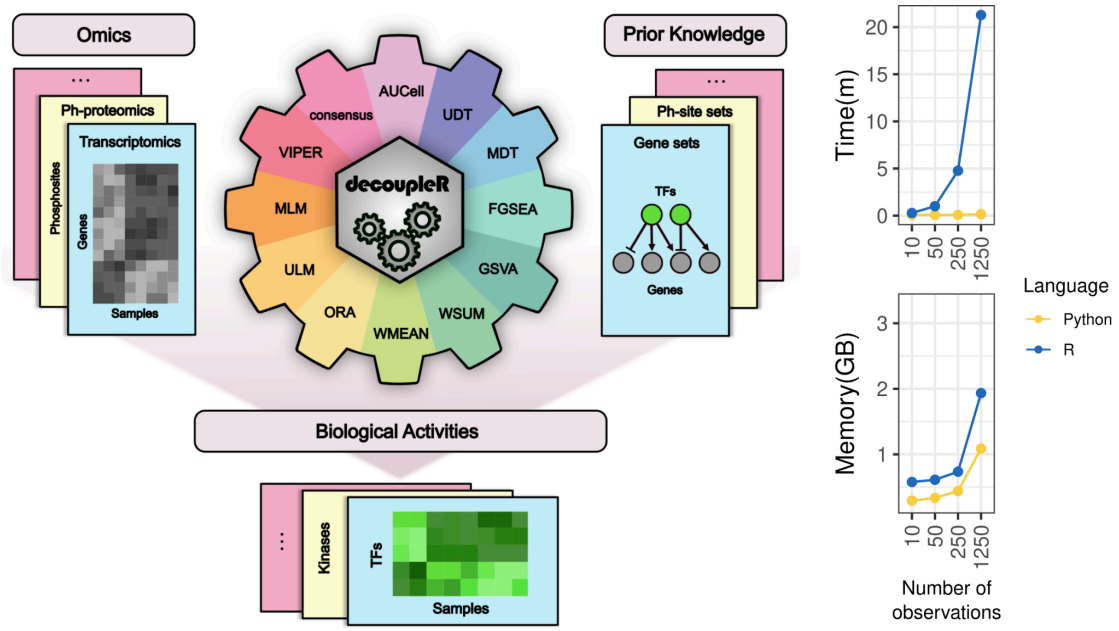

Figure 2.2. Inference of enrichment scores using the decoupleR workflow (left). Mean runtime and memory usage for R and Python were tested with a network of 250 sources and a matrix of 20,000 targets, using different numbers of observations on an Intel i7-8550U CPU @ 1.80 GHz (right).

Decoupler offers separate implementations in both R and Python. However, the R version is relatively slow and depends on several external packages, as it relies on the original method implementations. In contrast, I developed the Python version from the ground up, porting many of these methods for the first time into this programming language. To implement them I leveraged Numba[120], a Python accelerator that compiles functions to optimized machine code at runtime via the industry-standard LLVM compiler. This approach ensures efficient performance, enabling the methods to handle large-scale omics datasets, such as those at single-cell and spatial resolution. Indeed, methods in decoupler run relatively fast in R, but about three times faster in the Python version, with median runtimes of 1.44 ms in R and 0.44 ms in Python per observation and source (**Figure 2.2.**). A similar trend can be observed for the memory usage, with median memory of 1.53 GB in R and 0.62 GB in Python.

Decoupler includes ten different enrichment methods (**Table 2.2.**). Some can account for weights in the feature set, allowing them to consider the direction of change (positive or negative). This should not be confused with the numerical range

of the produced scores. For example GSEA can produce positive or negative scores but does not consider weights. This distinction is important when modeling sets that contain negative weights, such as a repressor TF that negatively regulates the expression of specific genes. If the genes linked to the repressor are underexpressed, it indicates that the repressor was active, resulting in a high positive score. In contrast, methods that do not account for weights would inaccurately predict a low positive score. Additionally, some methods test the enrichment scores they generate and provide a p-value.

Table 2.2. Enrichment methods available in decoupler.

Name	**Models weight**	**Tests p-value**	**Score range**	**Ref.**
Area Under the Cell (AUCell)			$\in [0, 1]$	[80]
Univariate Decision Tree (UDT)	✔		$\in [0, \infty]$	[105]
Multivariate Decision Trees (MDT)	✔		$\in [0, \infty]$	[105]
Gene Set Enrichment Analysis (GSEA)		✔	$\in [-\infty, \infty]$	[106]
Gene Set Variation Analysis (GSVA)			$\in [-1, 1]$	[121]
Weighted sum or mean (WSUM,WMEAN)	✔	✔	$\in [-\infty, \infty]$	[105]
OverRepresentation Analysis (ORA)		✔	$\in [0, \infty]$	[116]
Univariate Linear Model (ULM)	✔	✔	$\in [-\infty, \infty]$	[105]
Multivariate Linear Model (MLM)	✔	✔	$\in [-\infty, \infty]$	[105]
Virtual Inference of Protein-activity by Enriched Regulon analysis (VIPER)	✔	✔	$\in [-\infty, \infty]$	[107]
Consensus		✔	$\in [-\infty, \infty]$	[105]

Apart from classical enrichment methods, I also introduced a collection of methods that infer enrichment scores by predicting the observed omic vector using the set's feature weights. This formulation is very similar to the ones used to infer GRNs. The simplest of these is the Univariate Linear Model (ULM) (**Figure 2.3**). This approach uses the molecular features from one observation as the population of samples and it fits a linear model with a single covariate:

$$y \sim \beta_0 + \beta_1 x + \epsilon$$

where:

- y are the observed features, either molecular readouts or differential feature statistics,
- β_0 is the model's intercept,
- x is the vector of feature weights for a given set (if features do not belong to it, they are set to 0).
- β_1 is the coefficient of x,
- and ϵ is the error term.

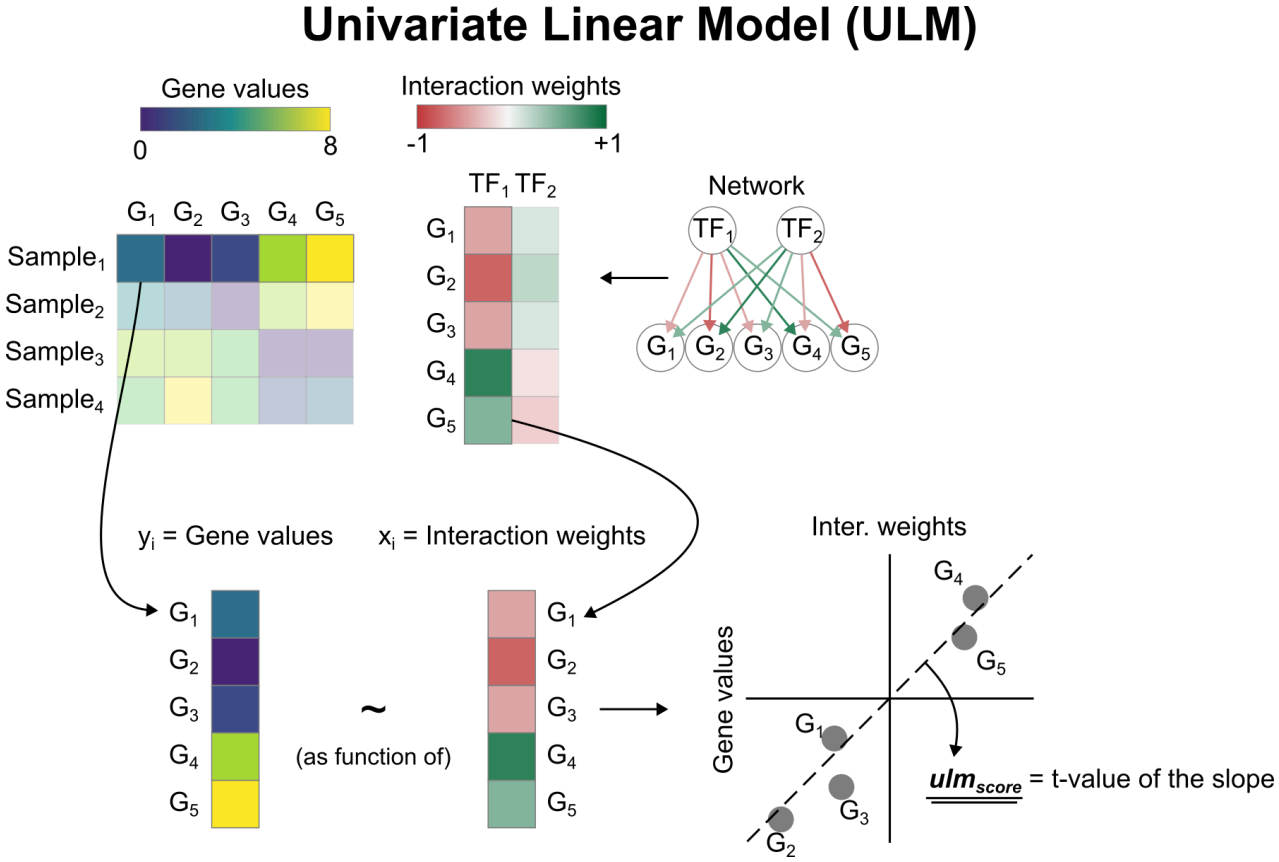

Figure 2.3. Univariate Linear Model (ULM) scheme. In this example, the observed gene expression of $Sample_1$ is predicted using the interaction weights of TF_1. Once the coefficient is fitted, its t-value is the enrichment score. Since the target genes that have negative weights are lowly expressed, and its positive target genes are highly expressed, the relationship between the two variables is positive so the obtained score is positive. Scores can be interpreted as active when positive, and repressive when negative.

In this case, the background distribution are the features not included in the set. Once the model is fitted, the t-value of β_1 is extracted since it encodes for both the direction of the slope with its sign (either positive or negative), and its significance with its magnitude. One can interpret the obtained score as the agreement between the observed features and the available prior knowledge of them. For example, if a

TF has high expression levels in its positive target genes, the score will be highly positive. The same applies if a TF's negative target genes are expressed at low levels. Conversely, if genes with positive weights are lowly expressed or if genes with negative weights are highly expressed, the resulting score will be negative.

A variation of this method is the Multivariate Linear Model (MLM), which fits all sets simultaneously rather than one at a time. This can lead to potentially better scores since it accounts for all sets together, rather than treating them independently. However, it may fail when there is collinearity between sets. Similar formulations include the Univariate Decision Tree (UDT) and its counterpart, the Multivariate Decision Tree (MDT), which use a non-linear approach with decision trees for modeling instead of linear regression.

Additionally, I implemented a consensus score that summarizes the results from all methods. For each method, scores are transformed into z-scores, separately for positive and negative values. First, values greater or less than zero are selected, mirrored to the opposite sign, and then a standard z-score is calculated. This transformation ensures comparability across methods while preserving the original sign (active or inactive). The final consensus score is the average of the z-scores across methods.

2.2.1. Comparison of enrichment methods

To compute scores and compare their similarities at the score and predictive performance level across methods, I used two collections of perturbation datasets. The first dataset consists of single-gene perturbation experiments[58], where a TF was perturbed, and the resulting molecular changes were measured, followed by the generation of contrast statistics[115]. The second dataset followed the same approach but used phosphoproteomics data, where kinases were perturbed instead[122]. For feature sets, I used the literature-derived GRN dorothea[123], and a kinase substrate network[122]. Then, for each contrast I computed enrichment scores across methods for each feature set available.

Methods displayed general similarities, with a median Spearman correlation of 0.52 for transcriptomics and 0.65 for phosphoproteomics (**Figure 2.4.**). There was also moderate agreement among methods in the top 5% of ranked regulators, with median Jaccard indexes of 0.23 and 0.21, respectively. Of note, UDT and GSVA were the two methods less similar to the rest, as seen by the hierarchical clusterings.

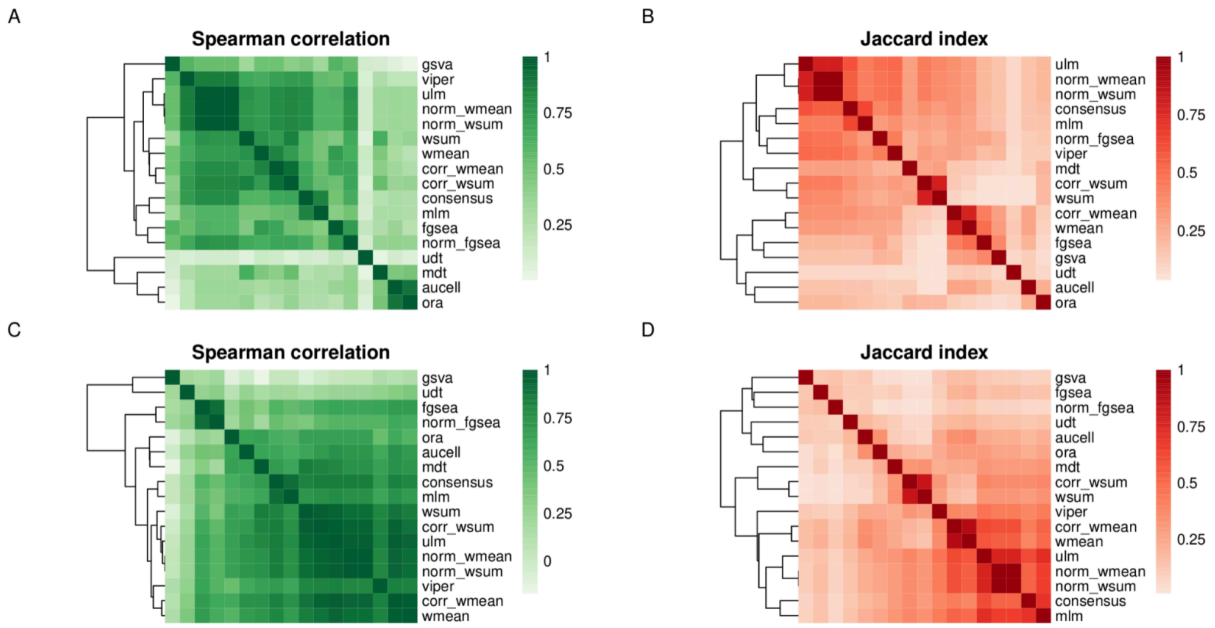

Figure 2.4. Spearman correlations between methods were calculated for the transcriptomics (**a**) and phosphoproteomics (**c**) datasets. The median Jaccard index between methods was assessed for the top 5% of TFs (**b**) or kinases (**d**), ranked by the absolute value of their enrichment scores.

To assess the robustness of inferred scores to noisy or missing information I randomly added or deleted a percentage of edges at each source in the gene set network (25%, 50% and 75%). For every mode (addition or deletion) and percentage, I generated five networks with different seeds, which were then used to infer scores from the perturbation data. Robustness was measured as the Spearman correlation between the obtained scores and those from the two original unmodified networks. Independently of the method, deleting edges had a greater impact than adding them, with median Spearman correlations of 0.84 for addition and 0.77 for deletion (**Figure 2.5.**)(p-value < 2.2e-16; one-sided Wilcoxon rank-sum test). Most methods showed similar correlations, except UDT and MLM, which consistently had a lower median correlation correlation than the rest of methods in both datasets.

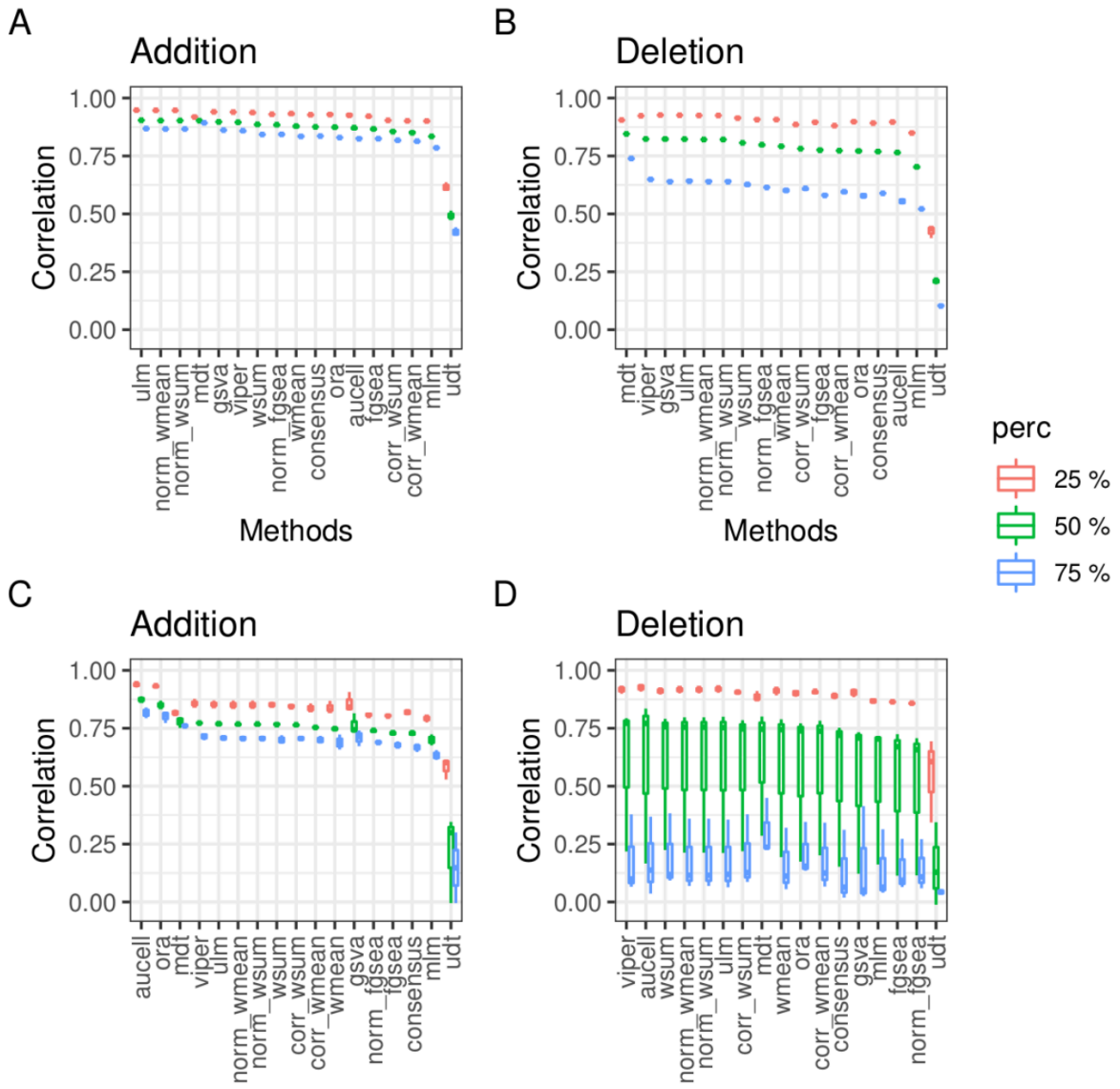

Figure 2.5. Correlations between the original enrichment scores and those obtained after adding or deleting a percentage of edges from the prior knowledge resource for the transcriptomic (**a, b**), and phosphoproteomic (**c, d**) datasets.

After comparing the methods at the score level, I developed a benchmarking pipeline to assess differences in their predictive performance. For a given collection of contrasts, it calculated enrichment scores for all methods using the provided source sets. The scores for each contrast were then adjusted by the sign of their perturbation, with a negative sign for knockouts and a positive sign for overexpressions. Next, the scores from each experiment were combined into a single vector. The evaluation task was to distinguish between perturbed sources (true positives) and unperturbed ones (true negatives). To address class imbalance across networks, a downsampling strategy was applied in the benchmark. In each permutation, an equal number of positive and negative classes were randomly

chosen to compute the area under the Receiver Operating Characteristic (AUROC) and Precision-Recall Curve (AUPRC). This procedure was repeated 1,000 times per network to generate performance metric distributions.

Although methods showed enrichment score similarities (**Figure 2.4.**), their predictive performance varied across both sets of perturbation experiments (**Figure 2.6.**). Surprisingly, commonly used methods like GSEA and GSVA performed poorly, as they do not account for feature set weights and therefore the direction of enrichment can be wrong. In contrast, ULM, MLM, ORA and MDT along with the consensus approach, achieved significantly higher median AUCs compared to other methods in both datasets (p-value < 2.2e-16; one-sided Wilcoxon rank-sum test). However, the interpretability of MDT and ORA is limited, as both return only positive enrichment scores, and ORA does not account for feature weights. Given that MLM may be impacted by collinearity and the consensus score depends on the methods used in its construction, I recommend using the ULM method for its robustness, predictive performance, simple formulation, and scalability.

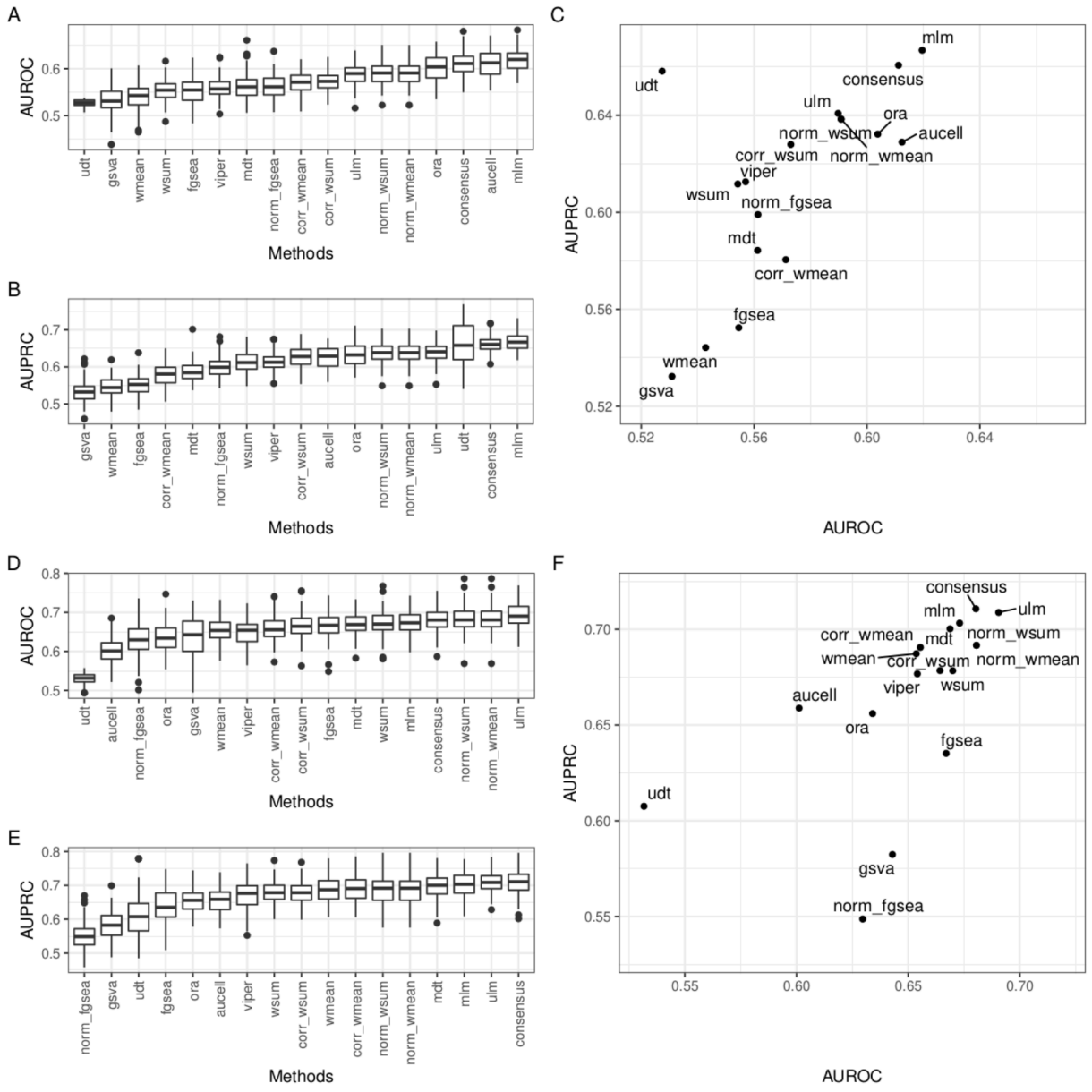

Figure 2.6. Distributions of AUROCs (**a**), AUPRCs (**b**), and their medians (**c**) for each method in the transcriptomics dataset. Similarly, distributions of AUROCs (**d**), AUPRCs (**e**), and their medians (**f**) for each method in the phosphoproteomics dataset.

2.2.2. Tool capabilities

In this section, I outline the main features that decoupler offers as a tool. While developing it, I adhered to best open-source programming practices such as continuous testing, comprehensive documentation, and distribution through standard tools. As a result, decoupler is part of the scverse ecosystem, a multi-institutional open-source project aimed at addressing storage and analysis needs for single-cell profiling omics data[124]. I have written extensive documentation

(https://decoupler-py.readthedocs.io/), with a detailed API to describe its functions, several vignettes showcasing how to use it and detailed release notes to inform about the changes across the different versions. Since November 2021, I have been developing and updating decoupler, resulting in over thirteen versions, more than 612 commits from eleven contributors, and a total of over 100,000 downloads from PyPI alone. Below, I describe some key functionalities of decoupler, based on the aforementioned vignettes I have written. For the full collection of vignettes, their description and interpretation I refer the reader to decoupler's documentation.

2.2.2.1. Basic usage

Decoupler requires Python version 3.6 or higher to run. There are two installation methods available. The first method involves installing it directly from PyPI, which provides a lightweight version with only the essential dependencies:

```
pip install decoupler
```

However, some functionalities require additional dependencies. To simplify the installation process, I also offer decoupler through the conda package manager, which can be installed using its faster implementation, mamba:

```
mamba create -n=decoupler conda-forge::decoupler-py
```

To use decoupler it needs to be imported:

```
import decoupler as dc
```

When decoupler is imported for the first time in a freshly installed environment, it may take a few seconds to load. This delay is due to Numba compiling decoupler's code. After the initial import, the compiled code is cached, making subsequent imports in the same environment to have normal loading times.

To showcase how to use the tool, I created a function to load a toy dataset consisting of a matrix of gene expression across several samples and a GRN:

```
mat, net = dc.get_toy_data()
```

These samples consist of two small populations, each one formed by 12 samples, with two gene expression patterns (**Figure 2.7a**), where the first four genes are highly expressed in the first population and the next four in the second population. Based on the example GRN (**Figure 2.7b**), TF1 is expected to be active in the first

population, as its positive target genes are highly expressed, and inactive in the second population due to their lack of expression. TF2 should be active in the first population because its negative targets are expressed at low levels, but inactive in the second population where they are overexpressed. Similarly, TF3 is expected to be active in the second population, TF4 slightly active in the first, and TF5 inactive in both populations.

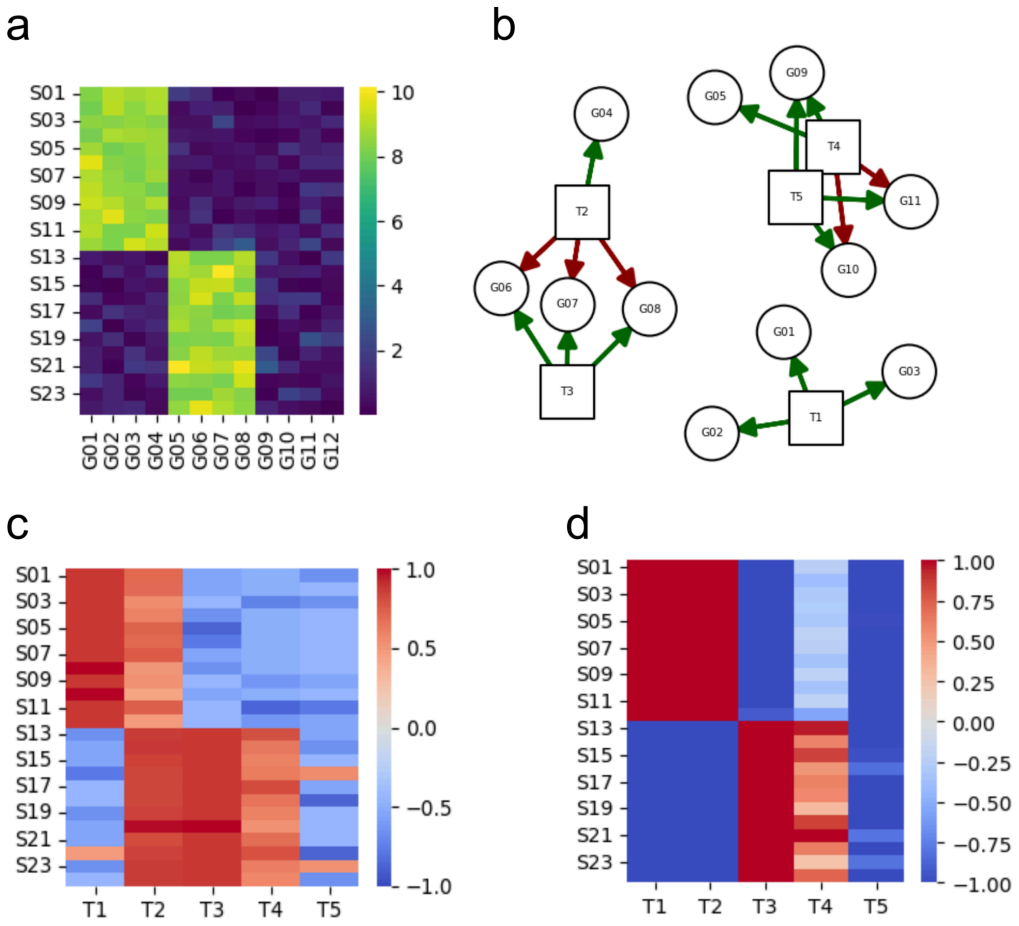

Figure 2.7. a. Matrix of gene expression of two populations of samples that express genes differently. **b**. GRN with five TFs and eleven target genes. Color indicates mode of regulation: green is positive and red is negative. **c**. Matrix of enrichment scores by GSEA. **d**. Matrix of enrichment scores by ULM.

As mentioned before, decoupler contains multiple enrichment methods, which can be queried by running the following function:

```
dc.show_methods()
```

Since decoupler is a unified framework, all of its available methods follow the same format in their arguments:

- mat: input matrix of molecular features. It can be either a pandas dataframe, an AnnData object, or a list containing a matrix with its column and row names.
- net: collection of weighted feature sets relating molecular features to sources. Non weighted sets are assumed to have a weight of one.
- source, target and weight: column names denoting the source, target and weight columns in the net dataframe.
- min_n: Minimum of target features per source (five by default). This filtering prevents noisy scores from feature sets with few target features present in mat.
- verbose: Whether to show progress of the calculation and display extra information.

This allows users to run any of the methods with one simple line of code. However, some methods may have specific arguments, and might provide more than one score.

As an example, here is how to run the most popular enrichment method, GSEA:

```
acts, norm_acts, pvals = dc.run_gsea(mat, net, min_n=0, times=100)
```

In this case, GSEA returns its running sum statistic, its normalized enrichment score and its p-values. While it correctly identifies TF1 as active in the first population, it incorrectly predicts that TF2 is also active (**Figure 2.7c**). This error arises from the method's inability to account for prior weights in its score calculation. On the other hand, when ULM is used it correctly predicts the direction of activity for all TFs (**Figure 2.7d**).

Finally, several methods can be run sequentially by using the decouple function, which returns a dictionary containing all the calculated scores and p-values:

```
results = dc.decouple(mat, net, min_n=0, methods="all")
```

The methods parameter allows users to specify which methods to run. Once executed, it also provides the consensus score across the selected methods.

2.2.2.2. Bulk analysis

Decoupler can perform classic enrichment analysis, such as for bulk transcriptomics data. In this context, transcriptomics is analyzed for samples from two conditions, followed by differential gene expression analysis and enrichment based on the

resulting contrast statistics. To demonstrate this with decoupler, I use a bulk dataset of hepatic stellate cells (GSE151251), where three samples were treated with transforming growth factor (TGF-□) and three served as controls.

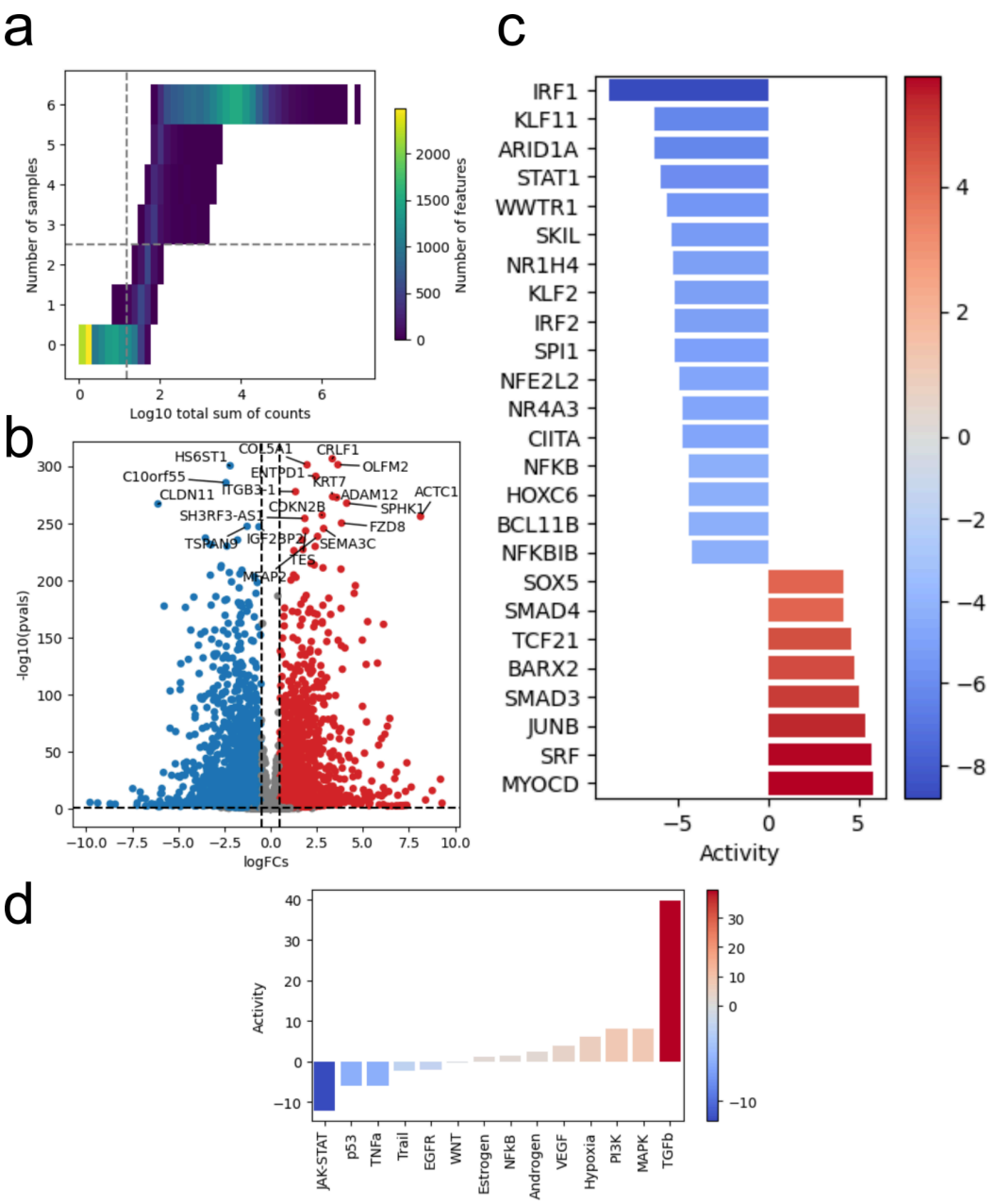

Figure 2.8. a. Bivariate histogram displaying the total sum of counts and the number of samples with expression for each gene. Following the gene filtering approach from edgeR, genes with enough total sum of counts and present in enough samples will be selected. **b**. Volcano plot displaying the results of pyDESeq2. **c**. Top activated or deactivated TFs from CollecTRI. **d**. Top activated or deactivated pathways from PROGENy.

After loading the data, feature selection is applied to eliminate noisy genes. I use the edgeR strategy[125], which selects genes with sufficient total counts and those expressed in a minimum number of samples. Contrary to the original implementation, a plotting function (**Figure 2.8a**) is provided to help determine appropriate thresholds. For differential expression analysis, I use pyDESeq2[126], an optimized version of the classic framework, along with custom plotting functions to explore the results (**Figure 2.8b**)[114]. The Wald statistics from pyDESeq2 are then used to infer TF activity with the collectri[42] GRN (**Figure 2.8c**) and pathway activity using the progeny[127] database (**Figure 2.8d**). As expected, fibrosis-related TFs, such as MYOCD[128], SRF[129] and JUNB[130], are active in this contrast, with the TGF-□ pathway being the most active.

2.2.2.3. Single-cell analysis

In single-cell data, enrichment scores can be inferred at the cell level or at the contrast level. To showcase cell-level enrichment, I use the classic 10X Genomics dataset, which includes 3,000 peripheral blood mononuclear cells from a healthy donor. After standard single-cell preprocessing, cell clusters are identified, followed by cell type annotation. Enrichment can assist in this process by using cell type marker gene collections. In the vignettes I illustrate this with decoupler using the panglaodb database[131], which contains marker genes for various cell types. By determining the top enriched cell type for each cluster, this information can guide cluster annotation (**Figure 2.9a**). However, while this automatic annotation provides a useful draft, review from domain experts is always advised to ensure accurate annotation for the tissue of interest.

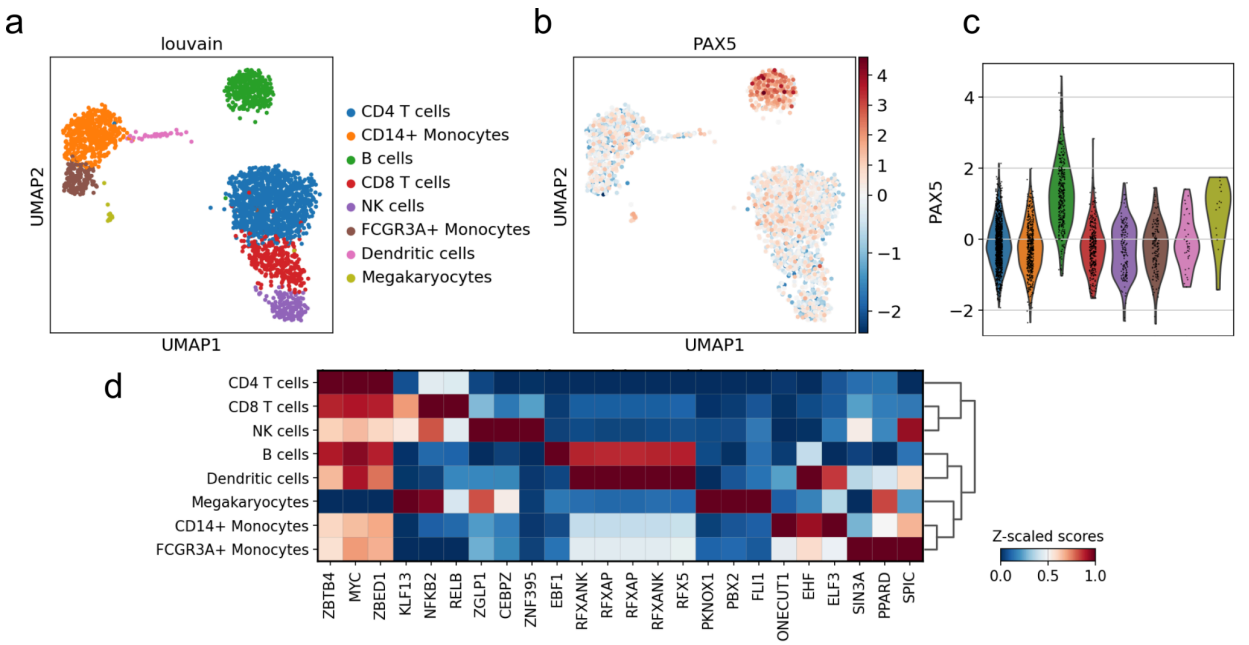

Figure 2.9. a. UMAP embedding of single-cells, color indicates cell type. **b**. Predicted activity of PAX5 per cell. **c**. Violin plots displaying the predicted PAX5 scores, color indicates cell types as in **a**. **d**. Heatmap of mean enrichment scores per cell type.

Once the clusters are annotated, TF and pathway inference can be performed. In this example I showcase how PAX5, a crucial TF for B cell identity and function[132], is active in B cells (**Figure 2.9b**). Another way to visualize this is to plot their distributions as violin plots (**Figure 2.9c**), or summarized in a heatmap (**Figure 2.9d**).

For contrast-level enrichment, I developed a function that simplifies generating pseudobulk profiles from single-cell data. This approach requires multiple samples or patients across at least two conditions. Pseudobulking, also named in-silico bulking, involves summarizing the counts of all cells from a specific cell type and sample into a single profile[133]. Typically, the counts are summed, but other aggregation methods like the mean or median can also be applied. Once the pseudobulk profiles are prepared, the bulk workflow described earlier can be followed for each cell type, including feature selection, differential testing, and enrichment based on contrast statistics.

2.2.2.4. Spatial analysis

Similar to cell-level enrichment analysis, it can also be applied to spatial data. To demonstrate this, I use a Visium slide from 10X Genomics of a human lymph node, which captures the expression of genes across spots, each containing a small population of cells. By applying standard preprocessing steps similar to those used

in single-cell analysis, spots can be clustered and annotated into tissue regions (**Figure 2.10a**). Despite capturing thousands of genes, expression in Visium slides can be noisy and sparse (**Figure 2.10b**).

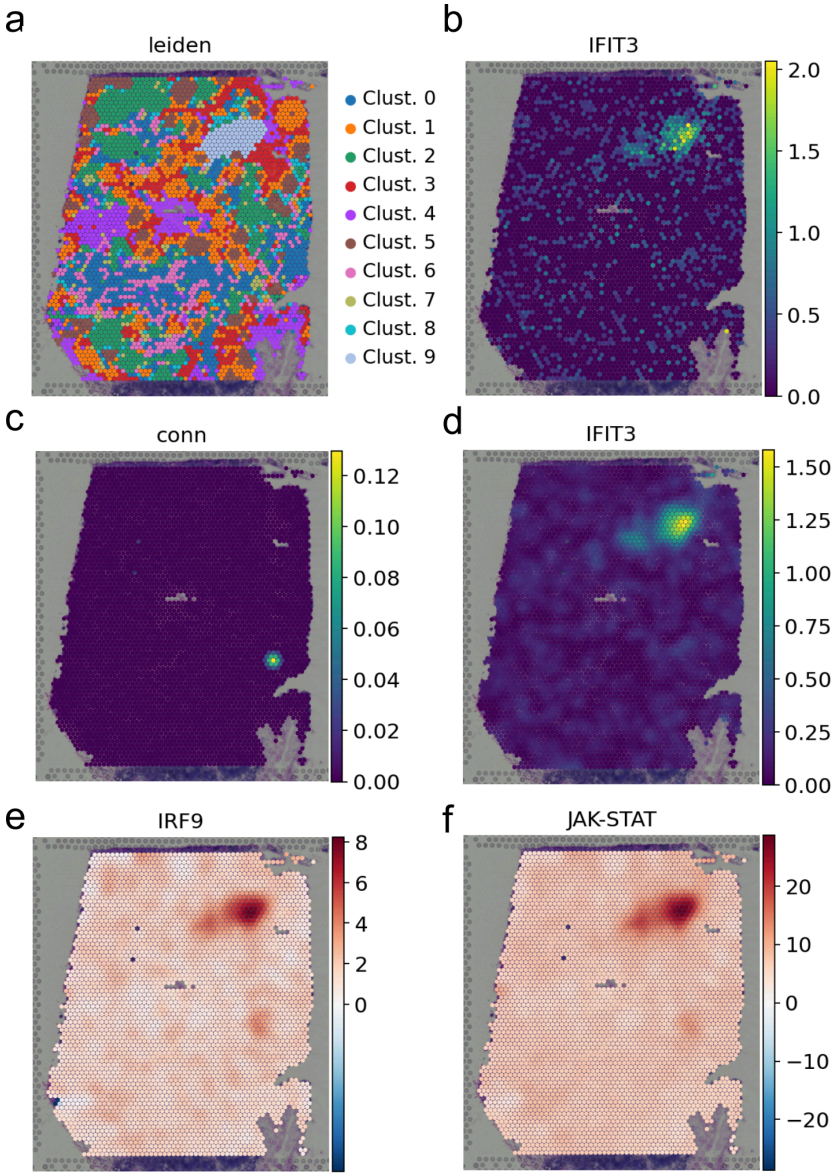

Figure 2.10. a. Visium slide of a human lymph node. **b**. Expression of the gene *IFIT3* across the slide. **c**. Spatial connectivity bandwidth. **d**. Transformed expression of the gene *IFIT3* across after weighting for the spatial connectivity. **e**. Spatial enrichment score for the TF IRF9. **f**. Spatial enrichment score for the JAK-STAT pathway.

To address the noise and incorporate the spatial component, I first transform the input spatial gene expression data using spatial connectivity weights from LIANA+[134] (**Figure 2.10c**). Spots near a given spot contribute more to the final gene value, while those further away contribute less. This spatial weighting smooths the

gene expression data, reducing noise and integrating the spatial aspect into the molecular readouts (**Figure 2.10d**). With the transformed gene expression data, spatially-informed TF and pathway activities can be inferred (**Figure 2.10e,f**).

2.2.2.5. Benchmark pipeline

I have also developed an extended and faster benchmark pipeline, similar to the one in the original decoupler manuscript. This pipeline can be used to evaluate enrichment methods, assuming the provided gene set or GRN is correct, or to assess a collection of GRNs if the enrichment method is considered reliable. For this benchmark, I processed KnockTF2[135], a recent and comprehensive database that compiles gene contrasts from TF knockout perturbation experiments. The pipeline supports the simultaneous testing of multiple methods and GRNs, generating metrics such as AUROC, AUPRC, mean rank, and mean quantiles (**Figure 2.11.**).

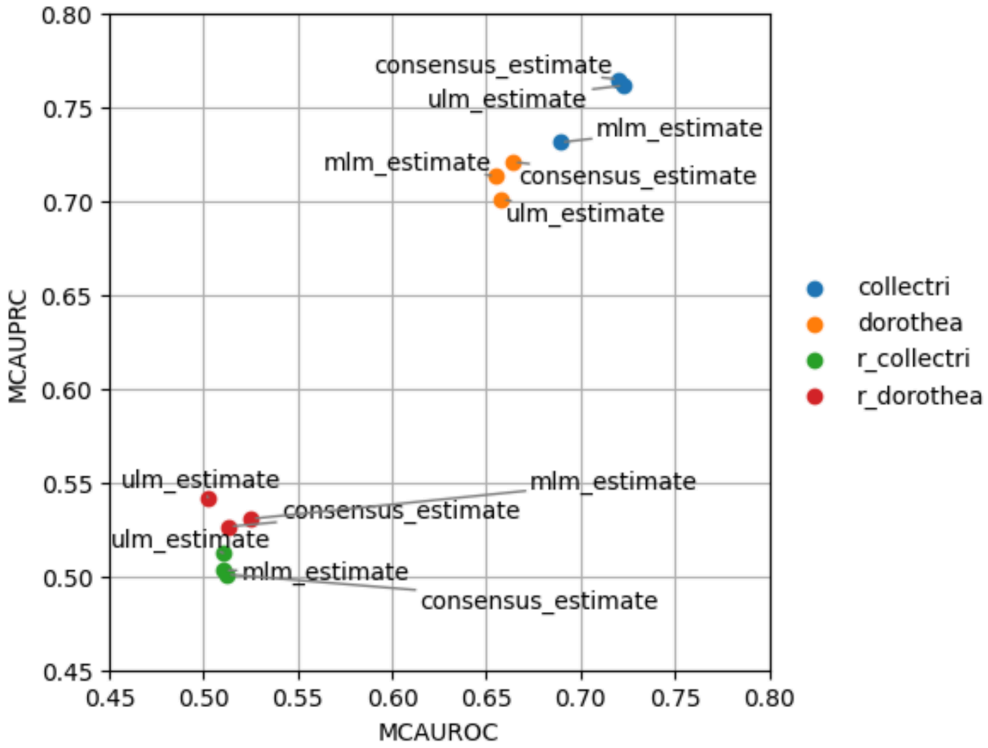

Figure 2.11. Example of evaluation results for two literature-derived GRNs (collectri and dorothea) and their random counterparts across the ULM, MLM and consensus enrichment methods.

2.2.2.6. Conversion to other organisms

Since most prior knowledge is derived from human data, it cannot be directly applied to the study of model organisms like flies, mice, and others, limiting its utility. To address this, homology conversion can be used to map gene names from

humans to other organisms. Decoupler handles this through OmniPath[136], a metadatabase that integrates multiple resources. This allows gene transformations to be performed with a simple one-liner command, for example:

```
mouse_grn = dc.translate_net(grn, target_organism = 'mouse')
```

2.3. Applications of regulation activity with decoupler

During the development of decoupler, I had the opportunity to collaborate with several laboratories where it proved useful. In this section, I discuss two of these projects and how decoupler contributed: molecular consequences of Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) liver tropism[137], and spatial cell type mapping of multiple sclerosis lesions[138].

2.3.1. Molecular consequences of SARS-CoV-2 liver tropism

Since the onset of the coronavirus disease 2019 (COVID-19) global pandemic, extrapulmonary manifestations of the disease have drawn attention due to their impact on clinical outcomes and potential long-term effects. SARS-CoV-2 shows tropism for various organs, such as the heart and kidneys, but whether it directly affects the liver remains uncertain.

In this study, Wanner et al. explored this hypothesis by analyzing three cohorts of COVID-19 patients hospitalized due to the disease. Although liver disease was rare among these patients (1.4%), a significant percentage exhibited elevated levels of aspartate aminotransferase and alanine aminotransferase, both markers of liver injury[139]. Notably, higher levels of these enzymes were linked to decreased survival rates, highlighting hepatic injury as a key clinical feature in hospitalized COVID-19 patients. To investigate if the virus could be causing this, the presence of SARS-CoV-2 in the liver was confirmed through RT-qPCR targeting the E gene in an independent cohort.

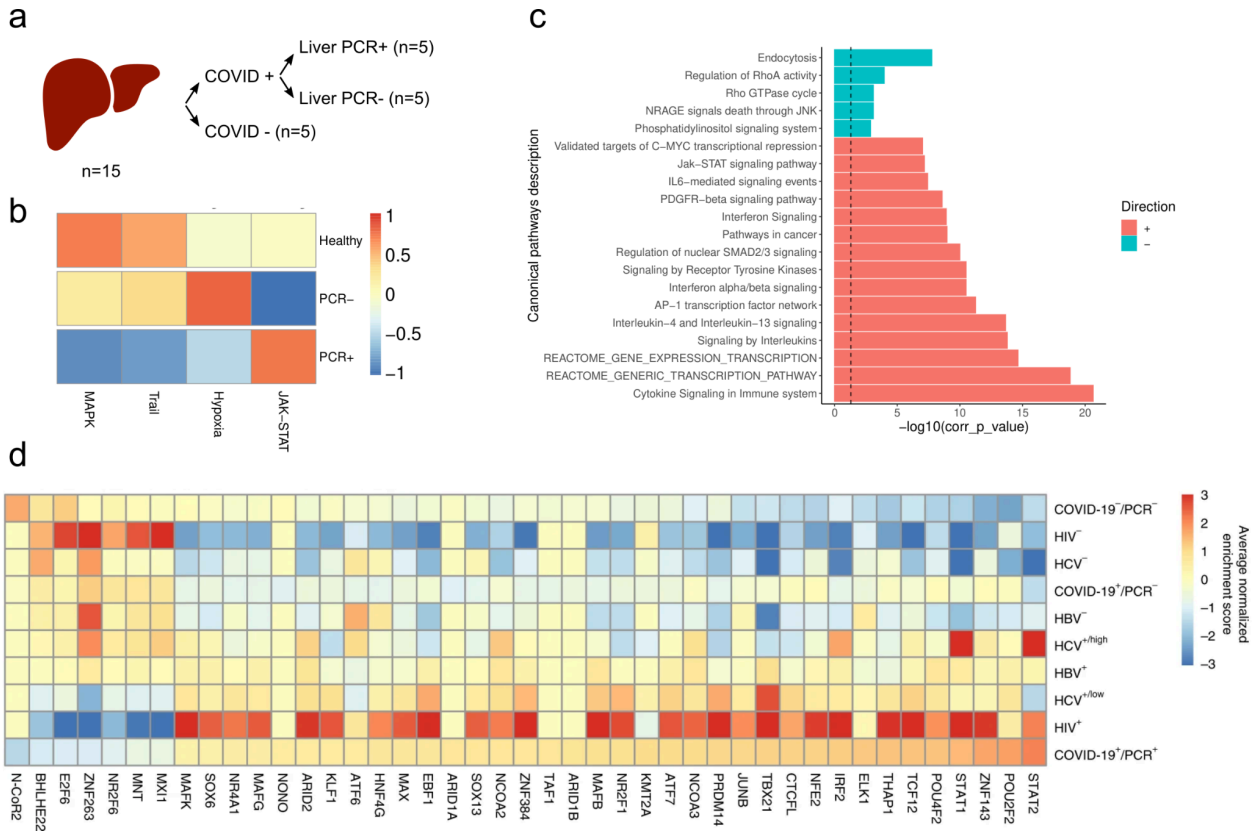

Figure 2.12. **a**. Experimental design for the RNA-seq profiling of patient liver samples. **b**. Mean pathway enrichment score across groups from progeny. **c**. Differential pathway enrichment scores for the PCR+,PCR- contrast for reactome gene sets. **d**. Mean TF enrichment scores across groups. HCV: hepatitis C virus. HIV: human immunodeficiency virus. HBV: hepatitis B virus.

To investigate the molecular changes associated with SARS-CoV-2 liver tropism, liver tissues from a subset of COVID-19 autopsies (n=10) were selected for transcriptomic analysis. Three comparison groups were defined a priori (**Figure 2.12a**): (1) patients with COVID-19 and SARS-CoV-2-positive livers (n=5); (2) patients with COVID-19 but SARS-CoV-2-negative livers (n=5) to account for the systemic effects of COVID-19; and (3) non-COVID-19-related deaths (n=5, i.e., controls) to account for unrelated effects from autopsy-related collection and storage.

Transcriptomic profiling confirmed the presence of key SARS-CoV-2 entry receptors and facilitators in the liver autopsy samples, including *ACE2*, *TMPRSS2*, *CTSL*, and *RAB7A*[140]. Notably, a comparison with publicly available dataset[141] showed that the liver had similar expression levels for these genes as the lung, suggesting that liver tissue is susceptible to SARS-CoV-2 infection. Additionally, transcriptional changes linked to SARS-CoV-2 liver tropism revealed a marked upregulation of type I and II interferon genes.

To better characterize the transcriptional changes, I used decoupler to infer pathway enrichment scores using the progeny[127] and reactome databases[142]. As expected, I observed an increase in immune related pathways such as JAK-STAT and cytokine signaling in the immune system (**Figure 2.12b,c**). Interestingly, modulation of JAK-STAT signaling has been proposed to intervene in SARS-CoV-2 viral entry and replication[143]. Next, I inferred TF enrichment scores with the dorothea[58] GRN (**Figure 2.12d**). Interestingly, STAT2, a key mediator of type I IFN signaling, was predicted to be activated in PCR+ samples, while NCOR2, which interacts with histone deacetylases to suppress basal transcription, was predicted to be inhibited.

Given that both clinical and molecular data indicated that SARS-CoV-2-mediated liver injury resembled that of previously characterized hepatotropic viruses, I compared publicly available bulk RNA-seq datasets from liver tissue samples of patients with hepatitis C virus[144], human immunodeficiency virus[145], and hepatitis B virus[146]. After inferring TF scores, I found similar patterns in livers from patients with severe COVID-19 and those with liver infections (**Figure 2.12d**). Therefore, this finding highlights a shared molecular signature between SARS-CoV-2 and several viruses known to cause liver injury.

Despite all these layers of evidence supporting SARS-CoV-2 hepatic tropism, the precise mechanism of infection remains unclear. Nevertheless, these findings clearly indicate that the human liver is susceptible to SARS-CoV-2 infection, having profound implications for patient management and treatment strategies, particularly for those with pre-existing liver conditions or complications related to COVID-19.

2.3.2. Spatial cell type mapping of multiple sclerosis lesions

Multiple sclerosis (MS) is a chronic inflammatory disease of the central nervous system characterized by the formation of multiple lesions[147]. These lesions follow specific temporal and spatial patterns of inflammation and tissue damage. Acute MS lesions begin with active myelin breakdown, after which they may undergo remyelination or progress to a chronic active stage, featuring an inflammatory rim and a clearly defined demyelinated core. Over time, inflamed lesions may become inactive, losing the rim inflammation and forming a dense glial scar. Although the progression of these lesions is known, their exact cellular composition, molecular dynamics, and signaling events within these niches remain poorly understood.

To better understand MS lesions, Lerma et al selected snap-frozen subcortical blocks from patient autopsies based on histological assessment, together with tissue

donors as controls. Samples that met sequencing quality standards were profiled for snRNA-seq and/or spatial RNA-seq, resulting in a study cohort of 7 controls, 8 chronic active lesions and 4 chronic inactive lesions (n=19), from which 5 controls, 6 chronic active lesions and 4 chronic inactive lesions were paired (n=15) (**Figure 2.13a**).

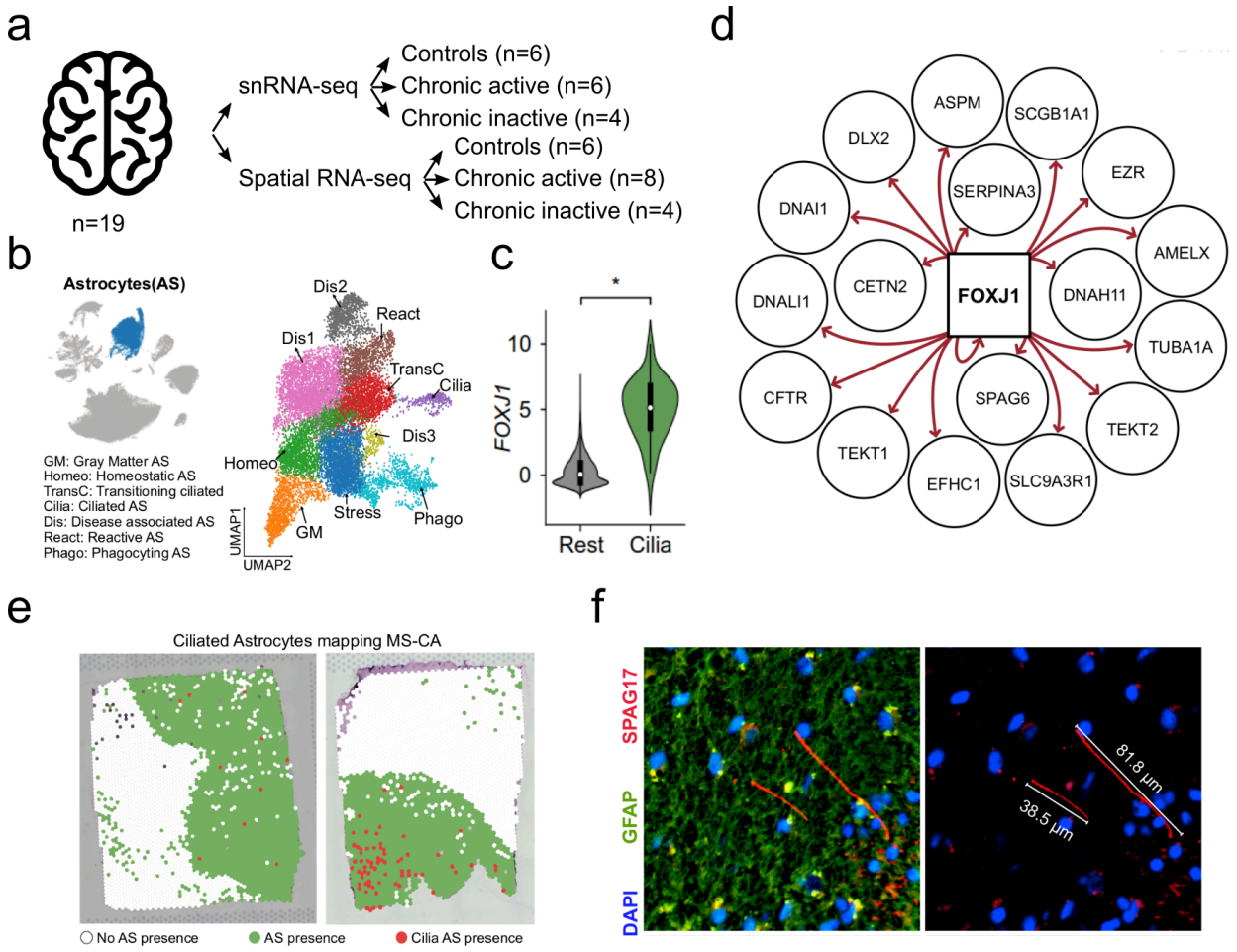

Figure 2.13. **a**. Experimental design for the joint snRNA-seq and spatial transcriptomics profiling of patient subcortical samples. **b**. UMAP of annotated astrocytes subpopulations. **c**. Predicted activity of the TF FOXJ1. **d**. Target genes of FOXJ1. **e**. Predicted spatial locations of ciliated astrocytes in two chronic active MS lesions. **f**. Immunofluorescence example showcasing the length of cilia of astrocytes in µm in the lesion core of chronic active MS lesions.

Together with Lerma's biological expertise, I have processed this atlas and annotated multiple cell-type subpopulations (**Figure 2.13b**). During this process I used decoupler to identify a novel subpopulation of ciliated astrocytes. After TF score inference, I observed that a small population of astrocytes showed a significantly higher enrichment score for the TF FOXJ1 compared to the rest (**Figure 2.13c**). Further exploration of FOXJ1 target genes from the collectri[42] GRN

revealed its key role in regulating cilia-associated genes, such as *DNAH11* and *CETN2*[148], leading us to name this group ciliated astrocytes (**Figure 2.13d**).

The next step was to identify their spatial location in MS lesions. To achieve this, I extracted marker genes from this subpopulation using differential expression analysis and performed enrichment analysis at the spot level across all slides with decoupler. Interestingly, spots predicted to contain ciliated astrocytes were located in the lesion core of chronic active lesions (**Figure 2.13e**), suggesting a potential role in lesion progression. Presence of this particular subpopulation was then validated by Lerma using immunofluorescence targeting astrocyte and cilia marker proteins (**Figure 2.13f**). Microscopy images revealed that these astrocytes produce abnormally large cilia structures, measuring up to 80 μm, compared to normal cilia length of around 2 μm.

Although the in vivo function of this subpopulation remains unclear, we hypothesize that their location and absence of a proinflammatory phenotype suggest a role in tissue remodeling and chronic scar formation. Alternatively, they may represent an aberrant response to stress, as astrocytes with abnormal motile cilia gene expression have been observed in the mouse cortex following induced mitochondrial dysfunction[149], a process known to occur in chronic MS lesions. Future research will aim to clarify the exact role this subpopulation plays in MS.

Chapter 3: Benchmark of gene regulation models

In the second chapter I wrote about how enrichment analysis could be used to infer transcription factor (TF) regulatory activities to discover novel biological insights, where I also showed that simple linear models outperform classical enrichment strategies. However, enrichment analysis also requires a gene regulatory network (GRN) assumed to be correct. In this chapter, I begin by examining and quantifying the differences in GRN inference across methods. This analysis is followed by the design of a benchmarking pipeline and the evaluation of these methods across various tasks to determine their effectiveness.

3.1. Comparison of multimodal GRN methods

Introduced in the first chapter, recent methods use multimodal RNA-seq and ATAC-seq for GRN inference. By including CRE information, they can focus on TF-Gene interactions that are more likely to be relevant, potentially reducing the number of false associations. While the tools have been validated in their respective publications, a comprehensive comparison is still lacking, and their similarity to classical methods has not been explored. In this section, I compare some of these new methodologies, celloracle[85], dictys[91], figr[84], granie[86], and pando[87], not only against each other but also against baseline approaches (**Figure 3.1.**). These include collectri[42] and dorothea[123], two literature-derived networks, scenic[80], the current state-of-the-art GRN method for transcriptomics, and a random network with a realistic GRN topology. For this comparison, I used a published, publicly available multiome dataset from a patient's pituitary gland that includes multiple cell types[150].

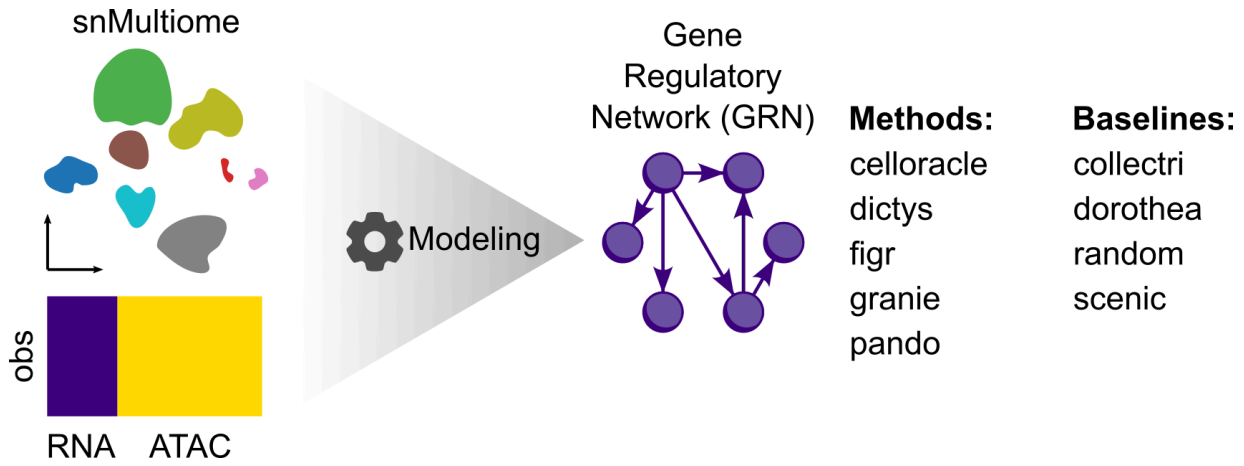

Figure 3.1. From a paired snRNA-seq and snATAC-seq dataset I have generated GRNs for several multimodal and baseline GRN inference methods.

The first test I conducted was to assess whether these methods were robust in reproducing consistent results. Stable methods were assumed to not drop their overlap coefficient drastically when gradually downsampling for either number of features or cells. To quantify this, I initially inferred a GRN for each method using the full pituitary dataset, and then I applied two downsampling strategies to it: one reduced the number of features while keeping the number of cells constant, and the other kept the number of features constant while randomly downsampling cells. For the feature downsampling, a predefined top number of highly variable features, both genes and CREs, were selected to maximize differences across cell types. The downsampling steps were repeated with progressively fewer features and cells, and each step was performed three times to account for randomness. The downsampled GRNs showed a substantial decrease in overlap coefficient when compared to their whole counterpart (**Figure 3.2.**), indicating that preprocessing single-nucleus data can greatly affect inference.

Notably, there was a sharper decline in overlap when fewer cells were used. When halving the number of cells to 8,192, I observed overlap coefficients lower than 0.5 for celloracle, dictys and figr. This is something concerning for single-nucleus technologies as they typically sample only a small fraction of cells, around 10,000 per sample, compared to the millions present in an organ. The two exceptions at this were granie and pando. Although their overlap coefficient also dropped, they achieved higher values compared to the other methods.

Regarding the downsampling of features, all methods were affected similarly, dropping their overlap coefficients below 0.5 after just one downsampling step. This behavior is also alarming, as for example celloracle recommends inferring their GRNs using only a subset of highly variable features.

This downsampling strategy also allowed me to quantify the computational cost of running these methods (**Figure 3.2.**), an important aspect to consider as it can limit their accessibility to a broader community. The majority of methods took less than two hours to run on 32 CPUs, the exceptions being dictys which also required a GPU and figr, which I observed that its computing time grew exponentially. Regarding memory usage, most methods required more than 16 GBs, the exceptions being dictys and granie. Additionally, methods were difficult to use, even when following their vignettes, as they require multiple inputs in various formats, which most of the time are left for the user to manage.

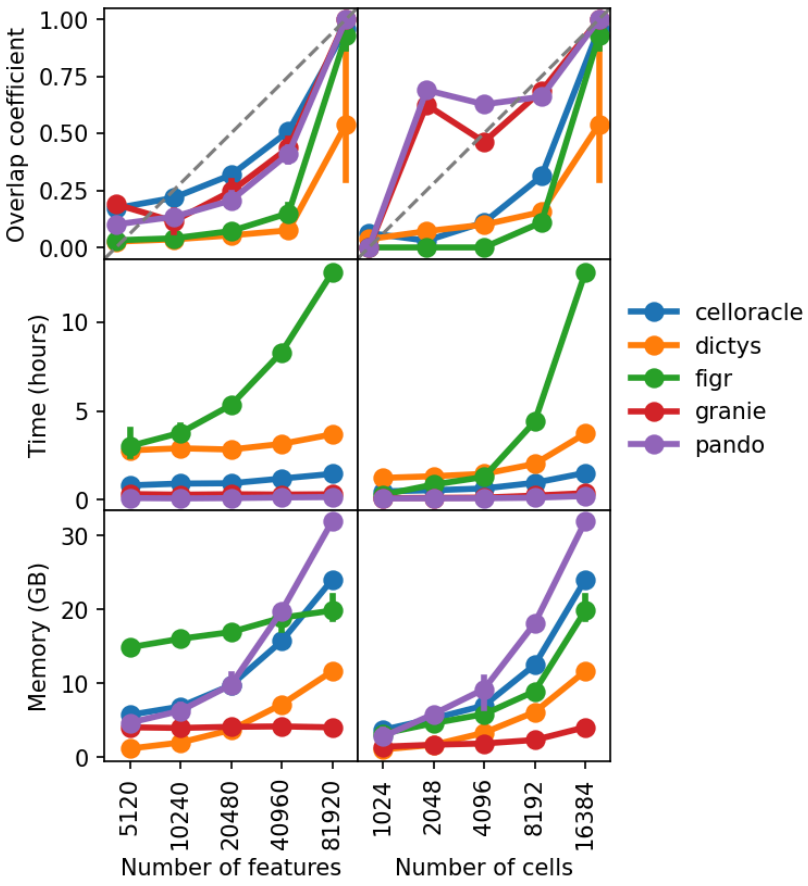

Figure 3.2. Stability and scalability analysis. Each dot represents the mean across three runs using different random seeds, while error bars represent one standard deviation from the mean in each direction.

Next, I compared the overlap of the different solutions proposed by these methods. Although they share a moderate number of TFs and target genes (mean TF overlap coefficient = 0.46; mean target gene overlap coefficient = 0.36), these nodes are predicted to be connected in completely different ways with surprisingly low overlap coefficients at the edge level between methods (mean edge overlap coefficient = 0.02) (**Figure 3.3a**). I also visualized this by selecting only the interactions shared across more than half of the methods and plotting them as subnetworks (**Figure 3.3b**). Even when selecting for shared interactions, it is clear that their topologies are quite different and that they potentially encode for distinct regulatory programs. Nonetheless, together they seem to agree on some interactions, such as the regulation of the target gene AREG. Interestingly, these AREG interactions are absent in literature-derived GRNs (**Figure 3.3c**), highlighting the potential of these methods to discover new biology. Scenic also captured these interactions as well, but predicting many more edges than multimodal GRN methods, as it relies solely on transcriptomics and may introduce more errors.

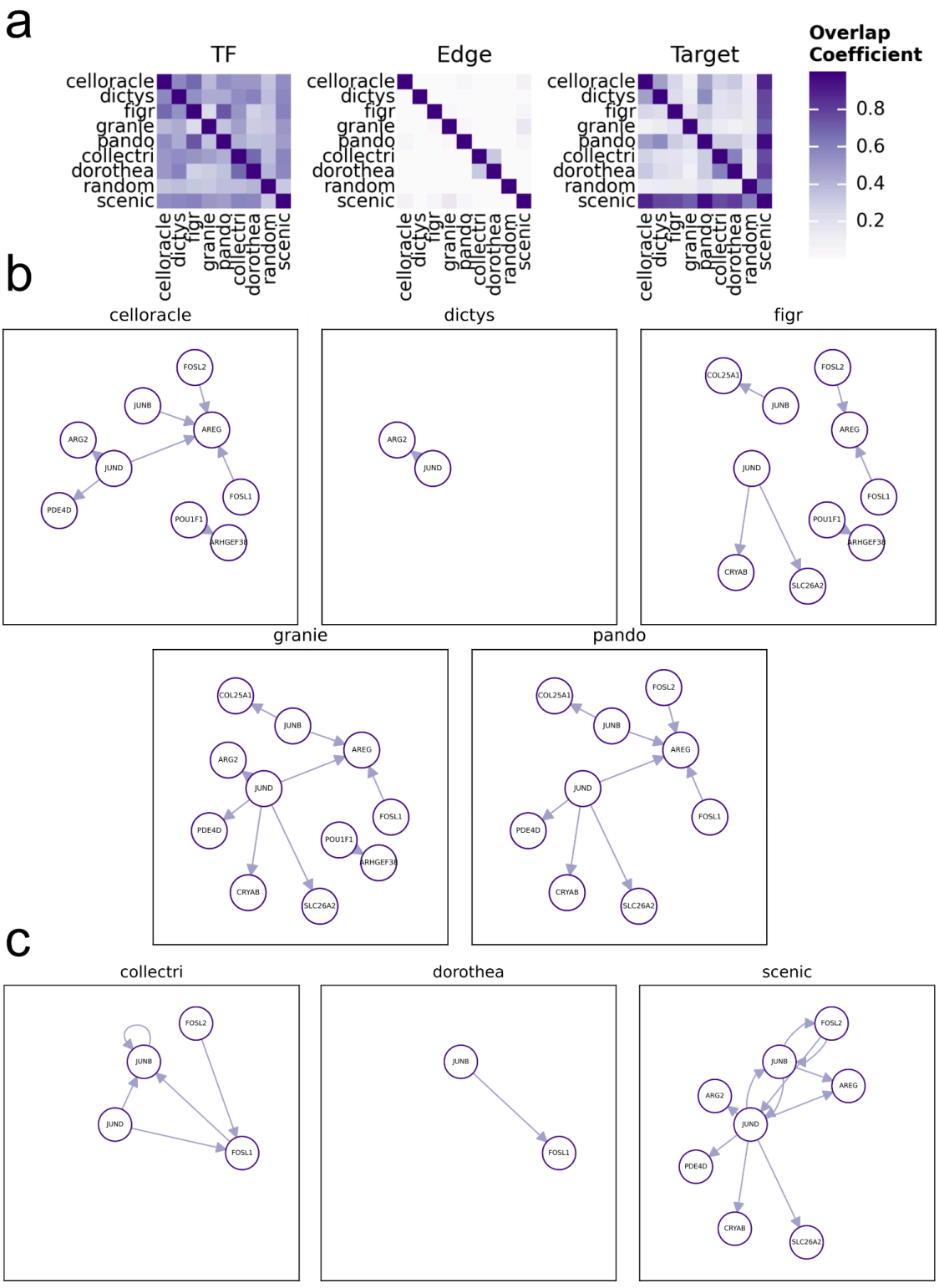

Figure 3.3. **a**, Overlap coefficients between GRNs of the pituitary dataset across multimodal methods and baselines at the TF, edge and target gene levels. **b**, network plot representing the shared TF-Gene interactions present in at least three multimodal methods. **c**, same but for the baselines (random is excluded since it did not capture any of these).

Because multimodal methods also identify which CREs can explain the inferred TF-Gene interactions, I decided to investigate the top TF interactions involving the target gene AREG, as many multiple TFs such as FOSL1, FOSL2 and JUND were predicted to regulate it by most methods. In this dataset, AREG, a member of the epidermal growth factor family[151], was particularly highly expressed in corticotropes. This signaling ligand for the epidermal growth factor receptor has been reported to impact cellular proliferation and differentiation across adjacent cells but its particular role in the pituitary gland remains unclear[152]. Interestingly, despite the methods predicting the same TF-Gene interactions for AREG, they used different CREs to justify how the gene is regulated (**Figure 3.4.**). This indicates that even when the methods agree on interactions, the underlying results differ and might be influenced by chance.

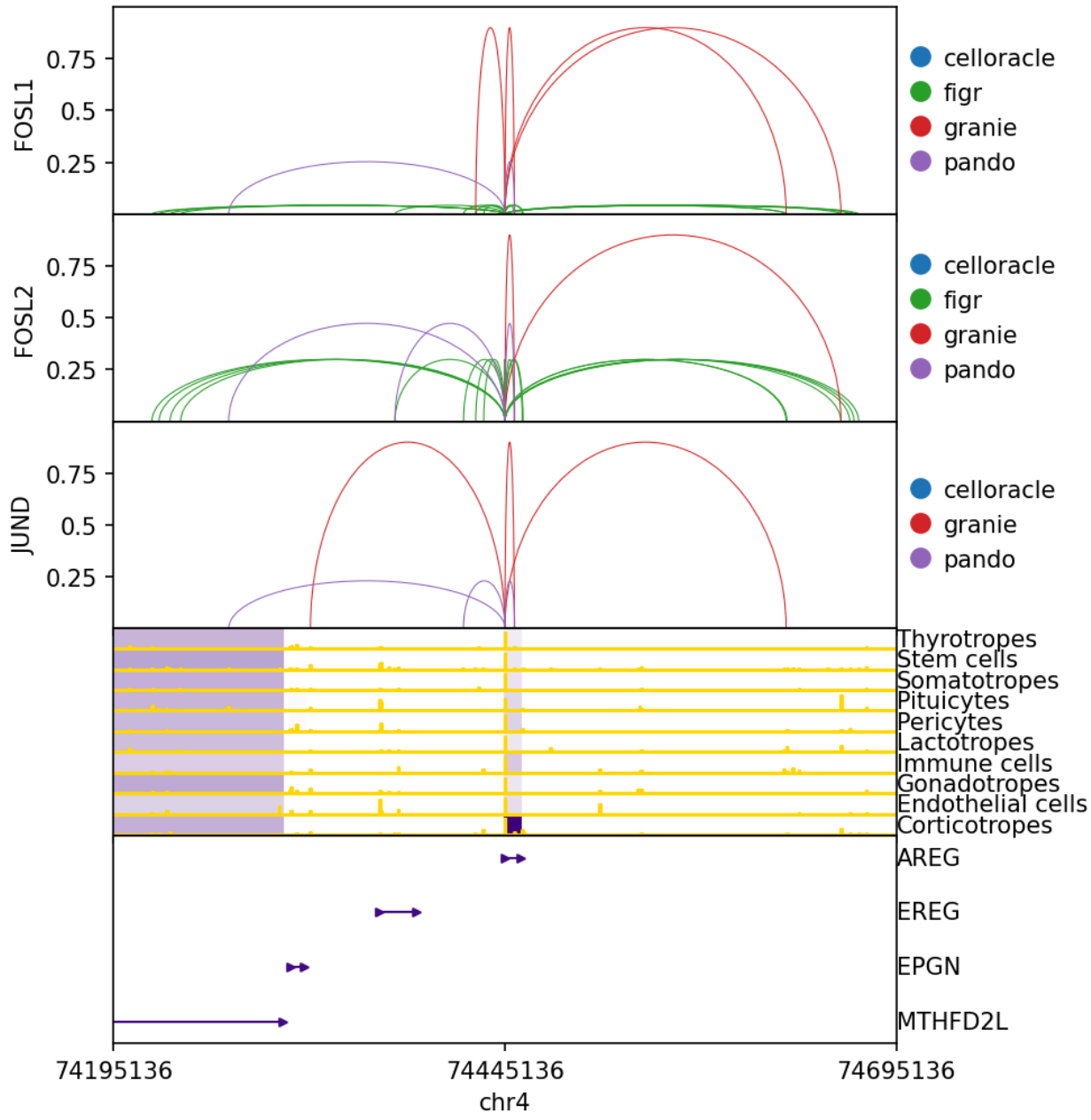

Figure 3.4. Genome plot. At the bottom, gene body annotations are shown, with arrowheads indicating the beginning and end and the strand direction of a gene. In the middle, mean gene expression and chromatin accessibility are displayed per cell type. The purple color represents log-normalized gene expression levels (gene bodies with no expression present are white), and yellow peak heights indicate log-normalized chromatin accessibility of CREs. At the top, inferred CRE-gene interactions are presented as arcs. The color represents the GRN inference method, while the height reflects the quantile of the final TF-gene interaction across the methods that have such interaction.

The previous results I have shown indicate that GRN inference is quite unstable, as they provide different solutions to the same dataset. Another potential source of disagreement in GRN inference is the nature of the multiome data, whether the

data is truly paired, representing readouts from the same cells, or unpaired, with readouts originating from different cells within the same tissue sample. As discussed in the first chapter, using unpaired data requires an integration step to computationally match cells from both omics. The impact of this integration on GRN inference is still not well understood but highly relevant, as currently most available datasets are unpaired due to their lower monetary cost compared to paired data[31]. The pituitary dataset I analyzed presents a unique opportunity, as it includes both paired and unpaired multiome data from the same patient's pituitary gland (**Figure 3.5a**). Initial comparisons at the population level showed cell abundance and molecular similarities between the two datasets (**Figure 3.5b,c**). However, this similarity does not confirm the accuracy of cell-level mapping during integration.

To address this, I created a synthetic paired dataset by integrating the paired multiome data as it were unpaired, allowing me to test the integration's efficacy since I knew the true cell identities (**Figure 3.5a**). For each predicted barcode of one modality to another, I computed its k-nearest neighbor and recorded which neighbor was its true barcode in that modality. I observed that pairing predictions were generally good, as the mean KNN position of the true barcode was 203.48 (**Figure 3.5d)**. When looking at the cell type of the predicted matching barcode, most cells were assigned a barcode of the same cell type (**Figure 3.5e**). However, the detailed molecular profiles, both at gene expression and chromatin accessibility level, differed significantly, as indicated by low Spearman correlations (mean correlation in chromatin accessibility = 0.03; mean correlation in gene expression = 0.14) (**Figure 3.5f**). This suggested that integrating unpaired data should be approached with caution, as it could introduce unwanted variability in the resulting GRN. To illustrate this, I compared GRNs derived from paired and unpaired pituitary data and, as anticipated, found notable differences between them (**Figure 3.5g**).

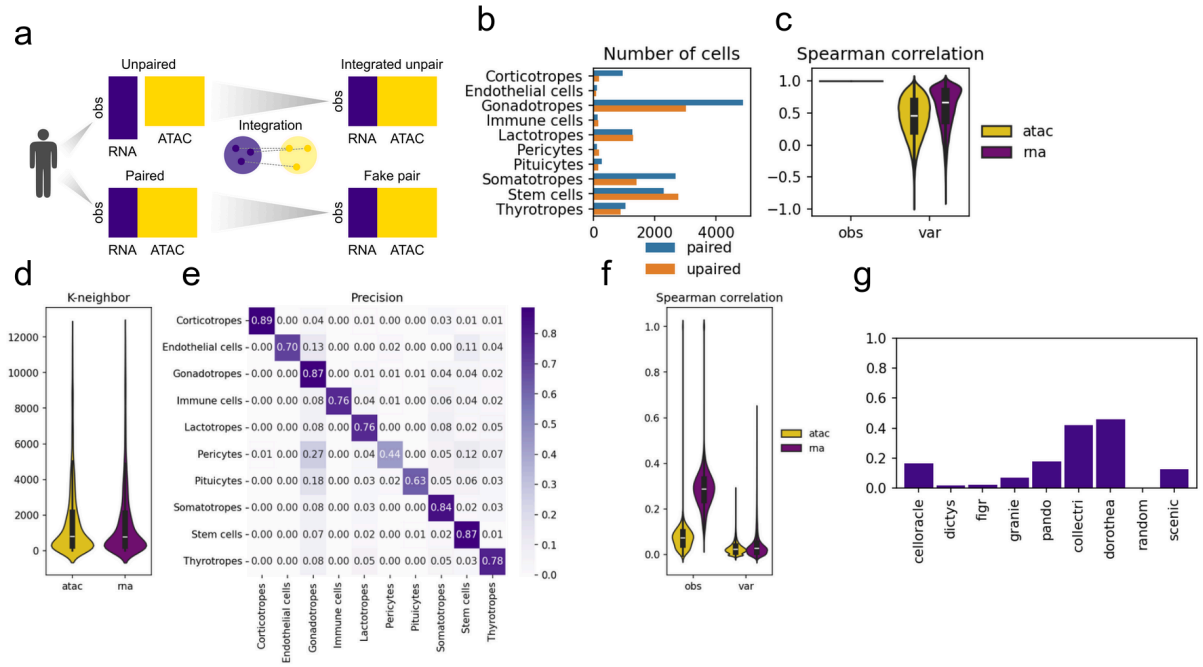

Figure 3.5. **a**, Paired and unpaired multiomics profiling were available for the pituitary dataset. To be able to use the unpaired data for GRN inference, I integrated cells from the two omics using figr's optimal transport approach. Unpaired data might contain a different number of cells for each modality but figr performs a 1-to-1 matching of cells. To evaluate how good this integration was, I also generated a synthetic paired dataset by integrating the paired dataset as if it were unpaired. **b**, Number of cells per cell type for both paired and unpaired datasets. **c**, Spearman correlations between the paired and unpaired pseudobulk profiles at the observation (one cell type at a time, comparing features) and feature (one gene or CRE at a time, comparing cell types) levels in both omics (obs: observations, var: features). **d**, Position of the real matching barcode for an anchor barcode in the KNN graph from the synthetic paired dataset. **e**, Precision of predicted barcodes from the synthetic dataset at the cell type level. **f**, Spearman correlations between the real and matched fake barcodes at different levels and omics (obs: observations, var: features). **g**, Overlap coefficient between GRNs inferred from the paired and the unpaired dataset.

Up until now, I have observed that these multimodal methods are quite unstable, necessitating a deeper understanding of their individual steps involved, which I explain in detail in chapter one: (1) preprocessing, (2) assignment of CREs to target genes, (3) TF binding prediction on CREs and (4) modeling. It could be the case that differences in results might stem from the final modeling step, or they could be due to the use of different motif databases, even if the modeling strategies themselves are similar. To exactly pinpoint the sources of this variability, I broke down the steps of the original methods into modular components, which I implemented in a reproducible Snakemake pipeline called GRETA (Gene Regulatory nETwork Analysis) (https://github.com/saezlab/greta_benchmark/). This modularity enabled

me to run any combination of methods, resulting in 4^N possible combinations, where N represents the number of GRN inference methods considered.

The first thing I tested was to fix the original pipeline for a given GRN method and observe which step had the most significant impact when replaced by the same step but from another method. I found that changing any step of one GRN method to another had a profound effect on the final inferred network (**Figure 3.6a**), the only two expectations being granie and pando in their CRE to gene assigment. Next, I wondered whether at least different runs would converge to a limited set of solutions. To quantify this, I clustered all the runs based on their pairwise overlap coefficients and found that all combinations of inferred GRNs were all quite distinct from one another as their overlap coefficient was just slightly higher than when compared to the random GRN (mean overlap coefficient of non-random comparisons = 0.071; mean overlap coefficient of random comparisons = 0.0026), suggesting that each combination produced a unique solution to the inference problem (**Figure 3.6b**). These two results support my earlier observation (**Figure 3.3a**), that currently these methods are quite unstable to processing and modeling choices.

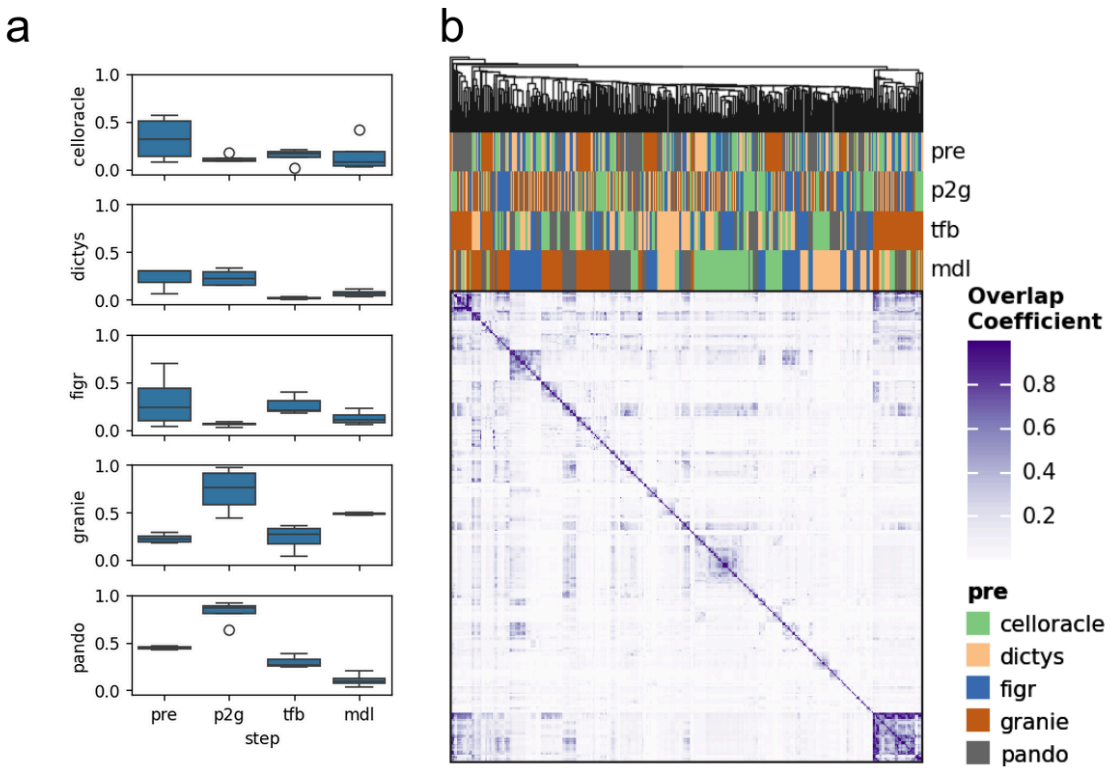

Figure 3.6. **a**, For each method, overlap coefficient between their original fixed pipeline with runs that only change one of the steps to another method's flavor (pre: preprocessing, p2g: CRE to gene, tfb: TF binding, mdl: modeling). **b**, Overlap coefficient across all combinations of runs.

3.2. Benchmark design

Up until now, this study has been purely descriptive. While the methods provide different solutions, critical questions remain: what can we trust? Are there formulations better than others? These are challenging questions, as there are no profiling technologies that directly measure GRNs, but rather capture their downstream effects. As a result, evaluating GRNs remains an open issue in the field and a comprehensive methodology to assess GRN performance is missing. To address these questions, I have designed an extensive set of evaluation metrics, divided into three main categories: mechanistic, predictive, and knowledge-based metrics (**Table 3.1.**). These metrics are aimed at testing the reconstruction of GRNs, where false positives are defined as regulatory interactions included in a GRN that either fail to recover effects following perturbations, fail to predict observed gene expression or chromatin accessibility, or are not supported by curated knowledge. For most metrics, I have parsed multiple external databases to be as comprehensive as possible as they have different coverages across them (**Figure 3.7**).

Mechanistic metrics assess whether GRNs are causal, meaning that if a TF is perturbed, its inferred target genes should show corresponding changes in expression. Additionally, these changes should propagate accordingly through the network. For this, I again used the KnockTF database[135], a curated collection of single TF perturbation experiments containing over 456 unique TFs and 907 experiments across various cell types and tissues. Specifically, I used their contrast statistics obtained from differential expression analysis between perturbed and unperturbed samples, often referred to as perturbation signatures.

In this category, I defined two metrics. The first, "TF activity", evaluated whether an inferred GRN could accurately predict TF perturbation based on the changes in expression of its target genes, using the ULM enrichment method from decoupler[105]. The second metric, "forecasting", tested whether the inferred network topology could replicate observed changes in target gene expression by simulating a TF perturbation with Markov chain propagation of gene expression through the network[85]. Predicted changes were then tested against the observed ones using Spearman correlation ($rho > 0.05$, $FDR < 0.05$).

For predictive metrics I defined two approaches. The first metric, "omics", involved splitting the original dataset used to infer the GRN into a training and testing set (80% train, 20% test). The goal was to assess how effectively the GRN topology can predict various omics features using XGBoost[153] regression from the original

observational data. Particularly, three different prediction tasks were employed: prediction of target gene expression based on accessibility of assigned CREs, prediction of CRE accessibility based on the expression of assigned TFs, and prediction of target gene expression based on expression of assigned TFs. Predictions were then tested with the test split using Spearman correlation (rho > 0.05, FDR < 0.05).

The second predictive metric, "Gene sets", evaluated whether the inferred GRN reflected the same gene sets detected at the single-cell level in the original data. Enrichment activity scores were first computed from the single-cell data. A gene set was deemed relevant if it was significantly enriched in a sufficient proportion of cells (ULM score > 0, FDR < 0.05, proportion > 20%). Subsequently, for each regulon overrepresentation of gene sets were tested using a one-sided Fisher exact test (FDR < 0.05). To maximize coverage, I used four different collections of gene sets: hallmarks[96] (4384 genes across 50 sets), kegg[154] (5244 genes across 186 sets), reactome[155](11290 genes across 1736 sets) and progeny[127](9499 genes across 13 gene sets).

For the knowledge-based metrics I used four assessment scores. The first metric, "TF markers", tested whether a GRN incorporated TFs that were critical for a given biological context. For example, a GRN defining the regulatory program of B-cells should contain PAX5 as it is a key TF to both induce and maintain B-cell lineage[156]. I have extracted such information from TF protein abundance located in the nuclei from the human protein atlas[157] (474 TFs across 117 labels), and from the database TF-Marker[158] (607 TFs across 302 labels).

The metric "TF binding" evaluated the capacity of a GRN to correctly recover previously measured TF-CRE binding events from ChIP-seq data. This evaluation has been the state-of-the-art approach to evaluate GRNs in past benchmarks[60,61], even when inferred only from transcriptomics where the exact CRE is not known and it is assumed to be the promoter region of the target gene. I extracted TF binding genomic regions from the chipatlas[159] (1055 TFs across 1032 labels), remap2022[160] (752 TFs across 130 labels) and unibind[161] (266 TFs across 983 labels) databases, only keeping binding peaks with less than 750 base pairs to remove potential artifacts.

The metric "CREs" measured if the CREs included in a GRN have been previously annotated as regulatory regions. This included regions with evolutionary conservation, promoter regions, disease-associated single nucleotide polymorphisms, and regions verified for regulatory activity through biochemical

assays. Evolutionary conserved regions have been reported to be functional as there is evidence of conserved enhancers subject to natural selection[162]. For this, I used the PhastCons regions collection[163] (2,243,811 CREs spanning 167 Mb). In a similar fashion, CREs that accumulate variants associated with disease can be assumed to be functional since their perturbation has consequences in the phenotype. Individual variants were extracted from the GWAS Catalog database[164] (295,333 CREs spanning 430,5 kb). Promoter regions are known for their regulatory activity[3], which I extracted from Ensembl through biomartR[165] (41,080 CREs spanning 82,2 Mb). Finally, I extracted annotated CREs from two enhancer databases, ENCODE[166] (1,060,118 CREs spanning 291,18 Mb) and Zhang21[167] (1,152,470 CREs spanning 463 Mb).

For the "Genes" metric, the objective was to determine whether the inferred GRN could accurately recover known CRE-Gene interactions. To achieve this, I leveraged the information extracted from expression quantitative studies which assign nucleotide variants to changes in gene expression. The data for these interactions was sourced from the eQTL catalog[168] (17,494 genes across 983 labels).

Because gene regulation is known to be highly context-specific[169], before applying most of these metrics, the aforementioned databases were filtered to include only tissue or cell-type information for the particular dataset being studied. For example, if a GRN was inferred from a hearth dataset, databases would be filtered to only contain cardiac information based on their available cell-type or tissue labels. Moreover, features such as genes or CREs not present in the original dataset were also removed from the databases.

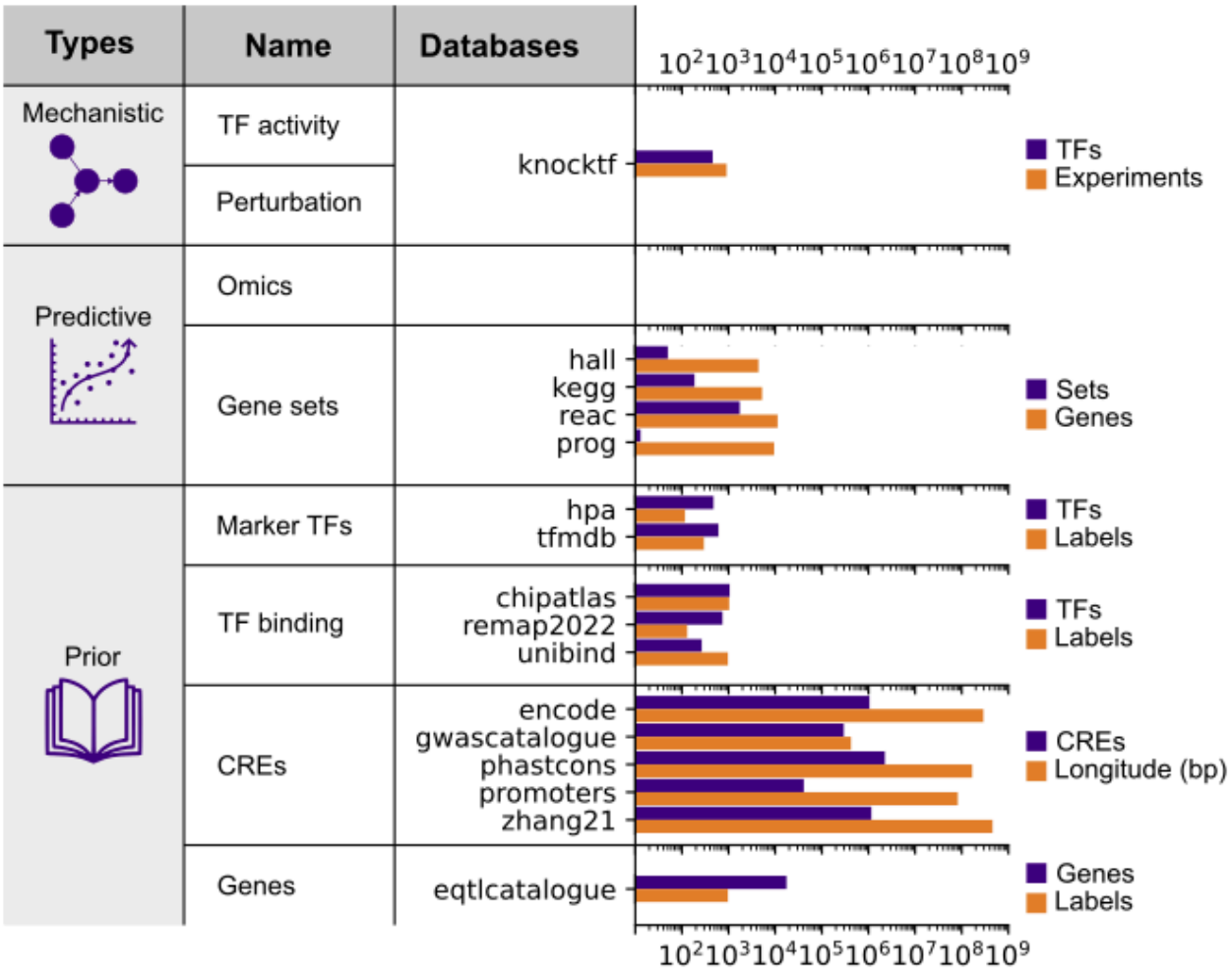

Figure 3.7. List of evaluation metrics, the databases employed to calculate evaluation scores and their statistics such as number of elements, or number of cell-type or tissue labels. The Omics does not use a database as it used the original single-cell data. Hallmarks (hall), reactome (reac), progeny (prog), human protein atlas (hpa).

Table 3.1. List of evaluation metrics used to evaluate multimodal GRNs together with their objective and definition. Metrics that mention context-specific information were tailored for each specific dataset. For example, when working with a heart dataset only entries that contain cardiac-related labels would be used. True Positives (TP), False Positives (FP) and False Negatives (FN).

Type	Name	Objective	Definition
Mechanistic	TF activity	Identify altered TFs in context-specific perturbation experiments	• TP = TF included in GRN with significant activity score for a perturbation experiment • FP = TF included in GRN but with insignificant activity score for a perturbation experiment • FN = TF not in GRN but in perturbation experiment
	Forecasting	Anticipate changes in gene expression after a context-specific TF perturbation simulation propagated through the GRN	• TP = TF targets included in GRN with significant Spearman correlation between predicted and observed changes in gene expression • FP = TF targets included in GRN but with insignificant Spearman correlation between predicted and observed changes in gene expression • FN = TF not in GRN but in perturbation experiment
Predictive	Omics	Independently predict omic feature readouts from other omic features from the original omics dataset. For example Gene ~ TF_1 + … + TF_n. The features to use are based on the inferred GRN topology	• TP = omic feature included in GRN with significant Spearman correlation between predicted and observed molecular readouts • FP = omic feature included in GRN but with insignificant Spearman correlation between predicted and observed molecular readouts • FN = omic feature not in GRN but in omic dataset
	Gene sets	Inclusion of gene sets in GRN that are predicted to be present in the original omics dataset	• TP = gene set in GRN and omics data • FP = gene set in GRN but not in omics data • FN = gene set not in GRN but in omics data
Knowledge	Marker TFs	Presence of context-specific TF markers in GRN	• TP = TF in GRN and marker collection • FP = TF in GRN but not in marker collection • FN = TF not in GRN but in marker collection
	TF binding	Presence of TF-CRE context-specific binding measurements in GRN	• TP = TF-CRE in GRN and in measurement collection • FP = TF-CRE in GRN but not in measurement collection • FN = TF-CRE not in GRN but in measurement collection
	CREs	Presence of annotated CREs in GRN	• TP = CRE in GRN and in annotation • FP = CRE in GRN but not in annotation • FN = CRE not in GRN but in annotation
	Regulated genes	Presence of context-specific CRE with annotated regulation to a neighboring gene in GRN	• TP = CRE-Gene in GRN and in annotation • FP = CRE-Gene in GRN but not in annotation • FN = CRE-Gene not in GRN but in annotation

Each metric's performance was evaluated using precision and recall scores. The final summary for each score was expressed as a weighted F-score, emphasizing precision ten times more than recall. This weighted approach is justified because while GRNs are unlikely to capture all possible interactions (reflected by recall), it is crucial to assess their ability to make accurate predictions (indicated by precision). Precision alone is insufficient, as a method that predicts only one interaction correctly would score perfectly on precision but poorly on recall. Therefore, recall was included but weighted with less importance.

Finally, GRN inference runs were ranked based on their performance for these metrics. To assess whether a particular method consistently ranked high across multiple runs in a given metric, I repurposed the GSEA enrichment method. Here, runs using the same step of a method were treated as a set, while the background consisted of all other runs not associated with that method. Empirical p-values were calculated by performing 1,000 random permutations to test for statistical significance.

3.3. Evaluation of multimodal GRN inference methods

To test the benchmarking capabilities of these metrics I required a multimodal dataset with enough context-specific information stored in the aforementioned databases. Unfortunately, the pituitary dataset was not suitable due to its low coverage in databases. Instead I used a dataset of human peripheral blood mononuclear cells available from 10X Genomics.

Similar to the pituitary dataset, I also generated all possible different combinations of GRN runs across methods using GRETA's pipeline. Then, each obtained GRN was evaluated by all the aforementioned metrics (**Section 3.2.**), constraining the databases to only contain human peripheral blood mononuclear cells labels whenever possible, as to make the evaluation as context-specific as possible. The resulting F_{01} evaluation scores were then summarized per metric type and step-method combinations by calculating their mean across databases for a particular task, and then summarized again by the mean at the metric type level (**Figure 3.8a**).

I observed that in general, all methods score relatively high in the predictive tasks (mean F_{01} score = 0.51). Runs significantly ranked higher than other strategies when using preprocessing from celloracle, figr or dictys (NES > 2, FDR < 0.05), CRE to gene assignment from dictys or pando (NES > 2, FDR < 0.05), TF binding predictions from dictys and modeling from celloracle or dictys (NES > 2, FDR <

0.05). Most methods performed moderately at knowledge-based (mean F_{01} score = 0.28), with runs significantly ranked higher than other strategies when using preprocessing of celloracle, figr or dictys (NES > 2, FDR < 0.05), CRE to gene assignment from celloracle (NES > 2, FDR < 0.05), and TF binding predictions from dictys or granie (NES > 2, FDR < 0.05). However, all methods struggled in the mechanistic tasks (mean F_{01} score = 0.02), with no strategy significantly ranked above the rest.

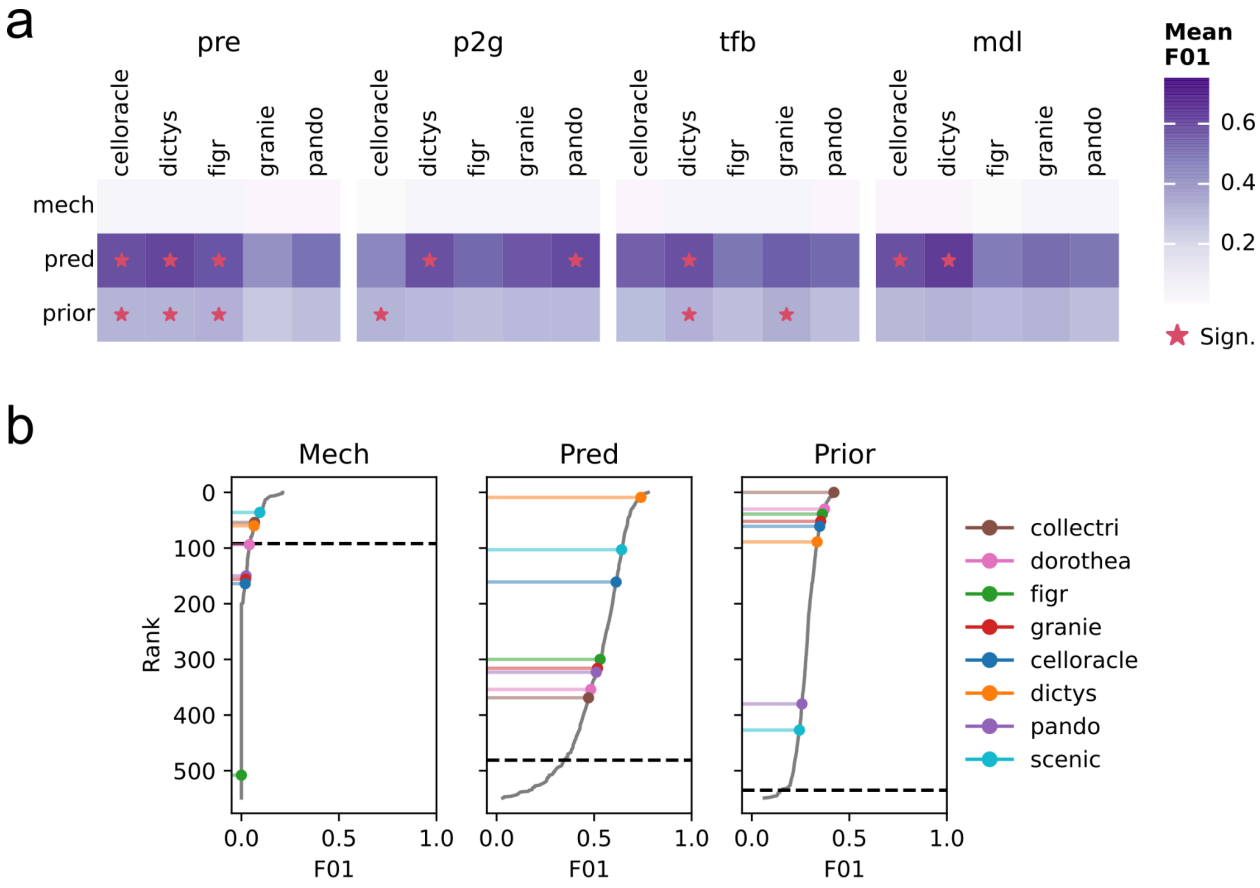

Figure 3.8. **a**, Mean F_{01} scores across methods and types of evaluation metrics for each inference step. Asterisks indicate significance of higher ranking for a particular step-method combination compared to the rest (FDR < 0.05). **b**, Ranked distributions of F_{01} for all possible combinations across types of metrics. Colored lollipop plots indicate the position in the ranking for the original fixed runs. Dashed black line indicates the position of the random network.

I also compared the ranking of the original methods with the literature-derived GRNs (collectri and dorothea), a data-driven method based solely on transcriptomics (scenic) and a random GRN (**Figure 3.8b**). Most multimodal GRN methods could not score better than the random baseline in mechanistic tasks, the exceptions being scenic, collectri and dictys. For the predictive metrics all methods outperformed the random baseline, with dictys, scenic and celloracle ranking higher

than the rest. Finally, all methods also were able to outperform the random baseline when recovering known biology. Literature-derived GRNs ranked better than any method for this metric, followed by figr, granie, celloracle and dictys.

Taken together, these results suggest that this new wave of multimodal GRN inference methods, although promising, fall short of their promises. I have shown that they are highly unstable to preprocessing and modeling decisions, and do not substantially outperform literature-derived or transcriptome-only-based GRNs. Considering also their high computational cost and difficult usage, these results suggest that it may be more convenient to use literature-derived GRNS. These can easily be fetched as lightweight dataframes and contextualize them by filtering for the observed expressed genes. Future studies will determine if new methods are able to surpass literature-derived GRNs.

Discussion and future perspectives

In this thesis, I have summarized the current state of the art in gene regulatory network (GRN) inference in the first chapter, highlighting the new generation of methods that integrate transcriptomics and chromatin accessibility data to potentially infer more accurate networks. I have identified and standardized the set of shared steps in multimodal GRN inference, providing method developers with a clearer understanding of the decisions they face. For example, I highlighted that many current methods infer transcription factor (TF) binding before assigning cis-regulatory elements (CREs) to genes, which can lead to an increased number of false predictions as numerous CREs will ultimately not be linked to genes.

There, I have focused on transcriptomic and chromatin accessibility methods that infer GRNs, but other modeling strategies exist that combine both omics to better understand gene regulation. For example, there is a collection of deep learning-based computational methods that model measured chromatin accessibility from sequence data at base pair resolution[170–172]. Through in silico mutagenesis of DNA sequence, these models identify context-specific impactful nucleotides. This information can then be used to better understand the effect of rare nucleotide variants in disease, or to perform context-specific TF binding predictions through motif matching.

Other layers of gene regulation exist that I have not explored. For example, MicroRNAs are small RNA molecules that can regulate gene expression by binding to complementary sequences of target mRNAs, leading to their degradation or translational inhibition[173]. This mechanism is crucial for cellular development and function. Approaches similar to gene co-expression analysis have been adapted to infer potential miRNA interactions with target genes[32,174]. However, such analyses require short-read RNA sequencing measurements, which are less commonly available than standard transcriptomics[175]. While there are some variations of miRNA profiling technologies at the single-cell level, these have yet to see widespread adoption[176].

Another layer of gene regulation that is often overlooked is alternative splicing, where a single gene can produce different mRNA transcripts from the same DNA sequence. TFs play a dual role in this case, as they have the ability to promote alternative splicing of mRNAs[177], but also because their own isoforms can yield completely different regulatory activities, some of them leading to diseases such as cancer[178]. However, dealing with transcript isoforms is a hard task, as it requires

long-sequencing profiling techniques and computational approaches to correctly identify and annotate valid isoforms[179,180].

It is also crucial to understand that gene regulation operates within a complex and interconnected cellular network. The classic example of the lac operon, where lactose metabolism activates gene regulation, illustrates that GRNs are intricately linked with broader cellular processes. Both metabolism and gene expression reciprocally regulate each other to maintain homeostasis and regulate cell growth, survival and differentiation[181]. Gene expression does it by controlling the output of enzymatic genes, and metabolism does it by modifying DNA and histones through enzymatic activity, which will have consequences in chromatin accessibility and thus gene expression.

Additionally, gene regulation depends on signaling cascades based on internal and external cues. When signaling receptors are activated by the binding of ligands, they trigger the activation of signaling proteins, which transmit the signal through protein modifications[182]. The most prevalent group of signaling proteins are kinases, proteins specialized in the transfer of phospho groups across molecules. These modifications propagate the signal, ultimately activating or inhibiting various TFs, which then impact gene regulation by modulating the expression of specific genes.

On top of that, these signaling cues rely heavily on the cellular context, influenced by the neighborhood in which a cell is located[183]. There are computational single-cell methods designed to predict ligand-receptor communication events. Even though they are widely used by the community, their reliability has been debated[134]. Recently, with the advent of spatial transcriptomics, there are refined approaches that incorporate spatial proximity to better infer cell-cell communication, enhancing the accuracy of predicted interactions[134].

An equally crucial aspect of gene regulation is its temporal dynamics. After a TF is transcribed, it may take hours before it becomes a functional regulator of other genes, highlighting the time-sensitive nature of regulatory processes. Classic GRN methods, such as those based on ordinary differential equations or boolean networks, have attempted to capture these dynamics. More recent approaches incorporate single-cell pseudotime to model regulatory changes over time[184]. Current technologies still face significant limitations in accurately resolving these temporal aspects, although lineage capture approaches are starting to be developed at the single-cell resolution[185].

The incorporation of chromatin state on top of transcriptomics into GRN inference has been a step in the right direction, but future modeling approaches should aim at

also incorporating these other key aspects of gene regulation to provide a more complete picture of how GRNs operate. This holistic approach could reveal more detailed regulatory dynamics and offer better insights into cellular behavior.

In the second chapter, I have showcased one of the most important applications of GRNs, TF activity scoring inference, where I showed that simple linear models outperform classic enrichment scoring approaches. Additionally, I have presented decoupler, a computational tool I developed that, by inferring TF activities, facilitates the discovery of new biological insights, as demonstrated in the COVID-19 and multiple sclerosis projects. Although decoupler's Python implementation is highly scalable compared to other enrichment frameworks, it still relies on CPU-based compiled code. With the increasing availability of single-cell and spatial transcriptomics data from initiatives like the Human Cell Atlas[169], involving millions of cells, parallel computing strategies, such as GPU-based processing, are becoming essential. Rapids single-cell[186], a GPU-accelerated analysis toolbox from the scverse[124], has already integrated decoupler in their framework, making it feasible to apply enrichment analysis to large-scale datasets.

In the third and last chapter, I developed the Gene Regulatory nETwork Analysis (GRETA) snakemake pipeline, a comprehensive comparison of these new multimodal GRN inference methods, with an extensive set of GRN evaluation metrics allowing me to rank these methods against baseline models. While this evaluation framework is, to my knowledge, the most exhaustive to date, it remains an indirect way of assessing GRN quality due to the lack of direct profiling technologies for regulatory interactions. Each metric presented relies on assumptions, for example, one of the knowledge-based CRE metrics assumes that CREs are evolutionarily conserved, which is not universally true. Despite these limitations, combining multiple evaluation metrics offers a more robust approach than traditional evaluation attempts, such as simply assessing overlaps with available ChIP-seq binding data. Currently, GRETA is restricted to human data, as the databases that I have processed were human-specific, but future work will aim to expand the framework to include model organisms such as the mouse or the fly. Moreover, the methods shown in this benchmark were selected above many others because of their easier implementation, but more are missing such as scenic+[70].

The main result of my benchmark is that while inferred GRNs perform well in predictive tasks and moderately reflect known biology, they lack mechanistic properties. This is expected because the process of trans-regulation involves numerous steps that are not fully captured by gene expression or chromatin accessibility data alone. For a TF to regulate a target gene, its transcript must first

exit the nucleus via nuclear pores, get translated into protein, undergo activation through post-translational modifications, re-enter the nucleus, interact with other nuclear TFs, and require accessible chromatin and favorable conditions at the target TSS to exert regulatory effects. This chain of events is further complicated by processes we may not yet fully understand or have not yet discovered.

One of the main limitations of all current inference methods is that they assume that TF expression variability is informative when it has been shown that TF expression measurements alone do not correlate well with their activity[187]. This issue is especially pronounced in single-cell datasets, which suffer from technical dropouts and general data sparsity, further complicating the inference of regulatory interactions.

In my benchmark, I have found that multimodal approaches sometimes perform comparably to, or worse, than literature-derived GRNs, suggesting that de novo GRN inference may be overly complex and that integrating prior biological knowledge in the inference pipelines could improve inference performance. For example, one could, instead of using TF expression, first infer TF activities using a prior GRN and then refine these interactions based on both observed chromatin accessibility and target gene expression.

In summary, my thesis serves as a comprehensive resource for understanding how gene regulation is modeled within systems biology. It demonstrates how GRNs can yield new biological insights in the context of human disease and evaluates the latest wave of multimodal inference methods. While these methods show promise, my work highlights significant caveats and raises the need for further advancements in the field of GRN inference.

Bibliography

1. Weidemüller, P., Kholmatov, M., Petsalaki, E. & Zaugg, J. B. Transcription factors: Bridge between cell signaling and gene regulation. *Proteomics* **21**, e2000034 (2021).
2. Crick, F. Central dogma of molecular biology. *Nature* **227**, 561–563 (1970).
3. Kim, S. & Wysocka, J. Deciphering the multi-scale, quantitative cis-regulatory code. *Mol. Cell* (2023) doi:10.1016/j.molcel.2022.12.032.
4. Zaret, K. S. & Carroll, J. S. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev.* **25**, 2227–2241 (2011).
5. Chen, H. *et al.* Dynamic interplay between enhancer-promoter topology and gene activity. *Nat. Genet.* **50**, 1296–1303 (2018).
6. Committee on the Review of Omics-Based Tests for Predicting Patient Outcomes in Clinical Trials, Board on Health Care Services, Institute of Medicine & Board on Health Sciences Policy. *Evolution of Translational Omics*. (National Academies Press, Washington, D.C., DC, 2012).
7. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
8. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
9. Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–322 (2008).
10. Kelly, T. K. *et al.* Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Research* vol. 22 2497–2506 Preprint at https://doi.org/10.1101/gr.143008.112 (2012).
11. Minnoye, L. *et al.* Chromatin accessibility profiling methods. *Nature Reviews Methods Primers* **1**, 1–24 (2021).
12. Yan, F., Powell, D. R., Curtis, D. J. & Wong, N. C. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biol.* **21**, 22 (2020).
13. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497–1502 (2007).
14. Kaya-Okur, H. S. *et al.* CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* **10**, 1930 (2019).
15. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
16. Marti-Renom, M. A. *et al.* Challenges and guidelines toward 4D nucleome data and model standards. *Nat. Genet.* **50**, 1352–1358 (2018).
17. Cha, J. & Lee, I. Single-cell network biology for resolving cellular heterogeneity in human diseases. *Exp. Mol. Med.* **52**, 1798–1808 (2020).
18. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).
19. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
20. Crosetto, N., Bienko, M. & van Oudenaarden, A. Spatially resolved transcriptomics and beyond. *Nat. Rev. Genet.* **16**, 57–66 (2015).
21. Rao, A., Barkley, D., França, G. S. & Yanai, I. Exploring tissue architecture using spatial transcriptomics. *Nature* **596**, 211–220 (2021).
22. Kleshchevnikov, V. *et al.* Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat. Biotechnol.* **40**, 661–671 (2022).
23. Bressan, D., Battistoni, G. & Hannon, G. J. The dawn of spatial omics. *Science* **381**, eabq4964 (2023).
24. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biol.* **21**, 31 (2020).
25. Argelaguet, R., Cuomo, A. S. E., Stegle, O. & Marioni, J. C. Computational principles and challenges in single-cell data integration. *Nat. Biotechnol.* **39**, 1202–1215 (2021).
26. Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.* **37**, 1452–1457 (2019).
27. Liu, L. *et al.* Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity. *Nat. Commun.* **10**, 470 (2019).
28. Ma, S. *et al.* Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* **183**, 1103–1116.e20 (2020).

29. Zhang, D. *et al.* Spatial epigenome-transcriptome co-profiling of mammalian tissues. *Nature* **616**, 113–122 (2023).
30. Davidson, E. H. *et al.* A genomic regulatory network for development. *Science* **295**, 1669–1678 (2002).
31. Badia-I-Mompel, P. *et al.* Gene regulatory network inference in the era of single-cell multi-omics. *Nat. Rev. Genet.* **24**, 739–754 (2023).
32. Lai, X., Wolkenhauer, O. & Vera, J. Understanding microRNA-mediated gene regulatory networks through mathematical modelling. *Nucleic Acids Res.* **44**, 6019–6035 (2016).
33. Du, J.-X. *et al.* Splicing factors: Insights into their regulatory network in alternative splicing in cancer. *Cancer Lett.* **501**, 83–104 (2021).
34. Statello, L., Guo, C.-J., Chen, L.-L. & Huarte, M. Gene regulation by long non-coding RNAs and its biological functions. *Nat. Rev. Mol. Cell Biol.* **22**, 96–118 (2021).
35. Carthew, R. W. Gene Regulation and Cellular Metabolism: An Essential Partnership. *Trends Genet.* **37**, 389–400 (2021).
36. Jacob, F. & Monod, J. Genetic Regulatory Mechanisms in the Synthesis of Proteins. *Molecular Biology* 82–120 Preprint at https://doi.org/10.1016/b978-0-12-131200-8.50010-1 (1989).
37. Davidson, E. H. & Erwin, D. H. Gene regulatory networks and the evolution of animal body plans. *Science* **311**, 796–800 (2006).
38. Su, E. Y., Spangler, A., Bian, Q., Kasamoto, J. Y. & Cahan, P. Reconstruction of dynamic regulatory networks reveals signaling-induced topology changes associated with germ layer specification. *Stem Cell Reports* **17**, 427–442 (2022).
39. Claringbould, A. & Zaugg, J. B. Enhancers in disease: molecular basis and emerging treatment strategies. *Trends Mol. Med.* **27**, 1060–1073 (2021).
40. Liu, Z.-P., Wu, C., Miao, H. & Wu, H. RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database* **2015**, (2015).
41. Han, H. *et al.* TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res.* **46**, D380–D386 (2018).
42. Müller-Dott, S. *et al.* Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities. *Nucleic Acids Res.* **51**, 10934–10949 (2023).
43. Mercatelli, D., Scalambra, L., Triboli, L., Ray, F. & Giorgi, F. M. Gene regulatory network inference resources: A practical overview. *Biochim. Biophys. Acta Gene Regul. Mech.* **1863**, 194430 (2020).
44. Galton, F. Regression towards mediocrity in hereditary stature. *J. Anthropol. Inst. G. B. Irel.* **15**, 246 (1886).
45. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
46. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring regulatory networks from expression data using tree-based methods. *PLoS One* **5**, (2010).
47. Moerman, T. *et al.* GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics* **35**, 2159–2161 (2019).
48. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **175**, 598–599 (2018).
49. Granger, C. W. J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37**, 424 (1969).
50. Wagner, A., Regev, A. & Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* **34**, 1145–1160 (2016).
51. Herring, C. A., Chen, B., McKinley, E. T. & Lau, K. S. Single-Cell Computational Strategies for Lineage Reconstruction in Tissue Systems. *Cell Mol Gastroenterol Hepatol* **5**, 539–548 (2018).
52. Specht, A. T. & Li, J. LEAP: constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering. *Bioinformatics* **33**, 764–766 (2017).
53. Papili Gao, N., Ud-Dean, S. M. M., Gandrillon, O. & Gunawan, R. SINCERITIES: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics* **34**, 258–266 (2018).
54. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
55. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
56. Bergen, V., Soldatov, R. A., Kharchenko, P. V. & Theis, F. J. RNA velocity-current challenges and future perspectives. *Mol. Syst. Biol.* **17**, e10282 (2021).
57. Zheng, S. C., Stein-O'Brien, G., Boukas, L., Goff, L. A. & Hansen, K. D. Pumping the brakes on RNA velocity by understanding and interpreting RNA velocity estimates. *Genome Biol.* **24**, 246 (2023).
58. Holland, C. H. *et al.* Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. *Genome Biol.* **21**, 36 (2020).
59. Marbach, D. *et al.* Wisdom of crowds for robust gene network inference. *Nat. Methods* **9**, 796–804 (2012).

60. Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A. & Murali, T. M. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat. Methods* **17**, 147–154 (2020).
61. McCalla, S. G. *et al.* Identifying strengths and weaknesses of methods for computational network inference from single-cell RNA-seq data. *G3* **13**, (2023).
62. Lee, T. I. *et al.* Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* **298**, 799–804 (2002).
63. Grosselin, K. *et al.* High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. *Nat. Genet.* **51**, 1060–1066 (2019).
64. Bartosovic, M., Kabbe, M. & Castelo-Branco, G. Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues. *Nat. Biotechnol.* **39**, 825–835 (2021).
65. Zambelli, F., Pesole, G. & Pavesi, G. Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. *Brief. Bioinform.* **14**, 225–237 (2013).
66. Pranzatelli, T. J. F., Michael, D. G. & Chiorini, J. A. ATAC2GRN: optimized ATAC-seq and DNase1-seq pipelines for rapid and accurate genome regulatory network inference. *BMC Genomics* vol. 19 Preprint at https://doi.org/10.1186/s12864-018-4943-z (2018).
67. Qin, Q. *et al.* Lisa: inferring transcriptional regulators through integrative modeling of public chromatin accessibility and ChIP-seq data. *Genome Biol.* **21**, 32 (2020).
68. Sonawane, A. R., DeMeo, D. L., Quackenbush, J. & Glass, K. Constructing gene regulatory networks using epigenetic data. *NPJ Syst Biol Appl* **7**, 45 (2021).
69. Weirauch, M. T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).
70. Bravo González-Blas, C. *et al.* SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nat. Methods* **20**, 1355–1367 (2023).
71. Kulakovskiy, I. V. *et al.* HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.* **46**, D252–D259 (2018).
72. Castro-Mondragon, J. A. *et al.* JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **50**, D165–D173 (2022).
73. Bruse, N. & van Heeringen, S. J. GimmeMotifs: an analysis framework for transcription factor motif analysis. (2018) doi:10.1101/474403.
74. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
75. Korhonen, J. H., Palin, K., Taipale, J. & Ukkonen, E. Fast motif matching revisited: high-order PWMs, SNPs and indels. *Bioinformatics* **33**, 514–521 (2017).
76. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).
77. Ambrosini, G., Groux, R. & Bucher, P. PWMScan: a fast tool for scanning entire genomes with a position-specific weight matrix. *Bioinformatics* **34**, 2483–2484 (2018).
78. Goldstein, E. *Encyclopedia of Perception*. (SAGE Publications, Inc., 2455 Teller Road, Thousand Oaks California 91320 United States, 2010).
79. Qin, J., Hu, Y., Xu, F., Yalamanchili, H. K. & Wang, J. Inferring gene regulatory networks by integrating ChIP-seq/chip and transcriptome data via LASSO-type regularization methods. *Methods* **67**, 294–303 (2014).
80. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
81. Ramirez, R. N. *et al.* Dynamic Gene Regulatory Networks of Human Myeloid Differentiation. *Cell Syst* **4**, 416–429.e3 (2017).
82. Starks, R. R., Biswas, A., Jain, A. & Tuteja, G. Combined analysis of dissimilar promoter accessibility and gene expression profiles identifies tissue-specific genes and actively repressed networks. *Epigenetics Chromatin* **12**, 16 (2019).
83. Johnson, J. S. *et al.* A Comprehensive Map of the Monocyte-Derived Dendritic Cell Transcriptional Network Engaged upon Innate Sensing of HIV. *Cell Rep.* **30**, 914–931.e9 (2020).
84. Kartha, V. K. *et al.* Functional inference of gene regulation using single-cell multi-omics. *Cell Genom* **2**, (2022).
85. Kamimoto, K. *et al.* Dissecting cell identity via network inference and in silico gene perturbation. *Nature* Preprint at https://doi.org/10.1038/s41586-022-05688-9 (2023).
86. Kamal, A. *et al.* GRaNIE and GRaNPA: inference and evaluation of enhancer-mediated gene regulatory networks. *Mol. Syst. Biol.* e11627 (2023).
87. Fleck, J. S. *et al.* Inferring and perturbing cell fate regulomes in human brain organoids. *Nature* (2022)

doi:10.1038/s41586-022-05279-8.
88. Gasperini, M. *et al.* A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell* **176**, 1516 (2019).
89. Bahr, C. *et al.* Author Correction: A Myc enhancer cluster regulates normal and leukaemic haematopoietic stem cell hierarchies. *Nature* **558**, E4 (2018).
90. Monahan, K., Horta, A. & Lomvardas, S. LHX2- and LDB1-mediated trans interactions regulate olfactory receptor choice. *Nature* vol. 565 448–453 Preprint at https://doi.org/10.1038/s41586-018-0845-0 (2019).
91. Wang, L. *et al.* Dictys: dynamic gene regulatory network dissects developmental continuum with single-cell multiomics. *Nat. Methods* **20**, 1368–1378 (2023).
92. Zuin, J. *et al.* Nonlinear control of transcription through enhancer–promoter interactions. *Nature* **604**, 571–577 (2022).
93. Kuppe, C. *et al.* Spatial multi-omic map of human myocardial infarction. *Nature* **608**, 766–777 (2022).
94. Argelaguet, R. *et al.* Decoding gene regulation in the mouse embryo using single-cell multi-omics. *bioRxiv* 2022.06.15.496239 (2022) doi:10.1101/2022.06.15.496239.
95. Duren, Z., Chen, X., Xin, J., Wang, Y. & Wong, W. H. Time course regulatory analysis based on paired expression and chromatin accessibility data. *Genome Res.* **30**, 622–634 (2020).
96. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417–425 (2015).
97. Duren, Z. *et al.* Sc-compReg enables the comparison of gene regulatory networks between conditions using single-cell data. *Nat. Commun.* **12**, 4763 (2021).
98. Xu, Q. *et al.* ANANSE: an enhancer network-based computational approach for predicting key transcription factors in cell fate determination. *Nucleic Acids Res.* **49**, 7966–7985 (2021).
99. Anderson, A. G. *et al.* Single nucleus multiomics identifies ZEB1 and MAFB as candidate regulators of Alzheimer's disease-specific cis-regulatory elements. *Cell Genomics* **0**, (2023).
100. Bachireddy, P. *et al.* Mapping the evolution of T cell states during response and resistance to adoptive cellular therapy. *Cell Rep.* **37**, 109992 (2021).
101. Thompson, D., Regev, A. & Roy, S. Comparative analysis of gene regulatory networks: from network reconstruction to evolution. *Annu. Rev. Cell Dev. Biol.* **31**, 399–428 (2015).
102. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
103. Lou, S. *et al.* TopicNet: a framework for measuring transcriptional regulatory network change. *Bioinformatics* **36**, i474–i481 (2020).
104. Zhang, S. *et al.* Inference of cell type-specific gene regulatory networks on cell lineages from single cell omic datasets. *Nat. Commun.* **14**, 3064 (2023).
105. Badia-i-Mompel, P. *et al.* decoupleR: ensemble of computational methods to infer biological activities from omics data. *Bioinformatics Advances* **2**, vbac016 (2022).
106. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).
107. Alvarez, M. J. *et al.* Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat. Genet.* **48**, 838–847 (2016).
108. Garcia-Alonso, L. *et al.* Transcription Factor Activities Enhance Markers of Drug Sensitivity in Cancer. *Cancer Research* vol. 78 769–780 Preprint at https://doi.org/10.1158/0008-5472.can-17-1679 (2018).
109. Walsh, L. A. *et al.* An Integrated Systems Biology Approach Identifies TRIM25 as a Key Determinant of Breast Cancer Metastasis. *Cell Rep.* **20**, 1623–1640 (2017).
110. Guan, X. *et al.* Androgen receptor activity in T cells limits checkpoint blockade efficacy. *Nature* **606**, 791–796 (2022).
111. Melms, J. C. *et al.* A molecular single-cell lung atlas of lethal COVID-19. *Nature* **595**, 114–119 (2021).
112. Goeman, J. J. & Bühlmann, P. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* **23**, 980–987 (2007).
113. Dugourd, A. & Saez-Rodriguez, J. Footprint-based functional analysis of multiomic data. *Curr. Opin. Syst. Biol.* **15**, 82–90 (2019).
114. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
115. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
116. Fisher, R. A. On the interpretation of $\chi 2$ from contingency tables, and the calculation of P. *J. R. Stat. Soc.* (1922).
117. Väremo, L., Nielsen, J. & Nookaew, I. Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic Acids Res.* **41**,

4378–4391 (2013).
118. Geistlinger, L., Csaba, G. & Zimmer, R. Bioconductor's EnrichmentBrowser: seamless navigation through combined results of set- & network-based enrichment analysis. *BMC Bioinformatics* **17**, 45 (2016).
119. Alhamdoosh, M. *et al.* Combining multiple tools outperforms individual methods in gene set enrichment analyses. *Bioinformatics* **33**, 414–424 (2017).
120. Numba: A llvm-based python jit compiler. in *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*.
121. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7 (2013).
122. Hernandez-Armenta, C., Ochoa, D., Gonçalves, E., Saez-Rodriguez, J. & Beltrao, P. Benchmarking substrate-based kinase activity inference using phosphoproteomic data. *Bioinformatics* **33**, 1845–1851 (2017).
123. Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D. & Saez-Rodriguez, J. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res.* **29**, 1363–1375 (2019).
124. Virshup, I. *et al.* The scverse project provides a computational ecosystem for single-cell omics data analysis. *Nat. Biotechnol.* **41**, 604–606 (2023).
125. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
126. Muzellec, B., Teleńczuk, M., Cabeli, V. & Andreux, M. PyDESeq2: a python package for bulk RNA-seq differential expression analysis. *Bioinformatics* **39**, (2023).
127. Schubert, M. *et al.* Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nat. Commun.* **9**, 20 (2018).
128. Sakai, T. *et al.* Myocardin regulates fibronectin expression and secretion from human pleural mesothelial cells. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **326**, L419–L430 (2024).
129. Yu-Wai-Man, C. *et al.* Local delivery of novel MRTF/SRF inhibitors prevents scar tissue formation in a preclinical model of fibrosis. *Sci. Rep.* **7**, (2017).
130. Cui, L. *et al.* Activation of JUN in fibroblasts promotes pro-fibrotic programme and modulates protective immunity. *Nat. Commun.* **11**, 2795 (2020).
131. Franzén, O., Gan, L.-M. & Björkegren, J. L. M. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database* **2019**, (2019).
132. Medvedovic, J., Ebert, A., Tagoh, H. & Busslinger, M. Pax5: a master regulator of B cell development and leukemogenesis. *Adv. Immunol.* **111**, 179–206 (2011).
133. Squair, J. W. *et al.* Confronting false discoveries in single-cell differential expression. *Nat. Commun.* **12**, 5692 (2021).
134. Dimitrov, D. *et al.* LIANA+ provides an all-in-one framework for cell-cell communication inference. *Nat. Cell Biol.* **26**, 1613–1622 (2024).
135. Feng, C. *et al.* KnockTF 2.0: a comprehensive gene expression profile database with knockdown/knockout of transcription (co-)factors in multiple species. *Nucleic Acids Res.* **52**, D183–D193 (2024).
136. Türei, D. *et al.* Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. *Mol. Syst. Biol.* **17**, e9923 (2021).
137. Wanner, N. *et al.* Molecular consequences of SARS-CoV-2 liver tropism. *Nat Metab* **4**, 310–319 (2022).
138. Lerma-Martin, C. *et al.* Cell type mapping reveals tissue niches and interactions in subcortical multiple sclerosis lesions. *Nat. Neurosci.* (2024) doi:10.1038/s41593-024-01796-z.
139. Thakur, S., Kumar, V., Das, R., Sharma, V. & Mehta, D. K. Biomarkers of hepatic toxicity: An overview. *Curr. Ther. Res. Clin. Exp.* **100**, 100737 (2024).
140. Harrison, A. G., Lin, T. & Wang, P. Mechanisms of SARS-CoV-2 transmission and pathogenesis. *Trends Immunol.* **41**, 1100–1115 (2020).
141. GTEx Consortium *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
142. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
143. Stebbing, J. *et al.* JAK inhibition reduces SARS-CoV-2 liver infectivity and modulates inflammatory responses to reduce morbidity and mortality. *Sci. Adv.* **7**, eabe4724 (2021).
144. Boldanova, T., Suslov, A., Heim, M. H. & Necsulea, A. Transcriptional response to hepatitis C virus infection and interferon-alpha treatment in the human liver. *EMBO Mol. Med.* **9**, 816–834 (2017).
145. Fourman, L. T. *et al.* Effects of tesamorelin on hepatic transcriptomic signatures in HIV-associated NAFLD. *JCI Insight* **5**, (2020).
146. Rico Montanari, N. *et al.* Transcriptomic analysis of livers of inactive carriers of hepatitis B virus with distinct expression of hepatitis B surface antigen. *J. Infect. Dis.* **225**, 1081–1090 (2022).

147. Lassmann, H. Multiple Sclerosis Pathology. *Cold Spring Harb. Perspect. Med.* **8**, (2018).
148. Wallmeier, J. *et al.* De Novo Mutations in FOXJ1 Result in a Motile Ciliopathy with Hydrocephalus and Randomization of Left/Right Body Asymmetry. *Am. J. Hum. Genet.* **105**, 1030–1039 (2019).
149. Ignatenko, O. *et al.* Mitochondrial dysfunction compromises ciliary homeostasis in astrocytes. *J. Cell Biol.* **222**, (2023).
150. Zhang, Z. *et al.* Single nucleus transcriptome and chromatin accessibility of postmortem human pituitaries reveal diverse stem cell regulatory mechanisms. *Cell Rep.* **38**, 110467 (2022).
151. Shoyab, M., Plowman, G. D., McDonald, V. L., Bradley, J. G. & Todaro, G. J. Structure and function of human amphiregulin: a member of the epidermal growth factor family. *Science* **243**, 1074–1076 (1989).
152. Bolitho, C., Moscova, M., Baxter, R. C. & Marsh, D. J. Amphiregulin increases migration and proliferation of epithelial ovarian cancer cells by inducing its own expression via PI3-kinase signaling. *Mol. Cell. Endocrinol.* **533**, 111338 (2021).
153. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, NY, USA, 2016). doi:10.1145/2939672.2939785.
154. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
155. Milacic, M. *et al.* The reactome pathway knowledgebase 2024. *Nucleic Acids Res.* **52**, D672–D678 (2024).
156. Cobaleda, C., Schebesta, A., Delogu, A. & Busslinger, M. Pax5: the guardian of B cell identity and function. *Nat. Immunol.* **8**, 463–470 (2007).
157. Uhlén, M. *et al.* A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol. Cell. Proteomics* **4**, 1920–1932 (2005).
158. Xu, M. *et al.* TF-Marker: a comprehensive manually curated database for transcription factors and related markers in specific cell and tissue types in human. *Nucleic Acids Res.* **50**, D402–D412 (2022).
159. Zou, Z., Ohta, T. & Oki, S. ChIP-Atlas 3.0: a data-mining suite to explore chromosome architecture together with large-scale regulome data. *Nucleic Acids Res.* **52**, W45–W53 (2024).
160. Hammal, F., de Langen, P., Bergon, A., Lopez, F. & Ballester, B. ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. *Nucleic Acids Res.* **50**, D316–D325 (2022).
161. Puig, R. R., Boddie, P., Khan, A., Castro-Mondragon, J. A. & Mathelier, A. UniBind: maps of high-confidence direct TF-DNA interactions across nine species. *BMC Genomics* **22**, 482 (2021).
162. Pennacchio, L. A. *et al.* In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499–502 (2006).
163. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
164. Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* **51**, D977–D985 (2023).
165. Drost, H.-G. & Paszkowski, J. Biomartr: genomic data retrieval with R. *Bioinformatics* **33**, 1216–1217 (2017).
166. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636–640 (2004).
167. Zhang, K. *et al.* A single-cell atlas of chromatin accessibility in the human genome. *Cell* **184**, 5985–6001.e19 (2021).
168. Kerimov, N. *et al.* A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat. Genet.* **53**, 1290–1299 (2021).
169. Regev, A. *et al.* The Human Cell Atlas. *Elife* **6**, (2017).
170. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
171. Avsec, Ž. *et al.* Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat. Genet.* **53**, 354–366 (2021).
172. Yuan, H. & Kelley, D. R. scBasset: sequence-based modeling of single-cell ATAC-seq using convolutional neural networks. *Nat. Methods* **19**, 1088–1096 (2022).
173. Lee, R. C., Feinbaum, R. L. & Ambros, V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**, 843–854 (1993).
174. Brennecke, J., Stark, A., Russell, R. B. & Cohen, S. M. Principles of microRNA-target recognition. *PLoS Biol.* **3**, e85 (2005).
175. Chi, S. W., Zang, J. B., Mele, A. & Darnell, R. B. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature* **460**, 479–486 (2009).
176. Hücker, S. M. *et al.* Single-cell microRNA sequencing method comparison and application to cell lines and

circulating lung tumor cells. *Nat. Commun.* **12**, 4316 (2021).
177. Boumpas, P., Merabet, S. & Carnesecchi, J. Integrating transcription and splicing into cell fate: Transcription factors on the block. *Wiley Interdiscip. Rev. RNA* **14**, e1752 (2023).
178. Belluti, S., Rigillo, G. & Imbriano, C. Transcription factors in cancer: When alternative splicing determines opposite cell fates. *Cells* **9**, 760 (2020).
179. Pardo-Palacios, F. J. *et al.* SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. *Nat. Methods* **21**, 793–797 (2024).
180. Pardo-Palacios, F. J. *et al.* Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. *Nat. Methods* **21**, 1349–1363 (2024).
181. Li, X., Egervari, G., Wang, Y., Berger, S. L. & Lu, Z. Regulation of chromatin and gene expression by metabolic enzymes and metabolites. *Nat. Rev. Mol. Cell Biol.* **19**, 563–578 (2018).
182. Garrido-Rodriguez, M., Zirngibl, K., Ivanova, O., Lobentanzer, S. & Saez-Rodriguez, J. Integrating knowledge and omics to decipher mechanisms via large-scale models of signaling networks. *Mol. Syst. Biol.* **18**, e11036 (2022).
183. Armingol, E., Baghdassarian, H. M. & Lewis, N. E. The diversification of methods for studying cell-cell interactions and communication. *Nat. Rev. Genet.* **25**, 381–400 (2024).
184. Burdziak, C. *et al.* scKINETICS: inference of regulatory velocity with single-cell transcriptomics data. *Bioinformatics* **39**, i394–i403 (2023).
185. Jindal, K. *et al.* Single-cell lineage capture across genomic modalities with CellTag-multi reveals fate-specific gene regulatory changes. *Nat. Biotechnol.* **42**, 946–959 (2024).
186. Nolet, C. *et al.* Accelerating single-cell genomic analysis with GPUs. (2022) doi:10.1101/2022.05.26.493607.
187. Essaghir, A. *et al.* Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data. *Nucleic Acids Res.* **38**, e120 (2010).