nature reviews genetics

https://doi.org/10.1038/s41576-023-00618-5

Review article

Check for updates

Gene regulatory network inference in the era of single-cell multi-omics

Pau Badia-i-Mompel[1], Lorna Wessels[1,2], Sophia Müller-Dott[1], Rémi Trimbour[1,3], Ricardo O. Ramirez Flores[1], Ricard Argelaguet[4] & Julio Saez-Rodriguez[1]

Abstract

The interplay between chromatin, transcription factors and genes generates complex regulatory circuits that can be represented as gene regulatory networks (GRNs). The study of GRNs is useful to understand how cellular identity is established, maintained and disrupted in disease. GRNs can be inferred from experimental data — historically, bulk omics data — and/or from the literature. The advent of single-cell multi-omics technologies has led to the development of novel computational methods that leverage genomic, transcriptomic and chromatin accessibility information to infer GRNs at an unprecedented resolution. Here, we review the key principles of inferring GRNs that encompass transcription factor–gene interactions from transcriptomics and chromatin accessibility data. We focus on the comparison and classification of methods that use single-cell multimodal data. We highlight challenges in GRN inference, in particular with respect to benchmarking, and potential further developments using additional data modalities.

Sections

[1]Heidelberg University, Faculty of Medicine, Heidelberg University Hospital, Institute for Computational Biomedicine, Bioquant, Heidelberg, Germany. [2]Department of Vascular Biology and Tumor Angiogenesis, European Center for Angioscience, Medical Faculty, MannHeim Heidelberg University, Mannheim, Germany. [3]Institut Pasteur, Université Paris Cité, CNRS UMR 3738, Machine Learning for Integrative Genomics Group, Paris, France. [4]Altos Labs, Granta Park, Cambridge, UK. e-mail: pub.saez@uni-heidelberg.de

Introduction

Cells regulate gene transcription to coordinate cellular activities in response to intracellular and extracellular signals. Transcription is largely regulated by transcription factors (TFs), proteins that bind to specific sequences of DNA (DNA binding sites) and can have positive or negative effects on the transcriptional rate of target genes[1]. Genomic DNA is tightly packed with structural proteins into complexes known as nucleosomes, which are the basic unit of chromatin, making most genes inaccessible to the transcription machinery. To enable transcription, the region near a gene transcription start site, known as the promoter, needs to be exposed by displacing tightly packed nucleosomes. Changes in DNA accessibility can be triggered by the binding of so-called pioneer TFs[2]. Other TFs can bind to distal *cis*-regulatory elements (CREs) of the DNA and, together with cofactors and other proteins, cooperatively enable the recruitment and stabilization of the RNA polymerase protein complex that synthesizes mRNA from the gene body DNA (Fig. 1a).

Gene regulatory networks (GRNs) are interpretable computational models of the regulation of gene expression in the form of networks, mathematically also defined as graphs. Multiple components of gene regulation, such as TFs, splicing factors, long non-coding RNAs, microRNAs and metabolites, can be incorporated in GRNs. Here, we focus on their simplest representation, which captures only the interplay between TFs and target genes, whereby the nodes of the GRN consist of genes, some of them being TFs, and the edges of the GRN represent regulatory interactions between the genes (Fig. 1b). Other possible GRN representations are discussed elsewhere[3–6]. Uncovering the topology and the dynamics of GRNs is fundamental to understanding how cellular identity is established and maintained[7], which has important implications for engineering cell fate[8] and for disease prevention[9].

Understanding GRNs is a long-standing quest in biology, as illustrated by the seminal work from the 1960s characterizing the bacterial lactose (lac) operon[10]. Reconstructing large-scale GRNs became a major focus of systems biology, leveraging various high-throughput experimental methods and computational algorithms[11–13]. Historically, GRNs have been commonly assembled from experimentally validated regulation events compiled in databases[14–17] or inferred de novo from gene co-expression in bulk transcriptomics data[18–20]. If sufficient transcriptomics data are available, GRNs can be inferred that are better contextualized for the biological question at hand than GRNs extracted from databases, which tend to be generalistic. However, transcriptomics data do not directly capture many underlying regulatory mechanisms, such as the TF protein abundance and DNA binding events, cooperation of TFs and cofactors, alternative transcript splicing, post-translational protein modification events and the accessibility and structure of the genome. The inclusion and measurement of these other aspects of gene regulation has the potential to generate GRNs that better represent gene regulation in vivo (Fig. 1b). For example, the inclusion of chromatin accessibility[21] data allows to fine-tune TF–gene links by considering whether genes are open and by including CREs in the inference of GRNs.

Furthermore, bulk profiling provides mixed measures across cell types in a tissue sample, and thus cannot disentangle regulatory programmes specific to particular cell types or cell states[22,23]. This limitation has been overcome by the use of single-cell technologies[24,25], allowing the inference of GRNs across different cell types, differentiation trajectories and conditions (Fig. 1c). For this reason, and with the introduction of multimodal profiling technologies[26–28], there has been a recent explosion of novel GRN inference methods.

In this Review, we outline general principles of GRN inference and their potential limitations. Furthermore, we describe how multimodal read-outs can be leveraged to infer more accurate GRNs, and we classify and briefly describe several novel tools that have been developed for this task. In addition, we highlight possible downstream GRN analyses and how to assess experimentally the obtained results. Finally, we discuss current challenges and future directions in the field.

Inference of GRNs

GRN inference refers to the process of summarizing gene regulation — a highly complex and dynamic process — into an interpretable network structure from data using computational methods. It is based on the assumption that the effects of a true underlying GRN can be observed and measured in molecular data[29] (Fig. 1b). Interactions in GRNs can be directed or undirected (denoting a causality relationship between genes or lack of it, respectively), signed (denoting the mode of regulation, positive or negative) and/or weighted (denoting the strength of the interaction).

From transcriptomics data

Methods in this category fit models that try to explain the observed variability of gene expression based on the expression of other genes. Weighted gene co-expression network analysis (WGCNA)[19] is one of the simplest and most popular approaches. It carries out pairwise correlations across the whole transcriptome to identify modules of co-expressed genes. The resulting network is commonly known as a gene co-expression network and its interactions are undirected owing to the symmetrical nature of correlations. Although this strategy is useful to identify gene modules in an unsupervised manner, the lack of causal regulatory links hinders its interpretability and typically yields a large number of false positive associations. To address these limitations, methods such as GENIE3 (ref. 20) and its faster implementation GRNBoost2 (ref. 30) first distinguish TFs from target genes based on previously reported regulatory activity[31] and then train models that predict the expression of target genes based solely on the expression of TFs, which markedly reduces the number of interactions to be considered. By doing so, undirected interactions are turned into directed connections and thus introduce putative causal relationships. Nevertheless, inference from transcriptomics data alone introduces false positives as many other mechanisms that are involved in gene regulation, such as chromatin accessibility, are ignored. Moreover, because many processes are required for a mRNA transcript encoding a TF to become a functional protein, transcript levels alone might not be informative enough[1,32]. These limitations may hinder the inference process, as it has been shown that, overall, these methods tend to have moderate success in accurately inferring GRNs[33–35].

From TF binding data or chromatin accessibility

Assays such as chromatin immunoprecipitation followed by sequencing (ChIP-seq)[36] and cleavage under targets and tagmentation (CUT&Tag)[37] enable TF binding to be measured across the genome. This information can be used to build GRNs directly by assigning TF binding sites to putative target genes[38]. However, despite some high-throughput alternatives[39–41], profiling of TF binding is still costly and limited to TFs for which antibodies are available. In addition, the use of TF binding data alone typically requires the assignment of bound TFs to their target genes by closest genomic proximity, ignoring possible distal interaction events that are known to be relevant in gene regulation[1]. By contrast, a pioneering study explored the integration of ChIP-seq

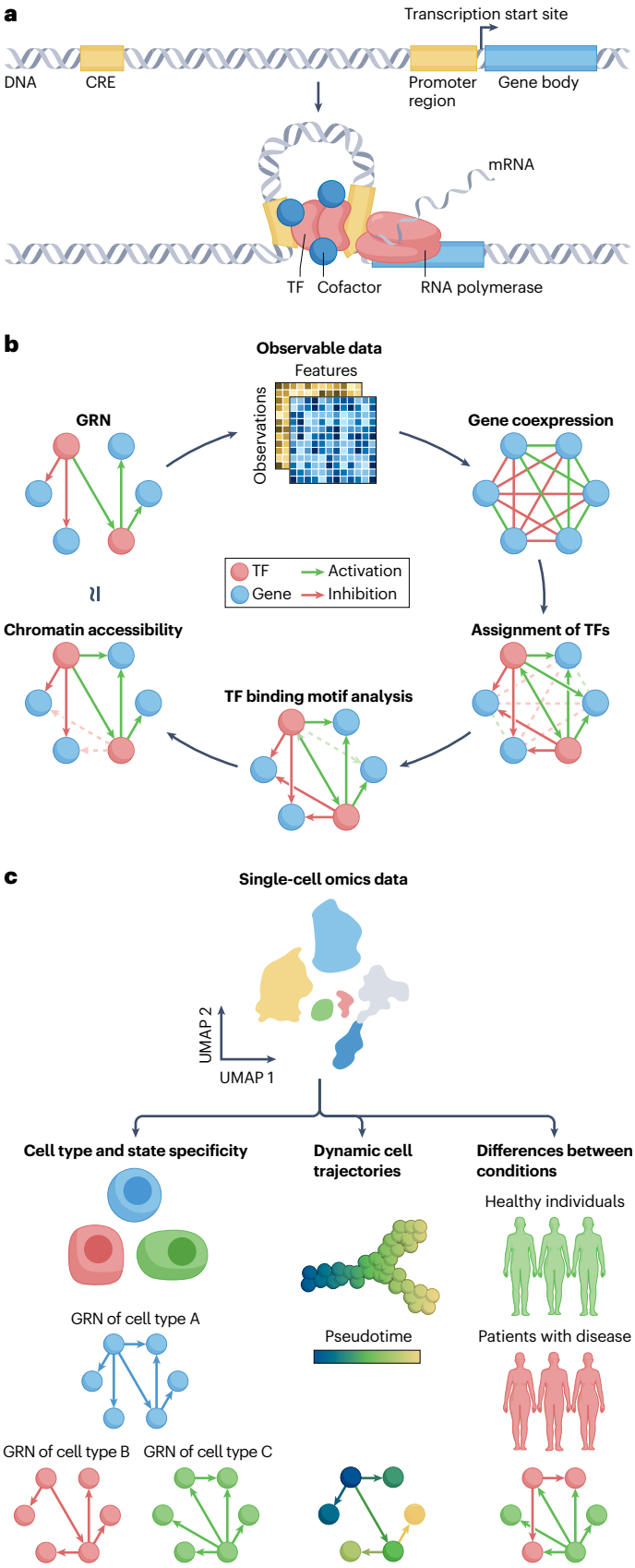

Fig. 1 | Principles of gene regulatory networks. a, Gene regulation and its key elements. Transcription factors (TFs) bind to promoter regions and *cis*-regulatory elements (CREs), displacing nucleosomes and making the transcription start site accessible. Cooperation between TFs, cofactors and other proteins allows for the recruitment and stabilization of the RNA polymerase protein complex, which synthesizes mRNA from the gene body DNA. **b**, Gene regulatory networks (GRNs) can be inferred from measured omics data and, through the modelling of additional information such as TF binding predictions or chromatin accessibility, can be refined to better resemble the true underlying GRN. The nodes of the GRN are TFs and their regulated genes, and the edges between nodes indicate the mode of regulation (activation or inhibition). **c**, GRNs generated from single-cell omics data allow to understand cell type and state specificity, explain the progression of dynamic trajectories and identify differences between conditions.

data and transcriptomics data to enable a more refined assignment of TFs to genes that did not depend on the closest gene[42].

An alternative approach is to use chromatin accessibility data to infer gene regulatory elements that are potentially targeted by TFs. The most commonly used technology owing to its simple and relatively cheap protocol is the assay for transposase accessible chromatin with sequencing (ATAC-seq)[21], but other technologies exist such as DNase-seq[43] and NOME-seq[44] (reviewed elsewhere[45]). Methods that leverage chromatin accessibility data split GRN inference into two steps: first, the assignment of TFs to gene regulatory elements (open chromatin regions, commonly referred to as peaks); and second, the assignment of these regulatory elements to genes (Fig. 2). For the first step, methods leverage TF binding motif databases and motif matcher algorithms to make binding predictions for TFs on accessible CREs (Box 1). For the second step, methods link accessible CREs to genes that are within a certain genomic distance. The distance cutoff is based on the observation that distal CREs such as enhancers or silencers generally interact with the promoter regions of genes at a typical distance[1]. Some examples of such inference methods include ATAC2GRN (ref. 46), LISA[47] and SPIDER[48]. These methods assume that if the promoter region of a gene is accessible, the gene is being transcribed, but that might not always be the case.

From single-cell transcriptomics data

GRN inference methods using bulk omics data have enabled the characterization of genome-wide regulatory events but, in the case of mixed samples such as tissues, they cannot capture the cell type or state specificity of GRNs[22,23]. In addition, GRN inference methods require large sample sizes to generate sufficient data, which can become prohibitively costly in bulk profiling.

With the emergence of single-cell technologies, particularly single-cell RNA sequencing (scRNA-seq), GRN reconstruction methods have been used to infer cell type-specific TF–gene interactions, together with the dynamic changes that occur in these GRNs across development and conditions[49] (Fig. 1c). One of the first GRN inference methods tailored to scRNA-seq data was SCENIC[50], an extension to the GRNBoost2 (ref. 30) method, which generates cell type-specific GRNs by exploiting TF–gene co-expression patterns and, in addition, prunes the edges of the GRN based on TF binding motif enrichment on gene promoter regions. The improved resolution of single-cell measurements also enables the identification of dynamic cell states and their transitions that may not be easily differentiated into distinct groups, such as during development, cell differentiation or disease progression[51,52]. Pseudotime ordering characterizes these continuous

changes and can be used to inform GRN inference. The resulting GRNs provide valuable insights into the complex processes involved in key fate decisions. LEAP[53] and SINCERITIES[54] are examples of GRN inference methods that leverage pseudotime ordering to infer the directionality between genes in the GRNs. The use of contrast-level statistics obtained after differential testing[55,56] is an effective means of identifying differences between conditions, such as between healthy individuals and a cohort of patients with disease. This strategy differs from computing differences between GRNs, as explained in the later section describing 'Downstream GRN analyses'.

Recent advances in single-cell chromatin accessibility profiling (such as single-cell ATAC-seq (scATAC-seq))[57], which can be carried out together with single-cell transcriptomics[26–28], have allowed for the refinement of GRN reconstruction at an unparalleled definition. Some early works inferred GRNs from unpaired multi-omics data to study human myeloid cell differentiation[58], mouse embryonic development[59] and HIV infection of dendritic cells[60]. However, they did not provide their method implemented as a tool for others to use. These were followed by an explosion of novel methods for GRN inference that leverage both scRNA-seq and scATAC-seq (Table 1 and Fig. 2; see Supplementary Box 1). The multimodal data used for GRN inference can be paired if both measurements come from the same cell or unpaired if they come from different cells. Some methods do not require matching chromatin accessibility and gene expression 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. By contrast, other methods model both modalities in the same cell simultaneously. In these 'simultaneous' methods, unpaired data can still be modelled if both modalities are matched using integration approaches[61]. To facilitate usage, some of these methods (for example, DeepMAPS[62], FigR[63], GLUE[64], scAI[65] and SOMatic[66]) implement their own integration approach.

Multimodal GRN inference methods use an extended framework to that used by single-modality methods to reconstruct GRNs. Specifically, they predict gene expression from TF gene expression, they assign TFs to accessible CREs using binding motif information and they associate CREs with target genes constrained by genomic distance (Fig. 2). For the prediction of TF binding events, different methods use different, highly heterogeneous TF binding motif databases and prediction algorithms (Table 1 and Box 1). As TF binding motif databases

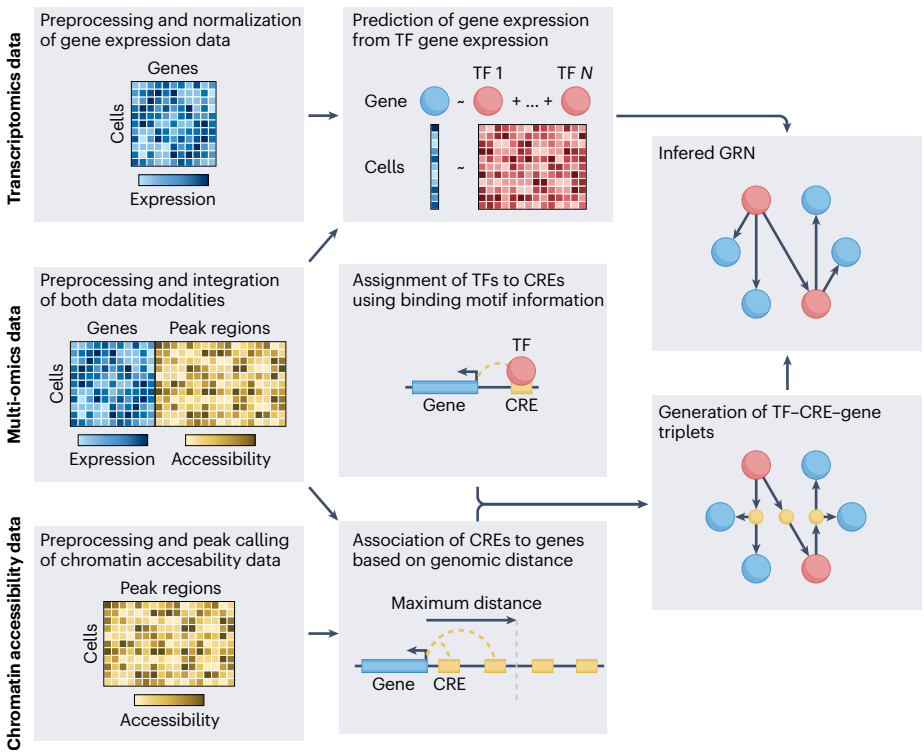

Fig. 2 | Flow chart of methods for gene regulatory network inference. Methods for gene regulatory network (GRN) inference involve different steps depending on the data modalities generated for the samples or cells being studied. Transcriptomics data are first preprocessed and normalized to build a gene expression matrix, containing the transcript levels of each gene across samples or cells. A list of known transcription factor (TF) genes is obtained from other sources to distinguish genes with regulatory capabilities. Interactions between TFs and target genes are then inferred by building models that try to predict the observed gene expression from TF transcript abundance, generating TF–gene associations. Finally, the obtained interactions are aggregated and represented as a GRN. Chromatin accessibility data are first preprocessed and peaks are called to build a peak accessibility matrix, containing binary information about the openness of *cis*-regulatory elements (CREs) across samples or cells. CREs are associated with genes based on genomic distance limits, and TFs are predicted to bind to CREs using TF binding motif databases and motif matcher algorithms. Together, this information is used to obtain TF–CRE–gene triplets. Finally, these interactions are simplified into TF–gene pairs and aggregated into a GRN. When samples are profiled by both transcriptomics and chromatin accessibility (multi-omics data), preprocessing of each modality is carried out and, if needed, the unpaired modalities are integrated. Having both modalities available, methods can simultaneously leverage the three aforementioned modelling steps to build TF–CRE–gene triplets, which then are simplified and aggregated into a GRN.

have different coverage of TFs, and prediction algorithms model binding differently, results between GRN inference methods might differ even if they use similar modelling strategies. The majority of methods allow for using different TF binding motif databases than their default, but most methods fix the motif matcher algorithm used – except for SCENIC+[67], which implements three algorithms, cisTarget[67], DEM[67] and HOMER[68]. In addition, GRN inference methods use different genomic distance cutoffs to assign open chromatin regions to target genes. Some consider close distances up to 10 kb, others medium distances up to 100 kb, others large distal effects up to 1,000 kb and others do not specify the distance cutoff either in the original publication or in the source code (Table 1). Given that functionally validated interactions are greatly enriched at the closest distances, and that they substantially fall off by 100 kb[1,69], differences in distance cutoffs will likely affect the resulting inferred GRN.

After carrying out the above steps (Fig. 2), multimodal GRN inference methods generate a candidate scaffold network made up of triplets of a TF associated with a CRE that is linked to a target gene. To generate a final GRN structure, different mathematical strategies are used. Some of these strategies assume a linear relationship between TFs, CREs and genes, and others assume a non-linear relationship (Table 1). Linear modelling assumes that one variable, for example gene transcripts, changes in direct proportion to another variable, for example TF transcripts or CRE openness. By contrast, non-linear modelling can accommodate more complex interactions between variables such as synergistic effects[70]. Although it is widely acknowledged that gene expression is a non-linear process[70], linear modelling of GRNs is often preferred owing to its simplicity in formulation and interpretation. Independently of the modelling strategy used, the significance of the obtained regulatory interactions can be assessed using either frequentist or Bayesian probability statistical frameworks (Table 1). A frequentist approach defines the probability of an event as the proportion of times that the event occurs in a large number of identical experiments, whereas Bayesian probability defines it as a measure of confidence in the occurrence of the said event based on both observed data and previous information. Bayesian methods can take into account available prior knowledge but they usually require larger computational resources than frequentist approaches, which can be a limitation when inferring genome-wide GRNs with large-scale single-cell data. In addition, the success of Bayesian inference depends on the quality of the prior knowledge used. Therefore, when no prior information is available or it is suspected to be inaccurate, frequentist inference might be more accurate.

Multimodal GRN inference methods can be grouped based on the combination of their modelling strategy and the types of input they accept (Table 1). The majority of methods are designed to model GRNs across distinct groups, usually cell types, by frequentist regression. FigR[63] and GRaNIE[71], among others, use frequentist linear regression; DIRECT-NET[72] and SCENIC+[67] use frequentist non-linear regression (random forest); and PECA[73] and Symphony[74,75] use Bayesian modelling. By contrast, CellOracle[76], Inferelator 3.0 (ref. 77) and Pando[78] offer multiple modelling strategies to the user. In case no distinct groups can be defined from the data owing to its continuous nature, for example in cell development, scMEGA[79] and IReNA[80] leverage trajectories to infer GRNs linearly and non-linearly, respectively. Also, Dictys[81], scMTNI[82] and TimeReg[83] use a combination of both cell type grouping and trajectory data to inform the GRN modelling, whereas CellOracle[76] and SCENIC+[67] use the latter to carry out downstream analyses. ANANSE[84], sc-compReg[85] and SCENIC+[67] build

Box 1

Binding motif databases and motif matcher algorithms

Generating genome-wide binding data for multiple transcription factors (TFs) requires laborious experiments, so methods for gene regulatory network (GRN) inference instead predict TF binding events on open genomic regions based on prior information. This information comes from a large collection of TF–DNA binding assays, such as chromatin immunoprecipitation followed by sequencing (ChIP-seq) experiments[36], that can be used to extract the most likely genomic sequence to which a given TF specifically binds, commonly known as a TF binding motif[208]. Several databases have collected such assays and generated TF binding motif collections for model organisms. Because coverage between databases may vary, they can be merged to increase the number of TFs considered in the GRN inference process. Moreover, several computational algorithms have been developed that leverage TF binding motifs to predict binding events, known as motif matcher algorithms. All of these algorithms are based on computing the probability of a TF binding event from the motif sequence and filtering the significant ones. Because the different methods model TF binding differently, results may vary between them and should be considered during GRN inference. The table lists the TF binding motif databases and motif matcher algorithms used across the reviewed methods.

Name	URL	Refs.
Binding motif databases		
CIS-BP	http://cisbp.ccbr.utoronto.ca/	209
cisTarget databases	https://resources.aertslab.org/cistarget/databases/	67
ENCODE	https://www.encodeproject.org/software/encode-motifs/	210
HOCOMOCO	https://hocomoco11.autosome.org/	211
JASPAR	https://jaspar.genereg.net/	212
TRANSFAC	https://genexplain.com/transfac/	213
UniPROBE	http://thebrain.bwh.harvard.edu/uniprobe/	214
Motif matcher algorithms		
FIMO	https://snystrom.github.io/memes-manual/	215
GimmeMotifs	https://gimmemotifs.readthedocs.io/	216
HOMER	http://homer.ucsd.edu/homer/motif/	68
MOODs (as implemented in motifmatchr)	https://github.com/jhkorhonen/MOODS https://github.com/GreenleafLab/motifmatchr	217, 218
motifanalysis (as implemented in reg-hint)	https://reg-gen.readthedocs.io/	219
PIQ toolkit	https://bitbucket.org/thashim/piq-single/src/master/	220
PWMScan	https://ccg.epfl.ch/pwmtools/pwmscan.php	221
pycisTarget	https://pycistarget.readthedocs.io/	67

Table 1 | Existing tools for inference of gene regulatory networks from multi-omics data

Tool[a]	Possible inputs	Type of multimodal data	Type of modelling	Type of interactions	Statistical framework	Default motif database/motif matcher	Default upstream/ downstream distance cutoffs	Language	Refs.
ANANSE	Groups, contrasts	Unpaired	Linear	Weighted	Frequentist	CIS-BP/ GimmeMotifs	100 kb/100 kb	Python	84
CellOracle	Groups, trajectories	Unpaired	Linear	Signed, weighted	Frequentist or Bayesian	CIS-BP/ GimmeMotifs	500 kb/500 kb	Python	76
DC3	Groups	Unpaired	Linear	Binary	Frequentist	Undefined/ HOMER	Based on Hi-C	Python	88
DeepMAPS	Groups	Paired or integrated	Linear	Weighted	Frequentist	JASPAR/ PWMScan	150 kb/150 kb or exon	Python	62
Dictys	Groups, trajectories	Unpaired/paired or integrated	Linear	Signed, weighted	Frequentist	HOCOMOCO/ HOMER	500 kb/500 kb	Python	81
DIRECT-NET	Groups	Paired or integrated	Non-linear	Binary	Frequentist	JASPAR/MOODs	250 kb/250 kb	R	72
FigR	Groups	Paired or integrated	Linear	Signed, weighted	Frequentist	CIS-BP/MOODs	50 kb/50 kb	R	63
GLUE	Groups	Paired or integrated	Non-linear	Weighted	Frequentist	JASPAR/cisTarget	150 kb/150 kb	Python	64
GRaNIE	Groups	Paired or integrated	Linear	Weighted	Frequentist	JASPAR, HOCOMOCO/ PWMscan	250 kb/250 kb	R	71
Inferelator 2.5	Groups	Unpaired	Linear	Signed, weighted	Frequentist or Bayesian	CIS-BP, ENCODE, TRANSFAC/FIMO	10 kb/10 kb	Python	203
Inferelator 3.0	Groups	Unpaired	Linear or non-linear	Weighted	Frequentist or Bayesian	JASPAR/FIMO	50 kb/2.5 kb	Python	77
IReNA	Trajectories	Unpaired	Linear	Signed, weighted	Frequentist	TRANSFAC/FIMO	250 kb/250 kb	R	80
MAGICAL	Groups, contrasts	Unpaired	Non-linear	Weighted	Bayesian	CIS-BP, ENCODE/ MOODs	Based on Hi-C	R/MATLAB	166
MICA	Groups	Unpaired	Non-linear	Signed, weighted	Frequentist	HOCOMOCO, JASPAR/ motifanalysis	Nearest transcription start site	R	204
Pando	Groups	Paired or integrated	Linear or non-linear	Signed, weighted	Frequentist or Bayesian	JASPAR, CIS-BP/ MOODs	100 kb/gene body	R	78
PECA	Groups	Paired or integrated	Linear	Weighted	Bayesian	JASPAR, TRANSFAC, UniPROBE/ HOMER	1,000 kb/1,000 kb	MATLAB	73
Regulatory Motifs	Groups	Paired or integrated	Linear	Signed	Frequentist	HOCOMOCO/ motifanalysis	5 kb/5 kb	MATLAB	205
RENIN	Groups	Paired or integrated	Linear	Signed, weighted	Frequentist	CIS-BP/MOODs	500 kb/500 kb	R	206
scAI	Groups	Paired or integrated	Linear	Weighted	Frequentist	CIS-BP/MOODs	250 kb/250 kb	R	65
sc-compReg	Groups, contrasts	Unpaired	Linear	Binary	Frequentist	Undefined/ undefined	Undefined	R	85
SCENIC+	Groups, contrasts, trajectories	Paired or integrated	Linear	Signed, weighted	Frequentist	cisTarget/ cisTarget	150 kb/150 kb	Python	67
scMEGA	Trajectories	Paired or integrated	Linear	Weighted	Frequentist	JASPAR/MOODs	250 kb/250 kb	R	79
scMTNI	Groups, trajectories	Unpaired	Linear or non-linear	Weighted	Bayesian	CIS-BP/PIQ	5 kb/5 kb	C++	82
SOMatic	Groups	Unpaired	Linear	Binary	Frequentist	HOCOMOCO FIMO	50 kb/50 kb	C++	66

Table 1 (continued) | Existing tools for inference of gene regulatory networks from multi-omics data

Tool[a]	Possible inputs	Type of multimodal data	Type of modelling	Type of interactions	Statistical framework	Default motif database/motif matcher	Default upstream/ downstream distance cutoffs	Language	Refs.
Symphony	Groups	Unpaired	Linear	Signed, weighted	Bayesian	Undefined/FIMO	Nearest transcription start site	Python	74,75
TimeReg	Groups, trajectories	Paired or integrated	Linear	Binary	Frequentist	Undefined/ HOMER	Undefined	MATLAB	83
TRIPOD	Groups	Paired or integrated	Non-linear	Signed, weighted	Frequentist or Bayesian	HOCOMOCO, JASPAR/MOODs	100 kb/100 kb	R	207

[a]Further detail regarding these tools and their methodologies is provided in Supplementary Box 1.

group-specific (for example, cell type-specific) GRNs but can also leverage gene contrast statistics during the inference process.

Downstream GRN analyses

Once GRNs have been inferred from any resolution and combination of omics data, they can be queried using various downstream analyses to provide novel biological insights (Fig. 3 and Box 2).

Topological analysis

Although GRNs are simple and interpretable models of gene regulation, they can still contain large numbers of genes and an even larger number of interactions between them. Network centrality measures can help identify which TFs or genes are more important for the connectivity or the information flow of the network (Fig. 3a). Some examples of network centrality measures include degree centrality, closeness centrality, betweenness centrality and eigenvector centrality. These measures have been useful to identify TFs that drive cell fate changes in diverse biological contexts, such as direct lineage reprograming[76], human myocardial infarction[86] and mouse development[87].

Another approach to characterize the topology of GRNs is using methods based on spectral graph theory, which explore the properties of a network when represented as a matrix. For example, non-negative matrix factorization applied to the adjacency matrices of GRNs has identified groups of TFs that cooperatively drive lineage transitions in mouse embryonic stem cells[83,88]. Similarly, clustering of GRN topology identified known regulators in human haematopoietic cell differentiation[70] and in the response of macrophages to interferon-γ[71]. The gene regulatory modules that are obtained can then be enriched for gene sets to characterize their potential biological functions[89].

Comparative analysis

Comparative analysis of GRNs can uncover the rewiring events that drive differences between cell types, cell states, disease states, treatment approaches and organisms (Fig. 3b). The easiest method for comparative analysis involves the pairwise subtraction of TF–gene interactions between GRNs. This methodology has identified key regulators in subpopulations of B cells in patients with lymphocytic leukaemia[85], groups of TFs for transdifferentiation of fibroblasts to different human cell types[84], candidate Alzheimer disease-specific *trans*-regulators[90] and cell state-specific regulators in human T cells[74,75]. It has also been used to assess evolutionary conservation of TF–gene interactions and adaptation of transcriptional regulation across species[91]. However, owing to the sparse and noisy nature of GRNs, direct comparison of TF–gene interactions is often not sufficiently robust. Topic modelling strategies such as latent Dirichlet allocation[92], an unsupervised Bayesian model that was originally developed for natural language processing, allow for generating dense, low-dimensional representations that filter noise in the structure of the GRN, and thus capture more robustly the differences in regulatory relationships. This strategy has been useful for predicting the survival of patients with cancer[93] and for identifying rewiring events in human haematopoiesis[82].

Inference of TF activities

GRNs can be coupled with enrichment methods to infer TF activities from transcriptomics data[15,50,94,95]. This approach allows for the observed gene expression to be integrated with the GRN topology to extract which TFs might have relevant roles in certain contexts (Fig. 3c). Common enrichment methods include GSEA[96], AUCell[50] and VIPER[94], among others[95]. In bulk studies, inference of TF activities through enrichment methods has enabled, for example, the identification of druggable oncoproteins[94], stratification of cell lines in response to drug treatment[97] and identification of a master regulator as a metastasis promoter in breast carcinoma[98]. In single-cell studies, enrichment methods have identified a novel mechanism of immunotherapy resistance in human T cells[99], regulators and inducers of oligodendroglioma[50], and potential druggable targets for pathological fibroblasts in patients with COVID-19 (ref. 100). These methods have also been recently applied to spatially resolved transcriptomics data, for example to suggest regulators involved in the functional transition of cardiomyocytes across the border zone that surrounds ischaemic lesions in human myocardial infarction[86].

Perturbation and prediction of cell fate

GRNs can be used to simulate gene expression values over time by propagating TF expression to target genes in an iterative manner. With this framework, in silico perturbations can be carried out by changing the expression of a candidate TF and then observing how it affects the resulting transcriptome after a given number of iterations (Fig. 3d). Afterwards, the simulated values can be compared with the gene expression of local neighbouring cells to estimate cell identity transition probabilities analogous to RNA velocity analysis[101]. First introduced by CellOracle[76], this strategy suggested the role of *Zfp57* in generating and maintaining mouse induced endoderm progenitors, which was later experimentally validated with in vitro perturbation experiments. SCENIC+[67] used a similar strategy to identify *RUNX3* as a potential driver of melanocytes to mesenchymal melanoma cells, showcasing the ability of GRNs to capture and model complex regulatory events.

Experimental assessment of GRNs

The connections predicted by GRN inference methods should be seen as hypothetical regulatory interactions that must be assessed by complementary information and/or experiments. In this section, we discuss common practices for this purpose (Fig. 4).

TF abundance and post-translational modification

The number of transcripts encoding a given TF is a limited proxy of its protein abundance, let alone its activity[102]. To this end, proteomics technologies can be used to measure the abundance of TFs. Targeted proteomics at single-cell resolution is still challenging but some

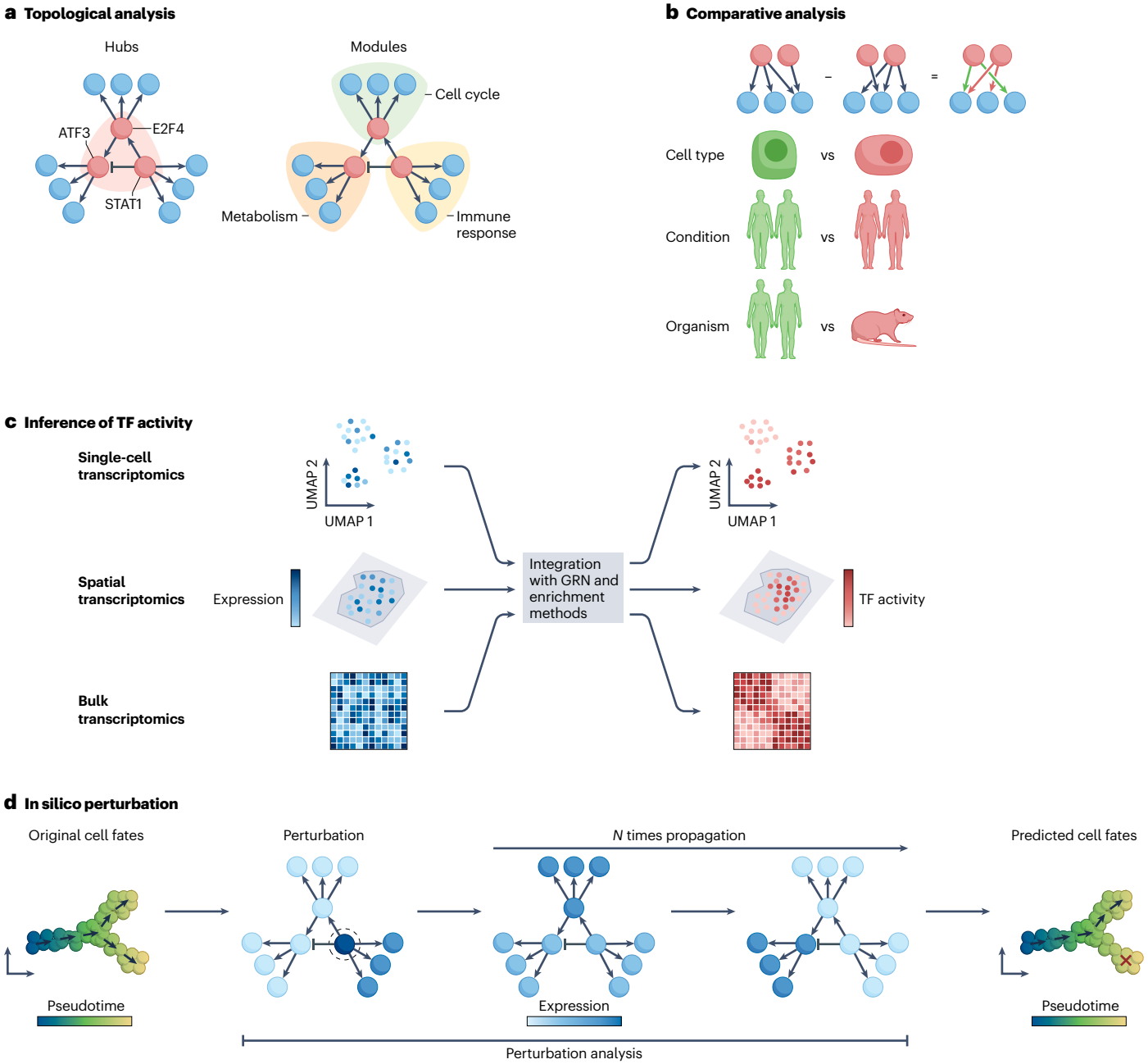

Fig. 3 | Applications of gene regulatory networks. a, Topological analysis. Network centrality measures can be used to identify hubs of transcription factors (TFs) or genes within a gene regulatory network (GRN) that are highly connected. Clustering of nodes based on their connectivity gives rise to sub-network modules that can be associated with biological functions. **b**, Comparative analysis. Comparison of the connectivities in different GRNs by the pairwise subtraction of TF–gene interactions between GRNs can provide insight into the rewiring of gene regulation between different cell types, individuals, conditions or organisms. **c**, Inference of TF activity. GRNs can be coupled to enrichment methods to infer which TFs might be functionally active from transcriptomics data. GRNs inferred from multi-omics data can then be used to infer TF activities in other contexts, such as independent single-cell, spatial or bulk transcriptomics data. **d**, In silico perturbation experiments. GRNs can be used to simulate perturbation experiments by propagating changes in gene expression through the network over short iterations. The obtained simulated gene expression profiles can then be used to infer cell fate decisions.

Box 2

Successful applications of gene regulatory networks derived from single-cell multi-omics data

Gene regulatory networks (GRNs) have been used in various contexts to answer a range of research questions; here, we summarize some recent examples.

- A multimodal time course of human brain organoids was generated using single-cell RNA sequencing (scRNA-seq) and single-cell assay for transposase accessible chromatin with sequencing (scATAC-seq)[78]. Using the GRN inference tool Pando, the authors predicted transcription factor (TF) binding sites and inferred a global GRN underlying organoid development. By making in silico predictions from the GRN followed by a CRISPR-based screen, they identified GLI3 as an essential TF for cortical fate establishment.
- A multimodal atlas of the fly brain was generated using scRNA-seq and scATAC-seq, and characterized developmental, reprogramming and maturation trajectories[222]. Using a deep learning model trained on the omics data, the authors inferred cell type-specific TF binding predictions and used this to decode the regulatory grammar of enhancer architectures that underlie neuronal diversity.
- CellOracle was introduced as a mathematical model to carry out in silico TF perturbations from GRNs trained using single-cell multi-omics data[76]. In the context of zebrafish development, the authors made systematic predictions of TF knockouts, which allowed the identification of new roles for key regulators of early zebrafish development, including *noto* and *lhx1a*.
- A multimodal atlas of mouse early organogenesis was built by profiling gene expression and chromatin accessibility from individual cells[87]. The authors developed in silico chromatin immunoprecipitation followed by sequencing (ChIP-seq), a method to predict TF binding sites, and used it to characterize the GRNs underlying the transition of neuromesodermal progenitors to somitic mesoderm. Using the CellOracle[76] framework for in silico predictions, followed by experimental validation, they characterized a role of Brachyury in priming *cis*-regulatory elements (CREs) for differentiation.
- scRNA-seq and scATAC-seq were carried out on cortical tissue from patients with Alzheimer disease[90]. By modelling relationships between TFs, CREs and target genes, the authors identified ZEB1 and MAFB as candidates involved in gene regulation and, potentially, disease progression in neurons and microglia.
- Inferelator 3.0 (ref. 77) was used to infer GRNs for several $CD4^+$ memory T cell populations from mice and the results were benchmarked by curating TF knockout and ChIP-seq data for 42 of the identified TFs[223]. The authors integrated the obtained GRNs with cell–cell communication networks, and functionally validated a regulatory circuit involving IL-6, MAF and CD153 in T follicular helper cells that is important for antibody-mediated vaccine responses in aged mice.

technologies such as mass spectrometry-based approaches or assays that use antibody–oligonucleotide conjugates are already available[103] (reviewed elsewhere[104]). Alternatively, databases such as The Human Protein Atlas[105,106] can be queried to confirm whether a candidate TF has been previously reported to be present at the protein level in a given tissue or cell type. In addition, post-translational modifications such as phosphorylation, ubiquitylation and methylation can affect TF localization, stability, activity and interaction with other proteins[107]. The most highly studied post-translational modification of TFs is phosphorylation, which can be informative as to whether a TF is in an inactive or active form[108].

TF binding and cooperativity

GRN inference methods rely on TF binding predictions based on binding motif analysis to assign TFs to open chromatin regions in the genome. It is known that this type of prediction produces many false positives as a large number of TF binding motifs have low specificity[109]. To this end, ChIP-seq[36], which as mentioned above measures the binding of TFs to DNA, can be used to test how many TF binding events were correctly predicted by the GRN inference method[76]. Databases such as ChIP-Atlas[110], EpiMap[111] and UniBind[112] compile large collections of ChIP-seq experimental data and organize them by organism, tissue and cell type, making them valuable resources for analysing GRN predictions. Because not all ChIP-seq peaks represent direct binding events of a TF, EpiMap[111] and UniBind[112] implement different strategies to curate the data, thus providing more reliable information regarding TF binding sites. Another alternative is single-molecule footprinting, a technique that jointly measures TF binding and nucleosome occupancy at single DNA molecule resolution[113,114]. It allows for checking the state frequency of each genomic region: bound by a TF, unbound but with open chromatin, or unbound and occupied by nucleosomes. The advantage of single-molecule footprinting over ChIP-seq is that it provides a dynamic and quantifiable state of TF binding instead of a binary description. GRN inference methods predict that several TFs bind to the same open genomic region, which is in keeping with the knowledge that TFs bind cooperatively to DNA to induce transcription[1,115,116]. Another approach to assess the obtained GRN is to check whether the network has recovered the cooperative binding of TFs. Technologies such as CAP-SELEX[117] enable cooperative interactions between selected TF pairs to be jointly profiled in the presence of DNA. Single-molecule footprinting can also be used for this purpose by checking the overlap of footprints. Other approaches include using protein–protein interaction assays[118] or checking databases of previously annotated interactions[119].

Regulatory activity of CREs

CREs can be in three different chromatin states: transcriptionally active, poised or repressive. Many of the reported open chromatin regions might not have a role in gene regulation, thereby increasing the number

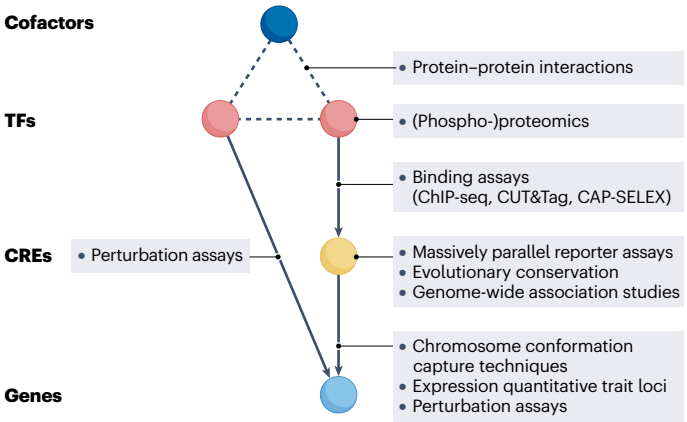

Fig. 4 | Experimental assessment of gene regulatory networks. Although there is no clear 'ground truth' for gene regulation, several experiments and analyses can be carried out to validate specific aspects of gene regulatory networks (GRNs). Interactions between transcription factors (TFs) can be queried in protein–protein databases for TFs that share large numbers of target genes and are assumed to interact. The presence of TF protein can be confirmed by proteomic assays, and the activation state of TFs can be assessed in targeted phosphorylation experiments. Links between TFs and *cis*-regulatory elements (CREs) can be confirmed by binding assays, such as chromatin immunoprecipitation followed by sequencing (ChIP-seq), cleavage under targets and tagmentation (CUT&Tag) and CAP-SELEX. Candidate CREs can be tested for regulatory capability with reporter assays or perturbation experiments. Alternatively, functional CREs can be assumed to be evolutionarily conserved or to be enriched in disease-associated loci as identified by genome-wide association studies. Links between CREs and genes can be evaluated experimentally by genome conformation assays such as Hi-C or super-resolution microscopy, or by CRISPR-based perturbation experiments. Alternatively, databases of expression quantitative trait loci may be used.

of false positives in GRN predictions. Therefore, experiments must assess whether candidate enhancers affect gene expression. The massively parallel reporter assay[120] is a technique that allows to test whether candidate genomic regions can induce gene expression in episomal vectors. Another strategy is to carry out pooled CRISPR-based perturbations of candidate CREs followed by RNA sequencing to identify which regions of the genome affect gene expression[69,121]. In addition, the ENCODE Consortium has catalogued, using various biochemical assays, more than one million candidate CREs with enhancer-like signatures that span ~16% of the human genome[36,122,123]. As functional enhancers are evolutionarily conserved[124,125], another approach is to check for genomic sequence similarity across different species. When studying diseases, CREs in open chromatin regions can be scanned for single-nucleotide polymorphisms (SNPs) that have been previously linked to diseases in genome-wide association studies. If candidate CREs are enriched for disease-associated SNPs, this suggests that those CREs are likely to be functional[90,126].

Linkage to target genes

Even if an open chromatin region is validated to have regulatory properties, it is necessary to test which specific genes it affects. Chromosome conformation capture techniques such as Hi-C[127–129] allow to measure the probability of contact between genomic regions and to identify topologically associating domains. Because gene regulation is based on sustained genomic interactions[130,131], if a genomic region is consistently in close contact with the promoter region of a gene, it may regulate the gene. To this end, super-resolution microscopy has also been used to validate candidate CRE–gene interactions[132], although with less throughput than Hi-C. Perturbation assays such as CRISPR-based screens coupled with whole-transcriptome analysis allow for deleting or activating specific CREs and observing how this changes gene expression[69,133]. In addition, databases of expression quantitative trait loci in human populations can be used to validate distal candidate CREs[134–136].

TF perturbation

A more direct approach to test whether a TF regulates a particular gene is to perturb TF expression and see how this affects gene expression. Technologies for pooled CRISPR (in)activation screens coupled with whole-transcriptome read out are already available[137–141]. The changes in gene expression upon TF perturbation can be used as the ground truth to check how many of the affected genes are identified as target genes in the GRN[35,95,142]. Also, one can see whether the estimated TF activities (see previous section on 'Inference of TF activities') correspond to the perturbation that has been carried out (low or high TF activity for knockout or overexpression of the TF, respectively).

Challenges and future directions

The accumulation of omics data sets, particularly single-cell multi-omics data, in recent years has enabled a new wave of improved GRN inference strategies. An improved understanding of GRNs should pave the way to use these models not only as means to understand the principles of gene regulation but also as tools to drive cell fate decisions for cellular engineering, enabling the generation of new cell types with new functions and the reprogramming of diseased cells to a healthy phenotype. The prospects are hugely promising, but many challenges remain regarding the modelling of GRNs and their use as predictive tools.

Integrating transcription and accessibility

The use of multi-omics, in principle, allows for a better representation of gene regulation but also comes with its own challenges. As highly interrelated processes, chromatin accessibility and transcription are temporally coordinated. Yet they have profoundly different kinetics and may be temporally shifted. These relationships are often not fully understood and it is typically assumed that paired chromatin accessibility and transcriptomics data in the same cell at a single time point are representative of the interplay between both processes[143,144]. This limitation is compounded in the case of unpaired data if inadequate integration results in mismatched scATAC-seq and scRNA-seq data that will mislead the downstream modelling of GRNs[145]. Novel integration strategies, such as that introduced by FigR[63], hold the promise to obtain better matching between cells. Among other factors, the temporal shift between chromatin accessibility and gene expression, as well as cooperative effects, gives rise to non-linear relationships. Some of the GRN inference methods that we have discussed use non-linear formulations to account for this, but they lose interpretability compared with linear models, and they often do not explicitly capture the sign of the interaction. For this reason, and for computational scalability, many methods still prefer to model gene regulation linearly. To improve the interpretability, SCENIC+[67] and IReNA[80] first infer regulatory interactions non-linearly using random forests, and then determine the sign of the interaction based on correlation analysis between TF and gene transcripts.

Glossary

Assay for transposase-accessible chromatin with sequencing
(ATAC-seq). A technique to identify accessible DNA regions using hyperactive Tn5 transposase.

Betweenness centrality
A network centrality measure representing the number of appearances of a node in the shortest path of any other two nodes in the network.

Chromatin
A higher-order filamentous structure of DNA–protein complex that can exist in a condensed or uncondensed state.

Chromatin immunoprecipitation followed by sequencing
(ChIP-seq). A technique to analyse protein interactions with accessible DNA regions using chromatin immunoprecipitation followed by DNA sequencing.

***cis*-Regulatory elements**
(CREs). Non-coding DNA regions that regulate the transcription of nearby genes upon binding of transcription factors (TFs). These include promoters, enhancers and silencers.

Cleavage under targets and tagmentation
(CUT&Tag). An antibody-based technique to analyse protein interactions with accessible DNA regions using transposase Tn5-mediated tagmentation followed by DNA sequencing.

Closeness centrality
A network centrality measure describing the average distance (length of the shortest path) of a node to all other nodes.

Degree centrality
A network centrality measure describing the number of edges (degree) of a node.

DNA binding sites
DNA sequences where transcription factors can bind to drive gene regulation, usually represented as nucleotide patterns known as motifs.

Eigenvector centrality
A network centrality measure describing the importance of a node in the network based on the centrality of its neighbours.

Enhancers
Distal regulatory DNA regions where transcription regulatory proteins can bind and activate transcription.

Expression quantitative trait loci
Genomic locations whose sequence variation is associated with changes in gene expression.

Gene regulatory networks
(GRNs). Network representations of molecular interactions between transcriptional regulators and target genes.

Genome-wide association studies
Analysis approach to identify frequently appearing single-nucleotide polymorphisms in the genome across a large cohort of individuals.

Hi-C
A technique to study chromatin conformation in three dimensions to identify genomic sequences that might be distal to each other in linear distance but closer in the 3D space.

Metacells
Groups of cells with a similar molecular profile that can be aggregated into a single omics profile to reduce sparsity of the data.

Motif matcher algorithms
String matching algorithms to detect transcription factor binding sites in DNA sequences.

Network centrality
A group of graph theory metrics that defines the relative importance of a node in a network.

Peaks
Regions of accessible chromatin that form the read-out of epigenetic sequencing techniques.

Promoter
A regulatory region in the genome located before the transcriptional start site of a gene.

Silencers
Distal regulatory DNA regions where transcription regulatory proteins can bind and repress transcription.

Single-nucleotide polymorphisms
(SNPs). DNA sequence variations caused by substitution of a single nucleotide in a specific position.

Topologically associating domains
Self-interacting genomic regions with high interaction frequency of sequences within the domain and relative isolation from neighbouring regions, forming a 3D chromosome structure.

Transcription factors
(TFs). Proteins that modify the rate of transcription by binding to specific DNA sequences.

Scale and sparsity of single-cell data

GRN inference methods require a large number of observations that capture the variability of the biological process being studied. These observations can be individual cells, samples or conditions. Single-cell technologies generate thousands of profiles for a given sample, making it easier to infer GRNs in a larger variety of biological contexts than for bulk profiling technologies. Nonetheless, cells from the same sample are not necessarily independent and cannot be considered true biological replicates[146]. For this reason, the inclusion of different samples might be needed to obtain meaningful GRNs. In addition, current single-cell GRN inference methods build an aggregate network across a population of cells and do not take into account that cells may come from different samples. A candidate approach to address this issue is LIONESS[147], which models the contribution of each sample when inferring GRNs and can generate sample-specific regulatory interactions. In addition, single-cell data are by nature sparse and noisy, particularly for data obtained by scATAC-seq, and proper filters need to be used to ensure a minimum quality[148,149]. For paired multi-omics technologies, a systematic benchmark comparing their varying coverage and sensitivity to their single-omic counterparts is missing[150]. Although sparsity is a known property of single-cell technologies[151,152], none of the GRN inference methods discussed here explicitly accounts for sparsity in its modelling. Some methods apply data transformations to counteract this limitation. Imputation methods can be used to reduce the number of 'dropouts' (caused by under-sampling of mRNAs or accessible DNA reads)[153–155], although it has been shown that they might have detrimental effects on GRN reconstruction[156]. Strategies that aggregate similar cells into pseudo-bulk profiles or metacells[146,157,158] have been reported to be beneficial[87]. Owing to their sparsity, most computational pipelines treat scATAC-seq data as binary data, assigning regions of the genome that are either accessible or closed for each cell. However, the true state of DNA accessibility is known to be more refined and can

involve regions of intermediate accessibility that fluctuate in a dynamic manner[159]. Thus, treating chromatin accessibility data as binary might be detrimental for downstream analyses[160,161], and methods that handle accessibility in a quantitative manner might improve GRN reconstruction[154,155,162,163].

The regulatory role of 3D genome structure

Current methods of GRN inference use arbitrary cutoffs based on genomic distance to assign CREs to genes. The aim of this filtering is to reduce the search space for each gene, requiring less computational resources, and to reduce the number of false positive interactions based on the fact that most genomic interactions are proximal[69]. However, there are some examples of interactions between CREs and genes separated by large distances, such as enhancers of the *MYC* gene located almost 2 Mb downstream of it[164]. Depending on the distance cutoff that is used, GRN inference methods might miss crucial CRE–gene interactions. In addition, some interactions occur across chromosomes, as reported during olfactory receptor selection[165], which current GRN methods are not able to consider. One solution to this problem is to use technologies based on 3D proximity, such as Hi-C[127,129], to assess whether a CRE might be regulating a gene. This strategy has been successfully applied by DC3 (ref. 88) and MAGICAL[166]. Despite some high-throughput alternatives[167,168], chromosome conformation capture techniques pose new challenges owing to their sparse nature[169] and the facts that they still require integration with other modalities and their protocol can be hard to reproduce[170,171]. Until they become more widely available, computational approaches have been used to predict the 3D structure of the genome based on accessibility data such as scATAC-seq data[172,173]. Their use in GRN modelling has the potential to overcome the limitations of using distance-based cutoffs.

Refinement of TF binding predictions

The current strategy used by GRN inference methods of assigning TFs to CREs relies on TF binding motif databases (Box 1). Each database has a different coverage of motif collections, which might bias the resulting predictions. Motif databases are based on data from previous binding experiments such as ChIP-seq. However, it is estimated that there are no available binding data for 10% of the approximately 1,600 sequence-specific TFs encoded in the human genome[31,109]. TFs without known binding motifs are excluded from GRN modelling, a factor that is exacerbated for non-model organisms as they tend to have fewer known TF binding motifs than other more well-studied organisms. One possible solution to incorporate missing TFs might be to leverage known protein–protein interactions during GRN inference. Moreover, current TF binding motifs are based on data from multiple tissues and cell types. It is known that TF binding is a highly context-specific process[1], and although the available motifs are still relevant for many tissues, cell type-specific motif models might help to increase the accuracy of TF binding predictions. Recent computational strategies based on deep learning allow for cell type-specific TF binding predictions[174,175]. These models are trained to predict cell type-specific DNA accessibility solely based on DNA sequence. Once trained, they identify which nucleotides are predicted to affect accessibility the most through in silico mutagenesis or by using strategies of interpretable machine learning such as SHAP[176]. To derive cell type-specific TF binding predictions, these methods combine the predicted nucleotide quantifications with binding motifs. Although these strategies have the potential to better contextualize GRN inference, they require pretrained models using large amounts of data and are still limited to known TF binding motifs. With the accumulation of high-quality cell atlases from consortium initiatives[49,177,178], we envision that these strategies could eventually replace classic TF binding motif predictions. Furthermore, current TF binding predictions are binary but a quantitative definition could be more informative. BANC-seq[179], a technology that measures quantitative TF binding affinities, has the potential to generate more accurate GRNs.

Emerging multi-omics for GRN inference

The paired profiling of transcriptomics and chromatin accessibility data has enabled the potentially more accurate inference of GRNs but is still a costly assay, limiting its widespread use. Newer alternatives such as ISSAAC-seq[180] enable multi-omics profiling at a much lower cost than the commercial 10× Multiome kit. Despite this, it might be the case that joint scRNA-seq and scATAC-seq data alone do not provide enough information to characterize gene regulation fully. In that case, advances in single-cell multi-omics profiling technologies that include more data modalities will be crucial[181]. Among such promising technologies is NEAT-seq[182], which simultaneously profiles intra-nuclear proteins, chromatin accessibility and gene expression, allowing to discard possible false positives in GRN modelling by including TF protein abundance. Another example is scChaRM-seq[183], which simultaneously profiles DNA methylation, chromatin accessibility and gene expression. Their joint profiling allows for TF assignment to CREs to be fine-tuned according to their methylation status. Moreover, ATAC-STARR-seq[184] can carry out massively parallel reporter assay and chromatin accessibility profiling simultaneously to test the transcribing capacity of open CREs. Advances in untargeted single-cell proteomics and phosphoproteomics may enable the profiling of functionally active TFs[185]. One example is Phospho-seq[186], a novel technology that profiles chromatin accessibility and phosphorylated proteins at the single-cell level. Genetic information is known to be heterogeneous among populations of individuals but most methods assume that they share the same genome[187]. scGET-seq[188], a technology that jointly profiles the genome and chromatin accessibility, has the potential to aid the inference of causal GRNs by testing how SNPs may affect chromatin accessibility owing to changes in TF binding affinities.

Benchmarking of GRNs

The benchmarking of GRNs is crucial to understand the accuracy of novel GRN inference methods, in particular those that leverage multi-omics data, which have not yet been evaluated systematically. Unfortunately, the validation of predicted GRNs is a complicated task as there is no clear 'ground truth' for gene regulation. One approach to benchmarking is to build in silico GRNs that allow us to assess GRN reconstruction against a known ground truth[34], yet one that might not well reflect true biological GRNs. As mentioned in the previous section, there are different methodologies that can be used to assess indirectly the quality of the predicted gene regulation events but these have certain limitations. Even if TF binding to a gene is observed, it does not necessarily mean that the TF regulates that gene as TFs bind stochastically into open regions of DNA and require cooperation with other molecules for effective regulation of transcription[159]. Chromosome conformation capture technologies provide contact information and define topologically associating domains. However, their resolution might not be sufficient to detect certain genomic interactions[189]. High-resolution Hi-C maps exist, such as Micro-C[190], but their cost becomes prohibitive when comparing many experimental conditions. To address this, machine learning approaches are being

used to impute higher-coverage Hi-C maps from lower-coverage data to increase their resolution[191]. Another possibility is to use super-resolution microscopy-based alternatives, but their throughput is rather limited[189]. TFs cooperatively drive gene expression but they do so mainly as a result of DNA-mediated interactions rather than protein–protein contacts[117]. Therefore, the evaluation of TF–TF interactions might be limited to particular cases only. The evaluation of GRNs through the use of perturbation experiments is a more promising approach owing to its inherent causality. However, perturbation screens are costly, sometimes do not work as expected and may be hindered by compensatory mechanisms and unaccounted for downstream effects. In addition to all of these limitations, as gene regulation is a time-dependent process, it might be the case that experiments contradict themselves because they captured a different time frame or because of experimental noise. Because the generation of a true 'gold standard' of gene regulation seems out of reach for the moment, we are more inclined to use these different assessment strategies as a collection of 'silver standards'. We envision that a computational tool that collects and distributes such information will be useful for the community to carry out quality control on the inferred GRNs and to benchmark novel GRN inference methods. Platforms such as the Open Problems for Single-Cell Analysis project[192] offer a suitable infrastructure to run and evaluate the large variety of GRN inference methods. These would also enable the evaluation of GRN inference methods in an unbiased manner through open competition, as was illustrated for GRN inference from bulk transcriptomics data by the DREAM challenges[33].

GRNs in the bigger picture

It is important to keep in mind that GRNs are not isolated. The classic example of the lac operon, whereby a metabolite (lactose) triggers gene regulation, highlights that GRNs are part of an entangled cellular machinery, including signalling and metabolic processes. The addition of single-cell phosphoproteomics and metabolomics[193] opens the possibility of linking gene regulation to cell signalling processes using context-specific network models[194].

Furthermore, cells rarely work as autonomous systems, and gene regulation is highly coordinated within tissues. Thus, another promising direction will be the integration of multimodal data with spatial information. In particular, we envision the integration of GRNs with intracellular and intercellular communication processes[195–197] into spatially aware models[198,199]. These strategies can help in understanding multicellular regulatory processes in time and space[200].

Conclusions

Advances in high-throughput, single-cell multimodal technologies together with computational methods are paving the way to increasingly accurate GRN inference models. The large scale of the data sets makes it increasingly possible to train deep learning methods to predict gene expression from sequencing data[175,201,202]. GRNs complement these approaches by giving a more interpretable model. Together, these different approaches might help us to better understand differences in gene regulation across cell types, organs, populations and species, and serve as tools to control cell fate decisions. In the biomedical field, such knowledge could enable the identification of novel drug targets that control pathophysiological processes in different diseases.

Published online: 26 June 2023

References

1. Kim, S. & Wysocka, J. Deciphering the multi-scale, quantitative *cis*-regulatory code. *Mol. Cell* **83**, 373–392 (2023).
   This extensive review covers the molecular basis of the *cis*-regulatory code.
2. Zaret, K. S. & Carroll, J. S. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev.* **25**, 2227–2241 (2011).
3. Lai, X., Wolkenhauer, O. & Vera, J. Understanding microRNA-mediated gene regulatory networks through mathematical modelling. *Nucleic Acids Res.* **44**, 6019–6035 (2016).
4. Du, J.-X. et al. Splicing factors: insights into their regulatory network in alternative splicing in cancer. *Cancer Lett.* **501**, 83–104 (2021).
5. 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).
6. Carthew, R. W. Gene regulation and cellular metabolism: an essential partnership. *Trends Genet.* **37**, 389–400 (2021).
7. Davidson, E. H. & Erwin, D. H. Gene regulatory networks and the evolution of animal body plans. *Science* **311**, 796–800 (2006).
8. 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 Rep.* **17**, 427–442 (2022).
9. Claringbould, A. & Zaugg, J. B. Enhancers in disease: molecular basis and emerging treatment strategies. *Trends Mol. Med.* **27**, 1060–1073 (2021).
10. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3**, 318–356 (1961).
    This seminal study delineates a gene regulatory system.
11. Ideker, T., Galitski, T. & Hood, L. A new approach to decoding life: systems biology. *Annu. Rev. Genomics Hum. Genet.* **2**, 343–372 (2001).
12. Davidson, E. H. et al. A genomic regulatory network for development. *Science* **295**, 1669–1678 (2002).
13. Snyder, M. & Gallagher, J. E. G. Systems biology from a yeast omics perspective. *FEBS Lett.* **583**, 3895–3899 (2009).
14. Han, H. et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res.* **46**, D380–D386 (2018).
15. 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).
16. 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**, bav095 (2015).
17. Keenan, A. B. et al. ChEA3: transcription factor enrichment analysis by orthogonal omics integration. *Nucleic Acids Res.* **47**, W212–W224 (2019).
18. Margolin, A. A. et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinforma.* **7**, S7 (2006).
19. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinforma.* **9**, 559 (2008).
20. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE* **5**, e12776 (2010).
21. 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).
22. Fiers, M. W. E. J. et al. Mapping gene regulatory networks from single-cell omics data. *Brief. Funct. Genomics* **17**, 246–254 (2018).
23. Cha, J. & Lee, I. Single-cell network biology for resolving cellular heterogeneity in human diseases. *Exp. Mol. Med.* **52**, 1798–1808 (2020).
24. Klein, A. M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
25. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
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. 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).
30. Moerman, T. et al. GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics* **35**, 2159–2161 (2019).
31. Lambert, S. A. et al. The human transcription factors. *Cell* **175**, 598–599 (2018).
32. 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).
33. Marbach, D. et al. Wisdom of crowds for robust gene network inference. *Nat. Methods* **9**, 796–804 (2012).
    This work is a crowdsourced benchmark for GRN inference from bulk transcriptomics data.
34. 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).

35. McCalla, S. G. et al. Identifying strengths and weaknesses of methods for computational network inference from single-cell RNA-seq data. *G3* **3**, jkad004 (2023).
36. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein–DNA interactions. *Science* **316**, 1497–1502 (2007).
37. Kaya-Okur, H. S. et al. CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* **10**, 1930 (2019).
38. Lee, T. I. et al. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**, 799–804 (2002).
39. Grosselin, K. et al. High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. *Nat. Genet.* **51**, 1060–1066 (2019).
40. 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).
41. Bartosovic, M. & Castelo-Branco, G. Multimodal chromatin profiling using nanobody-based single-cell CUT&Tag. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-022-01535-4 (2022).
42. 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).
43. Boyle, A. P. et al. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–322 (2008).
44. Kelly, T. K. et al. Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res.* **22**, 2497–2506 (2012).
45. Minnoye, L. et al. Chromatin accessibility profiling methods. *Nat. Rev. Methods Prim.* **1**, 1–24 (2021).
46. 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 Genom.* **19**, 563 (2018).
47. Qin, Q. et al. Lisa: inferring transcriptional regulators through integrative modeling of public chromatin accessibility and ChIP-seq data. *Genome Biol.* **21**, 32 (2020).
48. Sonawane, A. R., DeMeo, D. L., Quackenbush, J. & Glass, K. Constructing gene regulatory networks using epigenetic data. *NPJ Syst. Biol. Appl.* **7**, 45 (2021).
49. Tabula Sapiens Consortium et al. The Tabula Sapiens: a multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).
50. Aibar, S. et al. SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
This work presents the first bespoke method to infer GRNs at the single-cell level, introducing the use of TF binding motif information for the estimation of GRNs.
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. Wagner, A., Regev, A. & Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* **34**, 1145–1160 (2016).
53. 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).
54. 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).
55. 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).
56. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
57. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
This paper introduces single-cell assay for transpose-accessible chromatin (scATAC) technology.
58. Ramirez, R. N. et al. Dynamic gene regulatory networks of human myeloid differentiation. *Cell Syst.* **4**, 416–429.e3 (2017).
59. 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).
60. 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).
61. 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).
62. Ma, A. et al. Single-cell biological network inference using a heterogeneous graph transformer. *Nat. Commun.* **14**, 964 (2023).
63. Kartha, V. K. et al. Functional inference of gene regulation using single-cell multi-omics. *Cell Genom.* **2**, 100166 (2022).
This paper introduces FigR, which has a novel integration strategy for scRNA-seq and scATAC-seq data that can enhance GRN inference.
64. Cao, Z.-J. & Gao, G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat. Biotechnol.* **40**, 1458–1466 (2022).
65. Jin, S., Zhang, L. & Nie, Q. scAI: an unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles. *Genome Biol.* **21**, 25 (2020).
66. Jansen, C. et al. Building gene regulatory networks from scATAC-seq and scRNA-seq using linked self organizing maps. *PLoS Comput. Biol.* **15**, e1006555 (2019).
67. González-Blas, C. B. et al. SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. Preprint at *bioRxiv* https://doi.org/10.1101/2022.08.19.504505 (2022).
This study presents a large, curated collection of TF binding motifs and introduces a novel GRN inference method.
68. 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).
69. Gasperini, M. et al. A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell* **176**, 1516 (2019).
70. Zuin, J. et al. Nonlinear control of transcription through enhancer–promoter interactions. *Nature* **604**, 571–577 (2022).
71. Kamal, A. et al. GRaNIE and GRaNPA: inference and evaluation of enhancer-mediated gene regulatory networks. *Mol. Syst. Biol.* https://doi.org/10.15252/msb.202311627 (2023).
72. Zhang, L., Zhang, J. & Nie, Q. DIRECT-NET: an efficient method to discover *cis*-regulatory elements and construct regulatory networks from single-cell multiomics data. *Sci. Adv.* **8**, eabl7393 (2022).
73. Duren, Z., Chen, X., Jiang, R., Wang, Y. & Wong, W. H. Modeling gene regulation from paired expression and chromatin accessibility data. *Proc. Natl Acad. Sci. USA* **114**, E4914–E4923 (2017).
74. Burdziak, C., Azizi, E., Prabhakaran, S. & Pe'er, D. A nonparametric multi-view model for estimating cell type-specific gene regulatory networks. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1902.08138 (2019).
75. Bachireddy, P. et al. Mapping the evolution of T cell states during response and resistance to adoptive cellular therapy. *Cell Rep.* **37**, 109992 (2021).
76. Kamimoto, K. et al. Dissecting cell identity via network inference and in silico gene perturbation. *Nature* **614**, 742–751 (2023).
This work presents a novel GRN inference method from scRNA-seq and scATAC-seq data that also introduces an in silico TF perturbation strategy.
77. Skok Gibbs, C. et al. High-performance single-cell gene regulatory network inference at scale: the Inferelator 3.0. *Bioinformatics* **38**, 2519–2528 (2022).
78. Fleck, J. S. et al. Inferring and perturbing cell fate regulomes in human brain organoids. *Nature* https://doi.org/10.1038/s41586-022-05279-8 (2022).
79. Li, Z., Nagai, J. S., Kuppe, C., Kramann, R. & Costa, I. G. scMEGA: single-cell multi-omic enhancer-based gene regulatory network inference. *Bioinform. Adv.* **3**, vbad003 (2023).
80. Jiang, J. et al. IReNA: integrated regulatory network analysis of single-cell transcriptomes and chromatin accessibility profiles. *iScience* **25**, 105359 (2022).
81. Wang, L. et al. Dictys: dynamic gene regulatory network dissects developmental continuum with single-cell multi-omics. Preprint at *bioRxiv* https://doi.org/10.1101/2022.09.14.508036 (2022).
82. 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).
83. 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).
84. 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).
85. Duren, Z. et al. sc-compReg enables the comparison of gene regulatory networks between conditions using single-cell data. *Nat. Commun.* **12**, 4763 (2021).
86. Kuppe, C. et al. Spatial multi-omic map of human myocardial infarction. *Nature* **608**, 766–777 (2022).
87. Argelaguet, R. et al. Decoding gene regulation in the mouse embryo using single-cell multi-omics. Preprint at *bioRxiv* https://doi.org/10.1101/2022.06.15.496239 (2022).
88. Zeng, W. et al. DC3 is a method for deconvolution and coupled clustering from bulk and single-cell genomics data. *Nat. Commun.* **10**, 4613 (2019).
89. Liberzon, A. et al. The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
90. 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* **3**, 100263 (2023).
91. 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).
92. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
93. Lou, S. et al. TopicNet: a framework for measuring transcriptional regulatory network change. *Bioinformatics* **36**, i474–i481 (2020).
94. 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).
95. Badia-i-Mompel, P. et al. decoupleR: ensemble of computational methods to infer biological activities from omics data. *Bioinforma. Adv.* **2**, vbac016 (2022).
96. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
97. Garcia-Alonso, L. et al. Transcription factor activities enhance markers of drug sensitivity in cancer. *Cancer Res.* **78**, 769–780 (2018).
98. 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).